UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Community detection on sales data

Cedric Damen

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Koenraad VANHOOF

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2020
2021

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Community detection on sales data

Cedric Damen

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Koenraad VANHOOF

This master thesis was written during the COVID-19 crisis in 2020-2021. This global health crisis might have had an impact on the (writing) process, the research activities and the research results that are at the basis of this thesis.

Dutch summary

Doel, onderzoeksopzet, en methodologie

Retailers worstelen met de juiste informatie te ontginnen uit immense hoeveelheden data, voornamelijk omwille van het Big Data verhaal dat vandaag de dag heerst. Reeds wordt al market basket analysis gebruikt om koopgedrag van klanten te analyseren. Vooral associatieregels heersen in dit onderzoeksveld. Het nadeel hiervan is dat associatieregels zich alleen maar focussen op welke producten frequent samen gekocht worden in transacties. Het koopgedrag van klanten wordt hier niet goed mee aangetoond om twee redenen. Ten eerste, een retailer heeft duizenden verschillende producten en het is waarschijnlijk dat slechts enkele van deze producten frequent samen gekocht worden. Ten tweede, associatieregels worden meestal uitgevoerd op de volledige dataset, wat ervoor zorgt dat alleen veelvoorkomende regels gevonden worden en andere, kleinere patronen van het koopgedrag niet ontdekt worden. Dit resulteert in gemiste opportuniteiten voor de retailer om zijn winsten te verhogen. De literatuur stelt voor om community detection uit te voeren om klanten te segmenteren op bepaalde karakteristieken en het koopgedrag van deze segmenten te analyseren om dit probleem op te lossen. De onderzoeksvraag van deze masterproef luidt dus als volgt:

"Kunnen klanten succesvol en consistent gesegmenteerd worden om het onderliggende koopgedrag aan het licht te brengen en marketingstrategieën hierop te baseren?"

Deze studie heeft toegang gekregen tot een dataset van een Spaanse retailer die 67 783 transacties, verspreid over 1 652 klanten, bevat. Een klantennetwerk werd opgebouwd uit deze dataset door gelijkenissen te berekenen tussen alle klanten. Daarna werden meerdere community detection algoritmes uitgevoerd op dit klantennetwerk en de resultaten werden geanalyseerd en geëvalueerd. De klanten werden gesegmenteerd op basis van promotieaankopen, online aankopen, en merkaankopen.

Resultaten en waarde van het onderzoek

Er zijn vier klantennetwerken gecreëerd door vier verschillende gelijkenissen te gebruiken. Alle gelijkenissen zijn gebaseerd op de MInteraction similarity die

Vanhoof et al. hebben geïntroduceerd in 2018. De MInteraction similarity berekent hoe gelijkend twee klanten zijn door te verifiëren hoeveel producten in beide aankooporders gekocht zijn door beide klanten. Hoe meer producten door beide klanten aangekocht zijn in hun aankooporders, hoe hoger de gelijkenis tussen deze twee klanten. De drie andere gelijkenissen zijn een variatie op de MInteraction similarity en gaven een hogere score als dit soort producten respectievelijk online aangekocht werden, in promotie aangekocht werden, of van hetzelfde merktype waren. Drie community detection algoritmes werden uitgevoerd op deze vier klantennetwerken. Deze algoritmes zijn gebaseerd op het concept om een kwaliteitsfunctie te maximaliseren, genaamd de modularity, en de resulterende community structures en klantensegmenten zullen dus voor een deel hierop geëvalueerd worden. Daarnaast, nemen we de verdeling van de groottes van de klantensegmenten per structuur ook in acht. Als er alleen maar extreem grote en extreem kleine klantensegmenten ontstaan, dan wordt dit gezien als een slechte structuur aangezien we hier geen goed idee krijgen van andere patronen van het koopgedrag. De laatste evaluatiemetriek die dit experiment gebruikt, is de lift metriek. De lift is een bekende metriek in het data mining onderzoeksveld en meet hoe interessant een bepaald concept is door te verifiëren hoe zwaar het echte resultaat van een experiment afwijkt van een willekeurig gegenereerd resultaat. Deze studie gebruikt de lift om te verifiëren of klantensegmenten afwijken van willekeurig gegenereerde klantensegmenten in hun promotieaankopen, online aankopen, of merkaankopen. De lift wordt ook gebruikt om aan te tonen voor welke productcategorieën het voorgaande waar is. Als de lift verschilt van 1, dan wijken de resultaten af van willekeurig gegenereerde resultaten. Als de lift hoger is dan 1, dan versterkt de lift de hypothese. Als de lift kleiner is dan 1, dan verzwakt de lift de hypothese.

De resultaten tonen aan dat elk klantennetwerk succesvol gesegmenteerd kan worden in hun aankopen voor elk van de community detection algoritmes. De resultaten voor de lift tonen aan dat klanten die gevoelig zijn voor producten in promotie, voornamelijk voedingswaren kopen zoals broodproducten en vleesproducten die over de toonbank verkocht worden. Klanten die gevoelig zijn voor online aankopen, kopen producten voor persoonlijke verzorging, schoonmaakproducten, en vloeibare producten zoals drank. Er zijn drie merktypes in deze dataset: “MAR”, “SUP”, en “SIN”. De “MAR”- en “SUP”-merken zijn echte merken terwijl het “SIN”-merk een huismerk is en is dus goedkoper. Klanten die voornamelijk producten van het “MAR”-merk kopen, kopen dezelfde producten van dezelfde productcategorieën als klanten die gevoelig zijn voor online aankopen en kopen daarenboven non food-durable products. Het voorgaande is moeilijk te vertalen in het Nederlands, dus blijft het in het Engels staan. Klanten die gevoelig zijn voor het “SUP”-merk, kopen voornamelijk diepvriesproducten, producten voor persoonlijke verzorging, en drank. Tenslotte, klanten die gevoelig zijn voor het “SIN”-merk, kopen voornamelijk voedingswaren zoals broodproducten, groenten en fruit, en visproducten.

De dataset bevatte geen descriptieve data over de klanten, zoals leeftijd of gender. Echter waren de RFM-scores per klant wel beschikbaar zodat deze studie toch de klantensegmenten gedeeltelijks kon identificeren. RFM staat

voor recency, frequency, en monetary en is een bekende marketing techniek om klanten te identificeren op basis van hun transacties. Een klant wordt beoordeeld op elk van deze drie concepten op een schaal van 1 tot en met 5, waarbij 5 de hoogste waarde is. Recency duidt aan hoe recent de laatste aankoop van de klant plaatsvond, frequency duidt aan hoe frequent een klant zijn / haar aankopen doet in een bepaalde tijdsperiode, en monetary duidt aan hoeveel geld de klant heeft gespendeerd in een bepaalde tijdsperiode. Klanten die gevoelig zijn voor promotieaankopen en de "SUP"- en "SIN"-merken, hebben een hoge recency en frequency score en een lage monetary score. Klanten die gevoelig zijn voor online aankopen, hebben een lage recency en frequency score en een hoge monetary score. Tenslotte, klanten die gevoelig zijn voor het "MAR"-merk, hebben een lage frequency score en een hoge recency en monetary score.

De waarde van dit onderzoek ligt in het feit dat klanten succesvol gesegmenteerd konden worden op basis van verschillende klantenkenmerken dankzij het gebruik van community detection. Hierdoor kon het specifieke koopgedrag van elk klantensegment ontdekt worden, wat niet het geval zou zijn als men associatieregels zou gebruiken. Dit zorgt ervoor dat extra klantwaarde opgevangen kan worden door de retailer via marketingstrategieën, zoals gepersonaliseerd adverteren, wat kan zorgen voor verhoogde verkopen en dus verhoogde winsten voor de retailer.

Kritische beschouwingen

De resultaten voor de promotieaankopen lijken realistisch. Voedingswaren zijn vaak onderhevig aan promoties zoals "2+1 gratis" of gaan juist in promotie als hun houdbaarheidsdatum een paar dagen later zou verlopen om nog wat winsten te kunnen genereren en het product niet hoeven te dumpen. Ook is het logisch dat persoonlijke verzorgingsmiddelen, bijvoorbeeld parfum, online aangekocht worden aangezien dit bijvoorbeeld een goed cadeau kan zijn voor iemand maar het niet de moeite waard is om ervoor helemaal naar de supermarkt te rijden. Dat schoonmaakproducten en drank vaak online aangekocht worden, klinkt niet realistisch en moet dus verder onderzocht worden door de retailer zelf. De merkaankopen lijken ook realistisch. Klanten kopen producten waarvoor hoge kwaliteit verwacht wordt, zoals persoonlijke verzorgingsmiddelen en schoonmaakproducten, van echte merken terwijl ze producten waarvoor een goedkoop alternatief kopen een mogelijkheid is, zoals voedingswaren, kopen van een huismerk. De RFM-scores van klanten voor elk kenmerk bevestigen ook de resultaten.

Een voorbeeld van een grote beperking van het onderzoek is dat descriptieve klantendata niet beschikbaar was, zoals eerder aangehaald. Hierdoor kon dit onderzoek de klantensegmenten niet degelijk identificeren en kunnen er nog gemiste opportuniteiten zijn om marketingstrategieën te creëren.

Community detection on sales data

Cedric Damen

Universiteit Hasselt, Agoralaan Gebouw D, 3590 Diepenbeek

A problem that many retailers face is to gain useful insights from vast amounts of transaction data to increase profits. Especially in recent times where millions of transactions are recorded and stored in enormous databases every day. A well-known solution to this problem is to perform market basket analysis, where the most used technique is the apriori algorithm. With this algorithm, the retailer can identify which products are frequently bought together, which helps with relocating associated products closer to each other in the isles and bundling associated products together. While this method undoubtedly helps the retailer in increasing sales, nothing is known about the purchasing behaviours of certain types of customers. Segmenting customers in communities and identifying their most frequently bought products can create interesting marketing strategies that increase customer value and thus increase profits. This study aims to show the potential of segmenting customers in communities for marketing purposes by performing community detection on multiple customer networks. The study uses a real-life data set of a Spanish retailer containing 67 783 transactions distributed over 1 652 customers as an experiment to show this. The customers were related to each other by using multiple similarity measures and this information was mapped in multiple customer networks. Then, multiple community detection algorithms were executed on each customer network to search for the best community structure. The goal in this step is to create good communities of customers that frequently buy products in promotion, online, and of a specific brand type. Finally, each community structure is analysed for the purchasing behaviours of those customers. The results of this study look promising as the customers could be successfully segmented and interesting purchasing behaviours came to light.

Keywords: market basket analysis, community detection, customer network, customer segmentation, retail

1. INTRODUCTION

In the age of Big Data, it is quite difficult for businesses to search for the right information from vast amounts of data and convert it into actionable insights in an attempt to increase profit. This is especially the case for retailers that store millions of transactions each day [1]. Retailers use this information to identify patterns in the purchases of customers in order to better understand their customers and thus increase their overall sales. Around three decades ago, retailers started to store these transaction records in enormous databases and a faint concept of market basket analysis was introduced [1]. Market basket analysis is a field in data mining that studies the set of products that customers buy in a single order [2]. The most widely used data mining technique for performing market basket analysis, is called association rules [1, 3]. Within this field, the most well-known algorithm is the apriori algorithm. Essentially, in association rules, the items sets that are most frequently bought together are identified and used for promotion strategies [3]. Another method for identifying customer purchasing patterns is using clustering techniques such as K-means, K-mediod, SOM, hierarchical clustering, etc. [1, 3] However, these two techniques can cause some problems while analysing customer purchasing patterns. First, the computational complexity of clustering techniques can make them too time-expensive when executed on millions of customer transactions [3]. Second, a problem that is prevalent in both techniques, the product-level data is quite sparse [3]. More specifically, a retailer has thousands of products while only a few of those items are bought by a customer in a single transaction. This results in insights on the product-level that are not that useful in practice, such as product associations that are quite logical and do not give extra information about the purchasing behaviours of customers. Moreover, it could well be that some lesser-known, but important customer purchasing patterns stay hidden because it only applies to a relatively small group of customers and market basket analysis searches for frequent item sets for the entire data set [3]. Ignoring these patterns for small groups of customers results in lost opportunities to increase sales and thus decreases profit. Because of these shortcomings, this paper proposes to use community detection on the customer-level to identify customer

purchasing behaviours in a retail context. More specifically, multiple community detection algorithms are performed on multiple customer networks built from a data set from a Spanish retailer that solely stores information about customer orders. Four research questions are used in this paper to extract actionable insights from this data set. These are constructed as follows:

1. Are there well-defined community structures in the different customer networks?
2. Which product categories are bought more frequently by promotion-sensitive customers?
3. Which product categories are bought more frequently by online-sensitive customers?
4. Which product categories are bought more frequently by customers that are sensitive to a specific type of brand?

Section 2 discusses the related work on this topic. Section 3 describes the methodology for the analysis performed in this paper. It covers characteristics of the given data set, how the customer networks were built, and which tools were used for applying community detection algorithms on each customer network. Section 4 contains the results of applying community detection algorithms on each customer network. The results are written down in function of the previously stated research questions. Section 5 discusses the obtained results in section 4. Finally, section 6 concludes this analysis along with some challenges of this research in section 7.

2. LITERATURE REVIEW

This section describes the core of the literature on community detection that was used for this analysis. It mainly serves as a theoretical background about the concepts that will be used in the analysis in sections 4 and 5.

First, the methodology of searching for the appropriate literature is briefly explained.

Second, the core concepts of a network are explained. Community detection is a part of the research field known as network analysis, so it is necessary that different types of networks are clearly defined in this study. The type that is used in this study, is a customer network.

Third, networks, and especially customer networks, tend to have spurious associations between nodes. These associations need to be removed from the network before any analysis is performed on it. If these associations are not removed, they could have an undesirable impact on the results of the network analysis and bias the insights gained from it.

Fourth, we will go over the core concepts of the community detection research field and explain which community detection algorithms will be used in this research.

Finally, we will go over the evaluation criteria for the used community detection algorithms. The evaluation criteria help with identifying good community structures, which is necessary to have as there are more detailed analyses of the community structures in this study.

2.1. Research Methodology

This section describes the method of research. This includes the consulted databases to search for scientific papers, the used keywords that were entered in said databases, and other methods that were used to search for scientific papers. Only one academic database was consulted, namely Google Scholar. The keywords that were used in Google Scholar are described in table 1.

In addition to using Google Scholar in order to search for scientific papers, references of the found scientific papers were used to search for more scientific papers. This is also referred to as "backward reference searching". This method is especially useful with papers that describe the current state of the art in community detection. This way, more detailed papers about a specific algorithm, for example, can be found.

2.2. The Customer Network

As stated previously, this study uses a customer network to perform community detection on. A customer network can be seen as a more specific type of a social network. Each network $G = (V, E)$ consists out of a set of nodes, also called vertices, $V(G)$ and a set of edges $E(G)$ that connects each node to one or more other nodes [4, 5, 6, 7, 8, 9]. In a social network, the set of nodes represent a group of people, so each node is exactly one person, and the edges represent any kind of relationship that two persons have [4, 6, 10, 11]. A typical social network can be represented as a $M \times M$ adjacency matrix $A = (a_{ij})$, where M is the number of customers and a_{ij} is equal to 1 if customer i and j have a relationship with each other, and 0 otherwise [4, 5, 8, 9]. However, Kim et al. [8] and Huang et al. [9] use a customer-product network and map this type of network to an adjacency matrix. Fortunately, there are not that many differences between a customer network and a customer-product network, and the method of mapping them to an adjacency matrix stays relatively the same. The only difference is that the dimensions of a customer-product adjacency matrix are $M \times N$ where M is the number of unique customers and N is the number of unique products [8, 9]. Other aspects of social networks that are necessary to understand the customer network discussed in this study, is that social networks can be undirected or directed, and unweighted or weighted [6].

An undirected network is a network in which the direction of an edge between two nodes does not matter [6]. An example of an undirected social network is

Community detection	Market basket analysis	Social network analysis
Community detection	Market basket analysis	Social network
Community detection algorithms	Market basket analysis real case	Social network analysis
Community detection networks	Market basket analysis networks	/
Community detection social networks	Market basket analysis graphs	/
Community detection applications	/	/

TABLE 1. Search terms used in the Google Scholar database to search for relevant work.

Facebook, where it does not matter if person A made friends with person B or vice versa because in the end both A and B are friends and have a relationship [6]. An interesting effect of undirected networks is that their adjacency matrix is symmetrical [12].

A directed network is a network where the direction of an edge between two nodes does matter [6, 7]. Good examples of directed social networks are social media networks where there is an option to follow other users, such as Twitter [6].

An unweighted network is a network where the edges are not weighted [6, 12, 13]. More specifically, an edge between two nodes either exists and has a value of 1 in the adjacency matrix or does not exist and has a value of 0 in the adjacency matrix. Facebook is again a good example of this type of network as two persons are either friends with each other or not [6].

A weighted network is a network where the edges have specific weights, which can be positive integers, real numbers, and even negative integers [6, 7, 12, 13]. The customer network used in this study is classified as a rating network according to Kunegis et al. [6] as it will use real numbers as edge weights. Dong and Horvath, Holme et al., and Coscia et al. [7, 12, 13] simply use the term *weighted network*.

2.3. Removing Spurious Associations in Networks

A well-known problem for networks in community detection is that not all edges in the networks carry meaningful insights [1, 2, 14, 15]. Retaining these edges can have a major negative effect on the results of community detection algorithms as they are not relevant to the problem that needs to be analysed and can lead to a skewed view of reality. So, it is in the researcher's best interest to remove these edges from the network. The most common method of removing spurious associations in networks, is to use a threshold on the weight of the edges in a network [1, 2, 14, 15]. If the weight of an edge is below this threshold, then this edge will be removed from the network, and so a relationship between two nodes vanishes. There is no real standard of what this threshold must entail, the researcher is entirely free to choose the value of the threshold. What follows are a few examples of thresholds used in the literature about applying community detection algorithms on networks.

Videla-Cavieres and Ríos [1] use a threshold they call the Top three heavy edges threshold (tthet). This threshold is calculated by averaging the three heaviest weights in the network. After calculating this average value, the authors multiply this value with increments of 5 percent. So, a list is made with the values $\{tthet \times 0.05;\ tthet \times 0.10;\ ...;\ tthet \times 1\}$. Then, the researcher can experiment with these values to get the desired pruned network.

Raeder and Chawla [2] use an arbitrary threshold to filter their product network. Their product network is constructed of 2 248 nodes and almost 250 000 edges. Over 150 000 edges have a weight of 1 and over 235 000 edges have a weight of 10 or less. This statistic encouraged Raeder and Chawla to filter out all edges with a weight of 10 or less [2].

Faridizadeh et al. [14] chose to follow the same filtering process as Raeder and Chawla [2] for their product network and chose a weight threshold of 10, which reduced their product network to 38% of its original size. Faridizadeh et al. [14] also filtered their customer network. They assumed that a customer was loyal if the customer bought more than 1 product per month and filtered customers that did not meet the previous assumption out of the data set [14].

Vanhoof et al. [15] also use a similar approach as Raeder and Chawla [2] and Faridizadeh et al. [14]. First, They opted to use the average value of their metric for customer similarity. Afterwards, they experimented with a higher and lower value [15]. The paper does not explicitly state the threshold value that Vanhoof et al. ultimately chose for their customer network, although it is important to know that an arbitrary value for the threshold was chosen here as well.

In general, it seems that researchers tend to use their own arbitrary threshold which can be quite different from each other. This is somewhat logical as each network can be extremely different from each other in structure. Also, the purpose of a researcher's network analysis is mostly different from other researchers' network analyses. Thus, it is quite difficult to create a standard of pruning a customer network, making it much easier to use trial and error when searching for the optimal threshold value.

2.4. Community Detection

Community detection is generally defined as the process of detecting similar entities, or entities that interact

with each other more than other entities, within a network [4, 15]. This group of similar entities is then called a community [4, 15]. Another, more technical definition of a community which is commonly used, is when a subgraph in a network has more edges within that subgraph than edges pointing out of that subgraph [7, 16, 17, 18, 19]. Coscia et al. [7] call this definition the *density-based definition* as it is based on how densely packed the edges are in and between communities of nodes. The main idea behind a community is that there should be more edges in that community than edges pointing out of that community [7, 16, 17, 18, 19].

To find such community structures in networks, one must use one or more of the available community detection algorithms. Community detection algorithms divide a network into partitions, which are then seen as the communities in that network. These partitions can be overlapping or disjoint [7, 16, 19]. Overlapping communities are communities where the nodes can belong to multiple communities, which is in contrast with disjoint communities where each node can only belong to one single community [7, 16, 19]. This study only focuses on searching for disjoint communities since the characteristics we use to divide customers into communities do not allow for overlapping communities as is seen in research questions in section 1. Community detection algorithms find communities either by using an agglomerative method or a divisive method [16, 20, 10]. Agglomerative community detection algorithms iteratively merge clusters of nodes that are similar to each other until a given threshold is achieved [16, 19, 20, 10]. Divisive community detection algorithms iteratively split clusters of nodes that are not similar to each other until a given threshold is achieved [16, 20, 10, 21]. This threshold is a certain value of a quality measure that is used by a group of community detection algorithms. The most well-known example of such a measure, is the modularity [16]. The modularity will be explained later in this section. The earlier community detection algorithms, which were graph partitioning methods, mainly used divisive methods to divide a network into multiple communities. However, a major downside of this method is that the number of communities within that network must be known a priori, which is frequently not the case [16, 19]. So, these divisive algorithms did not look to achieve a certain threshold of a quality measure to end the algorithm. A well-known example of a divisive community detection algorithm, is the algorithm of Newman and Girvan [22]. Because of this shortcoming, researchers started to create and use agglomerative community detection algorithms and thus most of the traditional algorithms were agglomerative. Although, it seems that there recently was a renaissance of divisive algorithms because the a priori knowledge of the number of communities in a network is not necessary anymore [16].

As is stated previously, many different community detection algorithms exist for different purposes [16]. In order to know which community detection algorithm to use for our purpose, we looked up a few comparative analyses [23, 24, 25, 26]. In general, the literature suggests that the Louvain algorithm [26, 27] is the best to use in most of the cases [24, 25, 26]. A community detection algorithm that is not mentioned in the comparative analysis literature is the Leiden algorithm [28]. This algorithm is an improvement upon the Louvain algorithm [28], thus it could be interesting to compare the results of these two algorithms in this analysis. Other algorithms that were recommended in the comparative analysis literature [24, 25, 26], are the Fast & Greedy [29] and Eigenvector algorithm [30]. So, they will be included in this study aswell. The Infomap algorithm [26, 31] is another community detection algorithm that can outperform most other algorithms according to the comparative analysis literature [23, 24, 25, 26]. However, due to some coding problems, the Infomap algorithm could not be included in this study. This study uses the "CDlib" Python package [32] to perform these community detection algorithms on the different customer networks and some optional dependencies that are necessary to run the Infomap algorithm did not work. It seems that these packages were built for PCs that have Linux as its operating system, while the PC that was used in this study has Windows 10 as its operating system. We still recommend to use the Infomap algorithm in future researches. So, the community detection algorithms that are used in this study, are the Eigenvector algorithm [30], the Fast & Greedy algorithm [29], the Louvain algorithm [27], and finally the Leiden algorithm [28]. All these algorithms are modularity-based algorithms and thus try to find communities by optimising a quality measure called the modularity [16, 27, 28, 29, 30]. The modularity will also be used as an evaluation metric for the community structures resulting from performing each community detection algorithm on the customer networks and will be explained in the following section next to the other evaluation metrics.

2.5. Evaluation of community structures

There are many metrics to evaluate a community structure within the community detection research field and no real standard has ever been set up. This is mainly due to the abundance of different quality metrics that were introduced along with new community detection algorithms [16]. Evaluating the resulting community structure of a community detection algorithm is also a non-trivial task to perform as it highly depends on the filtering of the edges in the network. If a high threshold for filtering is chosen, only the strongest associations between the nodes remain and the density of the network decreases [15]. This causes the community detection algorithm

to find more smaller communities and in extreme cases a community detection algorithm is not even necessary as certain parts of the network can become disconnected, forming a community on its own [15]. If a low threshold for filtering is chosen, more spurious associations between the nodes are included in the network and the density of the network increases [15]. This causes the community detection algorithm to find a small amount of communities that are quite big [15].

As shown in the previous section, this study only uses community detection algorithms that try to optimise the modularity when constructing community structures [16, 27, 28, 29, 30], so this study will use the modularity to evaluate the community structures resulting from executing the community detection algorithms on the customer networks. Essentially, the modularity checks the difference between how many edges of a node are pointing towards other nodes in the same community versus nodes that are in another community, and this for every node. This result is compared to the result from a randomly generated graph and the modularity is calculated. The modularity can take values between -1 and 1. The higher the modularity, the stronger the community structure [14, 16, 27, 28, 29, 30]. However, we found in this study that solely using the modularity to evaluate a community structure is a narrow-minded perspective. This is shown further down this study in section 4.1.1. In short, we found that increasingly removing edges from a customer network, artificially increases the modularity score for each community detection algorithm. This is somewhat of a logical result as is explained in the previous paragraph and shown in the study of Vanhoof et al. [15]

Because of this downside, this study uses two other methods of evaluating community structures, namely manually inspecting the communities in the community structures and the lift measure used in data mining. First, manually inspecting the community structures seems cumbersome and inefficient, however it is necessary to know what the different sizes of each community in a community structure are. A community structure with extremely large and extremely small communities is less informative than a balanced community structure where the sizes of the communities are more equal to each other. For example, a customer network consisting of 1 000 nodes is separated into 5 communities by two community detection algorithms. The sizes of the communities resulting from the first algorithm are 500, 480, 8, 6, and 6 nodes respectively. The sizes of the communities resulting from the second algorithm are 325, 250, 200, 175, and 50 nodes respectively. In this case, the community structure of the first algorithm is less informative about the customer behaviour than the community structure of the second algorithm, because the three last communities of this community structure are too small to analyse for customer behaviour. The results of the analysis of these small communities would not be useful as it only applies to a few customers and thus it cannot be generalised to the entire customer population of the retailer.

Second, the lift interestingness measure is often used in data mining to measure the importance an association rule [33, 34, 35, 36]. This is essentially done by checking if the association rule occurs more frequently in the data set than is expected. This can also be used in community detection to evaluate community structures based on certain characteristics of customers. This is done by defining the lift as the real behaviour of the customer group divided by the expected behaviour of the customer group [15]. The real behaviour is the result of an aspect of a customer community in the community structure while the expected behaviour is the result of the same aspect of a randomly generated customer community of the same size as the real community. So, in more practical terms, a community detection algorithm is run on a customer network. Next, for example, the number of online purchases of a community is calculated. Then, a same-sized random sample of transactions is taken from the real retail data set to simulate a random community of customers. Afterwards, the number of online purchases in the random sample is calculated. Finally, the real number of online purchases is divided by the random number of online purchases and the lift value is calculated for that community. This is done for each community in each community structure and for online, promotion, and brand purchases. The lift measure is helpful here because we know exactly which customer communities are more sensitive to online, promotion, or brand purchases. These communities would have a lift value higher than 1 because their real results for a customer characteristic are higher than the expected results for that same characteristic. If the lift value is lower than 1, then we say that they are less sensitive for a certain customer characteristic as the real results are lower than the expected results. If the lift value is approximately 1, then the real results are approximately equal to the expected results and the customers in that community are indifferent for that characteristic. Afterwards, these customer communities can be analysed for their purchasing behaviours and marketing strategies can be created based on these purchasing behaviours. It also helps with evaluating community structures. If, for example, most of the communities in a community structure have a lift value of approximately 1 for most of the previously mentioned customer characteristics, then we can assume that this community structure is no different from a randomly generated community structure and thus that this is not a good community structure.

3. METHODOLOGY

This section describes the methodology that was used in this study. The methodology is divided into three sections. Section 3.1 analyses the given retail data set of the Spanish retailer to get a basic understanding of the data that will be used for segmenting the customers in communities. Section 3.2 describes how the different customer networks were created as these are a necessary input for the community detection algorithms. Finally, section 3.3 explains which tools were used to execute the chosen community detection algorithms on the different customer networks.

3.1. Data

The data was provided by a Spanish supermarket and contains anonymised transactions of customers over a two week period (26/02/2018 - 10/03/2018). The dimensions of the data frame are 67 783 rows and 21 columns. A single row represents one single transaction made by a certain customer on a certain date. The other columns give extra information about the transaction, such as: at which supermarket the product was bought, the quantity of the bought product (for example, if the product is water, then the quantity could be 2 liters), which product section the product belongs to (for example, water belongs to the section "LIQUIDOS", which translates to "fluids"), etc. The other columns will be used as meta data for the networks in order to explain the different communities in a certain context.

The data was analysed before it was transformed for building the different kinds of networks and it provided the following insights:

- There are 1 654 unique customers and 6 820 unique products.
- Most of the best selling products are part of the department "FRUTAS Y HORTALIZAS", which translates to "fruits and vegetables". A top 10 of the best selling products is shown in table 2.
- There are 18 different product departments in the dataset. The most commonly found product departments in customer orders are: "FRUTAS Y HORTALIZAS", "ALIMENTACION DULCE", and "ALIMENTACION SALADA". Products that belong to these 3 product departments compose around 45.09% of the whole dataset. A top 10 of the most bought product departments is shown in table 3.
- There are 2 364 product subcategories in the data set. A product subcategory is a level between the product departments and the individual product IDs that provides extra information about a specific product. For example, the most bought product subcategory is "AGUA SIN GAS MESA", which was bought in 898 transactions. This product subcategory is a part of the product department "LIQUIDOS" and individual products that belong to this product subcategory are different brands of still water.
- Out of the 67 783 transactions, 43 136 (63.64%) were bought offline while 24 647 (36.36%) were bought online. A bar plot of this information can be found in figure 1 in the appendix.
- out of the 67 783 transactions, 61 648 (90.95%) were not bought during a promotion while 6 135 (9.05%) were bought during a promotion. A bar plot of this information can be found in figure 2 in the appendix.
- Out of the 67 783 transactions, 35 848 (52.89%) are from the MAR brand, 16 031 (23.65%) are from the SIN brand, and 15 904 (23.46%) are from the SUP brand. A bar plot of this information can be found in figure 3 in the appendix.

3.2. Building the Customer Network

As stated in the previous subsection, the data set that was provided is in the form of a regular dataframe, where each row represents a single product that was bought by a single customer in an order of that customer. The Python package "CDlib" of Rossetti et al. [32] was used in order to split this dataframe into several communities. However, the community detection algorithms of this package require a network as input to function properly. Therefore, the Python package "NetworkX" [37] was used to transform the data set in the desired network form. This subsection describes on a high-level how the previous was achieved.

The customer network shows the relationships between the customers in the retail data set. The nodes represent the individual customers, they are extracted from the column "socio". Each edge represents a relationship between two customers. A relationship between customer i and customer j is based on the similarity between their orders. The similarity used in this paper, is calculated by using the MInteraction similarity, recently proposed by Vanhoof et al [15]. Here, the customers are represented as customer bags, where each customer bag consists of mutliple orders made by one single customer on different dates in a given time period. Each order is also considered as a multi-set in which the multiplicity of each product in the order is the number of times that the product was bought in that order [15]. A mathematical equation for the MInteraction is found in the following section.

Beside the MInteraction similarity metric proposed by Vanhoof et al. [15], this paper also introduces another similarity based heavily on the MInteraction similarity. The core of this paper is to examine which product categories are more frequently bought by customers who are more sensitive to a certain type of purchases. Here, online, promotional, and brand purchases are considered because this data is easily accessible in the available retail data set and the retailer could possibly change its marketing strategies

Articulo	Frequency	Product Department	Product Subcategory
26094	653	FRUTAS Y HORTALIZAS	PLATANO
242071	358	FRUTAS Y HORTALIZAS	FRESAS DE VERANO
991945	335	PANADERIA	PAN COMÚN
200632	260	LIQUIDOS	AGUA SIN GAS MESA
5526684	256	FRUTAS Y HORTALIZAS	CALABACINES
17200	220	FRUTAS Y HORTALIZAS	PERA ANUAL
56901	201	FRUTAS Y HORTALIZAS	PATATAS BLANCAS COMUNES
959935	201	PANADERIA	PANES RUSTICOS
7098385	191	HUEVOS	HUEVOS GRANDES L
53007	183	FRUTAS Y HORTALIZAS	AGUACATES

TABLE 2. The 10 most sold products in the retail store data set.

Product Department	Frequency
FRUTAS Y HORTALIZAS	10 509
ALIMENTACION DULCE	10 422
ALIMENTACION SALADA	9 635
CHARCUTERIA LIBRESERVICIO	6 442
LIQUIDOS	5 366
DROGUERIA	5 021
PANADERIA	3 589
LACTEOS REFRIGERADOS	3 530
PERFUMERIA	2 752
CARNE LIBRESERVICIO	2 647

TABLE 3. The 10 product departments with the most sales.

based on the results found from these purchases. The similarity in question will be called the *skewed MInteraction* from this point onwards. Essentially, the skewed MInteraction is the regular MInteraction, but altered in favour of the aforementioned purchasing types. So, when two customers purchase the same product online, for example, then their similarity score increases even more than what would be with the regular MInteraction. This skewed version of the MInteraction is described further down in this paper.

3.2.1. Regular MInteraction similarity

Vanhoof et al. first calculate the similarity between two customer orders of two different customer bags by using a modified Jaccard Similarity, called the $MSJaccard similarity$, which is defined as follows [15]:

$$MSJaccard(o_i, o_j) = \frac{|o_i \cap o_j|}{|o_i \cup o_j|} = \frac{\sum_{p_k \in |o_i \cap o_j|} min(multi_{p_k}(o_i)), multi_{p_k}(o_j)}{\sum_{p_k \in |o_i \cup o_j|} max(multi_{p_k}(o_i), multi_{p_k}(o_j))} \quad (1)$$

Where o_i denotes the $i-th$ order of a customer and o_j denotes the $j-th$ order of another customer.

As is shown in the equation above, the MSJaccard metric is similar to the normal Jaccard Similarity as it divides the intersection of the orders by the union of the orders. However, it is slightly modified to accommodate for the multiplicity of products in orders that are present in the given data set.

Then, the similarities of orders that belong to a certain customer bag need to be aggregated to the customer bag level in order to compare customer bags to each other. For this, Vanhoof et al. [15] introduce two extra metrics called the *MInteraction similarity* and the *Customer bag similarity* (*BagSim*). The *MInteraction similarity* is defined as follows in Vanhoof et al. [15]:

$$MInteraction(C_i, C_j) = \frac{min\{N_i, N_j\}}{max\{N_i, N_j\}} * \frac{BagSim(C_i, C_j) + BagSim(C_j, C_i)}{2} \quad (2)$$

$$BagSim(C_i, C_j) = \frac{\sum_{k=1}^{N_i} \mu_{C_j}(o_i^k)}{N_i} \quad (3)$$

$$\mu_C(o) = max_{x \in C}\{MSJaccard(x, o)\} \quad (4)$$

Where C_i and C_j denote the customer bag of customer i and j respectively, N_i and N_j are the number of orders for customer bag i and j respectively, and BagSim denotes the similarity between two customer bags based on the similarity between the orders in each bag. In summary, the BagSim measure uses the MSJaccard similarity

The customer network can be visualised in table form as is shown in table 4.

Source Customer	Target Customer	Similarity
2889	37484	0.015214
2889	54392	0.001899
2889	66276	0.012561
2889	88682	0.010909
2889	93298	0.001690

TABLE 4. The customer network visualised in table form, also called an edge list of a network. It is called an edge list because each row in this table represents an edge between two nodes in the customer network. The edge runs from the first customer (first column) to the second customer (second column) with a customer similarity that is stored in the third column.

3.2.2. Skewed MInteraction similarity

The only difference between the regular MInteraction similarity and the skewed MInteraction similarity is how the $MSJaccard(o_i, o_j)$ is calculated. The MSJaccard for the skewed MInteraction similarity is equal to the MSJaccard for the regular MInteraction plus a variable σ that makes it skewed to a certain purchasing type. The σ is calculated as the number of instances where the same product bought in both orders, is bought online, in promotion, or is the same type of brand in both orders divided by the number of unique products in both orders.

3.3. Tools for Identifying Communities

As previously stated, the Python package "CDlib", built by Rossetti et al. [32], was used to execute all the chosen community detection algorithms, except the Infomap Algorithm, on the customer network. It also helps with evaluating the community structures provided by the algorithms and comparing these with each other. The Infomap Algorithm from the "CDlib" package could not be performed on the networks because it is dependent on another Python package that is only available on PCs that have Linux installed as operating system, while this analysis used a PC that has Windows installed as operating system.

4. RESULTS

This section describes the results that were obtained by executing the previously discussed community detection algorithms on the different types of networks. The results are divided into four research questions, which make the results more interpretable and actionable in a business context.

4.1. Are there well-defined community structures in the different customer networks?

First, the network needs to be pruned to remove the more spurious associations between customers. We opted to use our own way of pruning the customer network as there is no standard for this in the literature, as mentioned in section 2.3. The customer network is pruned by keeping a top x% of the heaviest edges in the network. To find the most optimal value of x per community detection algorithm, we let x start at a value of 100 and decrease it by steps of 5 until the minimum value of 5 is reached. Here, the minimum value is 5 because a value of 0 would indicate that 0% of the heaviest edges in the customer network would remain after pruning, which results in an empty network. The maximum, or starting value of x is 100 as this indicates the full, unpruned customer network.

Next, each chosen community detection algorithm is performed on each iteration of the pruned customer network and the resulting community structures are evaluated as described in section 2.5. Here the Fast & Greedy Algorithm is not included for any similarity metric because of the consistent poor community structures that it provides at any percentage value of the top heaviest edges for each similarity metric. The algorithm would consistently construct community structures that include extremely large as well as extremely small communities and this is not a desirable result as is discussed in section 2.5. So the remaining community detection algorithms are the Eigenvector, Louvain, and Leiden algorithms.

The following sections describe the best community structures resulting from each community detection algorithm for each similarity metric.

4.1.1. Regular MInteraction similarity

An overview of the best split per community detection algorithm for the regular MInteraction similarity metric is given in table 5. The modularity of the best split of each community detection algorithm is also given in table 5. As previously stated in section 2.5, solely using the modularity to evaluate a community structure is not a good method and is quite narrow-minded. Why this is the case, can be seen in figure 4 in the appendix. This figure represents the relationship between a partition of the customer network left over after keeping a percentage of the heaviest edges in the customer network and the modularity score after performing a community detection algorithm on that partition. The graph is best read from right to left and the labels on the x-axis are best read as percentages instead of decimals, for example 0.2 is best read as 20%. Figure 4 shows that the modularity of any community structure can be artificially increased by incrementally removing weaker edges from the customer network. The previous statement sounds good in theory because a higher modularity score means a better community structure,

Community Detection Algorithm	% edges	Communities	Modularity score
Eigenvector Algorithm	10	6	0.3709
Louvain Algorithm	20	7	0.3058
Leiden Algorithm	25	7	0.2642

TABLE 5. General information about each community structure.

although problems arise in practice. According to the modularity, the smaller the network, the higher the modularity score and thus the better the community structure. For example, if we only focus on the modularity to evaluate a community structure, then performing the community detection algorithms on only 1% of the heaviest edges of the customer network would yield very high modularity scores (around 0.61 for the Eigenvector algorithm and around 0.71 for the Louvain and Leiden algorithms). However, at this point 403 customers (24.37%) are filtered out of the customer network and each community structure has a large amount of extremely small communities. These two points make it extremely difficult to make inferences out of the communities and generalise them to the entire population of customers in the retail data set. However, the modularity can still be useful to give an indication of the presence of a community structure within a network.

Table 6 shows the resulting community structure after performing the Eigenvector Algorithm on the pruned customer network. At first glance, It seems that communities 1, 4, 5, and 6 are the customer communities in which customers are more prone to buying products online. In these communities, the number of products bought online is around 50% of the number of products bought in each community. Customers in communities 2 and 3 seem to be more sensitive to buying products that are in promotion. The 2nd community has the highest number of products bought in promotion, however it also has the highest number of products bought of any community. So, it should be logical that this community has the highest number of products bought in a promotion. Although, if we look at the community with the second highest number of products bought in it, namely the 1st community, we see that while it only has around 1 500 products less than the 2nd community, the products that were bought in promotion are half of that of the 2nd community. Thus, it seems that the 2nd community can have interesting results for products bought in promotion. Almost the same logic applies for the 3rd community. It has around 2 000 and around 3000 products bought in it less than the 1st and 2nd community respectively, however, while it has around 500 products bought in promotion less than the 2nd community, it has around 600 products bought in promotion more than the 1st community. This makes the 3rd community also an interesting candidate for products bought in promotion. In order to calculate which brand type is more frequently bought in a certain community, we refer to the values in figure 3 and their corresponding percentage values, which are first calculated in the last bullet point in section 3.1. To summarise these values again, out of the 67 783 transactions, 35 848 (52.89%) are bought from the MAR brand, 16 031 (23.65%) from the SIN brand, and 15 904 (23.46%) from the SUP brand. Here, the same percentage values will be calculated for every community in all community structures. If a percentage value for a certain brand is higher than the corresponding percentage value previously summed up, we say that that specific brand type is bought more commonly by that community. This method can be a good start in identifying which communities buy which brand types more frequently as it is expected that a randomly generated community would follow the same brand type percentage value rules that were previously summed up. It seems that, based on these percentage values, customers in Eigenvector communities 2, 4, and 6 tend to buy products from the MAR brand, customers in Eigenvector communities 1, 3, and 5 do so for the SUP brand, and customers in Eigenvector communities 2 and 3 do so for the SIN brand.

To further reinforce the inferences made in the previous paragraph, the lift values for online, promotion, and brand purchases for each Eigenvector community are calculated. These lift values are shown in the appendix in figures 5, 6, and 7 respectively. Figure 5 shows that customers in communities 1, 4, 5, and 6 are indeed more prone to buying products online as the lift values of these communities lie between 1.3 and 1.5, which is quite high compared to a regular lift value of 1. Furthermore, figure 6 confirms that customers in communities 2 and 3 are more sensitive to promotion sales than the other communities, with both communities having a lift value of around 1.2, which is lower than the online communities, but still respectable. Finally, figure 7 confirms the assumptions made about the brand type purchasing behaviours at the end of the previous paragraph.

Table 7 shows the results for performing the Louvain Algorithm. The results suggest that customers in community 1, 3, 4, and 5 tend to buy more products online than the customers in the other communities. Around 50% of the total number of products bought in these communities, is bought online, which is a significantly higher percentage than for the other communities. It seems that only the customers in community 2 like to buy products in promotion more than the customers in the other communities. It has the highest amount of products bought during a promotion

	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Customers	490	404	270	210	154	123
Products	15696	17181	13934	9602	6926	4441
Online	8372	3357	2380	4514	3665	2359
Promotion	985	2053	1552	701	509	335
MAR brand	8004	9124	6737	5842	3393	2745
SUP brand	4180	3500	3432	1892	1959	941
SIN brand	3512	4557	3765	1868	1574	755

TABLE 6. Relevant information about each community in the Eigenvector community structure.

	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6	Com 7
Customers	684	518	170	112	95	44	29
Products	20052	27881	8241	4948	4035	1720	904
Online	9845	5048	4091	2842	1926	661	234
Promotion	1500	3213	572	314	281	142	113
MAR brand	10611	14155	5050	2344	2048	1074	564
SUP brand	4961	6238	1570	1482	1187	294	172
SIN brand	4480	7488	1621	1122	800	352	168

TABLE 7. Relevant information about each community in the Louvain community structure.

out of all communities, however one can counteract this statement by noting that this community also has the highest number of products bought out of all communities. It is rather logical that the community with the highest number of products bought also has the highest number of products bought during a promotion. Although, if we look at the percentage of products bought during a promotion divided by total number of products bought for each community, we can see that this percentage has a value of 6% - 8% for each community except community 2, where this percentage has a value of 11.52%. Furthermore, the second largest community in terms of total number of products bought and number of products bought during a promotion, is community 1 with 20 052 total number of products and 1 500 products in promotion. Comparing communities 1 and 2 to each other, we see that community 1 knows a decrease of 28% for the total number of products bought and a decrease of 53% for the number of products bought during a promotion. The previous statement further confirms that the customers in community 2 are more sensitive to promotions than the other communities. Finally, Louvain communities 1, 3, and 6 tend to buy products from the MAR brand, customers in Louvain communities 1, 4, and 5 do so for the SUP brand, and customers in Louvain community 2 do so for the SIN brand.

As with the Eigenvector community structure, the lift values for online, promotion, and brand purchases for the Louvain community structure are also calculated and are shown in figures 8, 9, and 10. The lift values for the online sales in figure 8 also confirm that customers in communities 1, 3, 4, and 5 like to buy products online more than customers in other communities. Community 6 also seems to be sensitive to online sales as it barely reaches the mark of having a lift value above 1. The lift values for the promotion purchases in figure 9 match the conclusions made from table 7 in the previous paragraph. The 2nd community has a lift value of around 1.25, confirming that customers in this community are more likely to buy products that are in promotion. Figure 9 also implies that the customers in the 7th community buy more products in promotion than other communities, having a lift value of around 1.4. However, the 7th community is an extremely small community, containing only 29 customers, so it will not be included in the further detailed analyses because its results could be counter-intuitive or contradictory to other results due to its small size. Furthermore, inferences made from this community cannot be generalised to the entire population of customers in the retail data set. Finally, figure 10 confirms the assumptions made about the brand type purchasing behaviours at the end of the previous paragraph.

Table 8 shows the results for performing the Leiden algorithm. It seems that customers in communities 1, 3, and 5 are more likely to buy products online. The number of products bought online for these communities is around 50% of the total number of products bought in these communities. For customer communities that are more sensitive to promotions, the results of the Leiden algorithm are almost the same as for the Louvain Algorithm. This is not surprising as the Leiden Algorithm is supposed to be an improvement of the Louvain Algorithm and is thus based on it. Here, community 2 is also the most likely candidate to be more sensitive to promotions. It has the highest number of products bought during a promotion out of all communities as well as its percentage of products bought during a promotion divided by total number of products bought has a value of 11.76%.

	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6	Com 7
Customers	640	466	170	125	113	94	44
Products	18892	22973	8163	7239	4980	3971	1563
Online	9692	4358	4004	1328	2909	1695	661
Promotion	1366	2701	579	729	316	326	118
MAR brand	10099	11955	4978	3477	2366	1979	992
SUP brand	4662	4947	1565	1774	1500	1191	265
SIN brand	4131	6071	1620	1988	1114	801	306

TABLE 8. Relevant information about each community in the Leiden community structure.

However, in contrast to the results of the Louvain Algorithm, this percentage is also relatively high for the fourth community with a value of 10.07%, introducing another candidate for communities that are sensitive to promotions. Finally, Leiden communities 1, 3, and 7 tend to buy products from the MAR brand, customers in Leiden communities 1, 4, 5 and 6 do so for the SUP brand, and customers in Leiden communities 2 and 4 do so for the SIN brand.

Figures 11, 12, and 13 show the lift values for online, promotion, and brand purchases for the Leiden community structure. Figure 11 confirms that customers in communities 1, 3, and 5 buy products online more than expected, especially community 5 with a lift value of around 1.6. Furthermore, customers in communities 6 and 7 seem to follow this trend as well, but less than the previous communities as their lift values lie around 1.2. Additionally, the results in figure 12 match the conclusions made in the previous paragraph about customers that are more sensitive to promotions, being that only customers in the 2nd community are more sensitive to that kind of sales. Although, it seems that customers in the 4th community also buy products during a promotion more than expected. However, the lift value for buying products in promotion for this community barely reaches a value greater than 1 and the lift value for not buying products in promotion in the same community also lies around 1. Finally, figure 13 confirms the assumptions made about the brand type purchasing behaviours at the end of the previous paragraph.

In conclusion, we can see that the customer network exhibits a well-defined community structure when the customer network is constructed by the normal MInteraction similarity. The modularity scores in table 5 give an indication that some form of a community structure is present in the customer network, however detailed information about the community structure is missing in this table. Table 6, 7, 8 and figure 5 to 13 help with this and show that the customer network can be split up in multiple communities of customers and that these communities can be more sensitive to a specific type of purchases, such as online purchases, promotion purchases, and purchases of a specific brand type.

4.1.2. Promotion skewed MInteraction similarity

Table 9 shows the best split for each community detection algorithm for the customer network based on the MInteraction similarity skewed to promotion purchases. It also provides the modularity score for each community structure. Curiously enough, it seems that the modularity scores are not that different from the ones of the community structures based on the normal MInteraction. This is easily shown by the modularity scores of the Eigenvector algorithm in both situations. The same percentage of heaviest edges is being left over, namely 10%, and the modularity scores are nearly identical. The modularity scores for the Louvain and Leiden algorithms are lower for the promotion skewed MInteraction, but this is probably due to the higher percentages of heaviest edges that are left in the customer network for these algorithms. Compared to the results from the normal MInteraction, each community structure here has 1 community less. So, it could be that the results from the community structures here are more concentrated per community.

Table 10 shows the number of customers, products bought, and products bought during a promotion in each community for each community structure for the promotion skewed MInteraction. The sizes of the communities of each community structure for the promotion skewed MInteraction are somewhat different than their normal MInteraction counterpart. The Eigenvector community sizes for the normal MInteraction are more evenly spread out, while here there is a bigger gap between the sizes of the biggest and smallest community. Although, this should not form a problem as the smallest community is still reasonably sized at 90 customers. In contrast, the Louvain and Leiden community sizes for the promotion skewed MInteraction are more evenly spread out than their normal MInteraction counterpart, which looks promising because this is a desirable result as explained in section 2.5.

As theorised in the previous paragraph, the customer communities are indeed more concentrated when it comes to promotional sales. For example, customers in community 1 in the Eigenvector community structure are vastly more sensitive to promotional sales than customers in other communities. This is shown by the fact that $\frac{4075}{36840} \approx 11.06\%$ of products bought in this community, are bought during a promotion. This

is not the case for the other Eigenvector communities where only approximately 6% or 7% of all products is bought during a promotion. This gives an indication that customers in the 1st Eigenvector community are more sensitive to promotions. The lift values in figure 17 in the appendix also reinforce this statement as the 1st Eigenvector community is the only community with a lift value higher than 1.

The Louvain community structure also has clear customer communities who are more sensitive to promotions, namely communities 1 and 5. Customers in the 1st community bought $\frac{3241}{29199} \approx 11.10\%$ of all their products during a promotion and customers in the 5th community did so aswell for $\frac{613}{5850} \approx 10.48\%$ of the total number of products bought in that community. The customers in the other Louvain communities bought approximately 6% to 8% of their total amount of products bought in a promotion, which is relatively low. The lift values in figure 19 also provide proof that only customers in communities 1 and 5 are more sensitive to promotions than the other Louvain communities as only their lift values are above 1, especially community 1 has a relatively high lift value.

The results from the Leiden community structure are quite similar to the Louvain community structure, probably because the Leiden algorithm is an improvement of the Louvain algorithm. Here, it looks like customers in community 1 and 5 are also more sensitive to promotions than customers in the other Leiden communities. Customers in communities 1 and 5 bought $\frac{3153}{27831} \approx 11.33\%$ and $\frac{649}{6203} \approx 10.46\%$ of their total number of products bought in a promotion respectively, while all the other Leiden communities bought approximately 6% to 8% of their total number of products bought, in a promotion. The lift values in figure 21 also confirm the previous statement as the lift values of communities 1 and 5 are the only ones that are higher than 1, especially community 1 has a relatively high lift value.

In conclusion, it seems that there is significant proof that each community detection algorithm provides a good community structure and thus we can assume that there is a well-defined community structure within the customer network when it is based on the promotion skewed MInteraction.

4.1.3. Online skewed MInteraction similarity

Table 11 shows the best split for each community detection algorithm for the customer network based on the MInteraction similarity skewed to online purchases. The modularity scores for each best split is also given in this table. Here, it also looks like the modularity scores are almost the same as the modularity scores of the community structures for the normal MInteraction. The same percentage of heaviest edges for the Leiden algorithm is left over for the online skewed and normal MInteraction and their modularity scores in both situations are almost the same. The Eigenvector and Louvain algorithm use more edges here than for the normal MInteraction similarity, so their modularity scores are also lower. A significant difference from the normal MInteraction results is that the Eigenvector community structure has 1 community less, and the Louvain and Leiden community structures have 2 communities less. So, as with the results from the promotion skewed MInteraction in the previous section, it could be that the number of products sold online are more concentrated in a few communities here because there are less communities.

Table 12 shows the number of customers, products bought, and products bought online in each community for each community structure for the online skewed MInteraction. The sizes of the communities in each community structure here are more evenly spread out than the sizes of the communities in each community structure for the normal MInteraction similarity, which is already a good sign because this is a desirable result.

Starting with the Eigenvector community structure, it is quite clear that community 2 is mostly populated by customers that buy their products online as $\frac{12832}{15276} \approx 84\%$ of all products in this community is bought online. The 2nd community is also quite sizeable, so the observations from this community can be easily generalised to the entire retail data set population. It also seems that community 5 is sensitive to online purchases because $\frac{4438}{7663} \approx 57.91\%$ of all products in this community is bought online while this is only the case for 10% to 20% of all the products for the other three Eigenvector communities. Figure 23 shows the lift values for the Eigenvector community structure based on the online skewed MInteraction. The lift values in this figure also reinforce the previously made claims. Both community 2 and 5 have a very high lift value, especially the 2nd community. So, we can conclude that customers in the 2nd and 5th Eigenvector communities are more sensitive to online purchases than the other Eigenvector communities.

The Louvain community structure also has a quite big customer community where a large portion of all the products is bought online. This is the 3rd Louvain community with a size of 425 customers and $\frac{14203}{17860} \approx 79.52\%$ of all its products bought online. The 4th community also looks like a good candidate as another Louvain community that is sensitive to online purchases. It is a much smaller community than the 3rd community at 163 customers, although it is still sizeable enough to be included in the analysis. Customers in this community bought $\frac{3903}{7160} \approx 54.51\%$ of all products in this community online, while the customers in the other Louvain communities did so for 7% to 18%. Figure 25 shows the lift values for the Louvain community structure based on the online skewed MInteraction. Here, it is also confirmed that customers in communities 3 and 4 are more sensitive to online purchases as is shown by their extremely high lift values.

Community Detection Algorithm	% edges	Communities	Modularity score
Eigenvector Algorithm	10	5	0.3750
Louvain Algorithm	35	6	0.2230
Leiden Algorithm	30	6	0.2467

TABLE 9. General information about each community structure for the MInteraction similarity skewed to promotional sales.

	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Eigenvector Customers	713	435	236	177	90	/
Eigenvector Products	36840	10095	9798	7482	3564	/
Eigenvector Promotion	4075	605	743	500	211	/
Louvain Customers	537	416	341	149	108	101
Louvain Products	29199	9724	12272	6711	5850	4025
Louvain Promotion	3241	553	1046	439	613	243
Leiden Customers	507	447	330	145	116	107
Leiden Products	27831	10782	12064	6423	6203	4478
Leiden Promotion	3153	617	996	408	649	312

TABLE 10. Relevant information about each community for each promotion skewed community structure

The Leiden community structure is actually very similar to the Louvain community structure for the online skewed MInteraction similarity, and the results are generally the same. Leiden communities 3 and 4 are the customer communities that are more sensitive to online purchases as $\frac{14023}{17253} \approx 81.28\%$ and $\frac{3793}{7271} \approx 52.17$ of their products are bought online respectively. Customers in other Leiden communities buy 7% to 19% of all products in their respective communities online. Figure 27 shows the lift values for the Leiden community structure based on the online skewed MInteraction. As stated before, the Leiden community structure is very similar to the Louvain community structure, and this is also the case for the lift values. The lift values in figure 27 also confirm that Leiden communities 3 and 4 are more sensitive to online purchases as they are extremely high.

In conclusion, it seems that there is significant proof that each community detection algorithm provides a good community structure and thus we can assume that there is a well-defined community structure within the customer network when it is based on the online skewed MInteraction.

4.1.4. Brand skewed MInteraction similarity

Table 13 shows the best split for each community detection algorithm for the customer network based on the MInteraction similarity skewed to specific brand types. The modularity scores for each best split is also given in this table. Remarkably, the best split for the Louvain algorithm occurs at 50% of the total amount of heaviest edges left over in the network, which is higher than any community structure for any other MInteraction similarity.

Table 14 shows the nmuber of customers, products bought, and products bought per brand type in each community for each community structure for the brand skewed MInteraction similarity. Here, we will be using the same method of analysing which community is more sensitive to which brand as is used for the normal MInteraction in section 4.1.1. To reiterate, a community is said to be more sensitive to the MAR brand if more than 52.89% of all products bought, are from the MAR brand. A community is said to be more sensitive to the SUP brand if more than 23.46% of all products bought, are from the SUP brand. Finally, a community is said to be more sensitive to the SIN brand if more than 23.65% of all products bought, are from the SIN brand. Starting with the Eigenvector community structure, it seems that the 1st (55.62%) and 5th (62.72%) communities are sensitive to the MAR brand; the 1st (24.94%), 3rd (24.73%), and 4th (28.41%) communities are sensitive to the SUP brand; and the 2nd (26.40%), 3rd (25.28%), and 6th (27.01%) communities are sensitive to the SIN brand. For the Louvain community structure, it seems that the 2nd (53.89%) and 4th (61.56%) communities are sensitive to the MAR brand; the 2nd (25.98%), 3rd (24.23%), 5th (24.25%), and 6th (30.52%) communities are sensitive to the SUP brand; and the 1st (25.63%) and 5th (28.11%) communities are sensitive to the SIN brand. Finally for the Leiden community structure, it seems that the 2nd (53.56%), 3rd (54%), and 4th (61.96%) communities are sensitive to the MAR brand; the 2nd (26.21%), 3rd (23.57%), 5th (25.41%), and 6th (30.50%) communities are sensitive to the SUP brand; and the 1st (26.09%) and 5th (27.68%) communities are sensitive to the SIN brand. The lift values in figures 29, 31, and 33 confirm the observations made from table 14.

In conclusion, there is a well-defined community structure present in the customer network when it is based on the brand skewed MInteraction similarity.

Community Detection Algorithm	% edges	Communities	Modularity score
Eigenvector Algorithm	15	5	0.3199
Louvain Algorithm	25	5	0.2766
Leiden Algorithm	25	5	0.2771

TABLE 11. General information about each community structure for the MInteraction similarity skewed to online sales.

	Com 1	Com 2	Com 3	Com 4	Com 5
Eigenvector Customers	444	376	369	289	174
Eigenvector Products	23779	15276	8962	12101	7663
Eigenvector Online	4046	12832	913	2418	4438
Louvain Customers	486	468	425	163	110
Louvain Products	26451	10257	17860	7160	6053
Louvain Online	4676	742	14203	3903	1123
Leiden Customers	473	461	418	166	134
Leiden Products	25391	10349	17253	7271	7517
Leiden Online	4828	760	14023	3793	1243

TABLE 12. Relevant information about each community for each online skewed community structure

4.2. Which product categories are bought more frequently by promotion-sensitive customers?

4.2.1. Regular MInteraction

Starting with the results from the Eigenvector Algorithm, figure 6 in the appendix shows that communities 2 and 3 have a lift value greater than 1 for promotion purchases, indicating that the customers within these communities are more sensitive to promotions than other customers and thus mainly buy products that are in promotion. These communities are quite sizeable with 404 customers in the second community and 270 in the third community. This will make it easier to generalise the insights made from these communities to the whole retail data set population.

Figure 14 in the appendix shows the lift values for each product category bought in each Eigenvector community from the community structure in figure 6. There is a lot of information present in figure 14, although we can still conclude that product categories "FRUTAS Y HORTALIZAS", "PANADERIA", and "PESCADO" are certainly most frequently bought during promotions. Why this is the case, will be explained in the following paragraphs.

Lets start with the easiest result to interpret, namely the product category "PANADERIA". We can see in figure 14 that the lift value for the product category "PANADERIA" is greater than 1 in communities 2 and 3, the only customer communities that are sensitive to promotions in the Eigenvector community structure. The lift value in community 3 is especially high with a value of almost 1.75. More importantly, all the other customer communities (being 1, 4, 5, and 6) are less sensitive to promotions and the lift value of the product category "PANADERIA" in each of those communities is extremely low. The previous statements indicate that the product category "PANADERIA" is more frequently bought in customer communities that are more sensitive to promotions, than expected. Thus, products of the "PANADERIA" product category are more likely to be bought in promotion by the customers in the retail data set. The same can be said about two other product categories, namely "FRUTAS Y HORTALIZAS" and "PESCADO". While these product categories are more often bought than expected only in community 2 as their lift values are significantly greater than 1 in this community, their lift values are less than 1 in all of the customer communities that are less sensitive to promotions and thus are bought less often than expected by these kind of customers. Therefore, products of the "FRUTAS Y HORTALIZAS" and "PESCADO" product categories are also more likely to be bought in promotion. Another strong product category candidate is "CARNE CORTE". Like product category "PANADERIA", the lift values for "CARNE CORTE" in both the 2nd and 3rd communities exceed the threshold of 1. However, this is also the case for the 5th community, albeit with a smaller amount than in the 3rd community and arguably the 2nd community. Customers in the 5th community are not sensitive to promotions as is shown in figure 6 in the appendix, but this should not be a problem because of the previous statements and thus "CARNE CORTE" can be considered as another product category that promotion-sensitive customers tend to buy.

Figure 9 in the appendix shows the same style of graph as figure 6, however this time it shows the results for the Louvain community structure. The Louvain communities that seem to be more sensitive to promotions are communities 2 and 7, although the results of the 7th community can be ignored as the size of this community is only 29 customers, which is considered to be too low in this analysis. A community that is too small in size can cause irrelevant results for the analysis, which is explained in the following

Community Detection Algorithm	% edges	Communities	Modularity score
Eigenvector Algorithm	5	6	0.4303
Louvain Algorithm	50	6	0.1583
Leiden Algorithm	25	6	0.2759

TABLE 13. General information about each community structure for the MInteraction similarity skewed to purchases of specific brand types.

	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6
Eigenvector Customers	454	396	261	247	148	142
Eigenvector Products	11305	20928	12707	10033	6387	6383
Eigenvector MAR brand	6288	10873	6353	4963	4006	3336
Eigenvector SUP brand	2820	4531	3142	2850	1231	1323
Eigenvector SIN brand	2197	5524	3212	2220	1150	1724
Louvain Customers	546	427	320	149	112	99
Louvain Products	29774	9936	12050	6516	5748	3758
Louvain MAR brand	15577	5355	6387	4011	2738	1779
Louvain SUP brand	6567	2581	2920	1295	1394	1147
Louvain SIN brand	7630	2000	2743	1210	1616	832
Leiden Customers	482	425	376	152	119	98
Leiden Products	25985	9786	14373	6814	6485	4338
Leiden MAR brand	13544	5241	7760	4222	3042	2037
Leiden SUP brand	5662	2565	3387	1319	1648	1323
Leiden SIN brand	6779	1980	3226	1273	1795	978

TABLE 14. Relevant information about each community for each brand skewed community structure.

paragraph. The 2nd community comprises of 518 customers and is sizeable enough to generalise the conclusions made from this community to the entire retail data set population.

Figure 15 in the appendix shows the lift values for each product category bought in each Louvain community from the community structure in figure 9. In figure 15, we find the main reason why the 7th community is not being considered in the analysis. The lift values for the product categories in the 7th community are very erratic when comparing these to the other communities. A great example of this behaviour can be seen with the "PERFUMERIA" product category. Community 2 indicates that this product category is not bought as frequently as expected by customers that are sensitive to promotions as its lift value is a little higher than 0.8. This is not the case for the 7th community where the lift value for "PERFUMERIA" significantly exceeds the threshold of 1 with a value of around 1.4, already contradicting the result in the 2nd community. Moreover, the customers in the communities that are less sensitive to promotions also tend to buy products from the "PERFUMERIA" product category more than is expected as the lift value of "PERFUMERIA" in each of these communities exceeds 1. This also contradicts the observations made in the 7th community as customers in this community are said to be more sensitive to promotions, but also tend to buy products from the "PERFUMERIA" product category. There are more examples, but for these reasons, Louvain community 7 is exempt from the analysis. Now that this decision is explained, let's analyse each product category in the 2nd community that has a lift value greater than 1.

The product categories that are most frequently bought by customers who are more sensitive to promotions are "CARNE CORTE", "PANADERIA", and "PESCADO". These categories all have a lift value greater than 1 only for the 2nd community, while they have a lift value less than 1 in all the customer communities that seem to be less sensitive to promotions. Especially "PANADERIA" excels here with a lift value of almost 1.4. Another product category candidate could be "FRUTAS Y HORTALIZAS" as its lift value is also higher than 1 in the 2nd community, but it also barely reaches this threshold in the 6th community. However, the lift value in the 2nd community is larger than the one in the 6th community. Furthermore, the 2nd community is an overall larger community of customers than the 6th community. Thus, the product category "FRUTAS Y HORTALIZAS" can also be considered as a product category that is frequently bought by customers that are prone to buying products in promotion in the Louvain community structure.

Finally, figure 12 in the appendix shows the lift values for the promotion purchases for each Leiden community. This figure clearly shows that the 2nd community is the community of interest here as the lift value for customers that buy products during promotions remarkably exceeds the threshold of 1. Community 4 barely reaches this threshold, so the results from the Leiden community structure will mainly focus on the observations found in the 2nd community. However,

the observations from the 4th community can still be used to check or support the observations found in the 2nd community. All the other customer communities are considered to be not sensitive to promotions as is shown by the corresponding lift values.

Figure 16 in the appendix shows the lift values for each product category bought in each Leiden community from the community structure in figure 12. We can conclude from figure 16 that products from product categories "CARNE CORTE" and "PANADERIA" are bought most frequently by promotion-sensitive customers. Both product categories have a lift value greater than 1 in communities 2 and 4, while they do not exceed this threshold in all the other communities. The lift values for these two product categories are in fact relatively low in the other communities. Furthermore, the lift value for product category "PANADERIA" in the 4th community is extremely high compared to the lift values for the other product categories. Other product category candidates could be "CHARCUTERIA LIBRESERVICIO", "FRUTAS Y HORTALIZAS", "HUEVOS", and "PESCADO" as they all have a lift value higher than 1 in the 2nd community. However, some problems arise with these product categories in other communities. For example, product categories "FRUTAS Y HORTALIZAS" and "PESCADO" have a relatively high lift value compared to "CHARCUTERIA LIBRESERVICIO" and "HUEVOS" and their lift values are also less than 1 for the customer communities that are not sensitive to promotions. However, the latter is also the case for community 4, which is somewhat sensitive to promotions. So, it seems that here is a connection between customers that are sensitive to promotions and them buying products from product categories "FRUTAS Y HORTALIZAS" and "PESCADO", but not as strong as the connection between this kind of customers and product categories "PANADERIA" and "CARNE CORTE".

In conclusion, when the results from all community structures are combined, "PANADERIA" is the most frequently bought product category by promotion-sensitive customers. It has a strong connection with these kind of customers in all three community structures. Another product category that promotion-sensitive customers like to buy, is "CARNE CORTE", but this less so than "PANADERIA" as is shown by comparing the lift values of these two product categories with each other. Other possible product categories are "CHARCUTERIA LIBRESERVICIO", "FRUTAS Y HORTALIZAS", "HUEVOS", and "PESCADO". These product categories all show up with a decent lift value in promotion-sensitive communities across the three community structures, although they also tend to not have a lift value higher than 1 in a promotion-sensitive community or to have a lift value higher than 1 in communities that are not sensitive to promotions. So, the main focus here should be on product categories "PANADERIA" and "CARNE CORTE", while the previously mentioned other 4 product categories should be an afterthought.

4.2.2. MInteraction skewed to promotion

Figures 17, 19, and 21 in the appendix show the lift values for the Eigenvector, Louvain, and Leiden community structures based on the promotion skewed MInteraction similarity respectively. As previously shown in section 4.1.2, customers in the 1st community in all three the community structures and the 5th community in the Louvain and Leiden community structures are more sensitive to promotions as the lift values for these communities are higher than 1.

Starting with the results from the Eigenvector community structure, figure 18 in the appendix shows the lift values for each product category for each community in the Eigenvector community structure. It is immediately apparent that product category "PANADERIA" is most frequently bought by customers in the 1st community as the lift value for this product category in this community is well above 1 while the lift values in all the other communities are significantly lower than 1. It also seems that product category "CARNE CORTE" is another candidate for product categories that are well-liked by customers that tend to buy products in promotion. Its lift value is also higher than 1 in the 1st community, although this is also the case for the 3rd community. However, the lift values for "CARNE CORTE" in the other communities are rather low at values around 0.7 to 0.9 and the lift value for the 3rd community is a little bit higher than 1. For these reasons, we can conclude that promotion-sensitive customers tend to buy products from product category "CARNE CORTE". Finally for the Eigenvector community structure, we have two edge cases in the form of product categories "COMIDA PREPARADA" and "FRUTAS Y HORTALIZAS". The lift value of "COMIDA PERPARADA" for the 1st community barely reaches the threshold of 1 and it barely misses that threshold for the 5th community. Furhtermore, the lift values of "COMIDA PREPARADA" for the other communities are quite low at lift values around 0.7 to 0.9. Almost the same applies for product category "FRUTAS Y HORTALIZAS", its lift values are slightly higher than 1 in communities 1 and 3, however also quite low in the other communities at lift values around 0.9. Thus for the Eigenvector community structure, promotion-sensitive customers tend to buy products from the "PANADERIA" and "CARNE CORTE" product categories most certainly, while it is also possible that they buy products from the "COMIDA PREPARADA" and "FRUTAS Y HORTALIZAS" product categories.

Figure 20 in the appendix shows the lift values for each product category for each community in the Louvain community structure. Here, it is also immediately apparent that promotion-sensitive

customers like to buy products from product category "PANADERIA" as its only lift values that are higher than 1, are its lift values in the 1st and 5th communities. Product category "COMIDA PREPARADA" only has a lift value higher than 1 in the 5th community, although it barely misses that threshold in the 1st community. Its lift values in the other communities are also lower than 1, so we can still assume that promotion-sensitive customers tend to buy products from this product category. Product category "CARNE CORTE" actually suffers from the same problem as "COMIDA PREPARADA". Furthermore, its lift value in the 6th community is decently higher than 1. However its lift value for the 5th community is significantly high, thus we also view this product category as a product category that could be frequently bought by promotion-sensitive customers. Other product categories that could be frequently bought by promotion-sensitive customers are "CARNE LIBRESERVICIO", "CHARCUTERIA LIBRESERVICIO", "CHARCUTERIA CORTE", and "FRUTAS Y HORTALIZAS". Product category "CARNE LIBRESERVICIO" has a lift value slightly lower than 1 in the 1st and 2nd communities and slightly higher in the 3rd and 6th communities, however its lift value in the 5th community is significantly higher than 1 and significantly lower than 1 in the 4th community. "CHARCUTERIA LIBRESERVICIO" and "CHARCUTERIA CORTE" both have lift values greater than 1 in both communities 1 and 5, although the lift value in the 1st community is only slightly higher than 1. Furthermore, the lift value for "CHARCUTERIA LIBRESERVICIO" is also slightly greater than 1 in the 3rd community and the lift value for "CHARCUTERIA CORTE" is decently higher than 1 in the 4th community. In their respective other communities, the lift values are lower than 1, although the lift values for both product categories in the 2nd community are only slightly lower than 1. "FRUTAS Y HORTALIZAS" only has a lift value higher than 1 in the 1st community, however its lift value in the 5th community is significantly lower than 1 at a value around 0.75. Despite the lift value problems that the previous four product categories face, they are still included as product categories that a promotion-sensitive customer would buy. However, the product category that still stands on top for the Louvain community structure, is "PANADERIA". Product categories "CARNE CORTE" and "COMIDA PREPARADA" form good runner-ups.

Figure 22 in the appendix shows the lift values for each product category for each community in the Leiden community structure. In short, the results from the Leiden community structure are generally the same as the results from the Louvain community structure. This is not a strange sight because the Leiden algorithm is an improvement of the Louvain algorithm and thus is heavily based on the Louvain algorithm. So, the product category that promotion-sensitive customers most frequently buy from, is the "PANADERIA" product category. Its lift value is higher than 1 in both communities 1 and 5 and significantly lower in all the other communities. Especially the lift value in the 5th community is extremely high. In contradiction to the Louvain community structure results, only product category "CARNE CORTE" will be seen as a runner-up here. "CARNE CORTE" still barely misses the lift value threshold of 1 in communities 1 and 3 and barely reaches it in community 4. However, the lift value in the 5th community is significantly higher than 1. It is even higher than its lift value in the 5th community of the Louvain community structure. The lift values for "CARNE CORTE" for the 2nd and 4th Leiden communities is also significantly lower than 1. "COMIDA PREPARADA" has moved more to the background in the Leiden community structure and has taken more of a role that "CARNE LIBRESERVICIO", "CHARCUTERIA LIBRESERVICIO", "CHARCUTERIA CORTE", and "FRUTAS Y HORTALIZAS" take in the Louvain community structure. The lift value for "COMIDA PREPARADA" barely misses the threshold of 1 in the 1st community and it barely reaches it for the 5th and 6th communities. Although, it has lift values significantly lower than 1 in communities 2 and 4. Nonetheless, we still decide to put "COMIDA PREPARADA" more on the background here than we did for the Louvain community structure. Finally, the lift values for product categories "CARNE LIBRESERVICIO", "CHARCUTERIA LIBRESERVICIO", "CHARCUTERIA CORTE", and "FRUTAS Y HORTALIZAS" in the Leiden community structure are not too different from the corresponding lift values in the Louvain community structure, so no further attention will be given to these product categories in the Leiden community structure.

To conclude this section, all three community detection algorithms agree that product categories "PANADERIA" and "CARNE CORTE" are most frequently bought by promotion-sensitive customers. Especially "PANADERIA" is well-liked by this type of customers. Other product categories which promotion-sensitive customers could be interested in, are "COMIDA PREPARADA", "CARNE LIBRESERVICIO", "CHARCUTERIA LIBRESERVICIO", "CHARCUTERIA CORTE", and "FRUTAS Y HORTALIZAS". They did not perform as good as product categories "PANADERIA" and "CARNE CORTE" in terms of their lift values, although each one of these five product categories still gave some kind of indication that promotion-sensitive customers would buy these more than expected.

4.3. Which product categories are bought more frequently by online-sensitive customers?

4.3.1. Regular MInteraction

Figure 5 in the appendix shows the lift values of online purchases for each community in the Eigenvector community structure. This figure shows us that customers in Eigenvector communities 1, 4, 5, and 6 are more likely to buy products online than customers in the other Eigenvector communities as is indicated by the values of the blue bars. Figure 14 in the appendix shows the lift values of each product category for each Eigenvector community. This figure will help us with identifying which product categories are more frequently bought by online-sensitive customers. The product categories that are certainly bought more frequently by online-sensitive customers are "PERFUMERIA" and "DROGUERIA". Other product categories that could be argued about, are "BAZAR" and "LIQUIDOS". The lift values for "PERFUMERIA" are all greater than 1 in the four online-sensitive Eigenvector communities and less than 1 in the other two communities, although the lift value in the 1st community is barely higher than 1. Despite that, we can still assume that online-sensitive customers tend to buy more products from product category "PERFUMERIA". Product category "DROGUERIA" only has lift values higher than 1 in communities 1, 4, and 6. The lift value in community 1 is slighty higher than 1 while the lift values in communities 4 and 6 are quite sizeable at a value around 1.25. Unfortunately, the lift value of "DROGUERIA" is smaller than 1 in community 5, but only slightly so. However, we can also conclude that online-sensitive customers buy products from this product category as its lift values in the 4th and 6th community are decently sized and the lift values in the 2nd and 3rd community are quite low at a value around 0.8. The lift values for product category "BAZAR" are higher than 1 in communities 3, 4, 5, and 6. Especially the lift value for the 6th community is extremely high at a value of almost 2. Unfortunately, the lift value for community 3 is around 1.25, which is almost the same as the lift value in the 4th community. Furthermore, the lift value for the 5th community is on the smaller side at a value around 1.15 and the lift value of the 1st community is slightly lower than 1, although this is also the case for the 2nd community. So, one can argue about whether online-sensitive customers tend to buy more products from the "BAZAR" product category or not. The only decisive factor that is presented here, is that the lift value for this product category is extremely high in the 6th community. The lift values for "LIQUIDOS" are also odd as the only lift values that are higher than 1 are the ones in communities 3, 4, and 5, where they are slightly higher than 1. The lift value for community 6 barely misses the threshold of 1 and the one for community 1 is slightly lower than 1. The latter is also the case for the 2nd community. So, one could say that "LIQUIDOS" is also a product category in which online-sensitive customers are interested in as the lift values for the 4th and 5th community are decently sized while the lift values for the 1st and 6th community are not extremely low. However, a counterargument can be made based on the lift values in the 2nd and 3rd communities, which are quite close to 1. This discussion whether to include "BAZAR" and / or "LIQUIDOS" as product categories that are bought more frequently by online-sensitive customers or not, is left open to the reader.

Figure 8 in the appendix shows the lift values of online purchases for each community in the Louvain community structure. This figure shows us that customers in Louvain communities 1, 3, 4, 5, and 6 are more sensitive to online purchases than customers in the other Louvain communities as is shown by their lift values, which are all higher than 1. Figure 15 in the appendix shows the lift values of each product category for each Louvain community. After analysing figure 15, it can be said that the results from the Louvain community structure are generally the same as the results from the Eigenvector community structure. Here, the products categories that are most frequently bought by online-sensitive customers, are "PERFUMERIA", "DROGUERIA", and "LIQUIDOS". Product category "PERFUMERIA" is again an obvious choice as its only lift values that are higher than 1 are in all the online-sensitive communities. The lift value in the 1st community is barely higher than 1, however the other lift values are between around 1.15 and 1.40, still giving a strong indication that "PERFUMERIA" is well-liked by online-sensitive customers. The lift values of product category "DROGUERIA" that are higher than 1, are in communities 1, 3, and 5. The lift value in the 4th community is barely lower than 1 and the lift value in the 6th community is slightly lower than 1 with a value around 0.9. However, the lift values in communities 1, 3, and especially 5 are decently sized and the lift value of "DROGUERIA" in the 2nd community is rather low at a value around 0.8. For these reasons, we can still assume that online-sensitive customers tend to buy products from the "DROGUERIA" product category more than customers that are not sensitive to online purchases. Product category "LIQUIDOS" has lift values higher than 1 in communities 3, 4, 5, and 6, which all have values between around 1.2 and 1.5. However, the lift value in the 1st community is rather low at a value around 0.8 and the lift value in the 2nd community is slightly lower than 1. Despite the previous, we can conclude that "LIQUIDOS" is bought more frequently by online-sensitive customers due to the relatively high lift values in the 5th and 6th communities and the decently sized lift values in the 3rd and 4th communities. Finally, we will discuss

the lift values for product category "BAZAR" as the Eigenvector community structure did not provide a clear answer for whether online-sensitive customers are interested in this product category or not. The lift values for "BAZAR" are higher than 1 in the 1st, 3rd, and 6th communities at a value around 1.2. However, the lift value in the 5th community is almost 1 and this is also the case for the 2nd community, which is not sensitive to online sales. Furthermore, the lift value in the 4th community is rather low at a value around 0.7. Because of these shortcomings, it becomes more clear that online-sensitive customers do not buy products from the "BAZAR" product category more than customers who are not sensitive to online sales.

Figure 11 in het appendix shows the lift values of online purchases for each community in the Leiden community structure. This figure shows us that customers in Leiden communities 1, 3, 5, 6, and 7 are more sensitive to online purchases than customers in the other Leiden communities as is shown by their lift values. Figure 16 in the appendix shows the lift values of each product category for each Leiden community. Here, it can also be said that the results from the Leiden community structure are generally the same as the Eigenvector or Louvain community structure. Product category "PERFUMERIA" only has lift values greater than 1 in the online-sensitive communities, although the lift value for the 1st community is barely higher than 1. The other lift values seem to be decently sized at almost 1.25 voor communities 3, 5, and 6 and 1.5 for community 7. "DROGUERIA" has lift values higher than 1 in communities 1, 3, 6, and 7, however the lift value in the 7th community is only barely higher than 1. While the lift value of "DROGUERIA" in the 5th community is lower than 1, the same is also true for communities 2 and 4, which are not sensitive to online sales. The lift values in these two communities is even lower than the lift value in the 5th community. So, for these reasons, we still include product category "DROGUERIA" in our list of online-sensitive product categories. The lift values of product category "LIQUIDOS" are higher than 1 in communities 3, 4, 5, 6, and 7. Fortunately, the lift value in the 4th community, which is not sensitive to online sales, is only slightly higher than 1 while the lift values in communities 3, 5, and 6 are rather high at a value of almost 1.25 and the lift value in community 7 is extremely high at a value around 1.6. While the lift value for "LIQUIDOS" is quite low at a value around 0.85 in the 1st community, this is also the case for the 2nd community. Nonetheless, because of the decent lift values in communities 3, 5, and 6 and the extremely high lift value in community 7, we can assume that online-sensitive customers also tend to buy products from the "LIQUIDOS" product category. Finally, we will discuss the lift values for product categories "BAZAR" and "MASCOTAS" as they can be misleading at first glance. Product category "BAZAR" has some decent lift values in communities 1, 3, 6, and 7, although this is also the case in the 4th community where the lift value is actually quite high at a value around 1.4. Furthermore, the lift value for "BAZAR" in the 5th community is quite low at a value around 0.75 and its lift value in the 2nd community is also almost equal to 1. So, at first glance it looks like product category "BAZAR" is more frequently bought by online-sensitive customers, but upon further inspection it does not seem to be the case. Thus, we exclude this product category from our list of online-sensitive product categories. The same can be said about product category "MASCOTAS". It has a quite high lift value of approximately 1.4 in the 1st community and a good lift value of approximately 1.1 in the 6th community. However, it has quite low lift values of around 0.7 in communities 3 and 7, and an extremely low lift value of around 0.5 in the 5th community. Hence, we can assume that online-sensitive customers do not buy products from the "MASCOTAS" product category more frequently than other customers.

In conclusion, if we combine the results from each community detection algorithm, we see that customers who buy products online more frequently than other customers, do so from the "PERFUMERIA", "DROGUERIA", and "LIQUIDOS" product categories. However, there was a problem present during the analysis of the online purchases. The problem in this section was that most of the communities in each community structure had a lift value higher than 1 and thus were sensitive to online purchases. This made analysing the lift values of each product categories per community structure quite difficult as it is easier to find contradictions between online-sensitive communities. This should not be a problem in the following section were the online skewed MInteraction similarity is used to construct the customer network instead of the normal MInteraction similarity. The previous is already shown by figures 23, 25, and 27 in section 4.1.3.

4.3.2. MInteraction skewed to online

Figures 23, 25, and 27 in the appendix show the lift values for the Eigenvector, Louvain, and Leiden community structures based on the online skewed MInteraction similarity respectively. As previously shown in section 4.1.3, customers in the 2nd and 5th Eigenvector communities are more sensitive to online purchases than customers in the other Eigenvector communities. Customers in the 3rd and 4th Louvain and Leiden communities are more sensitive to online purchases than customers in the other Louvain and Leiden communities respectively.

Starting with the results from the Eigenvector community structure, figure 24 in the appendix shows the lift values for each product category for each community in the Eigenvector community structure. As was the case with product category

"PANADERIA" for the promotion-sensitive customers, it is immediately apparent in figure 24 that product category "DROGUERIA" is frequently bought by online-sensitive customers. Its lift values in the 2nd and 5th communities are significantly higher than 1 and the lift values in all the other communities are lower than 1. Product category "PERFUMERIA" is also a good candidate to consider. It has a decent lift value that is higher than 1 in the 2nd community and a lift value that is significantly higher than 1 in the 5th community. Although, it has almost the same lift value in the 3rd community as in the 2nd community, and the lift value in the 4th community is also barely higher than 1. Despite these shortcomings, we still chose to include this product category because of its high lift value in the 5th community. Product category "LIQUIDOS" is located in a similar situation to that of "PERFUMERIA". Its lift values for communities 2 and 5 are higher than 1, although the lift value in the 2nd community barely reaches this threshold while the lift value in the 5th community is almost 1.2. Furthermore, It also has a lift value slightly higher than 1 in the 1st community, however the lift values in all the other communities are quite low at values around 0.8. Finally, there are two other product category candidates that might be of interest at first glance, these are "CONGELADOS" and "MASCOTAS". However, we do not consider these product categories to be bought frequently by online-sensitive customers for the following reasons. Starting with "CONGELADOS", which only has a lift value higher than 1 in the 2nd community and a slightly low lift value of around 0.9 in the 5th community. Also, its lift value in the 1st community barely misses the threshold of 1. While an argument could be made that online-sensitive customers tend to buy products from the "CONGELADOS" product category more than other types of customers, it does not make any practical sense. Normally, online products are ordered online and shipped to the customer, so this system would not work with frozen food products. The only way that this would work is that the customer would reserve these frozen foods online and come pick them up later, although this also seems to be not practical. Thus, product category "CONGELADOS" is not considered to be bought frequently by online-sensitive customers for the previous reasons. "MASCOTAS" is not considered because of the inconsistent lift values across the different Eigenvector communities. It has quite a high lift value of around 1.45 in the 2nd community, although also quite a low lift value of 0.75 in the 5th community, which is the lowest lift value for this product category in this community structure. Furthermore, it has a decently sized lift value in the 1st community and a lift value that barely misses the threshold of 1 in the 3rd community. For these reasons, product category "MASCOTAS" is also not considered to be bought frequently by online-sensitive customers.

Figure 26 in the appendix shows the lift values for each product category for each community in the Louvain community structure. In general, the results of the Louvain community structure are almost the same as the results of the Eigenvector community structure. Here, product categories "DROGUERIA", "PERFUMERIA", and "LIQUIDOS" are the product categories for which we are certain that they are frequently bought by online-sensitive customers. Product category "DROGUERIA" is an obvious choice here as its only lift values that are higher than 1, are those in communities 3 and 4, which are the only communities that are sensitive to online purchases. "PERFUMERIA" only has lift values higher than 1 in communities 2, 3, and 4. Although, the lift values in communities 2 and especially 3 are small, the lift value in community 4 is relatively high at a value of 1.37. Because of this high lift value, we include it in this online-sensitive product category group. Again, product category "LIQUIDOS" is in a similar situation as "PERFUMERIA". It only has a lift value higher than 1 in communities 3, 4, and 5. While the lift values in communities 3 and 5 relatively lower, the lift value in community 4 is quite high at a value of almost 1.25. So, the relatively smaller lift value in community 3 and the lift value in community 5 that is higher than 1, do not matter as much anymore due to the high lift value in the 4th community. Thus, we also include this product category in our online-sensitive product category group. Finally, as with the Eigenvector community structure, there are some product categories that seem that they are being frequently bought by online-sensitive customers at first glance, but they are not when looked at in more detail. For example, product category "BAZAR" has a decently high lift value of 1.25 in the 4th community, which is a community where customers are more sensitive to online purchases. However, the lift value for the 3rd community is actually quite low at almost 0.8. Furthermore, "BAZAR" has a significantly high lift value of around 1.5 for the 2nd community, which is absolutely not an online-sensitive community as is shown by its lift value in figure 25. So, because of these reasons, "BAZAR" is seen as a product category that online-sensitive customers would not frequently buy. There are some other examples, such as product categories "MASCOTAS", "CONGELADOS", and "CHARCUTERIA CORTE", however these examples will not be covered extensively as they are similar to "BAZAR".

Figure 28 in the appendix shows the lift values for each product category for each community in the Leiden community structure. The results from the Leiden community structure are actually almost identical to the results from the Louvain community structure, so they will not be extensively explained in this section. The product categories that are certainly bought more frequently by online-sensitive customers are "DROGUERIA", "PERFUMERIA", and "LIQUIDOS" for the same reasons as provided in

the previous paragraph about the Louvain community structure. "DROGUERIA" only has lift values higher than 1 in online-sensitive communities 3 and 4. "PERFUMERIA" does so aswell, although the lift value in the 3rd community is only slightly greater than 1 and this is also the case for the 2nd community. The decisive factor is the large lift value of "PERFUMERIA" in the 4th community, which is around 1.4. Finally, "LIQUIDOS" almost has lift values greater than 1 only in online-sensitive communities 3 and 4, however the lift value in community 5 barely reaches the threshold of 1. Although, because of the higher lift values in communities 3 and 4 and the lift value in community 5 being very close to 1, the lift value in community 5 can be ignored.

4.4. Which product categories are bought more frequently by customers that are sensitive to a specific type of brand?

4.4.1. Regular MInteraction

Figure 7 in the appendix shows the lift values for each brand type for the Eigenvector community structure. According to the lift values, it seems that customers in community 4 and 6 like to buy products from the MAR brand more than from the other brands, customers in community 1 and 5 tend to buy products from the SUP brand than from the others, and finally customers in community 2 seem to buy products from the SIN brand more than from the others. The 3rd community is an odd one in this community structure. Customers in this community seem to be more sensitive to products from the SUP and SIN brand. This makes it harder to make observations from this community for a single brand type, although it can still be used to support or contradict observations made from other communities. Figure 14 in the appendix shows the lift values for each product category bought in each Eigenvector community. Starting with the MAR brand products, it seems that there is a quite strong connection with the "DROGUERIA" product category as it has a lift value greater than 1 in both community 4 and 6, but this is also the case for community 1 albeit not as much as in the other two communities. The lift values for "DROGUERIA" are all lower than 1 in the other communities, which are more sensitive to other brands. Other product categories that customers buy mostly from the MAR brand are "BAZAR" and "PERFUMERIA". "BAZAR" has an extremely high lift value of almost 2 in the 6th community and also a decently sized lift value of around 1.30 in the 4th community. Although the latter is also the case for the 3rd community, where MAR brand products are not bought as much as expected. Furthermore, "BAZAR" also has a lift value of around 1.15 in the 5th community, which is also not interested in the MAR brand. However, the previous two observations can be somewhat ignored because of the higher lift values in community 4 and especially community 6. Product category "PERFUMERIA" also scores quite well in communities 4 and 6 with a lift value of around 1.30 in both communities. It also has a lift value of around 1.15 in community 5 in which products of the SUP brand are bought more frequently. Although, to counteract the previous statement, the lift values for "PERFUMERIA" in communities 1 and 3, which are more sensitive to products of the SUP and / or SIN brand, are quite low. The lift value in community 1 barely reaches the threshold of 1 and the lift value in community 3 is not that spectacular at a value of around 0.75. This gives an extra indication that customers that mainly buy MAR products, buy products from the "PERFUMERIA" product category. Only two product categories seem to be interesting for customers that tend to buy more products from the SUP brand, these are "CONGELADOS" and "CARNE LIBRESERVICIO". Both product categories only have lift values greater than 1 in communities 1, 3, and 5, which are more sensitive to the SUP brand than the other communities. Their lift values in community 1 and 3 are not that high, although their lift values in community 5 make up for this. At first glance it also seems that the product category "MASCOTAS" is bought more frequently by customers that buy more products from the SUP brand. It has decently sized lift values of around 1.25 in communities 1 and 3. However, the lift value for "MASCOTAS" in community 5 is quite low at almost 0.70 and quite high at almost 1.50 in community 6. So, because of the contradicting observations for "MASCOTAS", we cannot conclude that the SUP brand products of this product category is bought more frequently than the other brands. Finally for the Eigenvector community structure, the product categories that are most frequently bought by customers that like to buy products from the SIN brand, are "FRUTAS Y HORTALIZAS", "PANADERIA" and "PESCADO". Let's start with explaining the most simple product category, namely "PANADERIA". This product category has a lift value higher than 1 for both community 2 and 3, especially the lift value in the 3rd community is quite large with a value of almost 1.75. As previously stated, customers in the 3rd community like to buy products from the SUP brand as well as the SIN brand, so it could be that the high lift value for "PANADERIA" here refers to the SUP brand, although this is not the case for the following reason. The lift values for "PANADERIA" in communities 1 and 5, which are only sensitive to the SUP brand, are quite low at a value around 0.75. So, it is likely that the high lift value for "PANADERIA" in the 3rd community refers to the SIN brand. Now, let's go over the other two suggested product categories, namely "FRUTAS Y HORTALIZAS" and "PESCADO". The lift value for these two product categories are, unfortunately, only larger than 1 for the 2nd community. So, an argument can be made that these two product categories are

not representative for the SIN brand, because their lift values in the 3rd community are lower than 1. However, these product categories are actually the best fit, next to "PANADERIA", for the SIN brand when compared to the other product categories in the 2nd community that have a lift value higher than 1. The other product categories either have a lift value that is too close to 1, and thus lower than the lift values for "FRUTAS Y HORTALIZAS" and "PESCADO", or they also have a relatively high lift value in other communities that are not sensitive to the SIN brand.

Figure 10 in the appendix shows the lift values for each brand type for the Louvain community structure. The lift values for each brand type in each community here show that customers in communities 3, 6 and 7 seem to buy products from the MAR brand more than from the other brands, although the results from community 7 will not be ignored as the community size at 29 customers is too small to be considered a decent community. Customers in communities 1, 4, and 5 are more likely to buy products from the SUP brand. However, the lift value for the SUP brand in the 1st community is barely any higher than the threshold of 1. Furthermore, the lift value for the MAR brand in that community also hovers around the threshold of 1. So, for these reasons, the 1st community will be used as a support for observations made in communities 4 and 5. Finally, customers in community 2 tend to buy products from the SIN brand. Figure 15 in the appendix shows the lift values for each product category bought in each Louvain community and will be used as a basis for the observations made in the following paragraphs. First, The "BAZAR" and "CHARCUTERIA CORTE" product categories seem to be bought more frequently by customers that are more sensitive to the MAR brand. Starting with "BAZAR", it has a decent lift value in community 3 and even more so in community 6, but also in community 1. However, the 1st Louvain community is treated more as a support for SUP brand observations made in other communities and it also has a lift value of around 1 for MAR brand products. So, we will still regard "BAZAR" as a product category that is bought by customers who are sensitive to the MAR brand type. The product category "CHARTUCTERIA CORTE" also has a decently high lift value in the 3rd community, although it barely does not reach the threshold of 1 in the 6th community. Furthermore, the lift value in community 4 is also really close to reaching 1, while this community is not sensitive to the MAR brand type. Despite these problems, we still opt to include "CHARCUTERIA CORTE" in the MAR brand group, mainly because of the lower lift value in the 5th community. This indicates that the customers that are more sensitive to SUP brand products do not purchase products from the "CHARCUTERIA CORTE" product category as frequently as expected and it also decreases the impact from the observation made in the 4th community. The astute reader may have already noticed in figure 15 that product categories "DROGUERIA", "LIQUIDOS", and "PERFUMERIA" are also bought quite frequently by customers that are more sensitive to the MAR brand type. However, there are some other observations in figure 15 that contradict the previous statement. This will be explained in the following paragraph. Second, it is certain that product category "CONGELADOS" is more likely to be bought by customers that are more sensitive to the SUP brand. It has a decently sized lift value in the 5th community and a somewhat lower lift value in community 1 and 4, but still above the threshold of 1. The lift values of "CONGELADOS" is also lower than 1 in all other communities. As mentioned in the previous paragraph, determining in which brand type the product categories "DROGUERIA", "LIQUIDOS", and "PERFUMERIA" belong, seems to be a non-trivial task. For example, the lift values for "LIQUIDOS" in the 3rd and 6th community are relatively high, especially in the 6th community. However, this is also the case for communities 4 and 5. So, the conclusion that can be made from these observations is that customers who are more sensitive to the MAR and SUP brand tend to buy products from the "LIQUIDOS" product category more frequently. This forms a small problem, because there is not a single brand type on which the retailer could focus with the products in the "LIQUIDOS" product category. Although, we can still conclude that it does not have any connection to the SIN brand, which can be of some help. The same observations and conclusions also apply to the "DROGUERIA" and "PERFUMERIA" product categories. Although the lift value for "DROGUERIA" in community 4 and 6 is lower than 1 and higher than 1 in community 1, but the concept is still the same. Finally for the Louvain community structure, product categories "CARNE CORTE", "FRUTAS Y HORTALIZAS", "PANADERIA", and "PESCADO" are more likely to be bought by customers that are more sensitive to the SIN brand than to the other brands. Although, we can be more certain for product category "PANADERIA" than the other three product categories. Each of the previous four product categories have a lift value higher than 1 in community 2, but only "PANADERIA" has low lift values in all of the other communities. "CARNE CORTE" for example, has two lift values that almost reach the threshold of 1 in communities 1 and 4, which are not sensitive to the SIN brand. "PESCADO" has the same problem for communities 4 and 6 as well as "FRUTAS Y HORTALIZAS" for communities 1, 4, and 6. The lift value for "FRUTAS Y HORTALIZAS" in community 6 actually just reaches over 1. Due to these reasons, an argument could be made that these three product categories do not have any connection with the SIN brand. Despite these reasons, we still opted to include these product categories in this analysis.

Figure 13 in the appendix shows the lift values for each brand type for the Leiden community structure. The lift values in this figure indicate that customers in communities 3 and 7 are more likely to buy products from the MAR brand. Customers in communities 1, 4, 5, and 6 are more likely to buy products from the SUP brand. Finally, customers in communities 2 and 4 are tend to buy more products from the SIN brand. As was the case with the previous community structures, some communities will be used to support observations made in other communities. Here, this is the case for communities 1 and 4. Community 1 will serve as a support for observations made regarding the SUP brand due to the relatively low lift value for the SUP brand in this community when compared to the lift values for the SUP brand in communities 5 and 6. Community 4 will serve as a support for both the SUP and SIN brand as both their lift values are higher than 1 in this community. Figure 16 in the appendix shows the lift values for each product category bought in each Leiden community and will be used as a basis for the observations made in the following paragraphs. Starting with product category "PERFUMERIA", it has lift values greater than 1 in communities 3, 5, 6, and 7. This is somewhat of a problem because customers in the 3rd and 7th communities buy more products from the MAR brand while customers in the 5th and 6th communities buy more products from the SUP brand. However, the lift values for "PERFUMERIA" in the MAR-sensitive communites are higher than the lift values in the SUP-sensitive communities, so an argument can be made that only MAR-sensitive customers buy products from product category "PERFUMERIA" more frequently or both MAR-sensitive and SUP-sensitive customers buy products from product category "PERFUMERIA" more frequently. However, it is clear that SIN-sensitive customers do not buy "PERFUMERIA" more frequently because its lift values in SIN-sensitive communities are all lower than 1. What was said for the "PERFUMERIA" product category, can also be said for the "LIQUIDOS" product category. Finally, product categories "FRUTAS Y HORTALIZAS", "PANADERIA", and "PESCADO" are bought more frequently by customers that are sensitive to the SIN brand. The lift values for product categories "FRUTAS Y HORTALIZAS" and "PESCADO" are only higher than 1 in the 2nd community, which is only sensitive to the SIN brand. The lift values for product category "PANADERIA" are higher than 1 in communities 2 and 4, especially the lift value in the 4th community is extremely high. So, at first glance it seems that product category "PANADERIA" is bought by customers that are sensitive to the SUP and SIN brand. However, the lift values for product category "PANADERIA" are quite low or extremely low for all the other product categories, which are more sensitive to the MAR brand or the SUP brand. So, we can safely assume that SIN-sensitive customers buy products from the product category "PANADERIA" more frequently than other customers.

4.4.2. MInteraction skewed to specific types of brands

Figure 29 in the appendix shows the lift values for each brand type for the Eigenvector community structure based on the brand skewed MInteraction. According to the lift values, it seems that customers in communities 1 and 5 tend to buy more products from the MAR brand, customers in communities 1, 3, and 4 are more likely to buy products from the SUP brand, and customers in communities 2, 3, and 6 are more sensitive to the SIN brand. Figure 30 in the appendix shows the lift values for each product category bought in each Eigenvector community based on the brand skewed MInteraction. It is apparent from this community structure that MAR-sensitive customers tend to buy products from product categories "PERFUMERIA" and "DROGUERIA", SUP-sensitive customers are more likely to buy products from the product category "CONGELADOS", and SIN-sensitive customers buy more products from product categories "FRUTAS Y HORTALIZAS" and "PESCADO". The lift values for "PERFUMERIA" and "DROGUERIA" are higher than 1 in communities 1, 4, and 5. While community 1 is sensitive to both the MAR and SUP brand and community 4 is very sensitive to the SUP brand, the lift values for "PERFUMERIA" and "DROGUERIA" are the highest in the 5th community, which is very sensitive to the MAR brand. They are the second highest in community 1 and the third highest in community 4. Based on these observations, we can safely assume that MAR-sensitive customers tend to buy more products from the "PERFUMERIA" and "DROGUERIA" product categories than other customers. Product category "CONGELADOS" only has lift values higher than 1 in SUP-sensitive communities, so it is obvious that SUP-sensitive customers buy products from this product category more often than other customers. Finally for the Eigenvector community structure, the lift values for product categories "FRUTAS Y HORTALIZAS" and "PESCADO" are also only higher than 1 in SIN-sensitive communities, so here it is also obvious that SIN-sensitive customers more frequently buy products from these product categories than other customers.

Figure 31 in the appendix shows the lift values for each brand type for the Louvain community structure based on the brand skewed MInteraction. According to the lift values, it seems that customers in communities 2 and 4 are more likely to buy products from the MAR brand, customers in communities 2, 3, and 6 are more sensitive to the SUP brand, and customers in communities 1 and 5 tend to buy more products from the SIN brand. Figure 32 in the appendix shows the lift values for each product

category bought in each Louvain community based on the brand skewed MInteraction. According to the lift values, the MAR-sensitive customers buy more products from the "PERFUMERIA" product category, SUP-sensitive customers buy more products from the "PERFUMERIA" and "CONGELADOS" product categories, and SIN-sensitive customers are more likely to buy products from the "PANADERIA" product category. The lift values for "PERFUMERIA" are only higher than 1 in communities 2, 4, and 6, where community 2 is purely MAR-sensitive, community 6 is purely SUP-sensitive, and community 4 is sensitive to both brands, although more to the SUP brand. Product category "PERFUMERIA" has the a rather high lift value in the 4th community and decently sized lift values in the other two communities. So, while one can argue that SUP-sensitive customers also buy more products from "PERFUMERIA", here we chose to view "PERFUMERIA" as a product category that is frequently bought by MAR-sensitive customers. Product category "CONGELADOS" has lift values higher than 1 in communities 2, 5, and 6 which are all quite sensitive to the SUP-brand. So we can assume that SUP-brand customers tend to buy more products from the "CONGELADOS" product category. Finally for the Louvain community structure, the lift values for "PANADERIA" are only higher than 1 in communities 1 and 4, which both are quite sensitive to the SIN-brand. So, it is safe to assume that SIN-sensitive customers are more likely to buy products from the "PANADERIA" product category.

Figure 33 in the appendix shows the lift values for each brand type for the Leiden community structure based on the brand skewed MInteraction. According to the lift values, it seems that customers in the 4th community tend to buy more products from the MAR brand, customers in communities 2, 5, and 6 are more likely to buy products from the SUP brand, and customers in communities 1 and 5 are more sensitive to the SIN brand. Figure 34 in the appendix shows the lift values for each product category bought in each Leiden community based on the brand skewed MInteraction. It seems that the MAR-sensitive customers buy more products from the "PERFUMERIA" and "DROGUERIA" product categories, SUP-sensitive customers tend to buy more products from the "CONGELADOS" and (arguably) "PERFUMERIA" product categories, and SIN-sensitive customers buy more products from the "PANADERIA" and (arguably) "FRUTAS Y HORTALIZAS" product categories. The lift values for product category "PERFUMERIA" are higher than 1 in communities 2, 3, 4, and 6, although the lift value in the 3rd community is only barely higher than 1. The lift value for "PERFUMERIA" is the highest in the 4th community, which is a community which is purely sensitive to the MAR brand, however its lift values in communities 2 and 6 are also decently sized, where community 6 is purely sensitive to the SUP brand and community 2 almost so. So, an argument could be made whether only MAR-sensitive customers buy more products from the "PERFUMERIA" product category or both MAR- and SUP-sensitive customers do so. Because of the relatively high lift value in the 4th community, this research will assume that only MAR-sensitive customers tend to buy more products from the "PERFUMERIA" product category. The lift values for the "DROGUERIA" product category are only higher for communities 2 and 4. The 2nd community is a SUP-sensitive community and the 4th community is a MAR-sensitive community, however the lift values for "DROGUERIA" in other SUP-sensitive communities are rather low. Thus, we can assume that only MAR-sensitive customers tend to buy more products from this product category. The lift values for product category "CONGELADOS" are only higher than 1 in SUP-sensitive communities, so it is obvious that only SUP-sensitive customers buy more products from this product category. The same is true for the "PANADERIA" product category when it comes to SIN-sensitive communities. Finally, product category "FRUTAS Y HORTALIZAS" only has one lift value higher than 1 and this is in the 1st community, which is a purely SIN-sensitive community. Although, the lift value for "FRUTAS Y HORTALIZAS" in the 5th community is quite low, while this is also a heavy SUP- and SIN-sensitive community. So, an argument could be made whether SIN-sensitive customers buy more products from the "FRUTAS Y HORTALIZAS" product category or not. Here, we assume that this is the case.

5. DISCUSSION

This section discusses the results found in the previous sections and offers some actionable insights that the retailer could possibly use.

Let's start with the first research question, which checks whether there is a community structure in the different customer networks for each similarity, because this forms the basis for the other three research questions. Each community detection algorithm that was executed could provide a good community structure for each MInteraction similarity that was used to construct the customer network as is shown in section 4.1. Although, the communities in the community structures that are based on the skewed MInteraction similarities are more concentrated for the concept that they are based on. For example, most of the communities for all three community detection algorithms performed on the normal MInteraction based customer network, are sensitive to online purchases and had relatively high lift values of around 1.2 or 1.4. This was not the case for the online skewed MInteraction where there are only a few communities that were mostly populated by online

purchases. These communities also had lift values of 1.5 or even 2, which is rather high. Thus, it was easier to analyse the results for the online skewed MInteraction based customer network. Nevertheless, Each customer network exhibited some type of community structure as similar customers could be grouped together rather easily. This can be a huge advantage for the Spanish retailer, whose retail data set was used in this research, as now there is an opportunity to observe customer purchasing behaviours based on specific communities of customers and generalise these purchasing behaviours to the entire retailer's population. These generalised results can then be used to create specific marketing strategies in order to capture more customer value and increase profits in the long-run. The three research questions following the first research question in section 4.1 were constructed to help the retailer in this regard. The results of these research questions are discussed in the following paragraphs.

The second research question was based on identifying customer purchasing behaviours when focusing on promotional purchases. The results from the normal MInteraction similarity suggest that promotion-sensitive customers tend to buy products from the "PANADERIA", "CARNE CORTE", "PESCADO", "FRUTAS Y HORTALIZAS", "CHARCUTERIA LIBRESERVICIO" and "HUEVOS" product categories more frequently than customers that are not sensitive to promotions. Especially product categories "PANADERIA" and "CARNE CORTE" constantly had good results, thus these product categories should be more focused on while the others are more of an afterthought. The results from the skewed online MInteraction similarity suggest that promotion-sensitive customers are more likely to buy products from the "PANADERIA", "CARNE CORTE", "COMIDA PREPARADA", "CARNE LIBRESERVICIO", "CHARCUTERIA LIBRESERVICIO", "CHARCUTERIA CORTE" and "FRUTAS Y HORTALIZAS" product categories. As with the normal MInteraction, product categories "PANADERIA" and "CARNE CORTE" had consistently good results and thus should be more focused on than the other product categories. The results from the normal MInteraction similarity as well as the online skewed MInteraction are all food types and this is rather logical because the different methods of promoting these types of goods are quite easy. For example, the "buy one get one free"-promotion is often used for foods or their price could be marked down when their expiration dates are approaching and then those products could be promoted in order to still generate some sales instead of dumping the product. So, how can we best describe the promotion-sensitive customer in order to help with targeted advertising? The retail data set did not contain any descriptive customer data, however we could still use the RFM-scores for each customer here. Figures 36 and 37 show the distribution of the RFM-scores for promotion-sensitive customers based on the normal MInteraction and promotion skewed MInteraction respectively. We can conclude from these figures that promotion-sensitive customers can be identified by a relatively low monetary score of 2 and relatively high recency and frequency scores of 4 and 5. To support this claim, figure 35 shows the distribution of the RFM-scores for all customers in the retail data set. Here, we can see that each score is evenly distributed with the exception of a value of 3 for the recency score. This shows that the most frequent monetary score of promotion-sensitive customers, namely 2, is not due to the presence of more monetary score 2 values than any other monetary score values.

The third research question analysed the customer purchasing behaviours of online purchases. Surprisingly, the results from both the normal MInteraction and online skewed MInteraction are very similar to each other. This was unexpected as the majority of communities in each online community structure based on the normal MInteraction was sensitive to online purchases while this was not the case for each online community structure based on the online skewed MInteraction. This was somewhat of a problem as having more online-sensitive communities to analyse per community structure, makes it easier to find contradicting observations or irrelevant results for the online customer purchasing behaviours. Fortunately, this was not the case and the results were very similar. We found that customers who tend to buy products online more frequently than other customers, are mainly interested in products from the "PERFUMERIA", "DROGUERIA", and "LIQUIDOS" product categories as they had consistently high lift values across all the online community structures. Whether or not these results are logical or not, is hard to say. Buying products online from the product category "PERFUMERIA" seems the most logical choice here as this product category is mostly populated by luxury products that can be used as a gift to someone else for example. It seems quite logical to buy a perfume online as a quick gift for someone or for personal use because buying such a product could be easily forgotten while doing the regular shopping in a supermarket. Furthermore, because most people do not think it is worth it to go back to the supermarket for a single item, they tend to buy it online and let it be delivered to their homes. However, this seems to be more illogical and difficult to do for the "DROGUERIA" and "LIQUIDOS" product categories. So, we cannot help the retailer with these two product categories and the results of these should be taken with a grain of salt. As is already mentioned in the previous paragraph, there was no descriptive customer data available, although the RFM-scores for each customer and in which supermarkets they did their shopping were available. So we try to identify the online-sensitive customers based on these variables. Figures 38 and 39 show the distribution of the RFM-scores for online-sensitive customers based on the normal MInteraction and online skewed

MInteraction respectivley. Both graphs do not show immense deviations from the norm in figure 35 as was the case with promotion-sensitive customers in the previous paragraph. However, we can still conclude from both figures 38 and 39 that online-sensitive customers can be identified by relatively high monetary scores of 3, 4, and 5, and relatively low recency and frequency scores of 2. What can be more useful to know, is in which supermarkets the most transactions are made by online-sensitive customers. The products in these transactions are more likely to be bought online, thus it could be important for the retailer to know in which supermarkets these transactions will be made so that these supermarkets can optimise the process of handling online purchases. Figures 40 and 41 show the percentages of the transactions in the online-sensitive communities that are bought by each supermarket and this for the normal MInteraction and online skewed MInteraction respectively. Only the 10 highest percentages are shown in both figures. We can conclude from both graphs that the most online-sensitive transactions are made in the supermarkets with IDs 315, 323, 582, 656, and 661 as their percentages are vastly greater than the other 5 supermarkets that are shown in both graphs and thus also the other supermarkets that were not shown in these graphs. Finally, we have a recommendation for the retailer. It is possible to create a recommendation system for online purchases per product category. This will potentially increase the online sales, and thus profits made, because the online-sensitive customers will be more exposed to similar products that they have already bought and might be interested in.

The final research question focuses on the customer purchasing behaviours of specific brand types. For the normal MInteraction similarity measure, we concluded that customers who tend to buy more products from the MAR brand, mainly buy products from the "PERFUMERIA", "DROGUERIA", "LIQUIDOS", and "BAZAR" product categories. Customers who are more likely to buy products from the SUP brand, do so for the "CONGELADOS", "PERFUMERIA", and "LIQUIDOS" product categories. Finally for the normal MInteraction similarity measure, customers who are sensitive to the SIN brand are more likely to buy products from the "PANADERIA", "FRUTAS Y HORTALIZAS", and "PESCADO" product categories. For the brand skewed MInteraction similarity measure, it seems that customers who buy more products from the MAR brand, buy products from the "PERFUMERIA" and "DROGUERIA" product categories. Customers who tend to buy more products from the SUP brand, mainly buy products from the "CONGELADOS" and "PERFUMERIA" product categories. Finally for the brand skewed MInteraction similarity measure, customers who are sensitive to the SIN brand are more likely to buy products from the "PANADERIA" and "FRUTAS Y HORTALIZAS" product categories. As is mentioned before in section 4.4, product categories "PERFUMERIA" and "LIQUIDOS" are in an odd position as they tend to be bought by both MAR- and SUP-sensitive customers. So, we fail to make a clear distinction between the MAR and SUP brand for those product categories. However, we do know that they are not bought frequently by SIN-sensitive customers, which could still help the retailer. Unfortunately, there is not that much information available to check the validity of these results. However, we know from the retailer that the SIN brand mostly consists out of white label products that tend to be cheaper than the other brands. The product categories that were most frequently bought by SIN-sensitive customers seem to be "PANADERIA", "FRUTAS Y HORTALIZAS", and "PESCADO", which are actually product categories that were frequently present in promotion-sensitive communities in the second research question. Because promotion-sensitive product categories are more frequently bought by SIN-sensitive customers, it can be said that our results for the different brand types are somewhat realistic. Figures 42 and 43 show the distribution of the RFM-scores for customers that are sensitive to a specific brand type based on the normal MInteraction and online skewed MInteraction respectively. As was seen previously in section 4.4, some customer communities are sensitive to more than 1 specific brand type. These communities are ignored in the RFM-scores analysis to get a more crisp view of the customers that are sensitive to a specific type of brand. Although, this can cause a disparity in the counts between the brand types in figures 42 and 43. This is especially the case for figure 43 where the MAR and SUP brand counts are far lower than the counts for the SIN brand. This is because the size of the communities for the MAR and SUP brand are smaller than those for the SIN brand. We can conclude from figures 42 and 43 that MAR-sensitive customers have a relatively high monetary and recency score of 4 and 5, and a relatively low frequency score of 2. SIN-sensitive customers can be identified by a relatively low monetary score of 1 or 2 and a relatively high recency and frequency score of 4 or 5. Unfortunately, identifying the SUP-sensitive customers is not as easy as the other two brand types. On the one hand, figure 42 identifies SUP-sensitive customers as customers with monetary score 1, 2, and 4, and recency and frequency scores 2 and 4. On the other hand, figure 43 identifies SUP-sensitive customers as customers with relatively low monetary scores of 1 and 2, and relatively high recency and frequency scores of 4. So, it seems that SUP-sensitive customers can be identified by relatively low monetary scores and relatively high recency and frequency scores. However, some data in figure 42 needs to be ignored to come to this conclusion, so an argument can be made that SUP-sensitive customers cannot be identified with the given RFM-scores in figures 42 and 43.

6. CONCLUSION

This research recommended an extra method of executing market basket analysis on a retail data set, namely community detection. Community detection has been used before in market basket analysis, although mainly on product networks or customer-product networks [1, 3, 8, 9, 14, 36]. Unfortunately, not many researches exist where community detection is used on customer networks in a retail data set context, some examples can be found in the following studies: [9, 14, 15]. Although, community detection has been performed successfully on social networks, which the customer network essentially is. Javed et al. [20] list a few examples in their study such as social media networks and e-mail networks. Applying community detection on a customer network could give vast amounts of information on the customer purchasing behaviours within that retail data set, which can be translated to more enhanced marketing strategies to increase the profit of the retailer. A few examples of this, were given in this research in the form of the last three research questions. This is not the case for applying community detection on product and customer-product networks. Here, only the products that are most frequently bought together, are analysed, somewhat ignoring the purchasing behaviours of specific types of customers.

There is an abundance of community detection algorithms readily available in the literature. This research used three community detection algorithms, this being the Eigenvector, Louvain, and Leiden community. These algorithms were chosen because they are a few of the recommended algorithms in the literature and were able to split the retail data set into decent community structures.

There were four main research questions in this paper. The first question was to identify if there was a community structure present in the customer network based on various customer similarities. This seemed to be the case for all the introduced customer similarities, so we can safely conclude that customers in this retail data set can be grouped into communities based on their purchasing behaviours.

The second question identified which customer communities bought more products in promotion than other customer communities and determined which product categories were bought more frequently by these customers. The answer is that product categories "PANADERIA" and "CARNE CORTE" are most frequently bought by customers that are promotion-sensitive. These product categories translate to "bread" and "meat over the counter" respectively. Promotion-sensitive customers are identified by a relatively low monetary score and relatively high recency and frequency score.

The third question identified which customer communities bought more products online than other customer communities and determined which product categories were bought more frequently by these customers. It seems that product categories "PERFUMERIA", "DOGUERIA", and "LIQUIDOS" are bought more frequently by customers that tend to buy more products online. These product categories translate to "personal care", "household cleaning products", and "liquids" respectively. Online-sensitive customers are identified by a relatively high monetary score and a relatively low recency and frequency score.

The last question identified which customer communities bought more products of a specific brand type than other customer communities and determined which product categories were bought more frequently by these customers. The different brand types are the MAR, SUP, and SIN brand. We concluded that product categories "PERFUMERIA", "DROGUERIA", "LIQUIDOS", and "BAZAR" are most frequently bought by customers that tend to buy more products from the MAR brand. These product categories translate to "personal care", "household cleaning products", "liquids", and "non food-durable products" respectively. Product categories "CONGELADOS", "PERFUMERIA", and "LIQUIDOS" are bought more frequently by customers that are more sensitive to the SUP brand. These product categories translate to "frozen foods", "personal care", and "liquids" respectively. Finally, product categories "PANADERIA", "FRUTAS Y HORTALIZAS", and "PESCADO" are bought more frequently by customers that tend to buy more products from the SIN brand. These product categories translate to "bread", "fruits and vegetables", and "fish". Unfortunately, no clear distinction could be made for the "PERFUMERIA" and "LIQUIDOS" product categories when it comes to the MAR and SUP brand types. Although, the lift values for these two product categories are higher for the MAR brand than for the SUP brand. So, one could argue that product categories "PERFUMERIA" and "LIQUIDOS" are only bought by customers that are sensitive to the MAR brand. However, we did not chose that option in this study. MAR-sensitive customers are identified by a relatively high monetary and recency score and a relatively low frequency score. SUP-sensitive customers could not be identified that well, although they seem to have relatively low monetary scores and relatively high recency and frequency scores. Finally, SIN-sensitive customers are also identified by a relatively low monetary score and relatively high recency and frequency scores.

Our recommendations for further research on this retail data set mainly encompass the usage of extra similarity measures between the customers as this is the basis of the community detection algorithms. With other normal or skewed similarity measures between customers, more hidden information can be gathered about communities of customers and, eventually, more marketing strategies can be constructed to increase the profit of the retailer. Another recommendation is to

use associations rules in combination with community detection on retail data sets. This could be done in two ways. First, one could perform a community detection algorithm on a product network to identify product categories that are frequently bought together. Then, use frequent item sets from association rules to get a more detailed insight on products that are frequently bought together. Second, one could perform a community detection algorithm on a customer network to group similar customers together. Then, use frequent item sets to identify which products are bought frequently together in each community. This way, a recommendation system could be constructed for example.

7. CHALLENGES

This research also introduced a handful of challenges next to the interesting observations made from the retail data set. To conclude this paper, this section enumerates and explains the most difficult challenges to keep in mind for further researches.

First, there is no real standard of pruning customer network. Customer networks have a tendency to be overpopulated with weak edges between nodes. These weak edges are a burden in network analysis as they signify that there is a very weak similarity between two customers, which most researchers are not interested in. So, they are normally removed from the network as these weak edges could have a major negative effect on the results of the network analysis. Unfortunately, the literature does not suggest a real standard of removing these weak edges as is mentioned in section 2.3. It seems that each researcher is encouraged to create his or her own method of pruning their networks, probably because of the enormous differences between each type of network.

Second, evaluating a community structure is a non-trivial, although an essential task in the community detection research field. Community detection algorithms maximize a quality function to split a network into partitions, or communities, thus forming a community structure. So, researchers logically assume that the community structure with the highest value for that quality function, would be the best community structure. However, this is not always the case as is shown in this paper. Here, the main focus was on modularity optimizing community detection algorithms and it is shown that removing the weaker edges in a graph and / or subgraph artificially increases the modularity score for each community detection algorithm. So, in theory a subgraph containing only the 1% heaviest edges of the complete network would be the best basis for a community structure as it yields the best modularity score. However, such a subgraph would have little practical implications for two main reasons. First, the more edges are removed in the network, the more chance there is for nodes to become disconnected from the resulting subgraph. These disconnected nodes are then removed from the subgraph in order to perform community detection algorithms on that subgraph. In the extreme case, a vast amount of data tied to those removed disconnected nodes, will also be removed from the subgraph. This will have an impact on the created communities, which will now not accurately represent reality and thus the inferences made from these communities will not hold in practice. Second, the community structure resulting from an extremely small subgraph may have a high modularity score, but also a bad spread of nodes between all the communities. There could be extremely large as well as extremely small communities in this structure. As is already explained in section 2.5, this is not a desirable effect because the observations made from extremely small communities cannot be generalised to the entire data set, resulting in a useless community.

Third, descriptive customer data was not available for this research. Some examples of descriptive customer data include gender and age. These characteristics of customers can help with targeting specific customer groups in advertising product categories. An interesting example could be to identify whether younger people buy more products online and, if so, which product categories they are more interested in. However, while no vast amounts of descriptive customer data was available, the customers' RFM-scores were available in a later stage of this research, which can help in developing marketing strategies.

Fourth, analysing which product categories are frequently bought online was difficult for each community structure while using the normal MInteraction similarity between customers. Each community structure had 4 or 5 communities out of the 6 or 7 that were quite sensitive to online sales, which makes it difficult to find similarities in the purchasing behaviour of these 4 or 5 communities. However, this challenge was overcome by introducing the MInteraction that is skewed to online purchases between customers. Executing the community detection algorithms on the customer network based on this skewed similarity caused each community structure to only have 2 out of 5 communities that were sensitive to online sales, making it much easier to analyse the purchasing behaviours.

Finally, the implementation of the Infomap algorithm was more challenging than expected, hence the absence of the algorithm in this research. The "CDlib" Python package has a relatively easy method of implementing the Infomap algorithm, you only need to use one single function to execute the Infomap algorithm on a network. However, a problem kept persisting where the algorithm did not work. The problem was that the Python interpreter could not find the dependencies on which the Infomap algorithm was built, even after installing them. The main problem here was the "Wurlitzer" Python package, which translates C-level output into something that the Python interpreter

can understand. An alternative was to use the "Infomap" Python package by itself, however the Infomap algorithm in this package cannot be executed on a "Networkx" network because it uses its own method of constructing a network. Using the "Infomap" package was rather complicated and thus it seemed more time-saving to drop the algorithm and solely focus on modularity-based algorithms.

REFERENCES

[1] Videla-Cavieres, I. F. and Rios, S. A. Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications*, **41**, 1928–1936.

[2] Raeder, T. and Chawla, N. V. Modeling a store's product space as a social network. *2009 International conference on advances in social network analysis and mining*, pp. 164–169. IEEE.

[3] Zhang, L., Priestley, J., DeMaio, J., and Ni, S. A product affinity segmentation framework. , **?**

[4] Bedi, P. and Sharma, C. Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **6**, 115–135.

[5] Butts, C. T. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, **11**, 13–41.

[6] Kunegis, J. Konect: the koblenz network collection. *Proceedings of the 22nd international conference on World Wide Web*, pp. 1343–1350.

[7] Coscia, M., Giannotti, F., and Pedreschi, D. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **4**, 512–546.

[8] Kim, H. K., Kim, J. K., and Chen, Q. Y. A product network analysis for extending the market basket analysis. *Expert Systems with Applications*, **39**, 7403–7410.

[9] Huang, Z., Zeng, D. D., and Chen, H. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management science*, **53**, 1146–1164.

[10] Wang, C., Tang, W., Sun, B., Fang, J., and Wang, Y. Review on community detection algorithms in social networks. *2015 IEEE international conference on progress in informatics and computing (PIC)*, pp. 551–555. IEEE.

[11] Khatoon, M. and Banu, W. A. A survey on community detection methods in social networks. *International Journal of Education and Management Engineering*, **5**, 8.

[12] Dong, J. and Horvath, S. Understanding network concepts in modules. *BMC systems biology*, **1**, 1–20.

[13] Holme, P., Park, S. M., Kim, B. J., and Edling, C. R. Korean university life in a network perspective: Dynamics of a large affiliation network. *Physica A: Statistical Mechanics and its Applications*, **373**, 821–830.

[14] Faridizadeh, S., Abdolvand, N., and Harandi, S. R. Market basket analysis using community detection approach: A real case. *Applications of Data Management and Analysis*, pp. 177–198. Springer.

[15] VANHOOF, K., Frasquet, M., and HERRERA, I. F. Product affinity segmentation of multichannel grocery shoppers applying community detection. *Proceedings 5th Colloquium on European Research in Retailing (CERR20202)*, pp. 74–81. 5th Colloquium on European Research in Retailing.

[16] Fortunato, S. Community detection in graphs. *Physics reports*, **486**, 75–174.

[17] Fortunato, S. and Hric, D. Community detection in networks: A user guide. *Physics reports*, **659**, 1–44.

[18] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. Defining and identifying communities in networks. *Proceedings of the national academy of sciences*, **101**, 2658–2663.

[19] Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., and Samatova, N. Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, **6**, 426–439.

[20] Javed, M. A., Younis, M. S., Latif, S., Qadir, J., and Baig, A. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, **108**, 87–111.

[21] Roux, M. A comparative study of divisive hierarchical clustering algorithms. *arXiv preprint arXiv:1506.08977* , **?**

[22] Girvan, M. and Newman, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, **99**, 7821–7826.

[23] Lancichinetti, A. and Fortunato, S. Community detection algorithms: a comparative analysis. *Physical review E*, **80**, 056117.

[24] Mothe, J., Mkhitaryan, K., and Haroutunian, M. Community detection: Comparison of state of the art algorithms. *2017 Computer Science and Information Technologies (CSIT)*, pp. 125–129. IEEE.

[25] Wagenseller, P., Wang, F., and Wu, W. Size matters: A comparative analysis of community detection algorithms. *IEEE Transactions on Computational Social Systems*, **5**, 951–960.

[26] Yang, Z., Algesheimer, R., and Tessone, C. J. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, **6**, 1–18.

[27] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, **2008**, P10008.

[28] Traag, V. A., Waltman, L., and Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, **9**, 1–12.

[29] Clauset, A., Newman, M. E., and Moore, C. Finding community structure in very large networks. *Physical review E*, **70**, 066111.

[30] Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, **74**, 036104.

[31] Rosvall, M. and Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, **105**, 1118–1123.

[32] Rossetti, G., Milli, L., and Cazabet, R. Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, **4**, 1–26.

[33] Blattberg, R. C., Kim, B.-D., and Neslin, S. A. Market basket analysis. *Database Marketing*, pp. 339–351. Springer.

[34] Ordonez, C., Ezquerra, N., and Santana, C. A. Constraining and summarizing association rules in medical data. *Knowledge and information systems*, **9**, 1–2.

[35] Merceron, A. and Yacef, K. Interestingness measures for association rules in educational data. *Educational Data Mining 2008*.

[36] Raeder, T. and Chawla, N. V. Market basket analysis with networks. *Social network analysis and mining*, **1**, 97–113.

[37] Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J. (eds.), *Proceedings of the 7th Python in Science Conference*, Pasadena, CA USA, pp. 11 – 15.

[38] Newman, M. E. and Girvan, M. Finding and evaluating community structure in networks. *Physical review E*, **69**, 026113.

Appendices

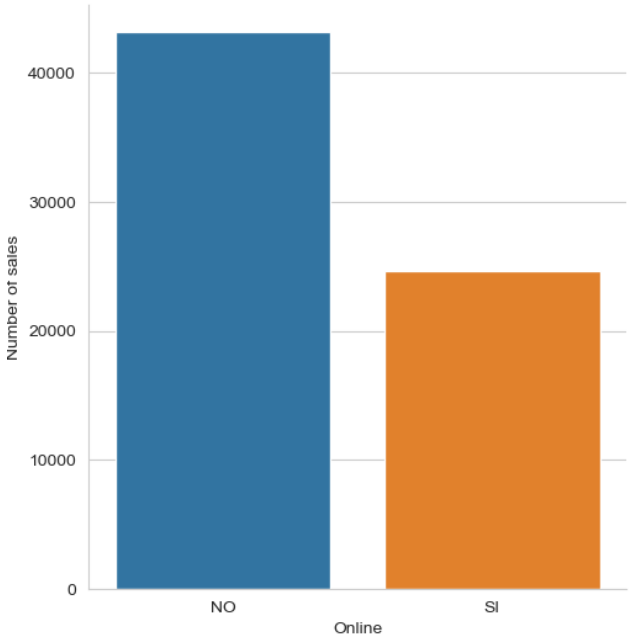

FIGURE 1. Online sales of the retail data set.

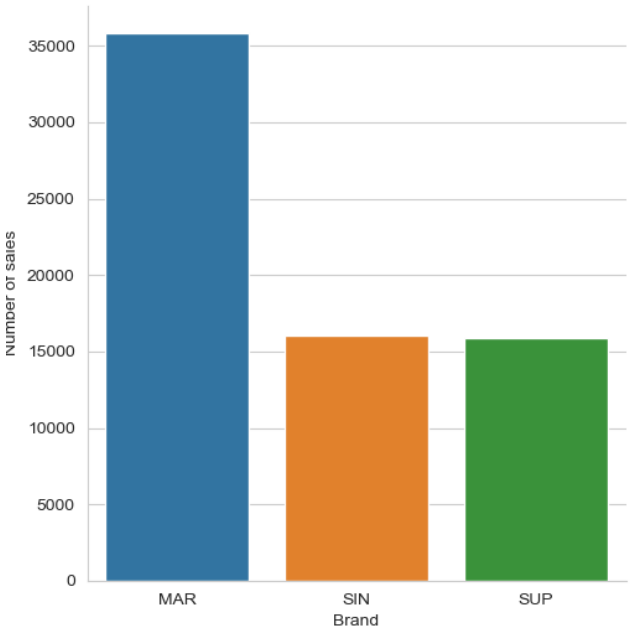

FIGURE 3. Brand sales of the retail data set.

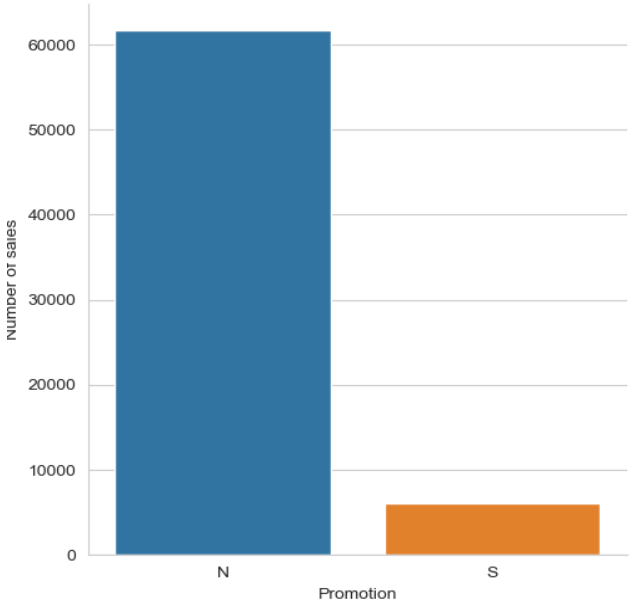

FIGURE 2. Promotion sales of the retail data set.

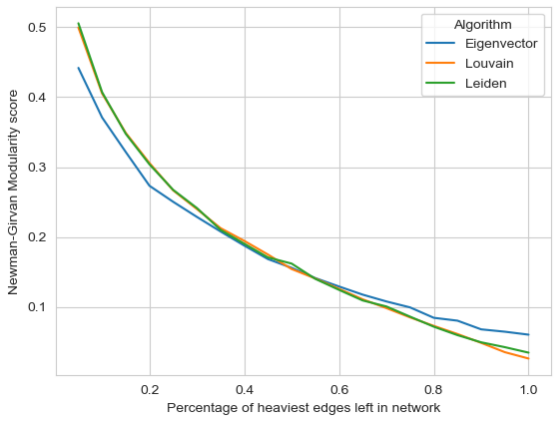

FIGURE 4. The relationship between a partition of the customer network left over after keeping a percentage of the heaviest edges in the customer network and the modularity score after performing a community detection algorithm on that partition. This graph shows that the modularity can be artificially increased by filtering out weaker edges in the network.

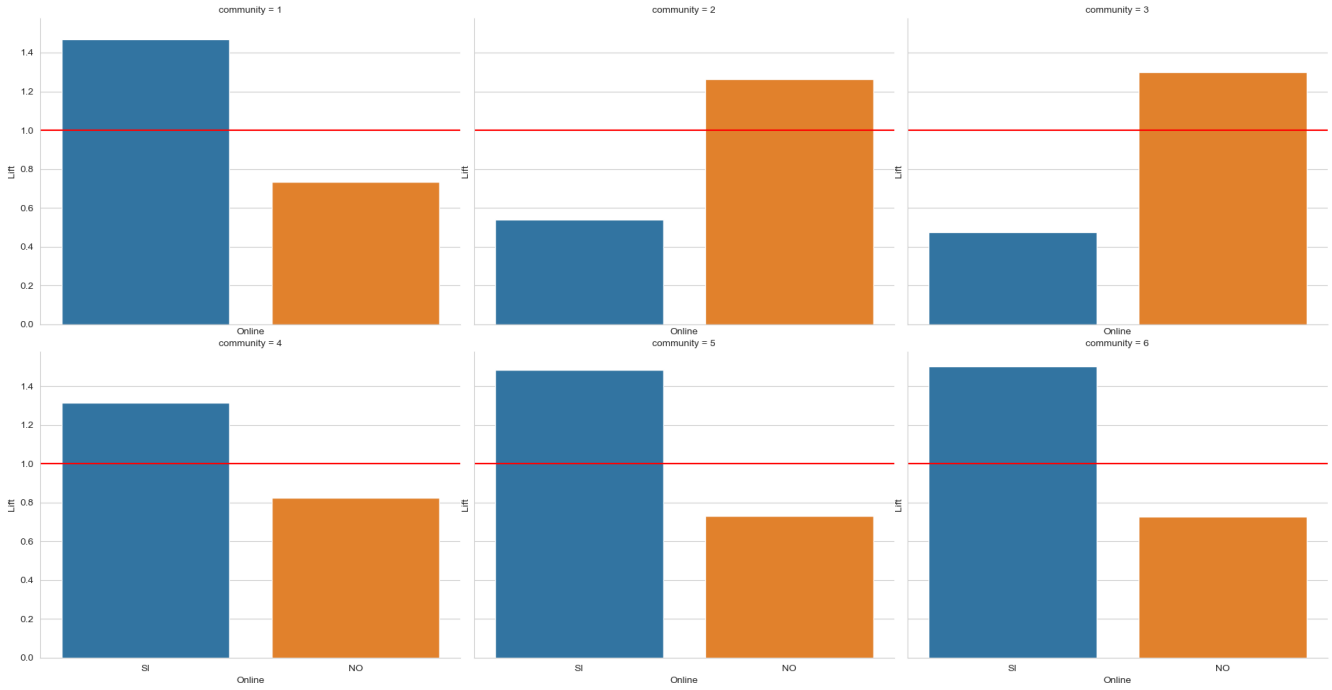

FIGURE 5. The lift values for online sales for each Eigenvector algorithm community based on the normal MInteraction.

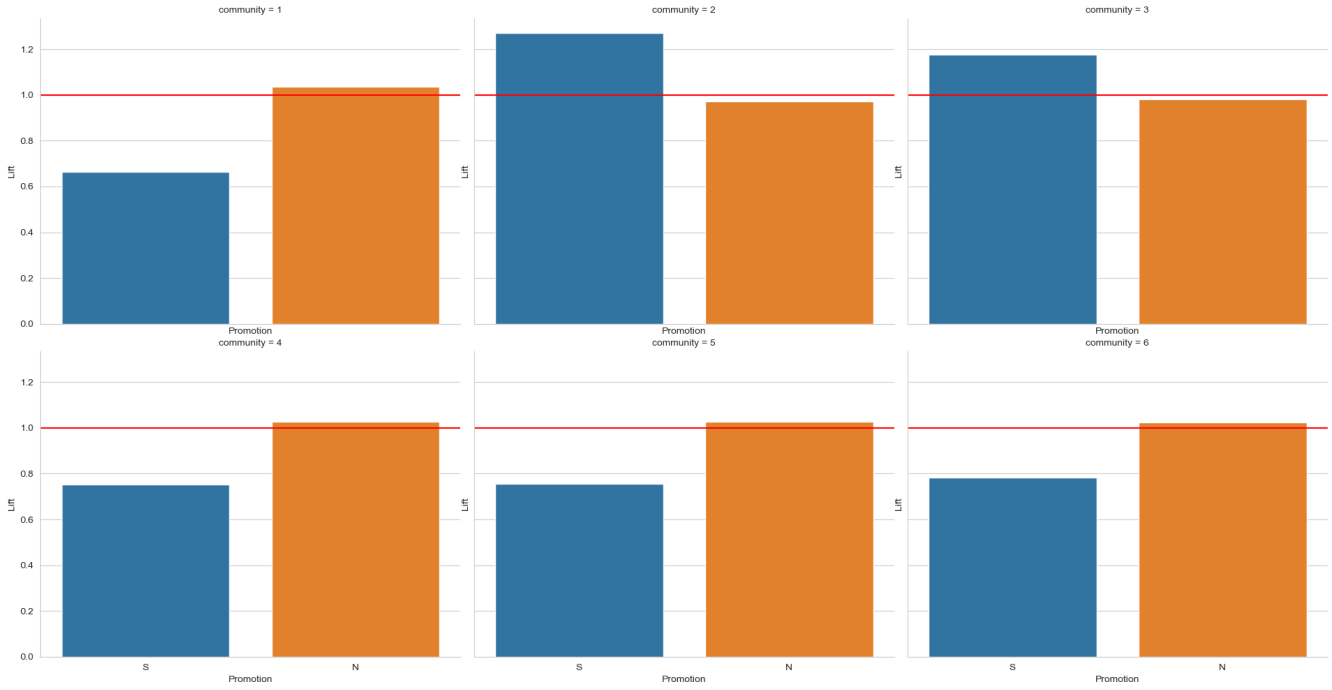

FIGURE 6. The lift values for promotion sales for each Eigenvector algorithm community based on the normal MInteraction.

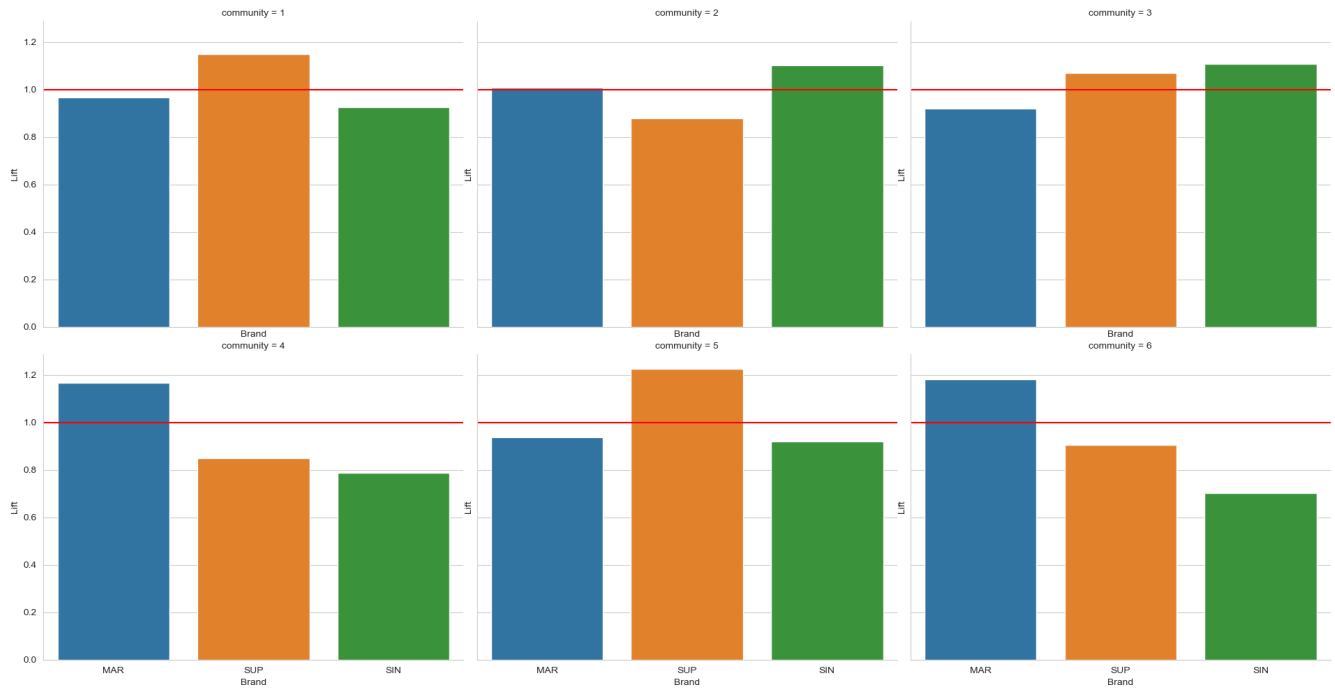

FIGURE 7. The lift values for each brand type for each Eigenvector algorithm community baseond on the normal MInteraction.

FIGURE 8. The lift values for online sales for each Louvain algorithm community based on the normal MInteraction.

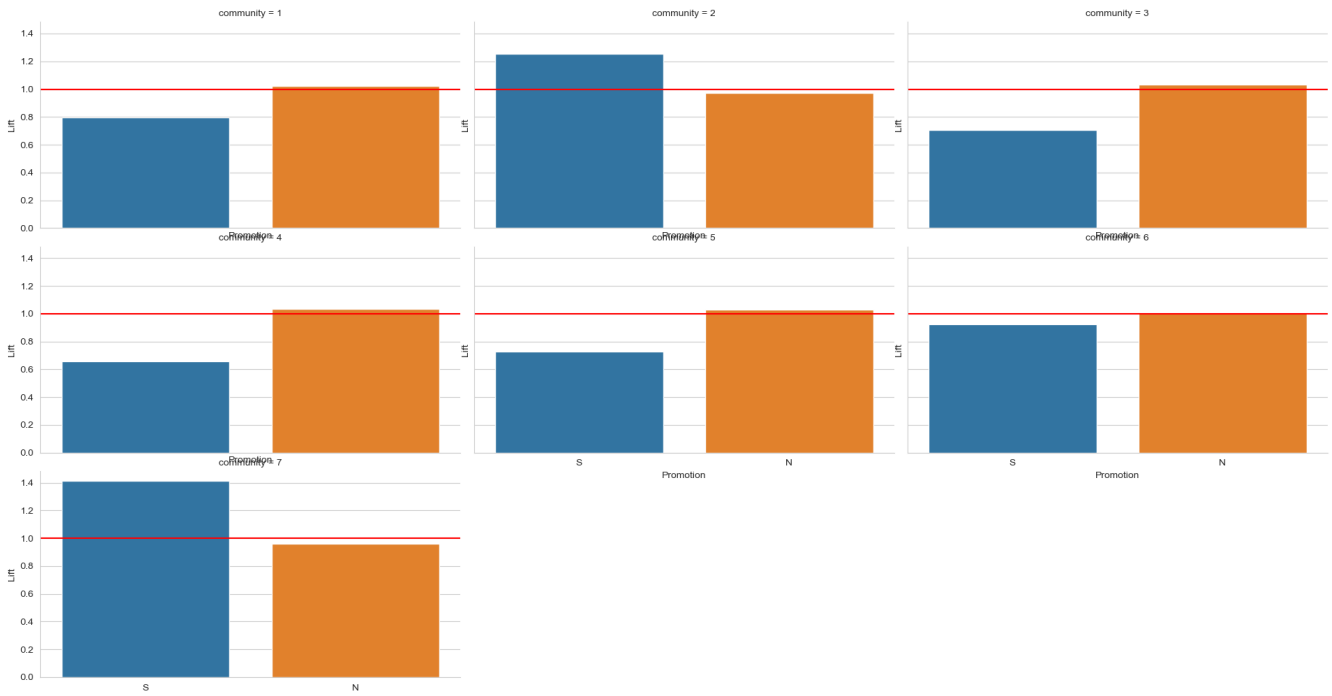

FIGURE 9. The lift values for promotion sales for each Louvain algorithm community based on the normal MInteraction.

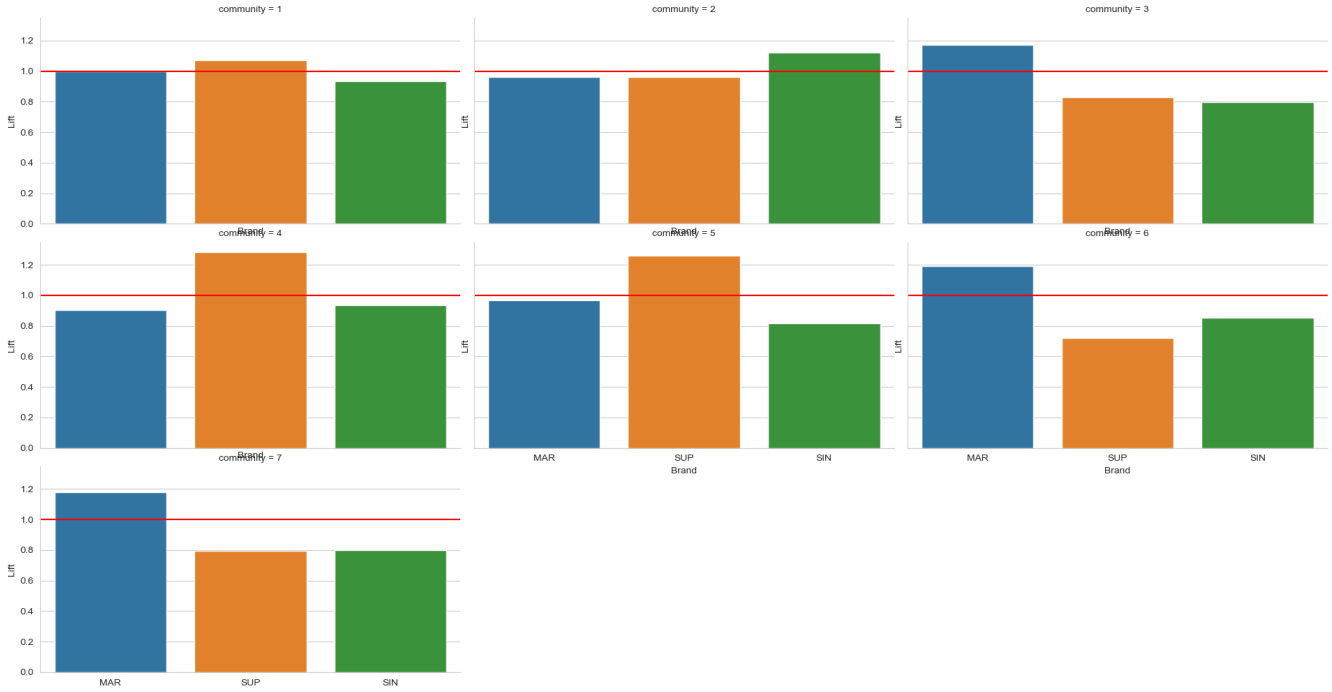

FIGURE 10. The lift values for each brand type for each Louvain algorithm community based on the normal MInteraction

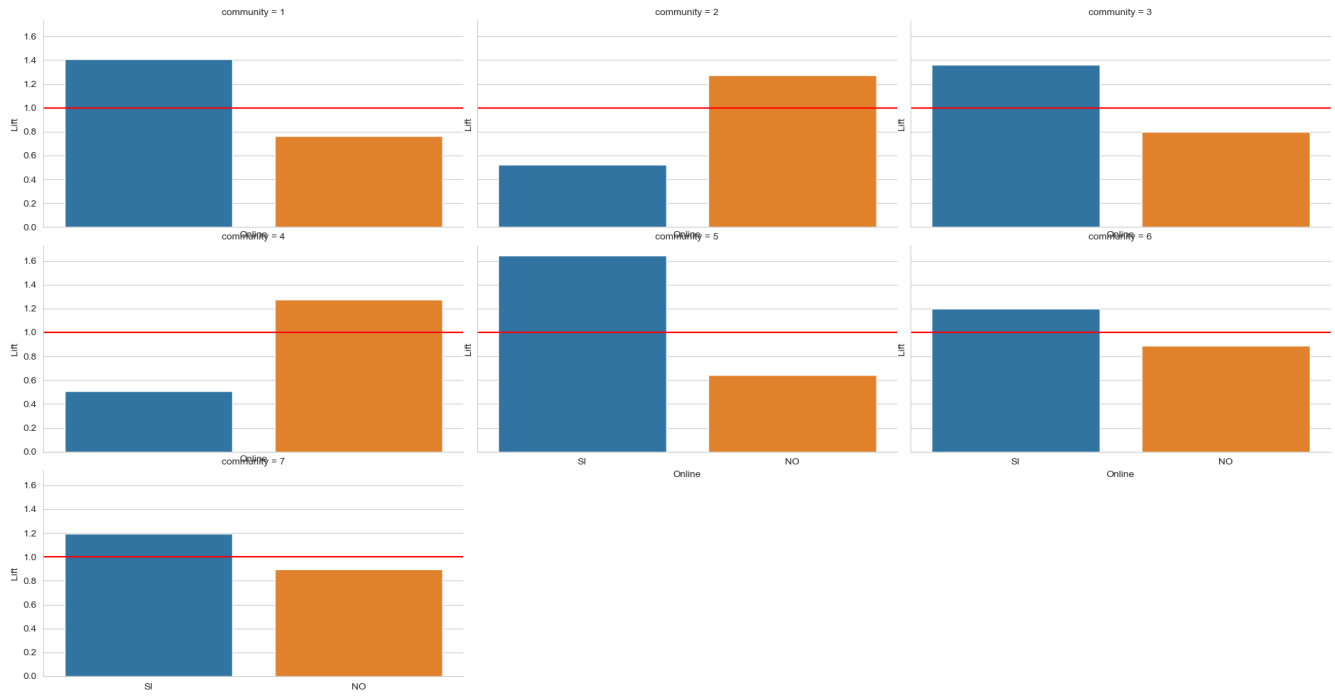

FIGURE 11. The lift values for online sales for each Leiden algorithm community based on the normal MInteraction.

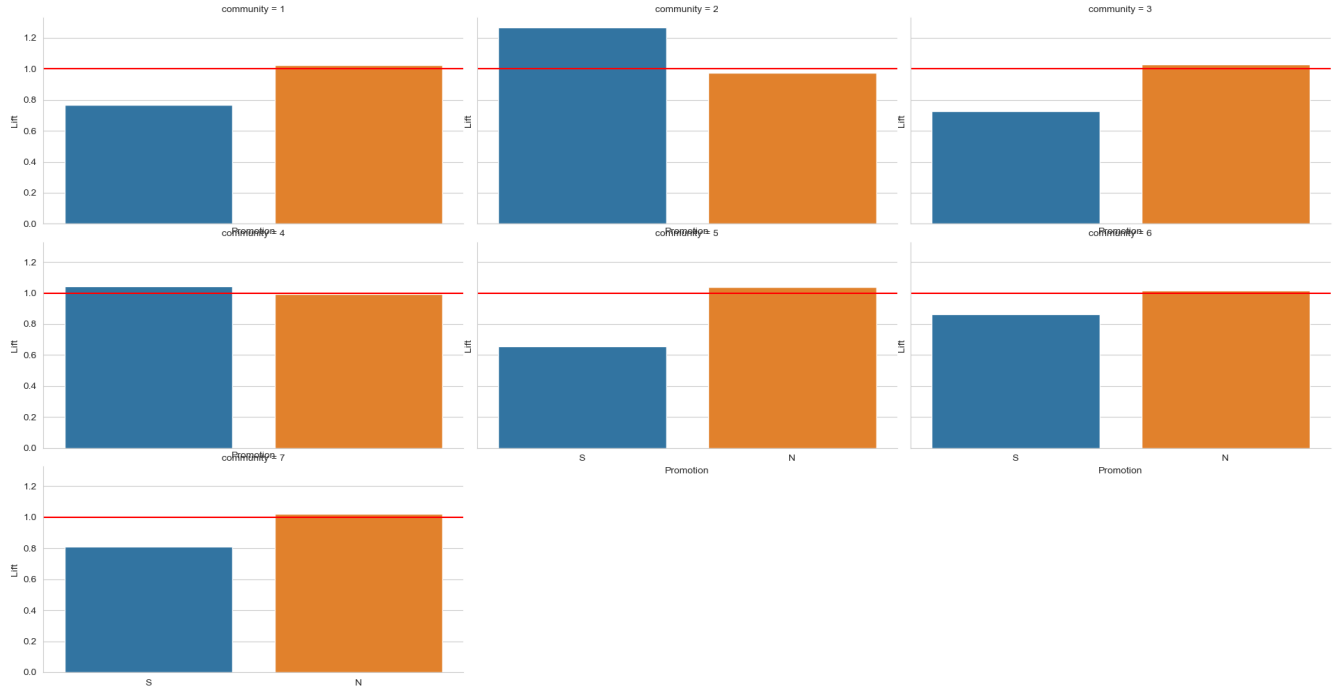

FIGURE 12. The lift values for promotion sales for each Leiden algorithm community based on the normal MInteraction.

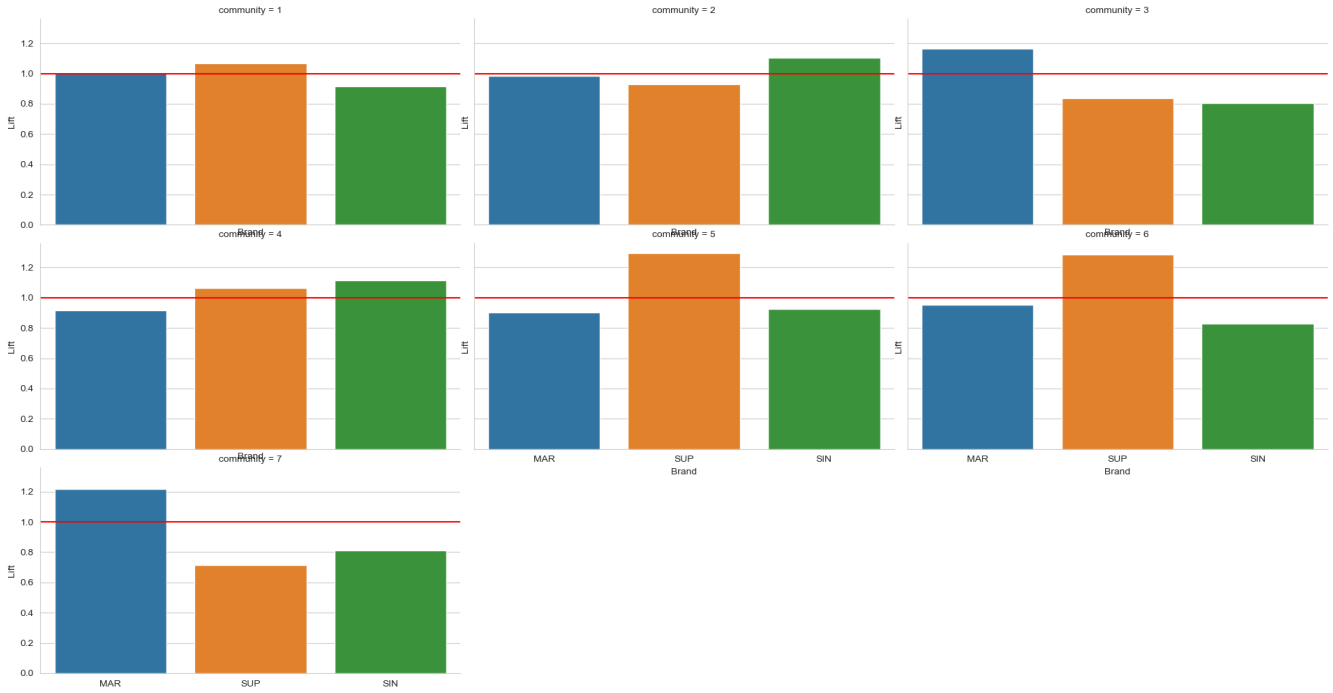

FIGURE 13. The lift values for each brand type for each Leiden algorithm community based on the normal MInteraction.

FIGURE 14. The lift values for each product category for each Eigenvector algorithm community based on the normal MInteraction.

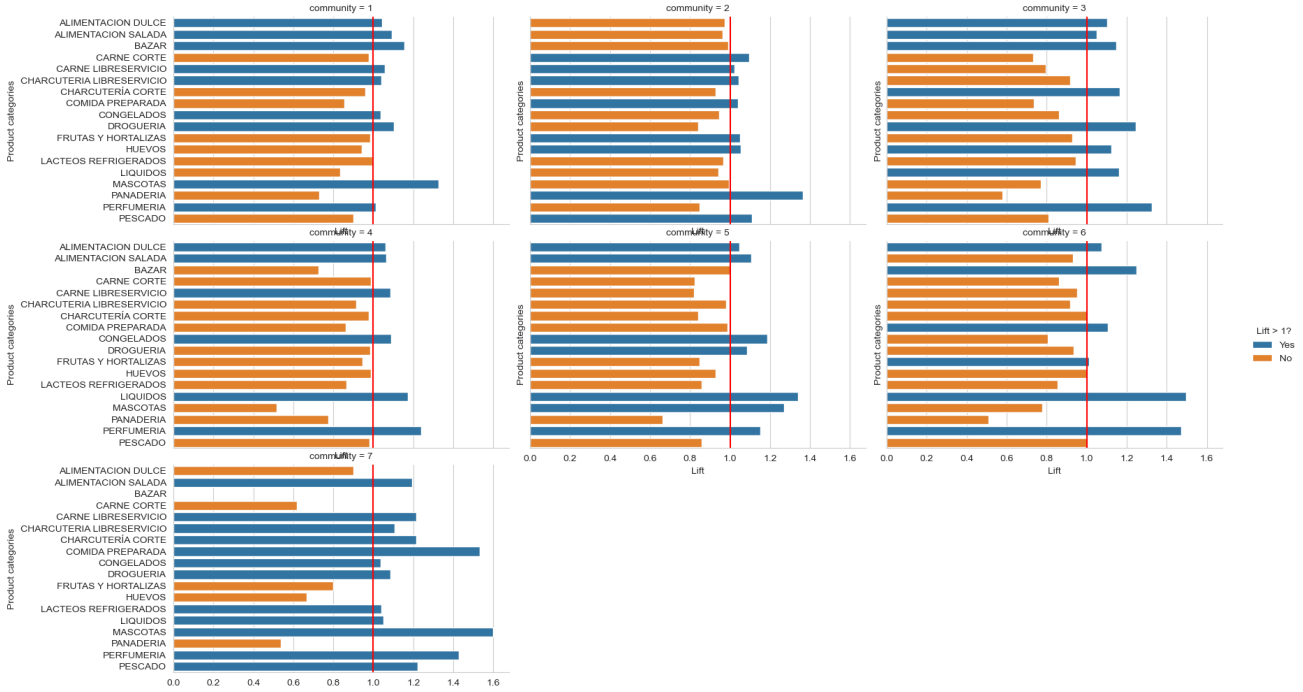

FIGURE 15. The lift values for each product category for each Louvain algorithm community based on the normal MInteraction.

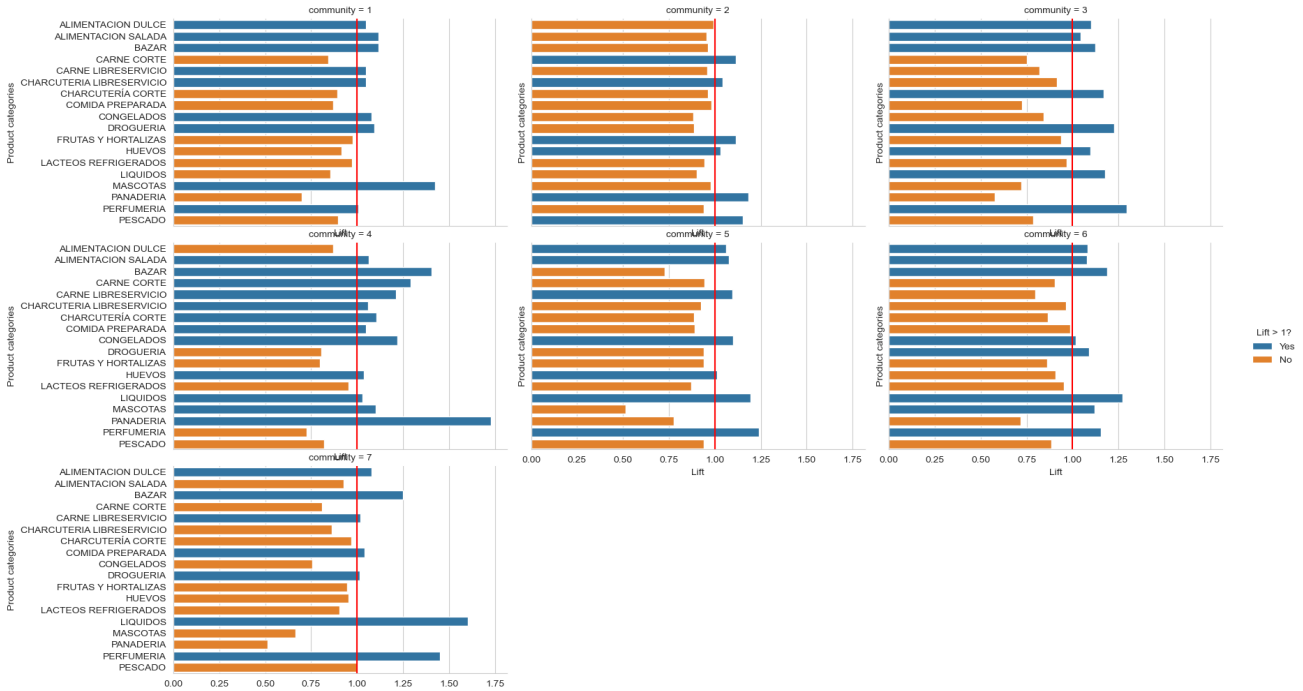

FIGURE 16. The lift values for each product category for each Leiden algorithm community based on the normal MInteraction.

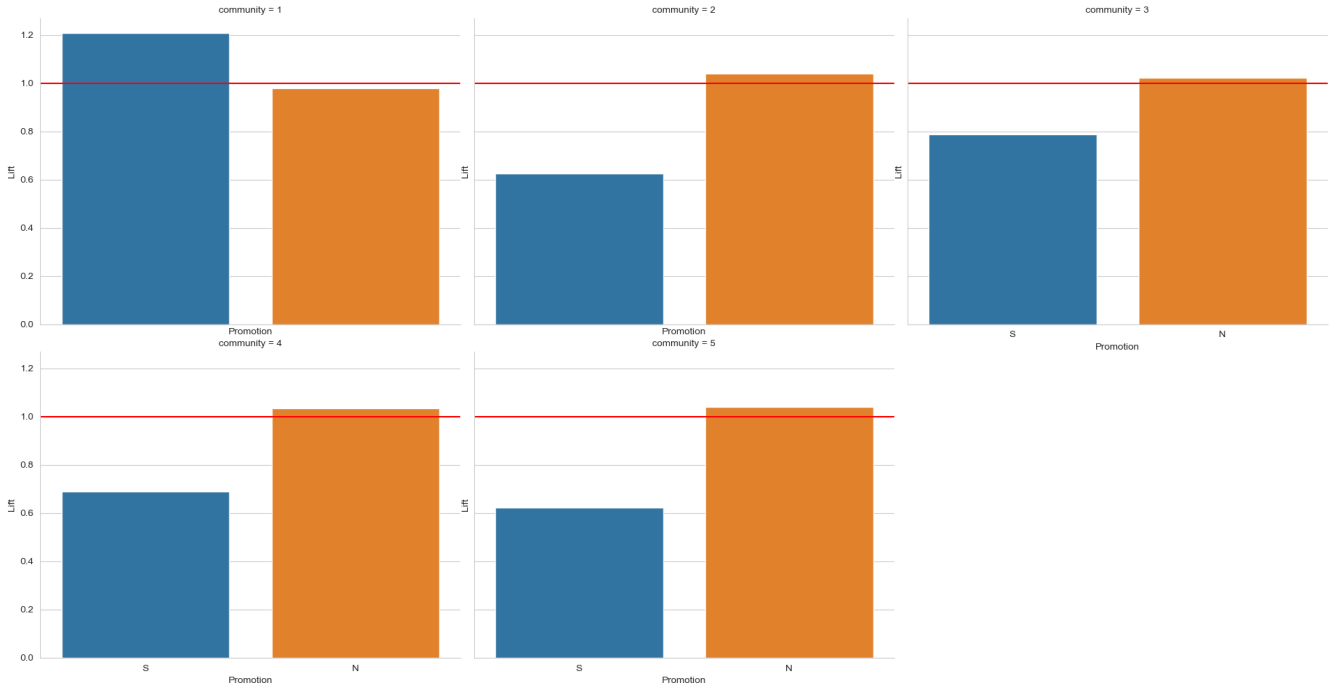

FIGURE 17. The lift values for online sales for each Eigenvector algorithm community based on the MInteraction skewed towards promotions.

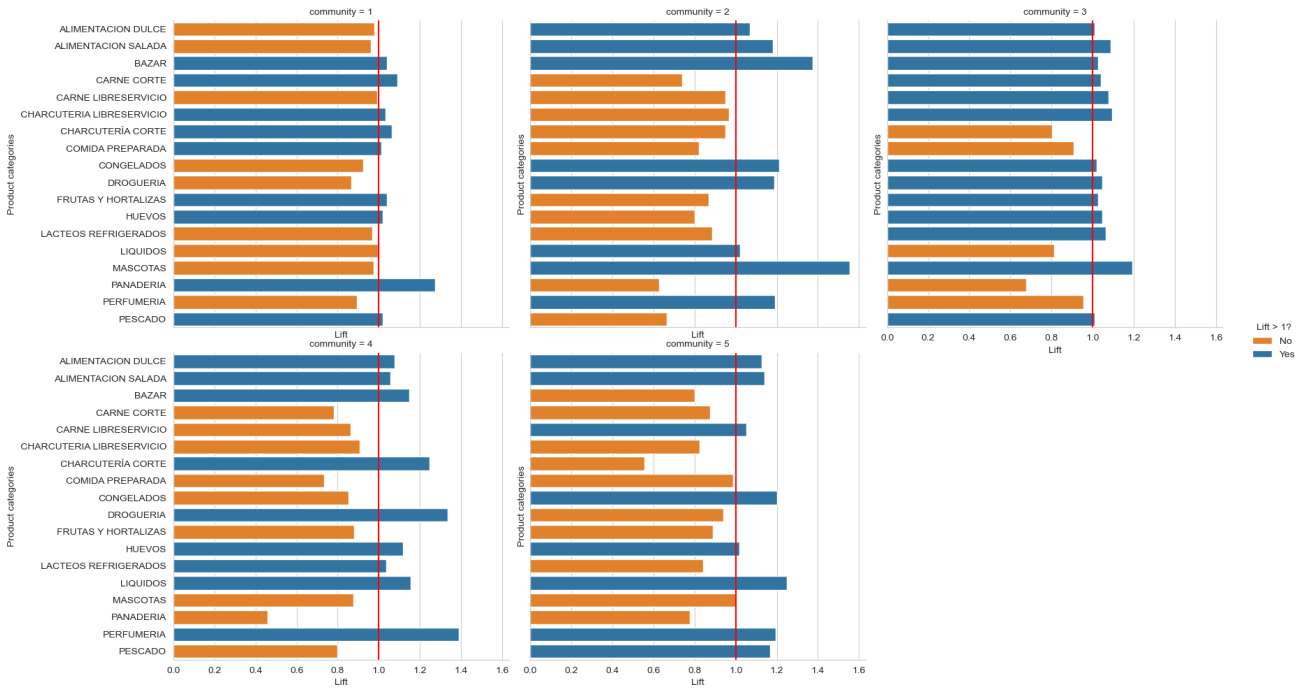

FIGURE 18. The lift values for each product category for each Eigenvector algorithm community based on the MInteraction skewed towards promotions.

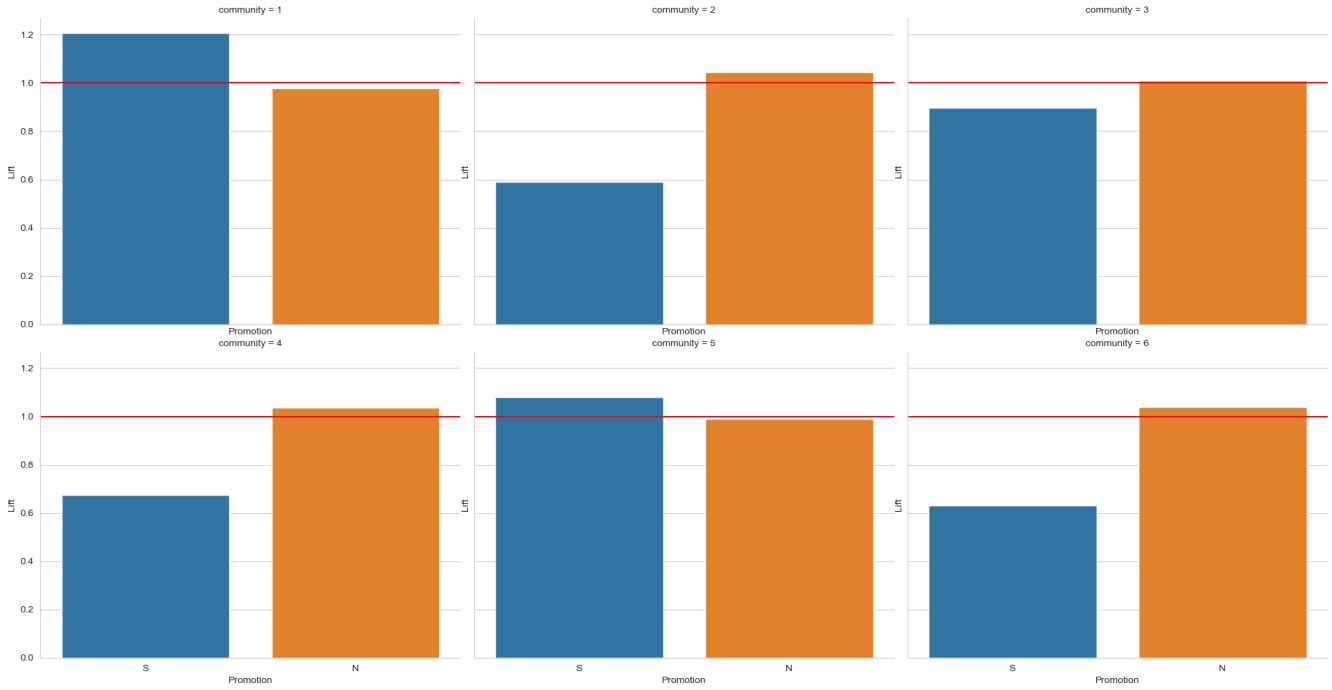

FIGURE 19. The lift values for online sales for each Louvain algorithm community based on the MInteraction skewed towards promotions.

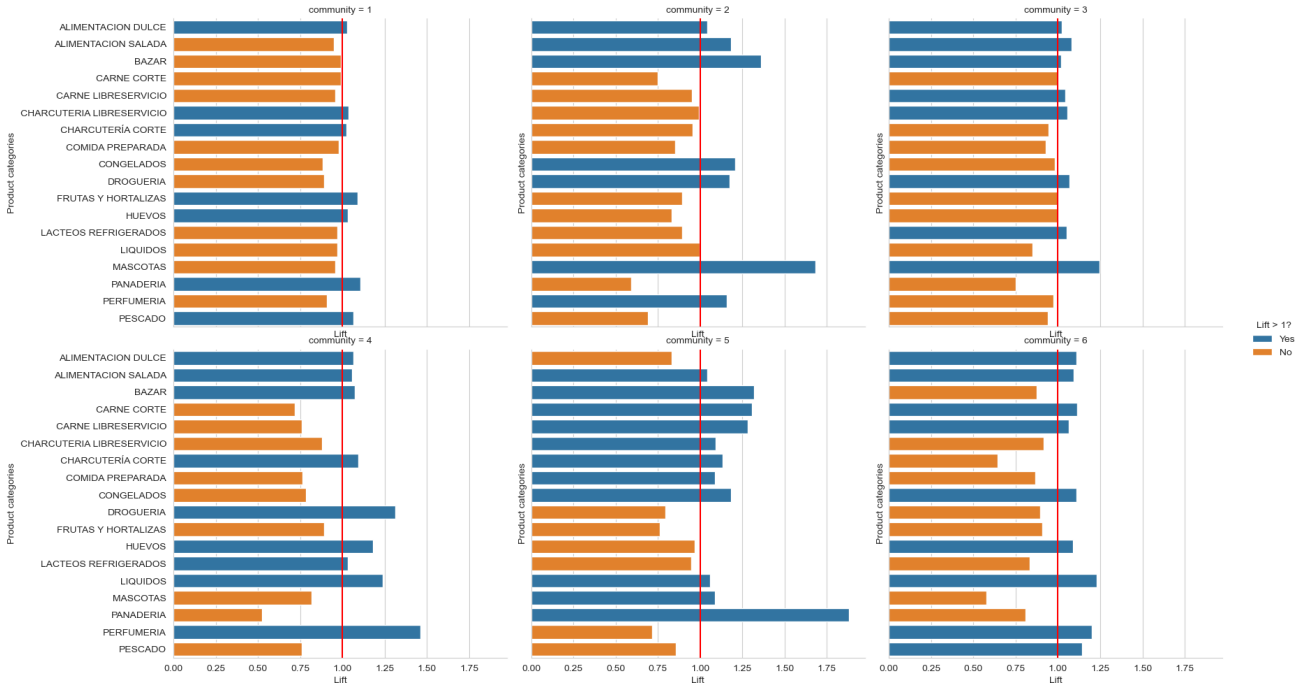

FIGURE 20. The lift values for each product category for each Louvain algorithm community based on the MInteraction skewed towards promotions.

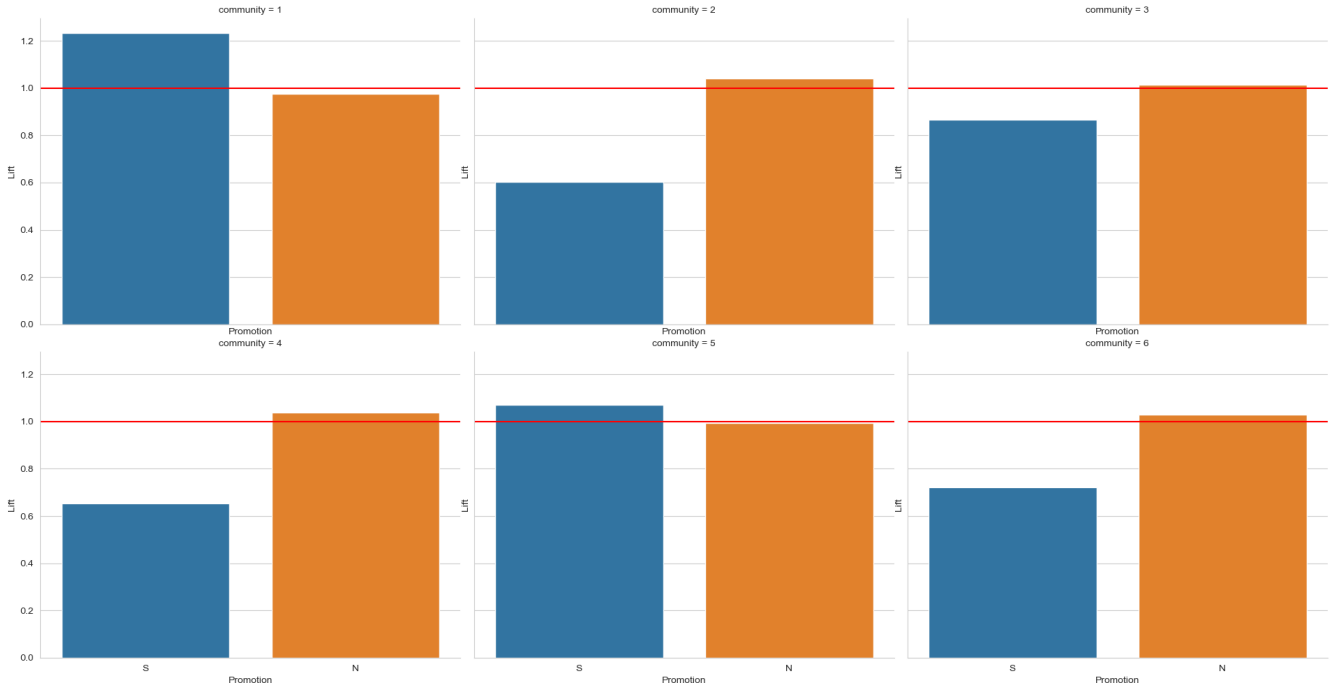

FIGURE 21. The lift values for online sales for each Leiden algorithm community based on the MInteraction skewed towards promotions.

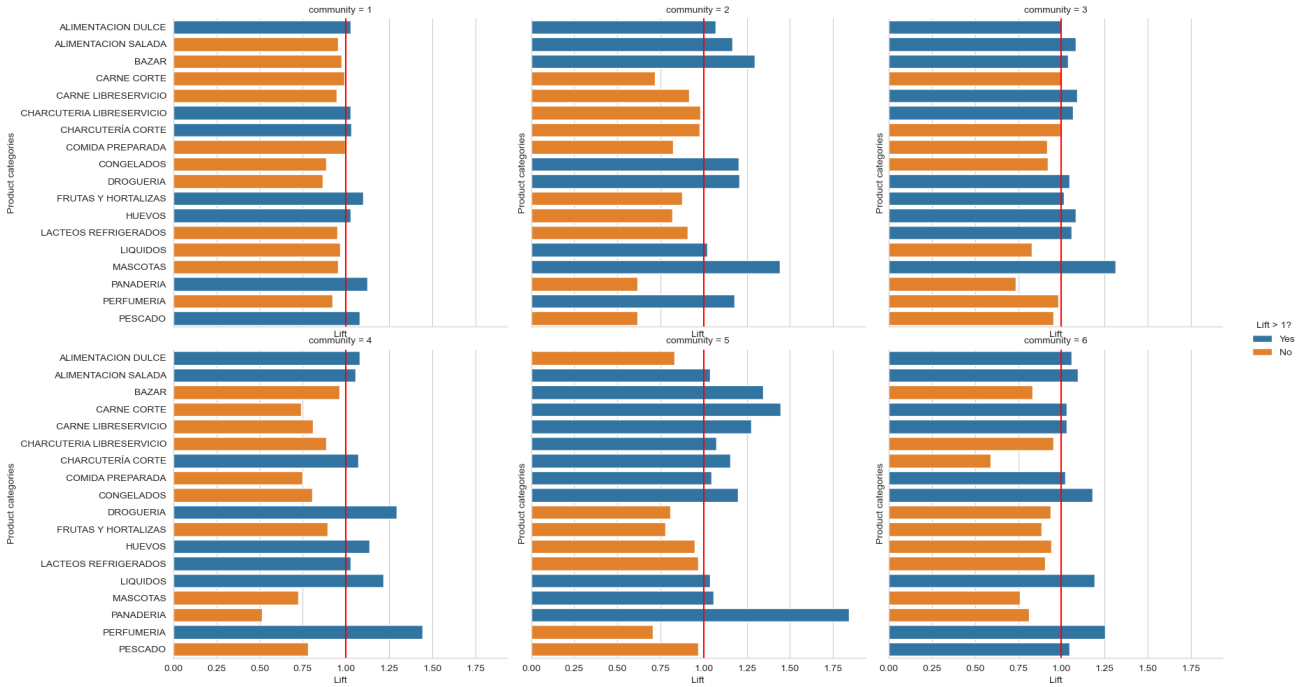

FIGURE 22. The lift values for each product category for each Leiden algorithm community based on the MInteraction skewed towards promotions.

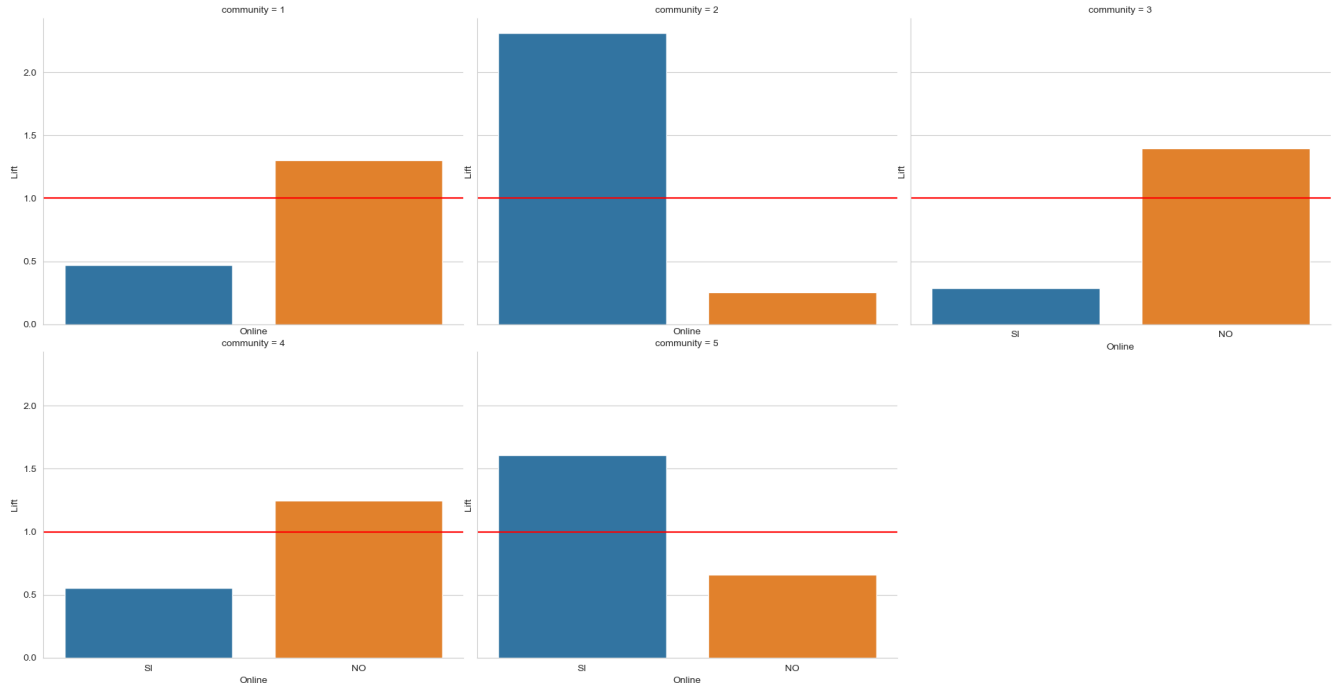

FIGURE 23. The lift values for online sales for each Eigenvector algorithm community based on the MInteraction skewed towards online purchases.

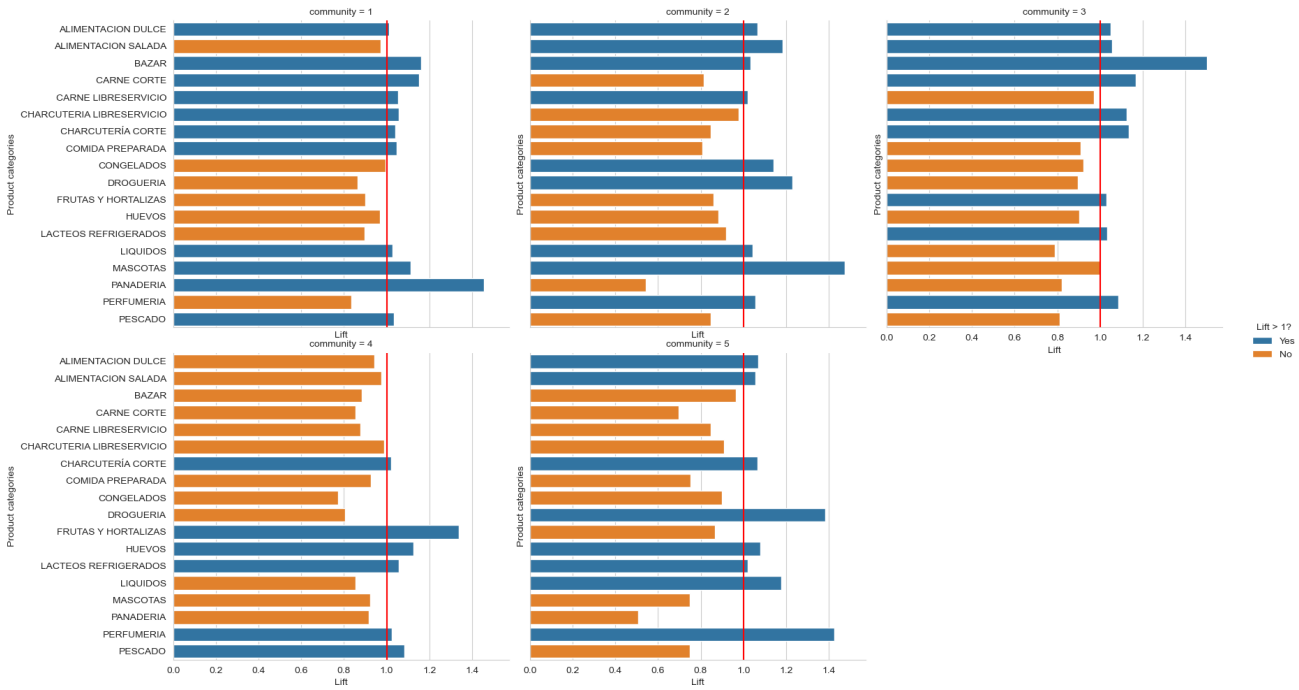

FIGURE 24. The lift values for each product category for each Eigenvector algorithm community based on the MInteraction skewed towards online purchases.

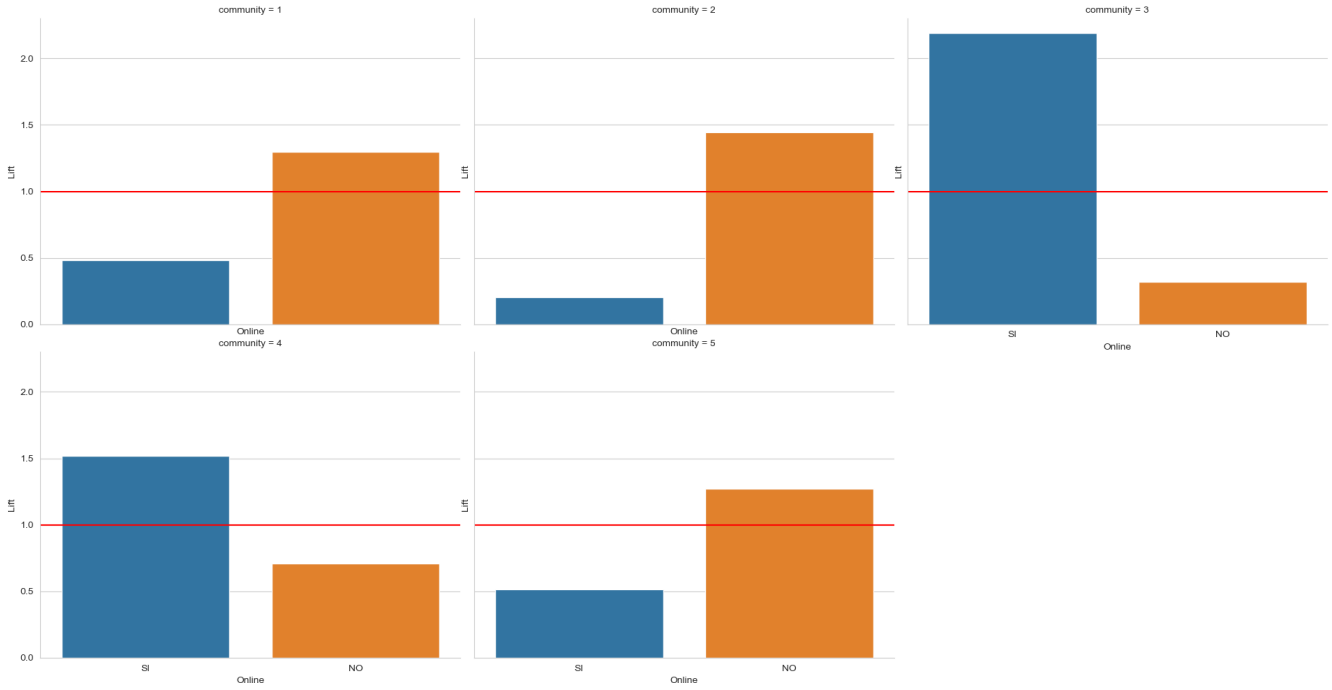

FIGURE 25. The lift values for online sales for each Louvain algorithm community based on the MInteraction skewed towards online purchases.

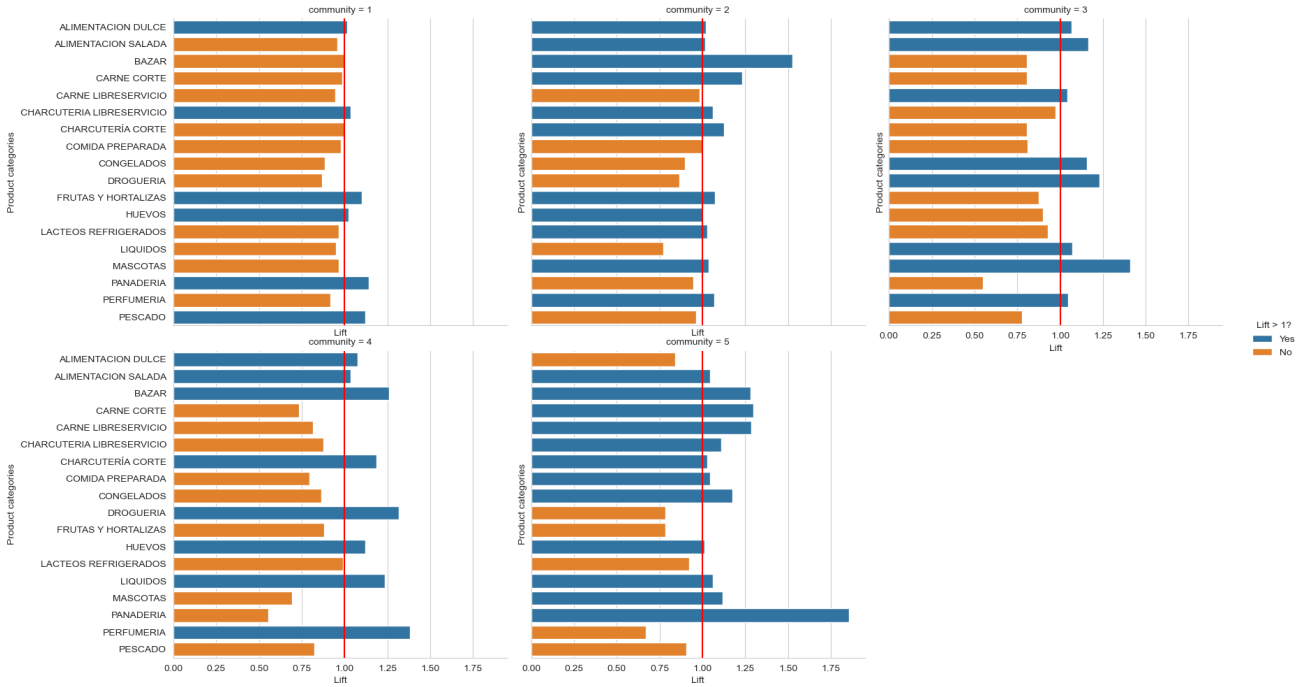

FIGURE 26. The lift values for each product category for each Louvain algorithm community based on the MInteraction skewed towards online purchases.

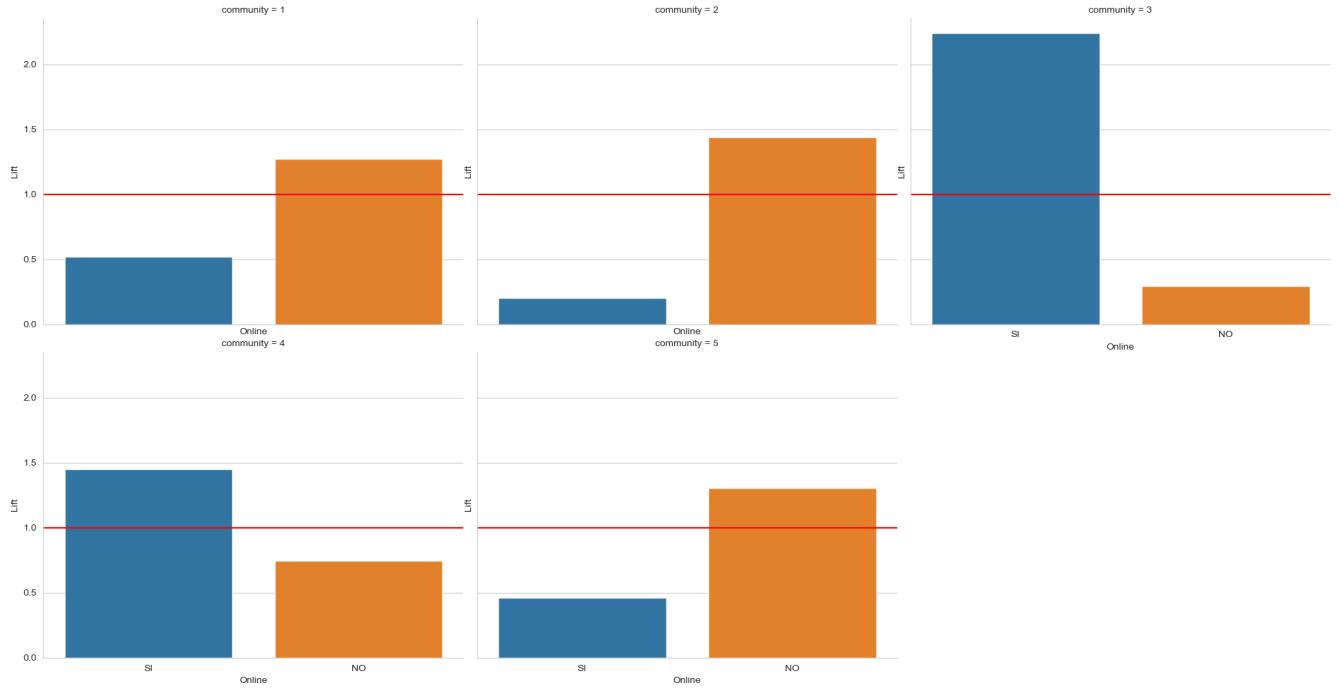

FIGURE 27. The lift values for online sales for each Leiden algorithm community based on the MInteraction skewed towards online purchases.

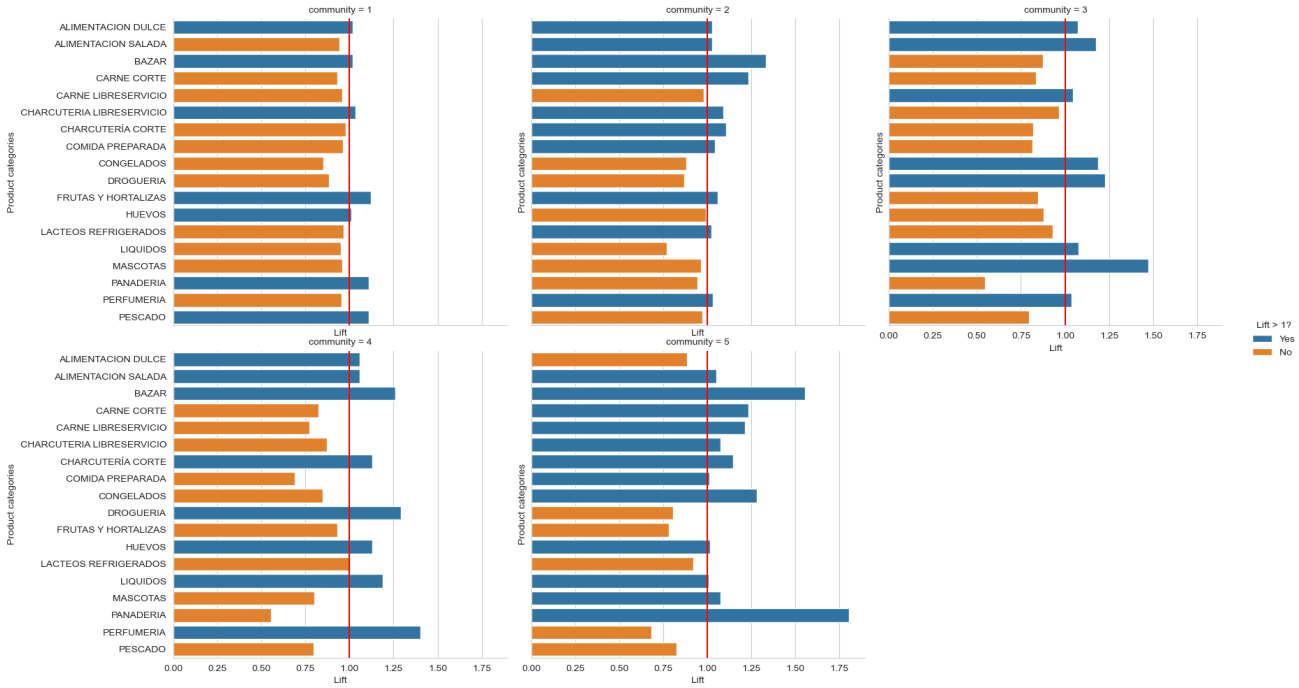

FIGURE 28. The lift values for each product category for each Leiden algorithm community based on the MInteraction skewed towards online purchases.

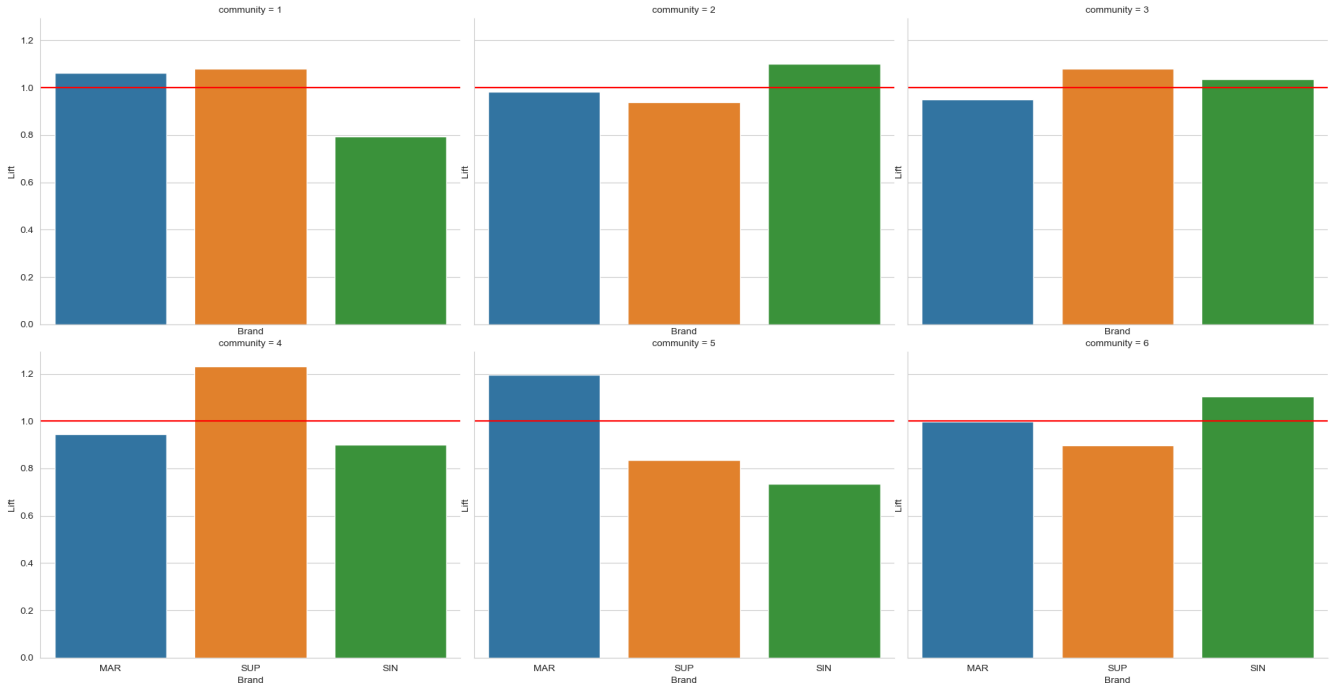

FIGURE 29. The lift values for online sales for each Eigenvector algorithm community based on the MInteraction skewed towards purchases of specific brand types.

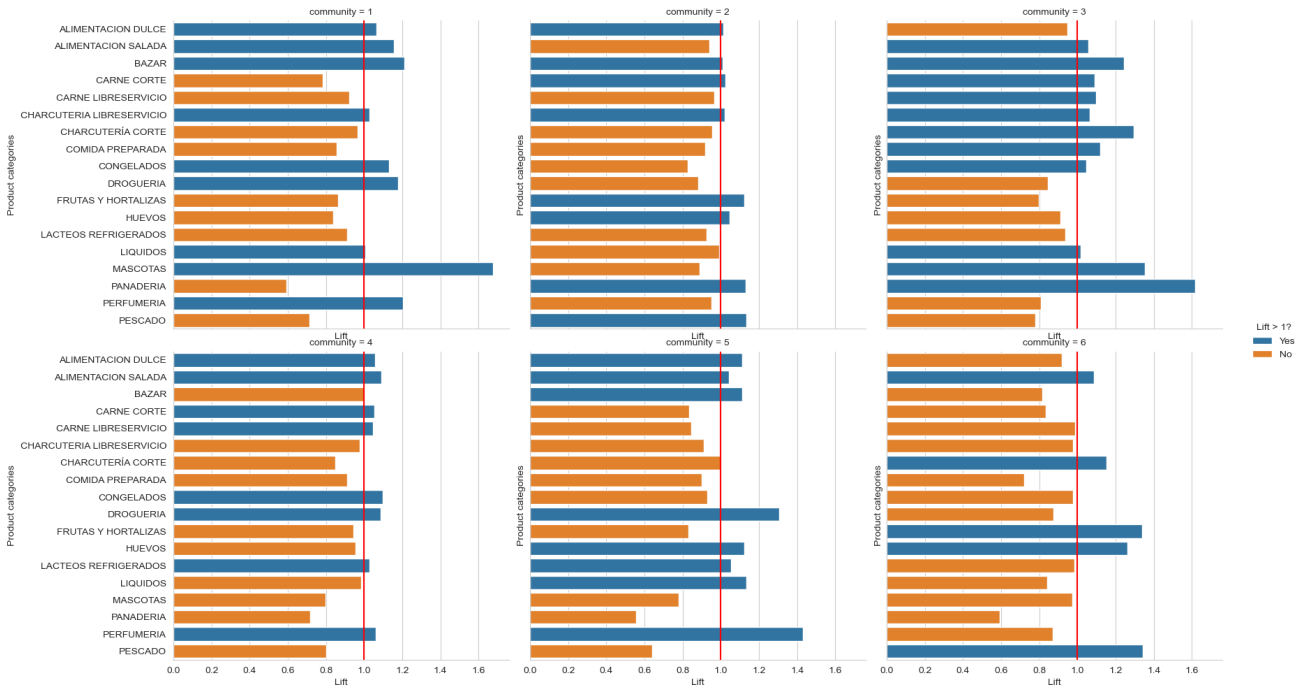

FIGURE 30. The lift values for each product category for each Eigenvector algorithm community based on the MInteraction skewed towards purchases of specific brand types.

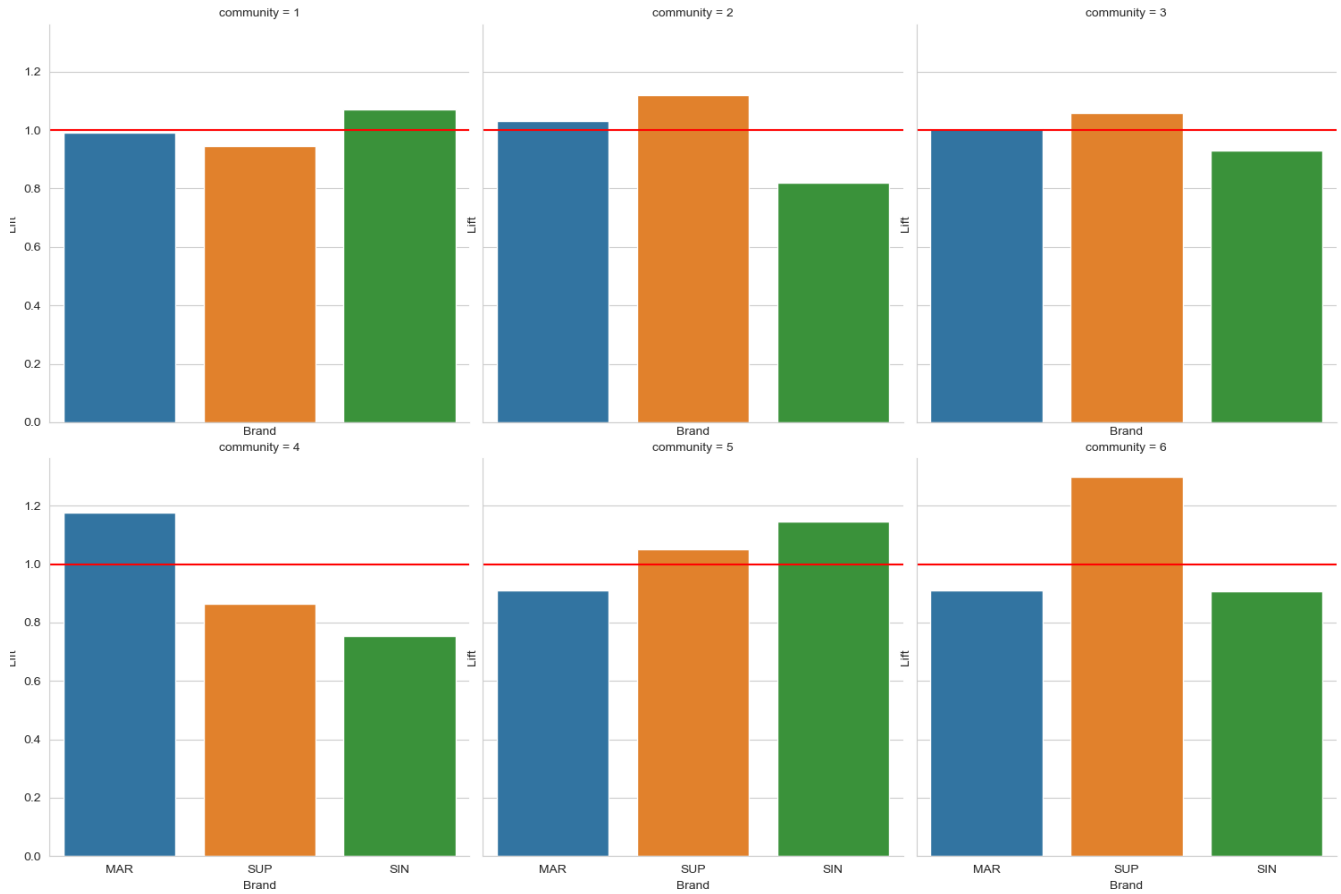

FIGURE 31. The lift values for online sales for each Louvain algorithm community based on the MInteraction skewed towards purchases of specific brand types.

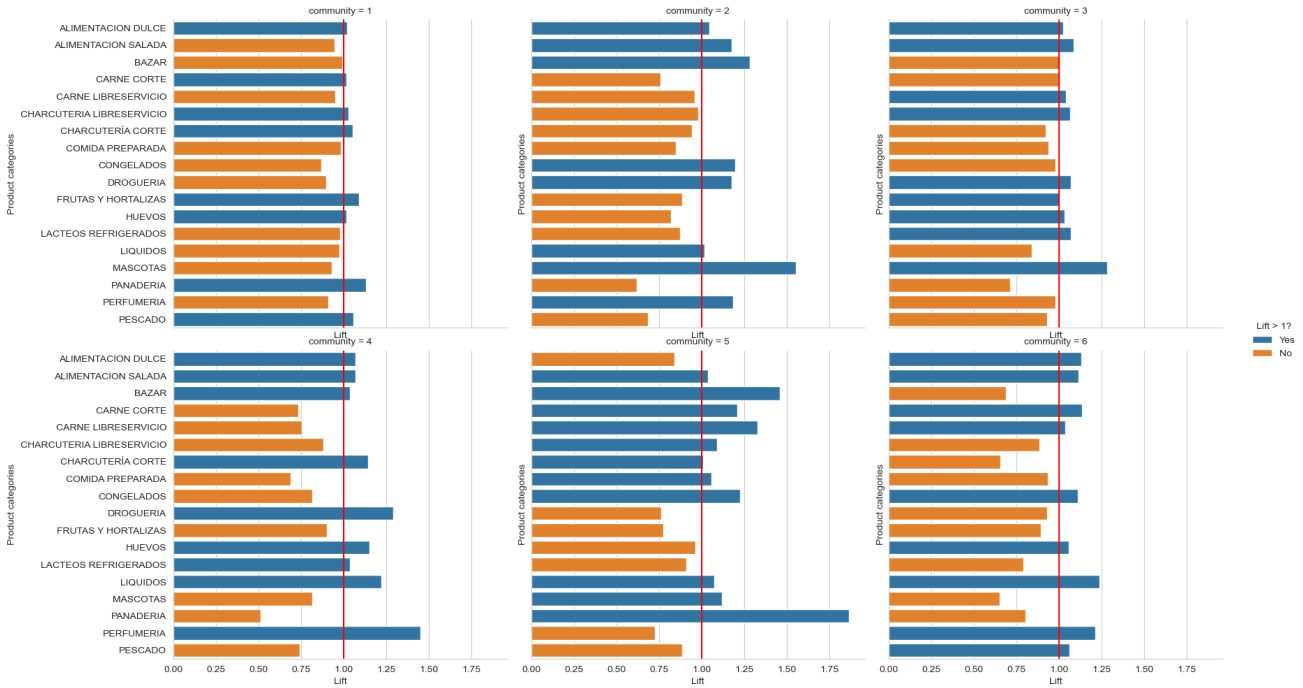

FIGURE 32. The lift values for each product category for each Louvain algorithm community based on the MInteraction skewed towards purchases of specific brand types.

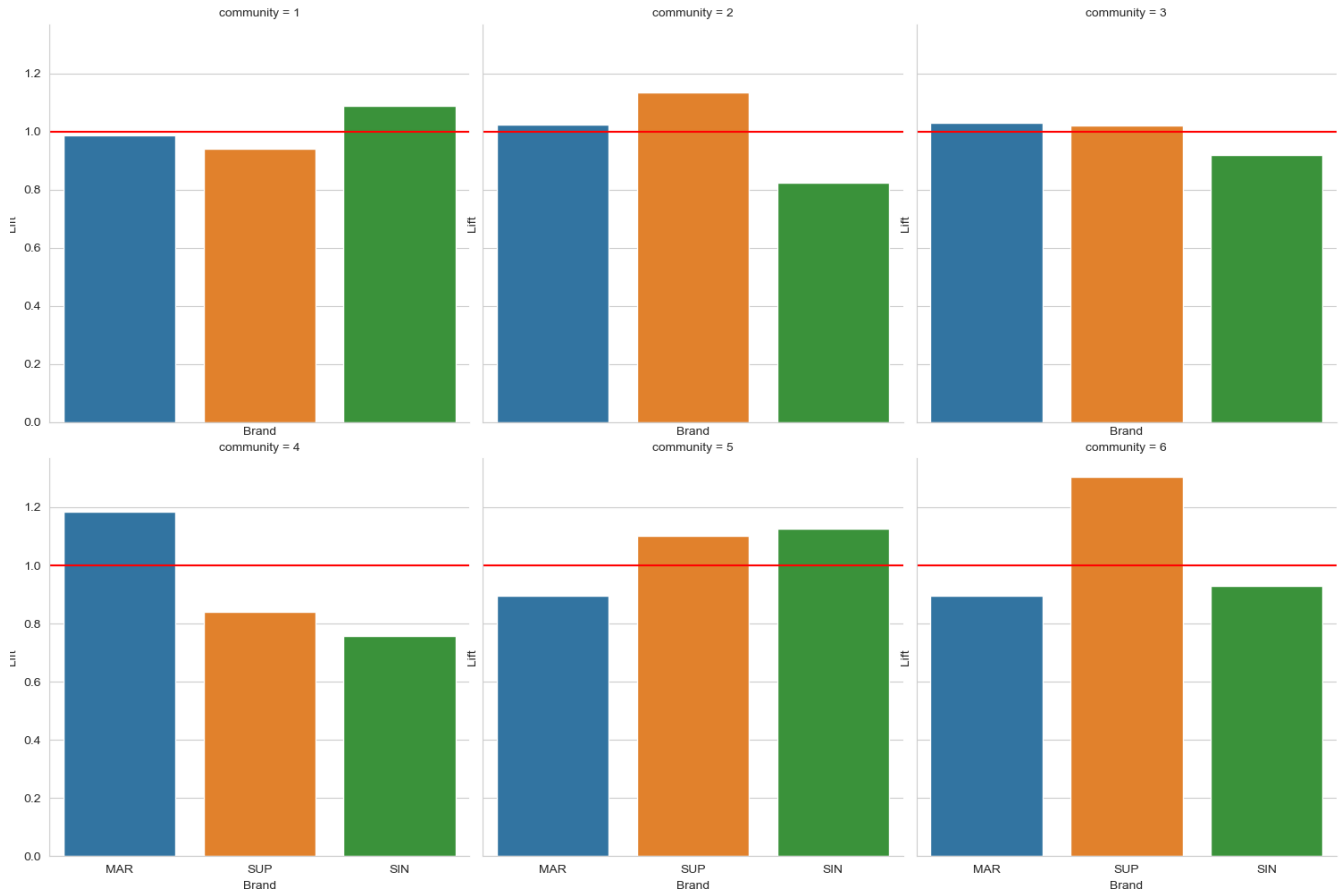

FIGURE 33. The lift values for online sales for each Leiden algorithm community based on the MInteraction skewed towards purchases of specific brand types.

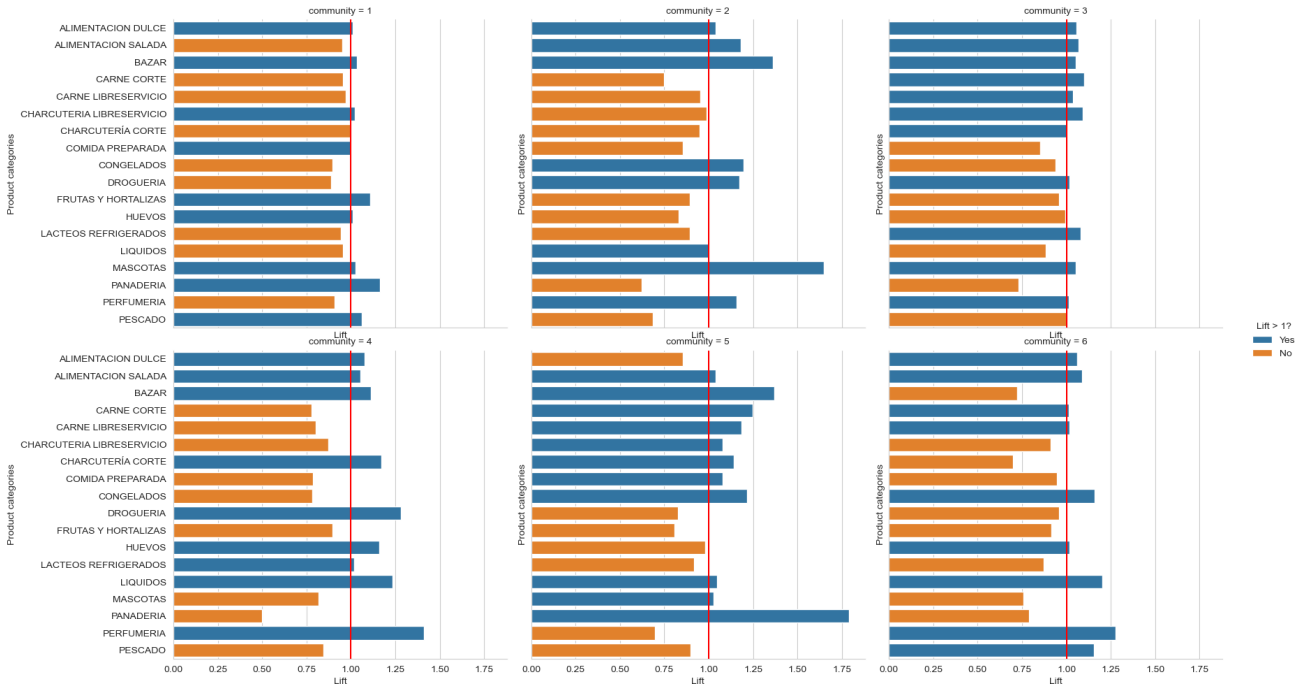

FIGURE 34. The lift values for each product category for each Leiden algorithm community based on the MInteraction skewed towards purchases of specific brand types.

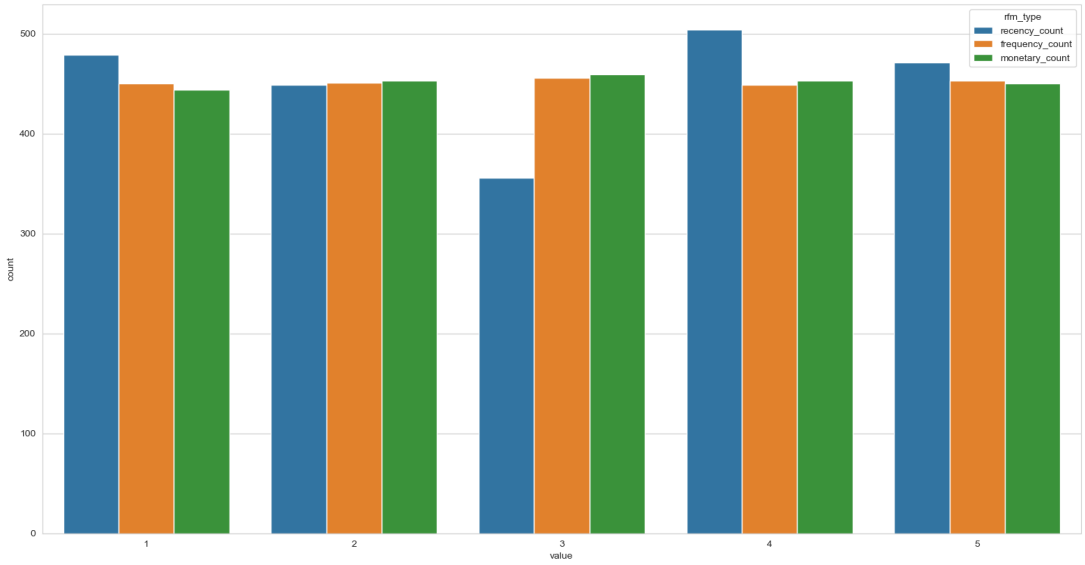

FIGURE 35. Distribution of the RFM-scores for all customers in the available retail data set.

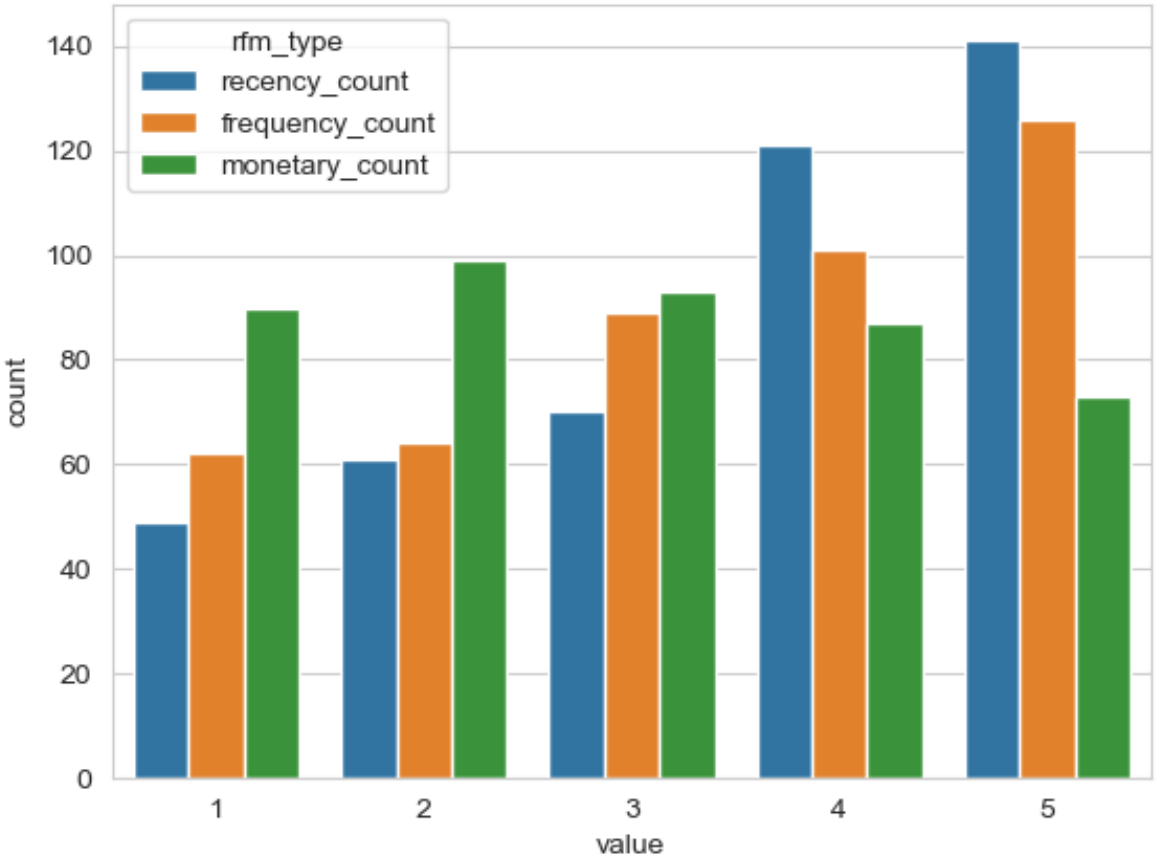

FIGURE 36. Distribution of the RFM-scores for promotion-sensitive customers based on the normal MInteraction similarity metric.

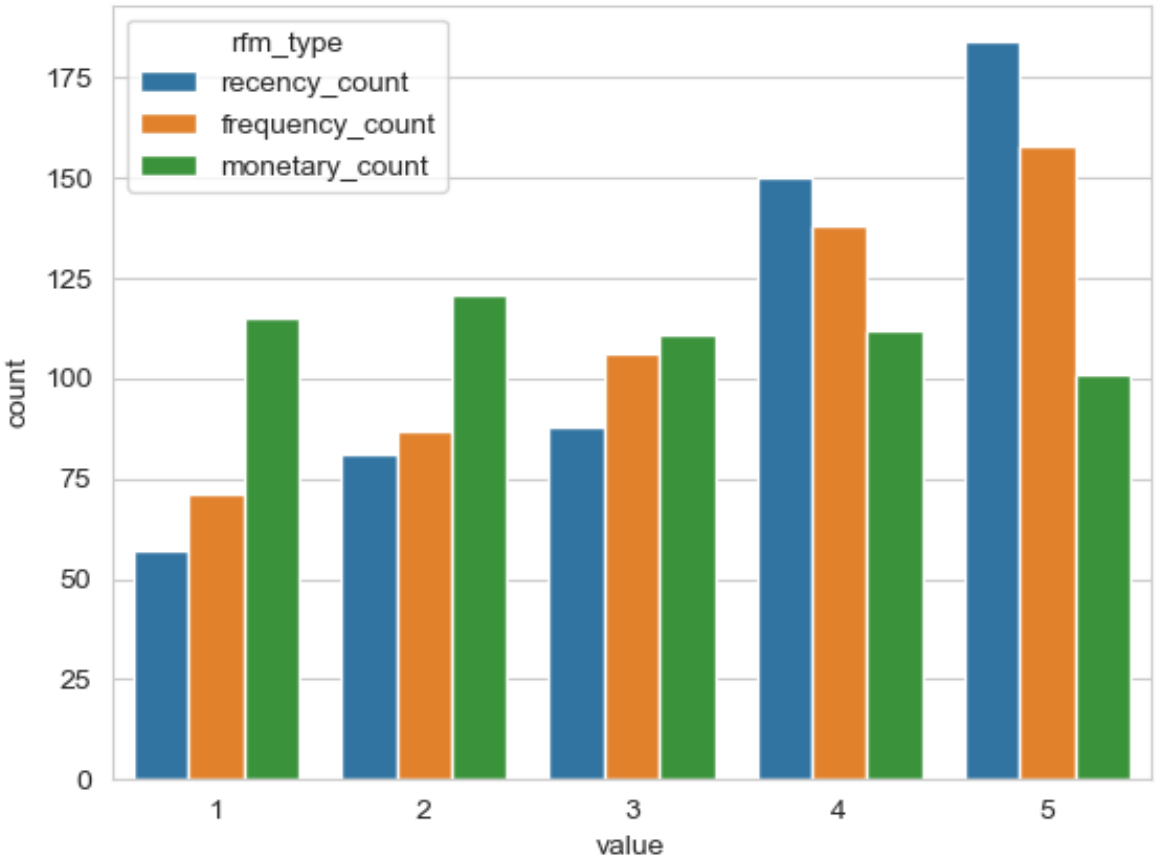

FIGURE 37. Distribution of the RFM-scores for promotion-sensitive customers based on the promotion skewed MInteraction similarity metric.

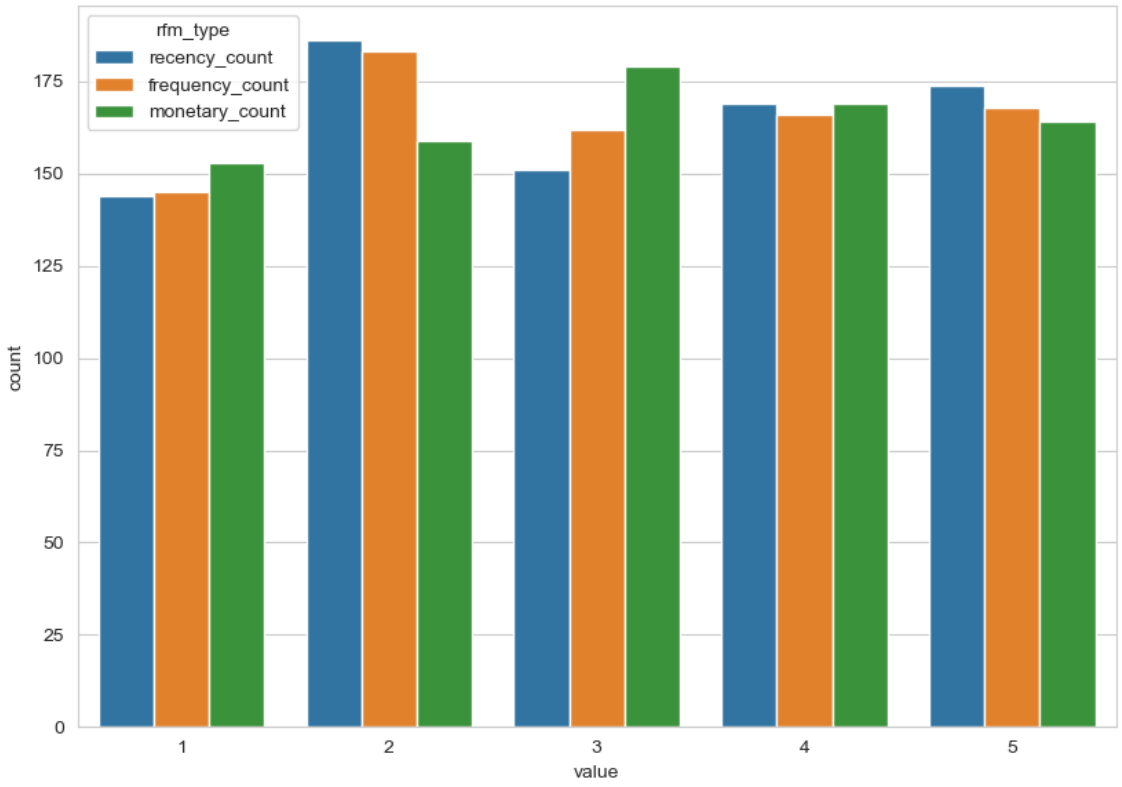

FIGURE 38. Distribution of the RFM-scores for online-sensitive customers based on the normal MInteraction similarity metric.

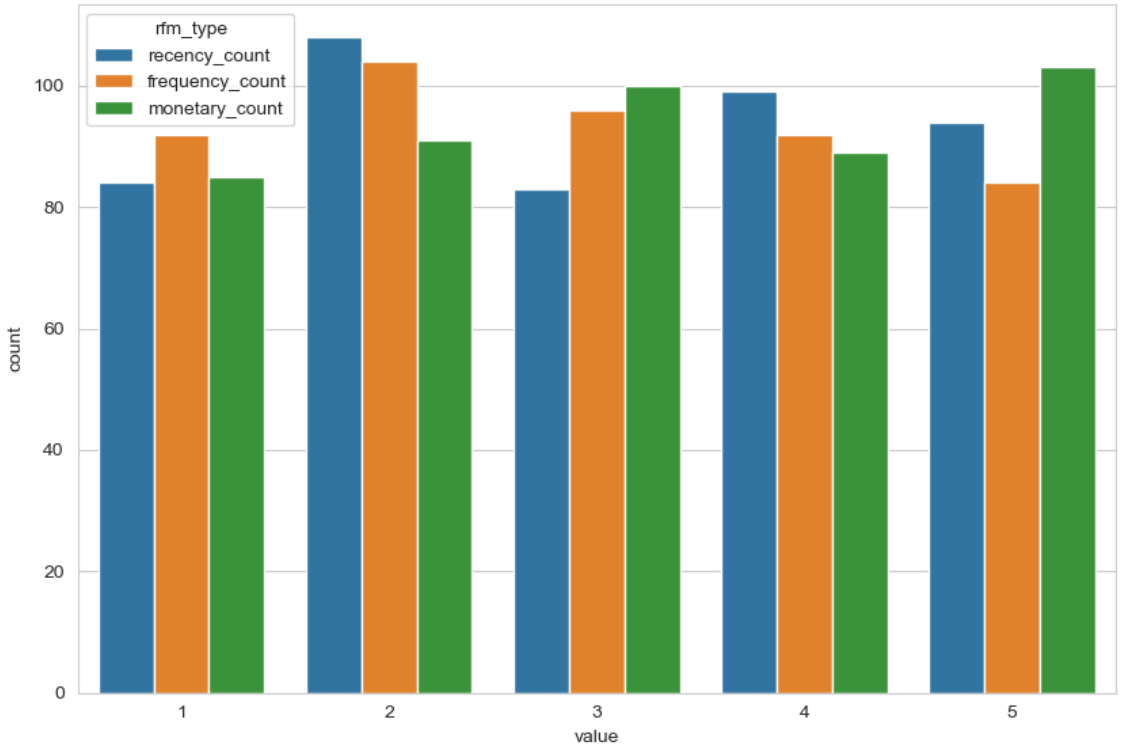

FIGURE 39. Distribution of the RFM-scores for online-sensitive customers based on the online skewed MInteraction similarity metric.

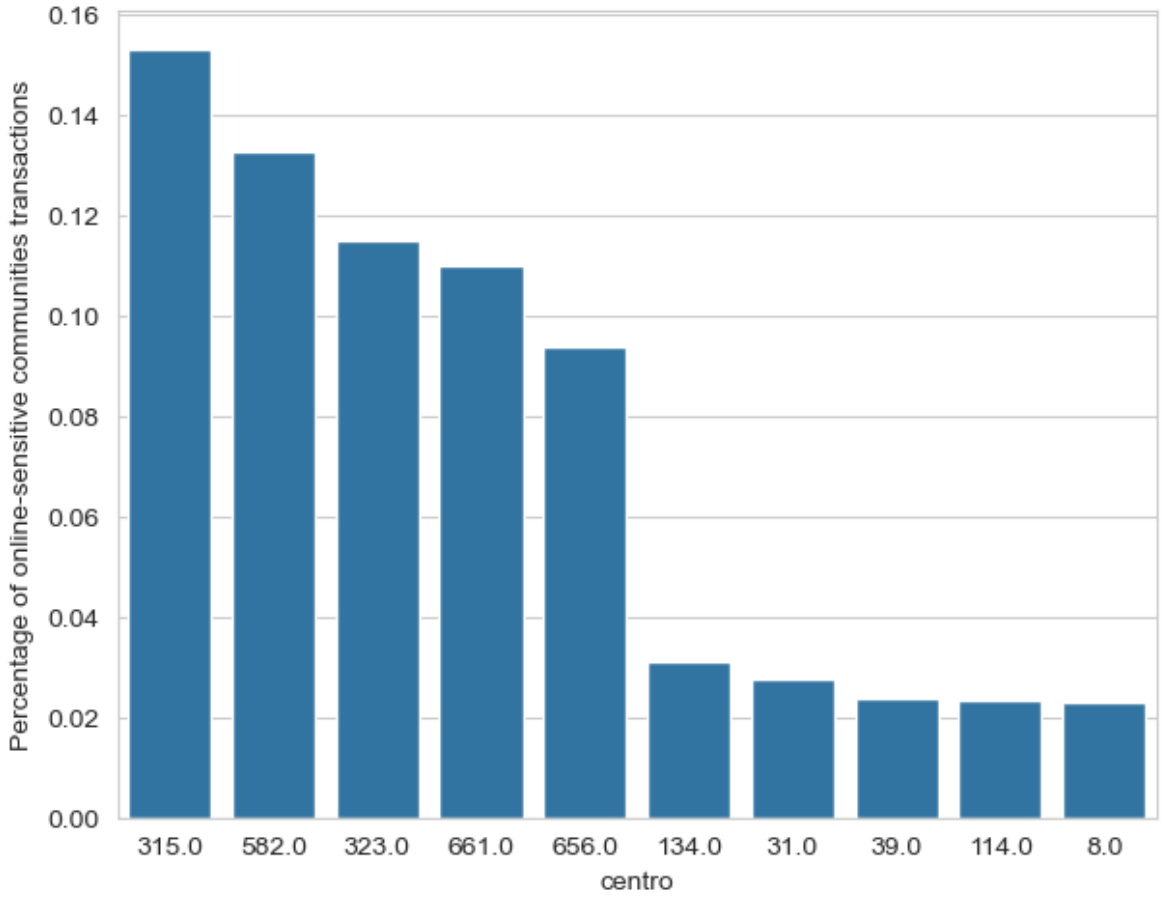

FIGURE 40. The 10 supermarkets with the most transactions of the online-sensitive communities based on the normal MInteraction.

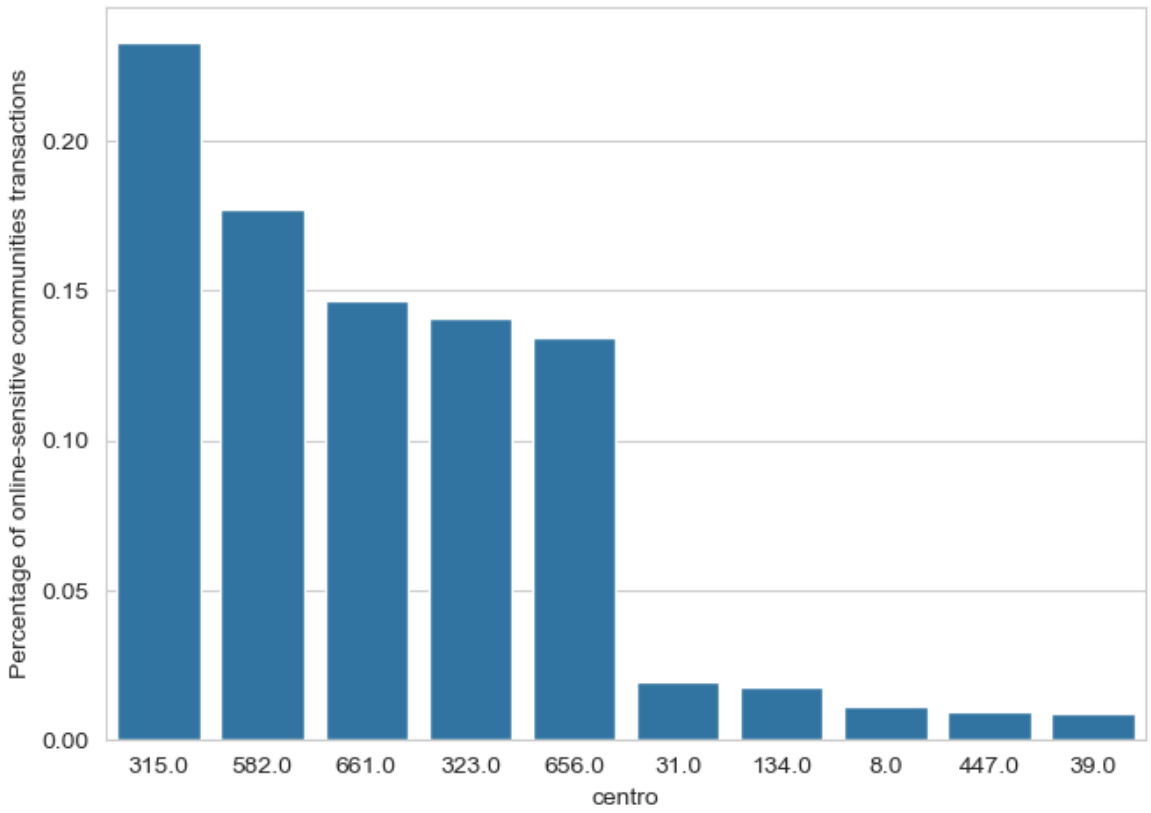

FIGURE 41. The 10 supermarkets with the most transactions of the online-sensitive communities based on the online skewed MInteraction.

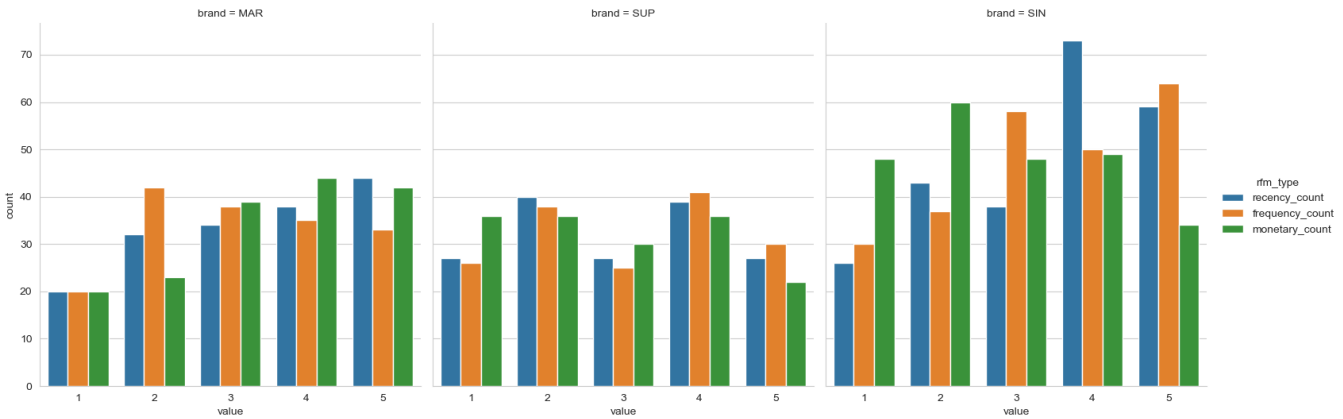

FIGURE 42. Distribution of the RFM-scores for customers sensitive to a specific brand type based on the normal MInteraction similarity metric.

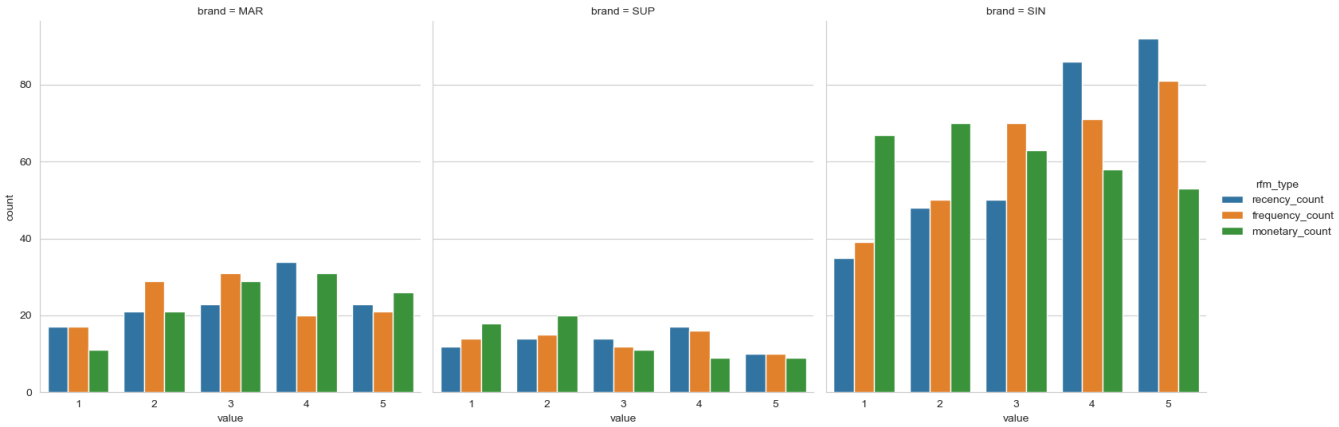

FIGURE 43. Distribution of the RFM-scores for customers sensitive to a specific brand type based on the brand skewed MInteraction similarity metric.