UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Identificeren van de stappen van een exploratieve data-analyse

Seppe Van Daele

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

PROMOTOR :

dr. Gert JANSSENSWILLEN

2020
2021

UHASSELT

KNOWLEDGE IN ACTION

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

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Identificeren van de stappen van een exploratieve data-analyse

Seppe Van Daele

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

PROMOTOR :

dr. Gert JANSSENSWILLEN

Deze masterproef werd geschreven tijdens de COVID-19 crisis in 2020-2021. Deze wereldwijde gezondheidscrisis heeft mogelijk een impact gehad op het schrijf- en verwerkingsproces, de onderzoekshandelingen en de onderzoeksresultaten die aan de basis liggen van dit werkstuk.

Identificeren van de Stappen van een Exploratieve Data-analyse

Seppe Van Daele, *Universiteit Hasselt, Handelsingenieur in de Beleidsinformatica*
Promotor dr. Gert Janssenswillen

Abstract - Welke stappen er doorlopen moeten worden om een exploratieve data-analyse uit te voeren is nog niet gedetailleerd beschreven in de literatuur. Door de stijging van de hoeveelheid data [5][17][18] wordt het analyseren van data echter wel een belangrijke vaardigheid in de huidige maatschappij [2][3]. Binnen deze thesis werd er aan de hand van een tweeledig experiment onderzocht welke stappen er genomen worden tijdens een exploratieve data-analyse, in welke volgorde deze handelingen genomen werden en welke verschillen er zijn tussen beginnende en gevorderde data analisten. Er werden achttien handelingen geïdentificeerd. De belangrijkste handelingen zijn het zoeken van de variabele, het voorbereiden van de data, het samenvatten van data en het opstellen van een grafiek. Tussen de gevonden handelingen werd geen eenduidige volgorde geïdentificeerd. Het uitvoeren van de data-analyse gebeurde volgens een eerder iteratief proces waarbij er stapsgewijs naar de oplossing werd toegewerkt. Op het vlak van de werkwijzen van beginnende en gevorderde data analisten werden er geen relevante verschillen gevonden.

1 Inleiding

Data wordt ook wel het nieuwe goud genoemd [19], maar kan veel beter vergeleken worden met goudrijke grond. Net zoals bij goud moeten er verschillende stappen doorlopen worden om bij de werkelijke waarde te komen. Door een sterke stijging in de hoeveelheid data [5][17][18] is het beheersen van deze data-analyse vaardigheid belangrijker dan ooit.

Hoe dit goud nu het beste uit data gehaald kan worden, is redelijk onbeschreven in de huidige literatuur. Er zijn reeds richtlijnen op welke manier het proces van data-analyse het beste uitgevoerd kan worden [18][22][26]. Echter beschrijven deze stappen vooral op een hoog niveau wat er moet gebeuren en wordt er niet concreet verteld hoe deze het best uitgevoerd kunnen worden. Exploratieve data-analyse is een van de stappen in het data-analyse proces waarbij er gezocht wordt naar waarde uit data [1]. Welke stappen er nu plaatsvinden tijdens een exploratieve data-analyse is tot op heden niet onderzocht.

In deze paper worden de handelingen geïdentificeerd die plaatsvinden tijdens het uitvoeren van een exploratieve data-analyse en wordt de basis gelegd voor een handleiding voor het exploratief analyseren van data. Verder worden deze handelingen ook gebruikt om de werkwijze van een data-analyse expert te vergelijken met deze van een beginner. Dit omdat er verwacht wordt dat een expert betere beslissingen kan nemen door beter ontwikkelde mentale voorstellingen. Deze mentale voorstellingen kunnen op hun beurt waardevol zijn bij het ontwikkelen van nieuwe onderwijsmethodes [23].

In de volgende sectie zal gerelateerde literatuur besproken worden. Sectie 3 beschrijft de probleemstelling en de toegevoegde waarde van het oplossen van dit probleem. In de vierde sectie wordt de gehanteerde methodologie besproken. Sectie 5 bespreekt de gevonden resultaten. Tot slot wordt er in de zesde sectie gereflecteerd op de resultaten met suggesties voor verder onderzoek en wordt er in sectie 7 afgesloten met de conclusie.

2 Gerelateerde literatuur

Binnen deze sectie zal er begonnen worden met te overlopen welke inzichten er reeds verworven zijn binnen het proces van data-analyse. Vervolgens zal er, in de subsecties *cognitieve belasting* en *deliberate practise & mentale voorstellingen*, onderbouwd worden waarom het identificeren van kleinere handelingen en de volgorde tussen deze handelingen voordelig kan zijn.

2.1 Data-analyse

Zoals aangehaald in de inleiding zijn er in de literatuur reeds een aantal stappen gedefinieerd die doorlopen moeten worden tijdens het uitvoeren van een data-analyse [18][22].

1) Het verzamelen van data
2) Het verwerken van data
3) Het cleanen van data
4) Een exploratieve data-analyse
5) Een predictieve data-analyse
6) Communicatie van de resultaten

Het verzamelen en verwerken van data zijn de stappen waarin aspecten uit de echte wereld gecapteerd worden en omgezet worden in toegankelijke ruwe data [18]. De

derde stap in dit proces is het *cleanen* van deze dataset ter voorbereiding van de verdere analyses [18]. Tijdens de exploratieve data-analyse wordt gebruik gemaakt van grafieken en statistieken om de data te verkennen en eerste inzichten te verwerven [18]. Vervolgens kan er in de predictieve fase een model gebouwd worden zoals een lineaire regressie. De afsluitende stap in dit proces is het communiceren van de gevonden resultaten [18]. In [26] wordt dit proces al iets gedetailleerder uitgewerkt en toegepast op de taal R. Hier begint het proces met het importeren van data om deze vervolgens te *cleanen* zoals in de derde stap hierboven. De werkelijke data-analyse stap wordt er beschreven door de cyclus van het transformeren, visualiseren en het modelleren van data en is dus reeds iets concreter dan de theoretische exploratieve en prescriptieve data-analyse. De afsluitende communicatie stap is wel dezelfde.

2.2 Cognitieve belasting

Het opdelen van een bepaalde handeling in kleinere stappen kan de cognitieve belasting verminderen bij het uitvoeren van de handeling [25]. Cognitieve belasting is de belasting die optreedt bij het verwerken van informatie [25]. Des te complexer deze informatie is, des te hoger de cognitieve belasting is [25]. Een te hoge cognitieve belasting kan het werkgeheugen overbelasten en zo het leerproces vertragen [24]. Door een handleiding op te stellen wordt er ingespeeld op *The Isolated Elements Effect* waarbij er een reductie plaatsvindt in de cognitieve belasting door stappen te isoleren en daarna pas het grotere geheel te bekijken [25]. In *The Structured Process Modeling Theory* werd deze theorie ook toegepast om de cognitieve belasting, bij het opstellen van een procesmodel, te verminderen [4]. Deelnemers die gestructureerd te werk gingen, en zo hun cognitieve belasting verminderden, maakten over het algemeen minder syntax fouten en creëerden betere procesmodellen [4]. Dat het volgen van gestructureerde stappen waardevol kan zijn werd ook reeds aangetoond in het domein van process mining [15]. In het, door deze paper [15], uitgevoerde experiment werd er gevraagd aan de deelnemers om een event log te bouwen. De testgroep kreeg de event log building handleiding uit [14] ter beschikking. Uit de resultaten bleek dat de, door de testgroep gebouwde, event logs op meerdere vlakken beter presteerden dan die van de controlegroep [15].

2.3 Deliberate practise & mentale voorstellingen

Een bijkomend voordeel van het identificeren van kleinere stappen is dat deze stappen gebruikt kunnen worden bij het opstellen van een *deliberate practise*. Een training is een *deliberate practise* wanneer er aan vier voorwaarden voldaan wordt [8][10]:

1) Taken met een gedefinieerde doelstelling
2) Directe feedback op de gemaakte taak
3) Mogelijkheid om deze taak meermaals te herhalen
4) Motivatie om ook werkelijk beter te worden

Karl Ericsson, een Zweedse psycholoog [10], bestudeerde wat de trainingen van verschillende experten in hun domein gemeen hadden [9]. Hieruit ontstond het concept *deliberate practise*. Dit concept werd bijvoorbeeld al succesvol toegepast in [17] waarbij de, naar *deliberate practise* herwerkte, fysica cursus resulteerde in een hogere aanwezigheid en betere cijfers. Indien er dus kleine taken geïdentificeerd kunnen worden in het data-analyse proces kunnen deze gebruikt worden voor het opstellen van een *deliberate practise*.

Naast het bestuderen van wat soort training experten gebruiken om hun expertise te verwerven, bestudeerde Karl Ericsson ook waarom experten beter zijn in een bepaald domein dan anderen. Hij concludeerde dat experten verfijndere mentale voorstellingen hebben waardoor ze in staat zijn om betere en/of snellere beslissingen te nemen [7][8][10]. Mentale voorstellingen zijn interne modellen over bepaalde informatie die verfijnder worden door te trainen [10]. Het identificeren van handelingen die genomen worden bij een data-analyse kan ook helpen bij het in kaart brengen van mentale voorstellingen van data-analyse experten. Dat kan gebeuren door de werkwijze van experten te vergelijken met beginners. Weten waarom een expert op een bepaald punt een bepaalde handeling uitvoert kan een positief effect hebben op de ontwikkeling van de mentale modellen van beginners. Het gebruiken van mentale voorstellingen van experten werd in [23] zelfs als een cruciale eerste stap beschouwd bij het opstellen van nieuwe onderwijsmethodes.

3 Probleemstelling

Het transformeren, visualiseren en het modelleren van data [26] zijn allemaal vrij ruwe stappen en geven weinig inzicht in wat er nu juist moet gebeuren. Het doel van deze thesis is om deze stappen te verfijnen en kleinere handelingen, die plaatsvinden tijdens dit proces, te identificeren. Binnen deze thesis werd er gezocht naar handelingen binnen de exploratieve data-analyse stap. De onderzoeksvraag luidt als volgt:

Welke stappen worden er doorlopen tijdens het uitvoeren van een exploratieve data-analyse?

Daarnaast werd er ook aandacht gegeven aan de volgorde waarin deze handelingen uitgevoerd werden. Doordat veel beginnende data analisten niet altijd even goed weten welke stap er moet en kan genomen worden, kan een handleiding zoals bij [15] meer structuur in dit proces brengen en bijkomend de cognitieve belasting verminderen [25].

Verder is er voorlopig slechts weinig onderzoek gedaan naar hoe data-analyse het beste aangeleerd kan worden. Eén manier om dit te bereiken is door middel van *deliberate practise*. Om aan de voorwaarden van *deliberate practise* te voldoen zijn kleinere stappen nodig [8][10]. Daarnaast kunnen deze kleinere handelingen ook gebruikt worden om de werkwijzen van beginners met experts te vergelijken en zo de mentale modellen van experts te identificeren. Een bijkomend doel van deze thesis is dan ook om reeds verschillen tussen beginners en experts te identificeren om zo op termijn betere onderwijsmethodes te kunnen ontwikkelen [23].

4 METHODOLOGIE

4.1 Experiment

Het analyseren van data is voor een deel een cognitieve taak. Niet alle stappen die genomen worden tijdens het analyseren zijn dus zichtbaar [27]. Het analyseren van cognitieve taken wordt in de literatuur Cognitive Task Analysis (CTA) genoemd [27]. CTA helpt erbij om verborgen stappen in de werkwijze van een deelnemer bloot te leggen.

Omdat bepaalde stappen als vanzelfsprekend gezien kunnen worden vanwege de door hen ontwikkelde mentale voorstellingen en automatismen [12] werd er gekozen om een experiment uit te voeren gecombineerd met een afsluitend interview. In dit experiment werd er gevraagd aan de deelnemers om enkele eenvoudige analyses te maken aan de hand van aangeleverde data en van dit proces een schermopname te maken. Vervolgens werd aan de deelnemers gevraagd om in het interview stap voor stap uit te leggen welke keuzes en stappen er genomen werden. Dankzij deze interviews kon er zo meer informatie geregistreerd worden dan zuiver de schermopname van de data-analyse zelf. Deze interviews werden ook opgenomen en vervolgens getranscribeerd.

Er werd gekozen om het interview te laten plaatsvinden na het uitvoeren van de data-analyse om niet te interfereren met de gebruikelijke werkwijze van de deelnemers. Zo had het stellen van vragen voor of tijdens de data-analyse ervoor kunnen zorgen dat deelnemers zouden gaan twijfelen, vertragen of zelfs andere keuzes zouden maken.

De beoogde doelgroep voor dit experiment waren deelnemers uit drie verschillende groepen: bachelorstudenten, masterstudenten en doctoraatsstudenten. Drie groepen met een verschil in ervaring laat hopelijk toe om reeds verschillen in mentale voorstellingen te identificeren tussen de deelnemers. Deze studenten werden gezocht binnen de richting handelsingenieur in de beleidsinformatica aan de Universiteit Hasselt. Van deze studenten is bekend dat ze een inleidende cursus ontvingen omtrent data-analyse in de eerste bachelor. Binnen dit vak werd er gewerkt met de taal R. Bijgevolg werd het gebruik van R bij het oplossen van de opdracht als randvoorwaarde meegegeven om consistente, vergelijkbare resultaten te bekomen.

Inleidende enquête
Alvorens de deelnemers aan de data-analyse begonnen, werd er gevraagd om een inleidende enquête in te vullen. Het doel van deze enquête was om inzicht te krijgen in hun eigen perceptie van hun data-analyse vaardigheid (in R). De gestelde vragen kunnen in Bijlage 1 gevonden worden.

Data-analyse opdracht
De exploratieve analyse gebeurde in de programmeertaal R. Er werd aan de deelnemers gevraagd hun beeldscherm te capteren gedurende de volledige data-analyse. Deze opname werd vervolgens tijdens het interview gebruikt om de genomen stappen te overlopen en om de tijdsduur van elke activiteit te registreren. De opdracht bestond uit drie onafhankelijke vragen over data van een huizenmarkt. De opdracht kan in Bijlage 2 gevonden worden.

De data-analyse opdracht werd ook beoordeeld. Deze beoordeling zal meegenomen worden in de analyse om na te gaan in welke stap er het meeste fouten werden gemaakt en of bepaalde groepen minder goed hebben gepresteerd dan anderen. Voor de beoordeling van elke vraag werd er gebruik gemaakt van twee binaire beoordelingscriteria: de juistheid van de input data en de juistheid van de analyse. De juistheid van de input data bepaalde of de subset die gebruikt werd als vertrekpunt (om een statistiek van te berekenen of een grafiek van op te stellen) juist was. De juistheid van de analyse beoordeelde de uitwerking van de analyse, of de berekende statistiek of grafiek overeenkwam met wat er gevraagd werd.

Afsluitend interview
In de laatste stap van het experiment werd er een interview georganiseerd. De interviews dienden om te achterhalen welke stappen er, volgens de deelnemers zelf, gezet werden. Er werd gevraagd aan de deelnemers om actief te vertellen welke handelingen er genomen werden en waarom. Deze interviews werden ook opgenomen zodat deze achteraf getranscribeerd konden worden.

4.2 Na het experiment

Transcriberen afsluitende interviews
Het transcriberen van de interviews gebeurde manueel. Doordat de meeste deelnemers actief vertelden over de genomen handelingen is er niet voor een vraag-antwoord structuur gekozen. Indien er toch een vraag gesteld werd, werd deze bij het transcriberen cursief geplaatst tussen twee streepjes.

Coderen en categoriseren
Voor het coderen van de transcripties van de interviews werd er in de eerste iteratie gebruik gemaakt van een combinatie van *descriptive* en *process coding*. Bij *descriptive coding* wordt er gekeken naar zelfstandige naamwoorden die de inhoud van de zin capteren [20]. *Process coding* probeert dan weer acties te registreren door voornamelijk actiegerichte (werk)woorden te coderen [20]. Deze codeertechnieken werden op de transcripties toegepast door de woorden en zinnen die hieraan voldeden te markeren. In een tweede iteratie werd gebruik gemaakt van *open coding* (ook wel *initial coding* genoemd) waarbij de gemarkeerde codes uit de eerste iteratie gegroepeerd werden met gelijkende, gemarkeerde codes [13][21]. Deze iteraties werden na elkaar uitgevoerd voor eenzelfde transcriptie alvorens te beginnen met de volgende transcriptie. Hieronder kan een voorbeeld gevonden worden van deze codeerstrategieën.

- *Descriptive coding*
  "... in die **data description** ben ik gaan opzoeken . . . "
- *Process coding*
  "... in die data description ben ik gaan **opzoeken** . . . "
- *Open/initial coding*
  Opzoeken variabele in de data description

Na de *open coding* iteratie werden de bekomen codes van al de transcripties samengevoegd. Deze codes waren de input voor het construeren van de categorieën. Hierbij werden de codes die hetzelfde doel hadden samengenomen en codes met een gelijkend doel werden gegroepeerd en kregen een overkoepelend begrip. Deze codeerstap wordt in de literatuur *axial coding* genoemd [13]. De uitkomst van deze stap vormde reeds een verzameling van handelingen die plaatsvonden en beantwoordde zo de centrale onderzoeksvraag. Verder was er ook interesse in de volgorde van de handelingen alsook in de verschillen in de werkwijze tussen de verschillende groepen. Om op deze vragen een antwoord te formuleren werd er een event log opgesteld en geanalyseerd.

Event log opstellen
De gevonden handelingen werden gebruikt om een event log op te stellen van de analyses van de deelnemers. Hierbij werd er gebruik gemaakt van zowel de schermopname als de transcriptie om zo elke genomen handeling juist te classificeren. Van elke handeling werden de volgende attributen opgeslagen [14]:

- *Case identifier*: elke deelnemer kreeg zijn eigen identifier.
- *Timestamp*: het tijdstip waarop de handeling eindigde, de duur van de handeling wordt bepaald door het tijdstip van de vorige handeling en de huidige handeling.
- *Question*: de vraag waarmee de deelnemer op dat moment bezig was (1-3). Andere activiteiten die los stonden van het oplossen van een vraag kregen een 0 voor dit attribuut.
- *Activity*: de activiteit die uitgevoerd werd.
- *Subactivity*: sommige activiteiten waren een overkoepelend begrip voor een groep activiteiten. In dit attribuut werd er een subactiviteit toegevoegd indien beschikbaar.

Een voorbeeld uit deze event log kan gevonden worden in Tabel 1. Daarnaast werd er, indien van toepassing, ook nog extra informatie opgeslagen om de data te verrijken zoals de locatie waar een bepaalde handeling werd uitgevoerd (bv. in de console, in de *environment* van de ontwikkelomgeving), wat er juist gebeurde in de handeling (bv. waarop er gefilterd werd) en dan nog een attribuut hoe dit gebeurde (bv. een variabele zoeken aan de hand van *control + f*).

Event log analyseren
Vervolgens werd deze event log geanalyseerd, voornamelijk om de frequenties en de volgorde van elke handeling te analyseren en te identificeren. Verder werden er ook verschillen gezocht tussen de analyses van de verschillende deelnemersgroepen. Voor deze analyse van de event log werd gebruik gemaakt van het R pakket bupaR [16]. Doordat er relatief weinig *cases* aanwezig waren in de event log bestond deze analyse ook voor een groot deel uit het kwalitatief analyseren van de individuele *traces*.

4.3 Deelnemers

De beoogde doelgroep werd reeds besproken. Uiteindelijk werden elf studenten overtuigd om deel te nemen aan dit experiment. 2 van de 11 deelnemers kwamen uit de bachelor handelsingenieur in de beleidsinformatica en hadden dus slechts beperkte ervaring met R. 4 andere deelnemers zaten in hun tweede master handelsingenieur in de beleidsinformatica en verwierven dus reeds een viertal jaar ervaring met R. De resterende 5 deelnemers zijn doctoraatsstudenten en hebben meer ervaring met data-analyse dan de andere twee deelnemende groepen. De elf deelnemers voerden elk de volledige analyse van drie opdrachten uit en dus werden er resultaten van 33 opdrachten verzameld.

5 RESULTATEN

In het eerste deel van de resultaten zullen de gevonden handelingen uitvoerig besproken worden. De handeling, wat deze handeling inhoudt en eventuele sub-handelingen komen aan bod. Verder wordt er ook gekeken naar de relatie tussen deze handelingen. Tot slot zullen in het tweede deel van deze sectie de verschillen besproken worden tussen de deelnemersgroepen enerzijds en de vragen anderzijds.

5.1 Handelingen

De stappen die doorlopen worden tijdens een exploratieve data-analyse kunnen opgedeeld worden in vier categorieën: de voorbereidende stappen, de analyse stappen, de debugstap en tot slot de handelingen die niet tot een categorie behoren, maar wel tijdens het hele analyse proces kunnen plaatsvinden. Welke handelingen zich in deze categorieën bevinden en de onderlinge relatie tussen deze handelingen worden in komende secties besproken.

5.1.1 Voorbereidende stappen

Handelingen zijn voorbereidende stappen indien ze plaatsvonden voorafgaand aan de werkelijke analyse zelf. Hierbij werden de handelingen geselecteerd die een hogere relatieve frequentie hadden bij de handelingen die voor de eerste vraag uitgevoerd werden dan tijdens de analyse. Enkel het lezen van de opdracht werd hier ook aan toegevoegd, vermits acht van de elf deelnemers hiermee begonnen en deze stap dus als een voorbereidende en als een analyse stap gezien kan worden.

Opdrachten lezen
Het lezen van de opdracht is een vrij voor de hand liggende stap, maar daarom niet minder belangrijk. Deze omvat het bestuderen van een opdracht of vraag. Het bekijken van data zelf of data gerelateerde bestanden zoals een data beschrijving behoort niet tot deze categorie. Deze activiteit werd door alle deelnemers uitgevoerd als voorbereidende stap en als de eerste stap van het oplossen van elke opdracht. Tien van de elf deelnemers voerde deze handeling uit vooraleer de data in te laden en te verkennen. Dat heeft als voordeel dat de data reeds iets gerichter verkend kan worden. De deelnemer die dit niet deed gaf aan dat het in dit specifieke experiment voordeliger (qua snelheid) was geweest.

Case	Time	Question	Activity	Subactivity
9	8:12	2	Data voorbereiden	Data filteren
9	8:35	2	Grafiek opstellen	
9	8:36	2	Code uitvoeren	

Tabel 1: Voorbeeld event log lijnen

Data inladen
Vooraleer er bewerkingen uitgevoerd kunnen worden op data moet er toegang zijn tot deze data. Analoog aan het ophalen van gegevens in bijvoorbeeld Excel dient de databron ook in R ingeladen te worden. Bij deze handeling werd er onder meer gekeken welk bestandstype de databron had, of er kolomnamen aanwezig waren, wat het scheidingsteken was indien van toepassing en in welke *directory* het databestand aanwezig was. Deze handeling komt overeen met het importeren van data uit [26].

Data controleren
Data controleren was de handeling die de deelnemers ondernamen om na te gaan of de data voldeed aan hun verwachtingen, of de data *tidy* was. Data is *tidy* wanneer elke rij een observatie is en elke kolom een variabele [26]. Ondanks dat dataset b niet voldeed aan de definitie van *tidy* data, werd er maar door één deelnemer teruggegaan naar de *data cleanen* stap uit [18]. De voornaamste stap die tijdens de data controleren handeling genomen werd was het controleren van de datatypes van de kolommen. Deze handeling vond zowel plaats in het begin van de data-analyse als tijdens het debuggen van een error. Het *tidy* maken van dataset b voor een specifieke opdracht wordt binnen deze categorisering toebedeeld aan de *data voorbereiden* handeling die verder nog besproken wordt.

Data verkennen
In deze stap werd er gekeken wat voor data er ter beschikking was. Deze handeling gebeurde door ofwel de data zelf te bekijken in, bijvoorbeeld, de ontwikkelomgeving of Excel, ofwel de *data description* te raadplegen. Het doel van deze handeling was voor de deelnemers vooral om voeling te krijgen met de data en te begrijpen rond welk onderwerp er data geanalyseerd moest worden. Waar *data controleren* echt het verkennen is van de kwaliteit van de data, wordt er bij de handeling *data verkennen* gekeken naar de inhoud van de data. Deze stap mag ook niet verward worden met de stap *exploratieve data-analyse* uit te voeren op het hoge niveau zoals beschreven in [18].

Pakket inladen
Een pakket is een verzameling van functies die niet behoren tot de basisfunctionaliteit van R. Deze zijn vaak ontwikkeld door derden en bieden zowel nieuwe functionaliteit als makkelijkere of snellere basisfunctionaliteit aan. In R moeten pakketten ingeladen worden vooraleer ze gebruikt kunnen worden. Deze handeling werd dan ook door alle deelnemers uitgevoerd. Het laden van een pakket gebeurde door zes van de elf deelnemers voorafgaand aan hun analyse, waarbij vijf van de zes kwamen uit de groep van de doctoraatsstudenten. Daarnaast vonden er nog twee andere pakket gerelateerde handelingen plaats bij de deelnemers uit dit experiment: het opzoeken en het installeren van pakketten. Sommige deelnemers wisten wel nog een functie, maar niet het bijbehorende pakket. Anderen wilden pakketten inlezen die nog niet geïnstalleerd waren en dus moesten deze eerst geïnstalleerd worden vooraleer deze ingeladen konden worden.

5.1.2 Analyse van de voorbereidende stappen

In Tabel 2 kan gezien worden hoeveel van de elf deelnemers deze handelingen voorafgaand aan de analyse uitvoerden. In de laatste kolom kan de absolute frequentie van deze handeling over alle deelnemers heen gevonden worden.

Figuur 1 visualiseert de relaties tussen de voorbereidende stappen. Hierin kan gezien worden in welke volgorde deze handelingen werden uitgevoerd door de deelnemers. Zoals zichtbaar op de figuur is er geen eenduidige volgorde tussen deze stappen aanwezig. In acht van de elf gevallen werd er wel begonnen met het lezen van de opdracht en was het inladen van de pakketten bij meer dan de helft van de deelnemers de laatst uitgevoerde handeling in de voorbereiding. Opvallend zijn wel de vele directe herhalingen van eenzelfde handeling bij onder andere *data inladen*, dat is echter logisch te verklaren doordat er twee databestanden waren die ingeladen moesten worden en deze apart gelogd werden.

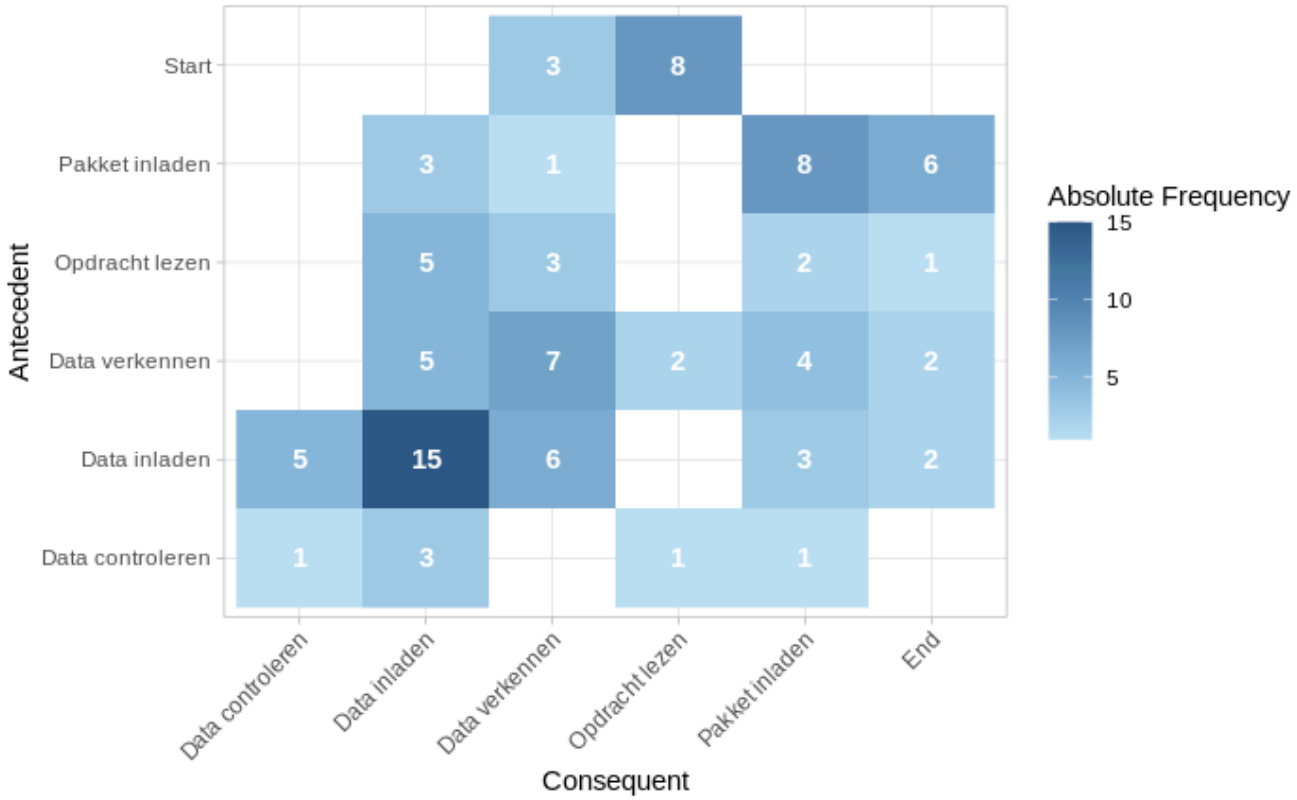
5.1.3 Analyse stappen

Eenmaal de voorbereiding gebeurd was, kon de analyse beginnen. De stappen die binnen deze categorie aan bod komen zijn handelingen die uitgevoerd kunnen worden om een specifieke opdracht te vervullen. Dit zijn handelingen rechtstreeks gerelateerd aan het oplossen van de data-analyse opdracht en niet bijvoorbeeld noodgedwongen handelingen die uitgevoerd moeten worden zoals het oplossen van een errormelding.

Opdracht lezen
Analoog aan het lezen van de opdracht voorafgaand aan de analyse werd deze ook door elke deelnemer een tweede keer gelezen voor het oplossen van de opdracht zelf. Dit is de eerste stap in het oplossen van een opdracht. In 27 van de 33 opdrachten was dit dan ook de eerste stap. De andere zes gevallen waren telkens bij de eerste opgave en toen zat mogelijks de opgave nog helder van het lezen van de opdrachten als voorbereidende stap. Er kan dus geconcludeerd worden dat er altijd best bij deze stap begonnen wordt. Het belang van deze stap valt ook niet te onderschatten vermits vier van de zes gemaakte fouten voortvloeide uit het niet volledig antwoorden op de vraag.

Handeling	Deelnemers	Minimum	Gemiddelde	Maximum	Totaal
Opdracht lezen	11	1	1	1	11
Data inladen	11	2	2,81	6	31
Data controleren	4	0	0,55	2	6
Data verkennen	11	1	1,82	4	20
Pakket inladen	6	0	1,64	4	18

Tabel 2: Samenvattende statistieken voorbereidende stappen



Figuur 1: *Precedence matrix* van de voorbereidende stappen

Variabele zoeken
Het zoeken naar een bepaalde gevraagde variabele in de data is cruciaal om vragen juist te kunnen oplossen. Deze handeling kwam vaak voor vermits de deelnemers niet bekend waren met de data. Deelnemers gingen in dit experiment ofwel in de description file ofwel in de data zelf op zoek naar de variabelen die ze nodig hadden om de opdrachten te kunnen vervullen. Eén deelnemer maakte in deze handeling een fout en identificeerde bij de tweede opdracht een foute variabele.

Data voorbereiden
De handeling data voorbereiden beslaat het voorbereiden van de data voor een specifieke opdracht. Deze voorbereiding kan bestaan uit verschillende sub-handelingen. In dit experiment werden er negen geïdentificeerd.

- *Data groeperen/aggregeren*: indien er voor de opdracht gekeken moet worden naar geaggregeerde statistieken of categorieën werd deze handeling uitgevoerd.
- *Data filteren*: deze handeling vond plaats als er aan een bepaalde conditie voldaan moest worden en komt overeen met het selecteren van rijen in de data. Er werd naar deze stap vaak gerefereerd als data verwijderen, waarbij verwijderen verwees naar de rijen die niet geselecteerd werden en zo uit de subset 'verwijderd' werden.
- *Data selecteren*: het selecteren van data is de handeling die overeenkomt met het selecteren van kolommen (in tegenstelling tot het selecteren van rijen bij data filteren). Deze handeling werd buiten het louter functioneel aspect ook vaak gebruikt om overzicht te verwerven door overbodige kolommen buiten beschouwing te laten.
- *Data pullen*: data pullen is een specifieke variant van het selecteren van data waarbij de geselecteerde kolom automatisch omgezet wordt in een vector. Deze vector kan op zijn beurt dan weer gebruikt worden om bijvoorbeeld te filteren aan de hand van de %IN% functie. De pull functie werd wel door slechts één deelnemer gebruikt.

- *Data joinen*: indien er meerdere datasets zijn kan het soms nodig zijn om deze datasets te koppelen bijvoorbeeld indien er gevraagd wordt naar relaties tussen kolommen uit verschillende datasets. Deze koppeling gebeurt aan de hand van een gemeenschappelijke kolom. Het zoeken naar deze gemeenschappelijke kolom werd dan ook enkele keren aangehaald als een genomen handeling.
- *Data spreaden en gatheren*: het spreiden en samennemen van een dataset is de handeling die uitgevoerd wordt om de data *tidy* te maken. In dit experiment kwam wel enkel het spreiden van de data voor vermits de *untidy* dataset b door middel van deze functie terug *tidy* gemaakt kon worden.
- *Data muteren*: is de activiteit waarbij er een kolom wordt toegevoegd met nieuwe inzichten of berekeningen die ontstaan uit de data. Bijvoorbeeld een som van twee andere variabelen.
- *Datatype veranderen*: bij deze handeling wordt het datatype van een kolom aangepast. Als een kolom een categorie is kan deze als een factor opgeslagen worden.
- *Variabele aanmaken*: dit is een handeling waarbij sommige deelnemers vertelden dat het niet noodzakelijk was, maar wel handig om structuur te creëren, maar ook om de oude variabele niet te verliezen. Het begrip variabele slaat hier op een variabele in R en niet in de data en kan bijvoorbeeld een subset zijn die daarna de input vormt voor een grafiek.

Het voorbereiden van data is de stap voordat er visualisaties worden gemaakt of statistieken worden berekend en is altijd een input voor een - eerder afwerkende - stap. *Data voorbereiden* bevat een grote analogie met *data cleanen* maar verschilt in het feit dat *data cleanen* handelingen zijn die toegepast worden op heel de dataset voorafgaand aan alle analyses [18]. Deze voorbereiding lijkt ook erg op de *transform* stap uit [26] alleen bevat deze transformeer stap ook het berekenen van statistieken, een stap die als aparte handeling *data samenvatten* in dit experiment geïdentificeerd werd.

Data samenvatten
In deze stap wordt er vertrokken van ofwel de dataset ofwel een subset die ontstond na de handeling *data voorbereiden*. Vertrekkende van deze dataset kunnen er statistieken berekend worden zoals het gemiddelde of de frequentie. Bij deze stap werd vaak de functie *summarise* gebruikt. Deze handeling werd door alle deelnemers gebruikt bij de derde opdracht waar het gemiddelde gevraagd werd.

Grafiek opstellen
In plaats van statistieken van de invoerdata te berekenen kunnen er ook grafieken opgesteld worden. Net zoals bij *data samenvatten* is het ook hier de bedoeling om informatie uit de data te halen. De voornaamste stappen in deze handeling zijn het bepalen van de assen en het type grafiek.

Grafiek opmaken
Deze handeling hoort grotendeels bij het opstellen van een grafiek. Deze handelingen kunnen dan ook niet volledig losgekoppeld worden, een grafiek kan niet opgemaakt worden als er geen grafiek opgesteld werd. Een grafiek kan daarentegen wel opgesteld worden zonder die daarna op te maken. De deelnemers beschreven het toevoegen van een opmaak als een optionele stap die de grafiek kan verduidelijken en zeker aangeraden is als deze gebruikt zou worden voor externe communicatie. Stappen die binnen deze handeling genomen werden waren onder andere het toevoegen van titels en assen, het aanpassen van labels en het aanpassen van de schaal.

Code uitvoeren
Het uitvoeren van de geschreven code is een vanzelfsprekende handeling, maar wel cruciaal om de werking van de code te verifiëren en de output te kunnen analyseren. Deze handeling beschrijft zowel het uitvoeren van een deel van de code als de hele code.

Resultaat evalueren
Na het verkrijgen van een resultaat in de vorm van een tussenstap, een statistiek of een grafiek kan een reflectie op dit resultaat plaatsvinden. Vragen die deelnemers zich stelden waren onder andere:

- *Is dit het resultaat wat ik verwacht?*
- *Beantwoordt het resultaat de vraag?*
- *Zijn er betere manieren om deze vraag te beantwoorden?*

5.1.4 Analyse van de analyse handelingen

In Tabel 3 kan gezien worden dat enkel het uitvoeren van de code door elke deelnemer bij elke vraag als handeling is uitgevoerd en dat buiten het opmaken van de grafiek alle handelingen wel eens over de drie vragen heen werden uitgevoerd door de deelnemers. Naast het identificeren van handelingen werd er geprobeerd om te bepalen in welke volgorde deze uitgevoerd werden. Hierbij werd elke opdracht als een individuele *case* binnen de event log gezien en werden er dus in plaats van 11, 33 *traces* bestudeerd. De relatie tussen de analysestappen kan gezien worden in de *precedence matrix* van Figuur 2. Er is geen eenduidige volgorde van handelen uit af te leiden. Elke *trace* was uniek. In 27 van de 33 gevallen werd er wel begonnen met het lezen van de opdracht. Hoeveel van deze 27 direct daarna op zoek gingen naar een variabele kan niet uit deze matrix afgeleid worden. Hiervoor werd Tabel 4 opgesteld.

Tabel 4 brengt de drie meest frequente handelingen die elkaar direct opvolgden in kaart. Deze opeenvolging van drie handelingen werd gezocht binnen de *traces* die gefilterd werden op zeven van de acht besproken analyse handelingen. De handeling *resultaat evalueren* werd uit deze analyse gelaten omdat deze 167 keer van de 177 keren dat deze handeling voorkwam, direct na het uitvoeren van de code uitgevoerd werd. Verder werden activiteiten die zichzelf opvolgden ook samengenomen om zoveel mogelijk de relatie en de gehanteerde volgorde tussen de verschillende handelingen te identificeren. Zo kan in de eerste strook van Tabel 4 gezien worden dat van de 27 gevallen die begonnen met het lezen van de opdracht er 19 direct daarna op zoek gingen naar de variabele.

Verder kunnen er nog een deel andere bevindingen in deze tabel gevonden worden. Zo valt het op dat na het lezen van de opdracht en het zoeken van de variabele, de data in 25% van de gevallen voorbereid wordt. Uit de vierde strook valt op dat het voorbereiden van data vaak iteratief gebeurde, waarbij de code van de voorbereiding uitgevoerd werd om dan of een herwerking of een verdere voorbereiding uit te voeren. In de praktijk kan dit bijvoorbeeld eerst het *joinen* van twee datasets zijn om vervolgens een filter toe te passen, waarbij tussendoor bekeken wordt of het samenvoegen van die datasets in het gewenste resultaat resulteerde. Dat deze handeling gemiddeld genomen meermaals uitgevoerd werd kan ook in Tabel 3 gezien. Zo werd deze handeling gemiddeld genomen bijna zes keer per opdracht uitgevoerd. Een andere bevinding die uit Tabel 4 gehaald kan worden is dat het uitvoeren van de code en dus het verifiëren van de code, bijna altijd de volgende stap is na het voorbereiden en samenvatten van de data alsook bij het opstellen en opmaken van grafieken. Er werd dus duidelijk stapsgewijs richting de oplossing gebouwd door de deelnemers. In Tabel 3 kan er waargenomen worden dat de handeling *Code uitvoeren* gemiddeld genomen zelfs tien keer per opdracht uitgevoerd werd.

De volgorde die binnen dit iteratief proces gebruikt werd was sterk verschillend bij de drie verschillende opdrachten (die gevonden kunnen worden in de Bijlage 2). Zo werd er bij de eerste twee opgaven een visualisatie gevraagd en bij de derde het berekenen van een gemiddelde. Zonder de handeling *code uitvoeren*, die vaak iteratief terugkwam, zien de processen voor de drie vragen er uit als in Figuur 3. Deze zijn gebaseerd op de proces visualisaties die in Bijlage 3 gevonden kunnen worden.

5.1.5 Debuggen

Debuggen is de derde categorie van handelingen die geïdentificeerd werd. Niet elke handeling verliep foutloos, zo resulteerde het uitvoeren van de code 77 keer van de 377 keer in een error. Debuggen is dan ook een noodgedwongen handeling (of reeks van handelingen) die genomen werd na het verkrijgen van een error of warning. De voornaamste errors en manieren om deze op te lossen zullen in deze sectie besproken worden. In Tabel 5 kunnen de vier errors gevonden worden die 60 procent van de verkregen errormeldingen omvatten.

- De *Object ∼ not found* error had hoofdzakelijk twee oorzaken binnen het experiment:
  - Ofwel was er een typefout aanwezig in het object waardoor deze niet werd gevonden. Bijvoorbeeld SalesPrice in plaats van SalePrice.
  - Ofwel werd het object niet meegegeven aan de methode die het object nodig had. Bijvoorbeeld select(x) %>% group_by(y).
- Ook de error *could not find function* ∼ had twee oorzaken binnen deze analyse:
  - Ofwel werd er gebruik gemaakt van een onbestaande functie. Bijvoorbeeld avg(x) in plaats van mean(x). In dit geval kan er best gezocht worden naar de juiste functie door bijvoorbeeld informatie te raadplegen op het internet.
  - Ofwel hoorde de functie die gebruikt werd tot een pakket dat nog niet ingeladen was. Bijvoorbeeld het *piping* symbool %>% behoort tot het pakket *magrittr* of *dplyr* en kan niet gebruikt worden vooraleer dit pakket is ingeladen.
- De derde meest voorkomende error was indien er een onverwacht teken aanwezig was in de geschreven code, bijvoorbeeld een haakje teveel. Het oplossen van deze error was dan ook meestal vrij eenvoudig.
- Als vierde werden enkele deelnemers geconfronteerd met de error ∼ *does not exist in current working directory* als het pad van de data die ze wilden inladen niet correct was. Bij het debuggen werd er gebruik gemaakt van de functie *getwd()* om te controleren wat de huidige working directory was om zo het bestandspad te kunnen aanpassen.

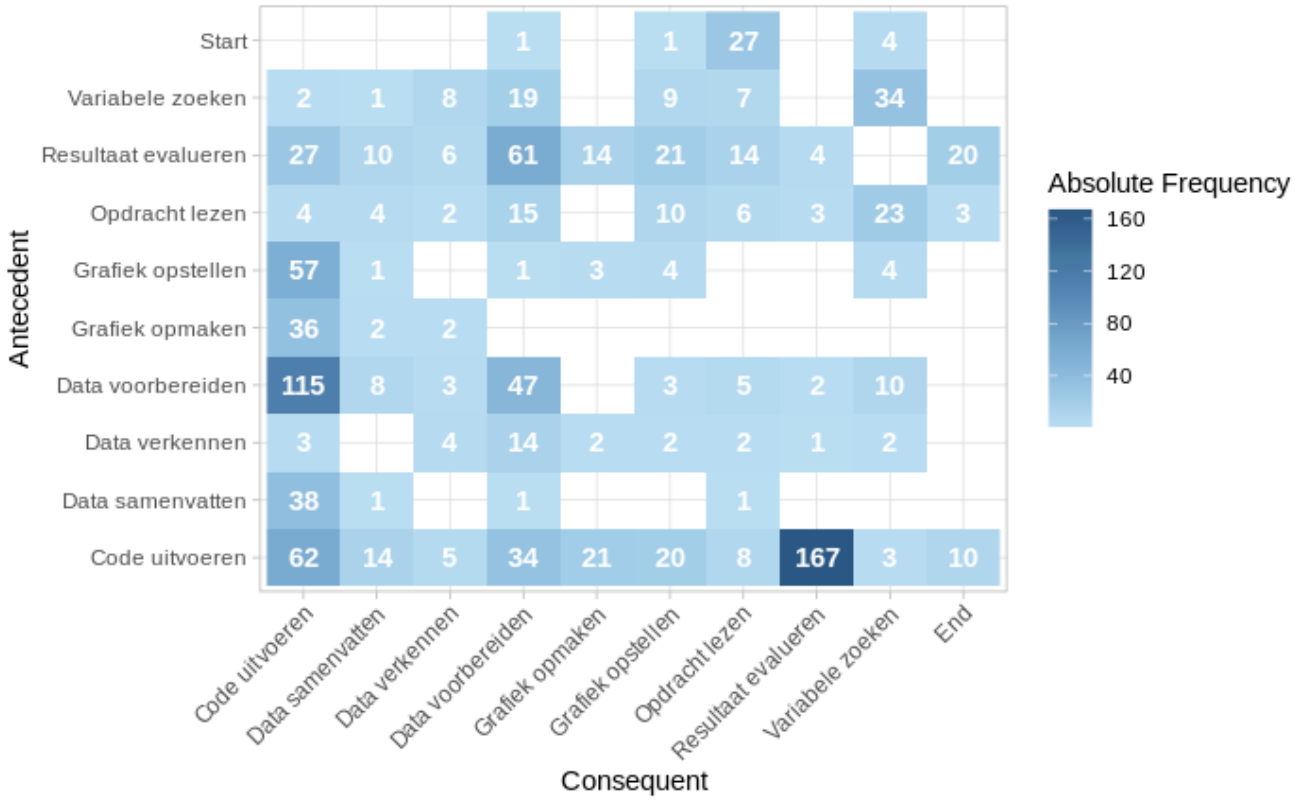
De meeste van deze errors waren vrij triviaal op te lossen. In twintig procent van de geregistreerde loglijnen tijdens het debuggen werd er echter wel bijkomende informatie geraadpleegd op bijvoorbeeld het Internet. Opvallend was wel dat er slechts eenmaal gezocht werd op de verkregen errormelding.

5.1.6 Andere handelingen

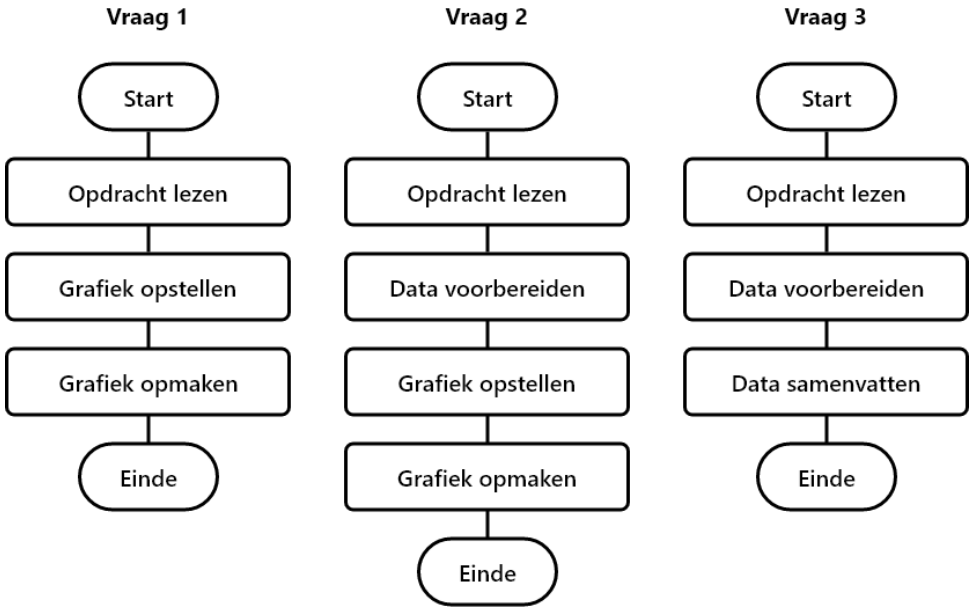
In deze laatste categorie zitten het toevoegen van structuur, nadenken, het nakijken van de opdrachten, het raadplegen van informatie en *trial-and-error*. Op het nakijken van de opdrachten na, die uitgevoerd werd na het vervolledigen van al de opdrachten, zijn deze handelingen redelijk onafhankelijk van de voorgaande handeling en werden ze dus op elk moment in de analyse uitgevoerd. In Tabel 6 kan de frequentie van deze handelingen gevonden worden. *Trial-and-error* werd niet expliciet geregistreerd in de event log en is dus niet aanwezig in deze tabel. Deze handeling zit impliciet vervat in iteraties van bepaalde handelingen.

Structuur toevoegen

Structuur brengen in het analyse bestand, de R-file in dit geval, is een handeling die misschien niet direct gerelateerd wordt aan het analyseren van data, maar wel gebeurde door alle deelnemers. Het doel van deze handeling was om structuur te creëren, overzicht te behouden en de kans op fouten te minimaliseren. Het belang van deze handeling is misschien wel groter dan verwacht: vier van de zes gemaakte fouten vloeiden voort uit onzorgvuldigheid in plaats van onkunde. Stappen die onder deze handeling vallen zijn het noteren van tussenstappen en *comments*, het kopiëren van de vraag (zodat deze altijd zichtbaar was vanaf de plek waar code geschreven werd) en het toevoegen van *chunks* (blokjes code om code van tekst te scheiden in een R markdown bestand). Eén deelnemer vertelde ook fysieke notities genomen te hebben en zo structuur toe te voegen aan de uitvoering van de analyse. Het toevoegen van structuur kan ook helpen om de cognitieve belasting te verminderen [25].



Figuur 2: *Precedence matrix* van de analyse stappen



Figuur 3: Proces per vraag zonder de handelingen *Code uitvoeren* en *Resultaat evalueren*

Handeling	Deelnemers	Vragen	Minimum	Gemiddelde	Maximum	Totaal
Opdracht lezen	11	28	0	2,12	6	70
Variabele zoeken	11	30	0	2,42	6	80
Data voorbereiden	11	23	0	5,85	27	193
Data samenvatten	11	23	0	1,24	6	41
Grafiek opstellen	11	22	0	2,12	11	70
Grafiek opmaken	8	13	0	1,21	8	40
Code uitvoeren	11	33	1	10,42	32	344
Resultaat evalueren	11	32	0	5,36	16	177

Tabel 3: Samenvattende statistieken van de analyse stappen

Eerste	Tweede	Derde	Absolute frequentie	Relatieve frequentie
Start	Opdracht lezen	Variabele zoeken	19	58%
Start	Variabele zoeken	Grafiek opstellen	4	12%
Start	Opdracht lezen	Data voorbereiden	4	12%
Opdracht lezen	Variabele zoeken	Data voorbereiden	15	25%
Opdracht lezen	Data voorbereiden	Code uitvoeren	11	19%
Opdracht lezen	Grafiek opstellen	Code uitvoeren	7	12%
Variabele zoeken	Grafiek opstellen	Code uitvoeren	9	33%
Variabele zoeken	Opdracht lezen	Data voorbereiden	3	11%
Variabele zoeken	Data voorbereiden	Variabele zoeken	3	11%
Variabele zoeken	Data voorbereiden	Data samenvatten	3	11%
Data voorbereiden	Code uitvoeren	Data voorbereiden	47	44%
Data voorbereiden	Code uitvoeren	Opdracht lezen	11	10%
Data voorbereiden	Code uitvoeren	Data samenvatten	11	10%
Data samenvatten	Code uitvoeren	Data voorbereiden	9	24%
Data samenvatten	Code uitvoeren	Einde	9	24%
Data samenvatten	Code uitvoeren	Data samenvatten	6	16%
Data samenvatten	Code uitvoeren	Opdracht lezen	4	11%
Grafiek opstellen	Code uitvoeren	Grafiek opstellen	18	31%
Grafiek opstellen	Code uitvoeren	Grafiek opmaken	10	17%
Grafiek opstellen	Code uitvoeren	Einde	8	14%
Grafiek opstellen	Code uitvoeren	Data voorbereiden	7	12%
Grafiek opmaken	Code uitvoeren	Grafiek opmaken	13	46%
Grafiek opmaken	Code uitvoeren	Einde	7	25%
Grafiek opmaken	Code uitvoeren	Grafiek opstellen	3	11%
Code uitvoeren	Data voorbereiden	Code uitvoeren	56	34%
Code uitvoeren	Grafiek opstellen	Code uitvoeren	28	17%
Code uitvoeren	Grafiek opmaken	Code uitvoeren	20	12%
Code uitvoeren	Data samenvatten	Code uitvoeren	19	12%
Data samenvatten	Code uitvoeren	Einde	9	27%
Grafiek opstellen	Code uitvoeren	Einde	8	24%
Grafiek opmaken	Code uitvoeren	Einde	7	21%
Code uitvoeren	Opdracht lezen	Einde	4	12%

Tabel 4: De sequenties van drie handelingen met een relatieve frequentie van 10% of hoger

Error	#
Object ∼ not found	17
Could not find function ∼	16
Unexpected ∼	8
∼ does not exist in current working directory	6

Tabel 5: De vier meest voorkomende fouten

Nadenken

Het nadenken over het uitvoeren van een opdracht werd ongetwijfeld uitgevoerd door alle deelnemers. Slechts zeven deelnemers haalden actief aan na te denken op bepaalde momenten tijdens de analyse. Binnen deze stap wordt er nagedacht over bijvoorbeeld welke oplossing het beste past voor een desbetreffende vraag.

Handeling	Deelnemers	Minimum	Gemiddelde	Maximum	Totaal
Structuur toevoegen	11	3	6,27	10	69
Nadenken	7	0	1,55	4	17
Opdrachten nakijken	8	0	1,18	3	13
Informatie raadplegen	11	4	20,82	63	229

Tabel 6: Samenvattende statistieken van de resterende handelingen

Opdrachten nakijken
Na het vervolledigen van alle opdrachten controleerden de meeste deelnemers nog al hun oplossingen en of deze voldeden aan de opdrachten.

Informatie raadplegen
Bij de handeling informatie raadplegen werd er telkens een vorm van hulpmiddel geconsulteerd. In dit experiment kwamen er vier soorten voor:

- *Het raadplegen van documentatie*: indien het gebruik van een functie (de syntax, het doel) onduidelijk was, of er gezocht werd naar bepaalde parameters, werd er gebruik gemaakt van de documentatie van R. Dit gebeurde zowel in de ontwikkelomgeving als online.
- *Het raadplegen van voorbeelden*: soms werd er ook gezocht naar voorbeelden om, ofwel de werking van een bepaalde functie nog te verduidelijken, ofwel om op zoek te gaan naar code die reeds een gelijkaardig probleem had opgelost. In dit laatste geval werd deze code regelmatig gekopieerd en aangepast naar de verkregen dataset.
- *Het teruggrijpen naar voorgaande analyses*: dit ligt in lijn met het zoeken naar voorbeelden gelijkend aan de opdracht. De motivatie hierachter was dat de deelnemers wisten wat ze wilden doen en dat ze het reeds eens eerder hadden gedaan, maar dat ze niet meer de precieze syntax wisten. Deze werd dan teruggezocht in voorgaande analyses.
- *Het raadplegen van cursusmateriaal*: analoog aan de redenen waarom er gebruik gemaakt werd van documentatie of voorbeelden grepen sommige deelnemers ook terug naar cursusmateriaal van de R-cursus die elke deelnemer in zijn eerste bachelorjaar ontving.

Deze laatste twee soorten kwamen wel enkel voor bij de deelnemers uit de bachelor die nog het meest recent deze R-cursus hebben gehad.

Trial-and-error
Op bepaalde momenten gaven sommige deelnemers aan dat ze aan het proberen of experimenteren waren. Soms was dit lukraak enkele dingen uitproberen andere keren werd bijvoorbeeld de uitkomst van de verschillende soorten *joins* met elkaar vergeleken. Deze handeling komt overeen met *trial-and-error*. *Trial-and-error* is niet apart gecodeerd in de event log, maar Tabel 7 toont wel hoe dit patroon zich in de event log weerspiegelde. Hier werden binnen een kort tijdsinterval meerdere *joins* met elkaar vergeleken. Het hanteren van een *trial-and-error* werkwijze zou wel in minder leerwinst resulteren dan wanneer er stilgestaan wordt bij een probleem en er nagedacht of gereflecteerd wordt alvorens opnieuw een handeling uit te voeren [6]. *Trial-and-error* is eerder een werkwijze dan een handeling apart, hierdoor kwam deze ook niet onder deze vorm voor in de event log en dus ontbreekt deze in de frequentietabel.

5.2 Andere inzichten

Naast de gevonden handelingen en de relaties tussen deze handelingen werd er nog gezocht naar andere inzichten. Er werd gekeken naar de juistheid, de duur, de frequentie, de structuur en de herhaling van de handelingen. Hierbij werd er gekeken naar mogelijke verschillen over de verschillende vragen en deelnemersgroepen heen.

5.2.1 Juistheid

Om de juistheid van de gemaakte analyses te beoordelen werd er, zoals aangehaald in de methodologie, gebruik gemaakt van twee binaire beoordelingscriteria: de juistheid van de gebruikte input data en de juistheid van de analyse zelf. De score voor een volledig juist gemaakte analyse bedraagt twee punten. In totaal maakten elf deelnemers elk drie opgaven en werden er dus maximaal 66 punten uitgedeeld. Over alle deelnemers heen werden er zes fouten gemaakt, bij deze zes fouten werd niet de juiste invoerdata gebruikt, maar was de analyse zelf wel correct. In Tabel 8 kunnen de gemiddelde resultaten gevonden worden van deze beoordeling. In de laatste kolom is zichtbaar dat iedereen de eerste opdracht correct wist op te lossen en dat de tweede opgave het minst goed gemaakt werd. Vijf van de zes fouten werden gemaakt bij de tweede opgave waarbij vier van deze vijf gerelateerd waren aan het onzorgvuldig lezen van de vraag.

Van de doctoraatstudenten werd er verwacht dat deze beter zouden presteren dan de masterstudenten en deze op hun beurt beter zouden presteren dan de bachelorstudenten dankzij het hebben van beter ontwikkelde mentale voorstellingen. Een betere prestatie kan zich zowel uiten in het sneller bereiken van de oplossing als in het bereiken van een juistere oplossing. Uit de resultaten van het experiment lijkt er een positief verband te zijn tussen de behaalde resultaten en de expertise van de deelnemersgroepen. Echter is het aantal deelnemers te laag om te kunnen spreken van statistisch significante verschillen. De behaalde resultaten liggen wel in lijn met de, door de deelnemers aangegeven, zelfperceptie (op 20) die berekend werd uit de inleidende enquête.

5.2.2 Duur

Binnen deze sectie wordt de duur van het experiment bekeken, voornamelijk de verschillen tussen de deelnemersgroepen en de vragen, zowel op het niveau van het

Case	Time	Question	Activity	Subactivity	What
7	30:39	3	Data voorbereiden	Data joinen	Left join
7	30:40	3	Code uitvoeren		
7	30:45	3	Resultaat evalueren		
7	30:50	3	Data voorbereiden	Data joinen	Right join
7	30:51	3	Code uitvoeren		
7	30:58	3	Resultaat evalueren		
7	31:02	3	Data voorbereiden	Data joinen	Full join
7	31:03	3	Code uitvoeren		
7	31:12	3	Resultaat evalueren		
7	31:29	3	Data voorbereiden	Data joinen	Inner join
7	31:30	3	Code uitvoeren		
7	31:37	3	Resultaat evalueren		

Tabel 7: Voorbeeld *trial-and-error* patroon in de event log

Vraag	Bachelor	Master	Doctoraat	Gemiddelde
1	2	2	2	2
2	1	1,5	1,8	1,55
3	2	2	1,8	1,91
Gemiddelde score /2	1,67	1,83	1,87	1,82
Zelfperceptie /20	12	12,5	16,4	14,18

Tabel 8: Gemiddelde score per vraag en per deelnemersgroep

gehele experiment als op het niveau van de individuele vragen. In Tabel 9 kan de duur en de standaardafwijking per vraag en per deelnemer gevonden worden, uitgedrukt in minuten. Eén van de opmerkelijke bevindingen is de grote standaardafwijking over de hele duur van het experiment heen bij de masterstudenten. Deze bedraagt 24,7 minuten en is beduidend groter dan de standaardafwijking bij de bachelor en doctoraatsstudenten met respectievelijk een standaardafwijking van 1,36 en 4,27 minuten. De standaardafwijking van de masterstudenten is zo groot doordat zowel de twee snelste, als twee traagste deelnemers binnen deze deelnemersgroep zat. Verder valt het op dat de bachelorstudenten duidelijk minder tijd besteden aan de voorbereidende stappen, alsook het nakijken van de opdrachten. Dit verschil heeft mogelijks een verband met de behaalde resultaten en kan een representatie in de data zijn van een mentale voorstelling van de deelnemers met een hogere expertise. Deze tijdsinvestering voorafgaand aan de werkelijke analyse weerspiegelde zich wel niet in bijvoorbeeld een snellere totaaltijd, zo zijn de bachelorstudenten gemiddeld genomen zelfs de snelste deelnemersgroep binnen het experiment. Zoals reeds uit de standaardafwijking en de totaaltijden bleek, werd er niet direct een verband gevonden tussen de deelnemersgroepen en de benodigde tijd om het experiment uit te voeren.

5.2.3 Frequentie

Bij de vorige sectie werd er gevonden dat de bachelorstudenten die deelnamen aan het experiment minder tijd besteedden aan de voorbereidende stappen. Binnen deze sectie wordt gekeken of dit en andere patronen zich ook weerspiegelen in de gemiddelde hoeveelheid uitgevoerde handelingen. Deze informatie kan gevonden worden in Tabel 10. Deze frequenties liggen in lijn met de tijd die de deelnemers nodig hadden. Er werd dus niet meer tijd besteed aan de handelingen bij bijvoorbeeld de tweede vraag, er werden gewoon ook meer handelingen uitgevoerd. Er waren meer stappen nodig om de oplossing te bereiken. Het enige merkwaardige element uit de verhouding van Tabel 10 en 9 is het feit dat de doctoraatsstudenten wel beduidend meer handelingen uitvoerden per minuut (4,11) dan de master (3,65) en bachelorstudenten (3,52). Dit kan een indicatie zijn van betere mentale modellen [7][8][10].

5.2.4 Structuur

In hoeverre de deelnemers dezelfde structuur gebruikten kan bepaald worden aan de hand van de structuur-metriek zoals geformuleerd in [11]. Deze wordt berekend door het aantal voorgekomen gedragingen te delen door het totaal aantal mogelijke gedragingen en die verhouding van 1 af te trekken [11]. Gedrag is hier de opeenvolging van twee handelingen. Des te hoger deze metriek is, des te gestructureerder er te werk is gegaan. In Tabel 11 kan de structuur per vraag en per deelnemersgroep gevonden worden. Over de drie deelnemersgroepen heen valt op dat de bachelorstudenten het meest gestructureerd te werk gingen over de drie vragen heen, al zijn de verschillen wel klein. Op het niveau van de vragen is zichtbaar dat de tweede vraag het minst structureel verliep, zo werd er gemiddeld genomen bijna dubbel zoveel uniek gedrag vertoond in vergelijking met de eerste vraag. Een bevinding die in lijn ligt met het feit dat er het meeste tijd en handelingen aan deze vraag werd gespendeerd over alle deelnemers heen.

5.2.5 Herhaling

Het herhalen van handelingen kan erop wijzen dat de voorgaande uitvoering van de handeling niet volledig succesvol was en dus opnieuw moest gebeuren. Deze herhaling kan zowel direct als later in de opdracht opnieuw gebeuren, naar het eerste geval zal in de resterende tekst

Vraag	Bachelor	Master	Doctoraat	Gemiddelde
Voor	1,42 (0,51)	4,18 (2,24)	6,37 (4,38)	4,67 (3,59)
1	4,85 (0,85)	9,14 (6,13)	4,51 (2,39)	6,25 (4,34)
2	19,3 (3,35)	14,1 (10,2)	19,8 (4,59)	17,6 (6,97)
3	11,4 (1,87)	12,4 (6,66)	7,65 (4,29)	10,0 (5,14)
Na	0,16 (0,26)	4,51 (3,67)	1,3 (1,20)	1,98 (2,30)
Totaal	37,1 (1,36)	42 (24,7)	39,7 (4,27)	40,09 (13,92)

Tabel 9: Gemiddelde tijd in minuten (met standaardafwijking) per vraag en per deelnemersgroep

Vraag	Bachelor	Master	Doctoraat	Gemiddelde
Voor	6 (0)	15,75 (4,92)	18,4 (10,67)	15,18 (8,66)
1	16 (2,83)	34,5 (18,95)	19,6 (9,48)	24,36 (14,52)
2	67,5 (31,82)	49,75 (23,17)	85,4 (20,50)	69,18 (26,71)
3	40,5 (6,36)	48,5 (13,67)	36,6 (7,23)	41,64 (10,62)
Na	0,5 (0,71)	4,75 (6,60)	2,8 (0,84)	3,10 (3,99)
Totaal	130,5 (36,06)	153,25 (43,92)	163 (10)	153,55 (30)

Tabel 10: Gemiddeld aantal uitgevoerde handelingen (met standaardafwijking) per vraag en per deelnemersgroep

Vraag	Bachelor	Master	Doctoraat	Gew. gemiddelde
1	0,955	0,933	0,957	0,948
2	0,901	0,910	0,885	0,897
3	0,932	0,912	0,930	0,924
Gemiddelde	0,930	0,918	0,924	0,923

Tabel 11: Structuur van de werkwijzen per vraag en per deelnemersgroep

verwezen worden als een *self-loop*. Op basis van de expertise wordt er verwacht dat een data-analyse expert minder vaak op zijn stappen moet terugkomen dan een beginnende data analist. In Tabel 12 en 13 kunnen de absolute en relatieve herhaling die plaatsvond bij de deelnemers gevonden worden. Tabel 12 geeft de herhalingen doorheen de volledige opdrachten weer en Tabel 13 brengt de *self-loops* in kaart. Opvallend is hier dat de master en doctoraatsstudenten zeker niet minder vaak een handeling moesten hernemen dan de bachelorstudenten en dat er geen duidelijke verschillen aanwezig zijn. Tussen de drie vragen valt wel op dat er relatief gezien minder herhaling en *self-loops* werden uitgevoerd bij de tweede dan bij de eerste en derde vraag. Desondanks dat er wel meer handelingen uitgevoerd werden, resulteerde dit niet in een relatief hogere herhaling. Dat impliceert dat er meer verschillende handelingen nodig waren om het eindresultaat te bereiken en dat de hogere tijd en de hogere absolute frequentie niet gelinkt is aan het feit dat de deelnemers meer handelingen opnieuw moesten doen.

5.2.6 Besluit

Er werden verschillende metrieken en aspecten geanalyseerd, maar de gevonden resultaten verschilden vaak nauwelijks en waren regelmatig zelfs tegenstrijdig met de verwachting dat een data analist met meer expertise beter, sneller of meer gestructureerd te werk zou gaan dan een beginnende data analist. Er is, op basis van de verzamelde data, geen eenduidig antwoord op de vraag of er verschillen aanwezig zijn tussen de verschillende deelnemersgroepen. Een van de verschillen die gevonden werd was het feit dat de bachelorstudenten minder tijd hadden besteed aan de voorbereidende stappen en het nakijken van de opdrachten. Verder consulteerden bachelorstudenten, bij het raadplegen van informatie, eerder voorgaande analyses en cursusmateriaal, waar de andere deelnemers sneller het internet of de ingebouwde documentatie opzochten.

Op het niveau van de verschillende vragen waren de verschillen iets groter en consistenter over de verschillende analyses heen. Zo hadden de deelnemers voor de eerste vraag het minste tijd, handelingen en herhaling nodig en werd deze vraag door alle deelnemers correct beantwoord. De tweede opgave werd duidelijk als moeilijkste ervaren. Niet alleen werden vijf van de zes gemaakte fouten bij de tweede opgave gemaakt, maar ook werd er het meeste tijd aan gespendeerd, werden er gemiddeld meer handelingen uitgevoerd en was er het minste structuur.

6 Discussie

De gevonden handelingen en hun onderlinge relatie kan een goed houvast zijn voor startende data analisten. Zowel voor inzicht te krijgen in welke stappen er genomen moeten en kunnen worden bij het uitvoeren van een data-analyse, alsook in welke volgorde deze uitgevoerd kunnen worden. Door de exploratieve data-analyse stap verder op te splitsen wordt het eenvoudiger om stapsgewijs te werk te gaan en zo mogelijks betere analyses te bekomen [4]. Verder kunnen deze handelingen gebruikt worden om *deliberate practise* te ontwerpen. Bij *deliberate practise* worden kleine taken met een vooropgesteld doel meermaals herhaald

Vraag	Bachelor	Master	Doctoraat	Gew. gemiddelde
1	3,5	5,5	4,8	4,82
2	9	7,25	10	8,82
3	6	8	6,2	6,82
Gemiddelde absoluut	6,17	6,92	7,00	6,82
1	0,211	0,126	0,283	0,212
2	0,139	0,163	0,120	0,139
3	0,150	0,168	0,171	0,166
Gemiddelde relatief	0,167	0,152	0,191	0,172

Tabel 12: Herhaling van een handeling binnen één opdracht, zowel het absoluut aantal herhaalde handelingen als de verhouding van de herhaalde handelingen ten opzichte van al de genomen handelingen

Vraag	Bachelor	Master	Doctoraat	Gew. gemiddelde
1	1	4	1,4	2,27
2	4,5	4,25	6	5,09
3	3,5	5,25	2,6	3,73
Gemiddelde absoluut	3	4,5	3,33	3,70
1	0,072	0,121	0,077	0,092
2	0,073	0,079	0,072	0,075
3	0,089	0,111	0,069	0,088
Gemiddelde relatief	0,078	0,103	0,072	0,085

Tabel 13: Directe herhaling (*self-loop*) van een handeling binnen één opdracht, zowel de absolute als relatieve frequentie

[8][10]. Voorafgaand aan dit onderzoek waren er nog geen handelingen gedefinieerd op voldoende laag niveau om zo een training te kunnen construeren.

Een beperking van dit experiment is dat er geen echte experten deelnamen en dat de gevraagde analyse een relatief eenvoudige en voorgekauwde oefening was. Hierdoor wordt er verwacht dat de gevonden lijst van handelingen niet exhaustief is. Verder zal het gebruik van alleen R en niet andere talen zoals Python er ook voor gezorgd hebben dat enkele handelingen allicht specifiek gerelateerd zijn aan R. R werd gekozen omdat alle deelnemers een basiskennis hadden van R door een inleidende cursus die ontvangen werd in de eerste bachelor. Deze cursus kan mogelijks ook reeds een bepaalde methodologie aangeleerd hebben. Dat zou een mogelijke oorzaak kunnen zijn voor het niet vinden van duidelijke verschillen tussen de werkwijze van bachelor, master en doctoraatsstudenten. Een bijkomende reden voor het niet vinden van deze verschillen is het feit dat de deelnemers uit de bachelor vrijwillig deelnamen en zich dus misschien comfortabeler voelen bij het uitvoeren van een data-analyse in R dan hun medestudenten.

Er wordt dan ook aangeraden verder onderzoek te doen naar zowel de handelingen, de volgorde van deze handelingen alsook de werkwijzen van experts en beginners. Door gebruik te maken van meer heterogene deelnemers, een moeilijkere opgave en verschillende programmeertalen wordt er verwacht dat er nog bijkomende handelingen geïdentificeerd kunnen worden. Verder kunnen er ook mogelijke verschillen in werkwijzen opduiken tussen experts en beginners. Deze verschillen kunnen gebruikt worden om de mentale voorstellingen van experts te identificeren en deze kunnen op hun beurt dan weer ingezet worden om nieuwe onderwijsmethodes op te stellen [23]. Daarnaast zou een analyse op het niveau van de subactiviteiten inzichten kunnen verschaffen over frequenties en een volgorde op een lager niveau, zoals in welke volgorde de subactiviteiten bij de handeling *data voorbereiden* meestal uitgevoerd werden.

7 Conclusie

De stappen die doorlopen worden tijdens een exploratieve data-analyse kunnen opgedeeld worden in vier categorieën: de voorbereidende stappen, de analyse stappen, de debugstap en tot slot de handelingen die niet tot een categorie behoren, maar wel tijdens het hele analyse proces gebruikt kunnen worden.

In de voorbereidende stappen werd de opdracht en de data verkend. De handelingen die hier geïdentificeerd werden zijn het lezen van de opdracht, het inladen van de data, het controleren van de data en het verkennen van de data. Tot slot behoort het inladen van een pakket ook nog tot de voorbereidende stappen, al kan deze ook uitgevoerd worden tijdens de analyse wanneer de nood voor een bepaald pakket opduikt.

Bij de analyse categorie werd er bijna altijd begonnen met nogmaals de opdracht te lezen. Daarna werd er gezocht naar de variabele in bijvoorbeeld de data beschrijving of de data zelf. Eenmaal deze gevonden was, werd er meestal gekeken of de data reeds voldeed om de vraag te beantwoorden. Als dit niet het geval was, werd de handeling *data voorbereiden* uitgevoerd. Denk hierbij aan

onder meer het filteren, groeperen en samenvoegen van data. Zodra de invoerdata aan de vereisten voldeed werd er, afhankelijk van de vraag, een grafiek of samenvattende statistiek opgesteld. Tot slot werd het resultaat geëvalueerd eenmaal de code uitgevoerd was.

De debugstap is de handeling die noodgedwongen uitgevoerd werd nadat er een error werd verkregen. Hierbij werd er geen werkwijze geïdentificeerd waarmee de meeste errors verholpen werden. Vaak was deze stap vrij triviaal binnen het uitgevoerde experiment en werden de errors relatief eenvoudig verholpen. Als dit niet het geval was werd er bijkomende informatie geraadpleegd.

Tot slot werden er nog vijf andere handelingen geïdentificeerd die uitgevoerd werden tijdens de analyses: nadenken, informatie raadplegen, structuur toevoegen, trial-and-error en de opdrachten nakijken. Waarbij *trial-and-error* eerder een werkwijze is dan een handeling. Al deze handelingen hebben geen rechtstreeks effect op het resultaat, maar kunnen wel helpen om het aantal fouten te reduceren, meer mogelijkheden te overwegen en dus indirect een robuuster eindresultaat te bereiken.

Met betrekking tot werkwijzen werd er geen duidelijke volgorde geïdentificeerd. Het data-analyse proces dat uitgevoerd werd door de deelnemers bleek eerder een iteratief proces waarbij stapsgewijs naar de oplossing werd toegewerkt. Verder werd er ook gekeken of er verschillen waren tussen de werkwijzen van beginnende en gevorderde data analisten. Er werd gekeken naar de juistheid, frequentie, herhaling, duur en structuur van de werkwijzen. Bij geen van deze aspecten was er een relevant verschil.

De handelingen die binnen deze masterproef zijn geïdentificeerd kunnen ingezet worden bij het ontwikkelen van *deliberate practise* om zo sneller een hoger niveau van expertise te bereiken op het vlak van het exploratief analyseren van data. Daarnaast kunnen ze, samen met de onderlinge relaties tussen de handelingen, de basis vormen voor een handleiding met betrekking tot het uitvoeren van een exploratieve data-analyse.

Dankwoord

Ik zou graag mijn promotor dr. Janssenswillen willen bedanken voor zijn begeleiding en waardevolle suggesties om deze masterproef naar een hoger niveau te tillen. Verder zou ik graag ook Benoit Depaire bedanken voor zijn feedback en ideeën, alsook de deelnemers van het experiment voor hun bijdrage aan deze masterproef.

Referenties

[1] Behrens, J. T. (1997). *Principles and procedures of exploratory data analysis.* Psychological Methods, 2(2), 131–160.

[2] Bradford, L. (2018). Why All Employees Need Data Skills In 2019 (And Beyond). Geraadpleegd van https://www.forbes.com/sites/laurencebradford/2018/10/11/why-all-employees-need-data-skills-in-2019-and-beyond/?sh=13de41de510f

[3] Bridgwater, A. (2020). Does Your Company 'Speak' Data Yet? Geraadpleegd van https://www.forbes.com/sites/adrianbridgwater/2020/02/24/does-your-company-speak-data-yet/?sh=6131bbbf4a22

[4] Claes, J., Vanderfeesten, I., Gailly, F., Grefen, P., & Poels, G. (2015). *The structured process modeling theory (SPMT) a cognitive view on why and how modelers benefit from structuring the process of process modeling.* Information Systems Frontiers, 17(6), 1401-1425.

[5] Cukier, K., & Mayer-Schoenberger, V. (2013). *The rise of big data: How it's changing the way we think about the world.* Foreign Aff., 92, 28.

[6] Edwards, S. H. (2004). *Using software testing to move students from trial-and-error to reflection-in-action.* In Proceedings of the 35th SIGCSE technical symposium on Computer science education (pp. 26-30).

[7] Ericsson, K. A. (2006). *The influence of experience and deliberate practice on the development of superior expert performance.* The Cambridge handbook of expertise and expert performance, 38(685-705), 2-2.

[8] Ericsson, K. A. (2008). *Deliberate practice and acquisition of expert performance: a general overview.* Academic emergency medicine, 15(11), 988-994.

[9] Ericsson, K. A., & Towne, T. J. (2010). *Expertise.* WIREs Cognitive Science.

[10] Ericsson, K. A., & Pool, R. (2017). *Piek* (2de ed.). Houten, Nederland: Spectrum.

[11] Günther, C. W. (2009). *Process mining in flexible environments.* Technische Universiteit Eindhoven.

[12] Hinds, P. J. (1999). *The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance.* Journal of experimental psychology: applied, 5(2), 205.

[13] Holton, J. A. (2007). *The coding process and its challenges.* The Sage handbook of grounded theory, 3, 265-289.

[14] Jans, M. (2017). *From relational database to valuable event logs for process mining purposes: a procedure.* Technical report, Hasselt University

[15] Jans, M., Soffer, P., & Jouck, T. (2019). *Building a valuable event log for process mining: an experimental exploration of a guided process.* Enterprise Information Systems, 13(5), 601-630.

[16] Janssenswillen, G., Depaire, B., Swennen, M., Jans, M., & Vanhoof, K. (2019). *bupaR: Enabling reproducible business process analysis.* Knowledge-Based Systems, 163, 927-930.

[17] Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences.* Sage. xv - xvii

[18] O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline.* O'Reilly Media, Inc.. 4-5, 41-42, 66-68

[19] Preter, W. (2019). *Zijn data een hype of worden ze echt het goud van de 21ste eeuw?* Retrieved from https://www.tijd.be/dossiers/het-datatijdperk/zijn-data-een-hype-of-worden-ze-echt-het-goud-van-de-21ste-eeuw/10162167.html

[20] Saldaña, J. (2014). *Coding and analysis strategies.* In The Oxford handbook of qualitative research.

[21] Saldaña, J. (2021). *The coding manual for qualitative researchers.* SAGE Publications Limited., 100-102

[22] Saltz, J. S., & Shamshurin, I. (2015). *Exploring the process of doing data science via an ethnographic study of a media advertising company*. In 2015 IEEE international conference on big data (Big Data) (pp. 2098-2105). IEEE.

[23] Spector, J. M., Polson, M. C., & Muraida, D. J. (Eds.). (1993). *Automating instructional design: Concepts and issues.* Educational Technology. 219

[24] Sweller, J. (1988). *Cognitive load during problem solving: Effects on learning.* Cognitive science, 12(2), 257-285.

[25] Sweller, J. (2011). *Cognitive load theory. In Psychology of learning and motivation* (Vol. 55, pp. 37-76). Academic Press.

[26] Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data.* O'Reilly Media, Inc..

[27] Yates, K. A., & Clark, R. E. (2012). *Cognitive task analysis. International guide to student achievement.* New York: Routledge, 1(1.1).

BIJLAGEN

7.1 Bijlage 1: inleidende enquête

In Tabel 14 kunnen de vragen gevonden worden die gebruikt werden voor de introductie enquête omtrent de zelfperceptie van de deelnemers omtrent hun data-analyse capaciteiten.

7.2 Bijlage 2: data-analyse opdracht

In deze bijlage kan de data-analyse opdracht gevonden worden. Alle deelnemers hadden drie bestanden ter beschikking. Twee datasets en een description file waarin de kolommen van de twee datasets beschreven werden. Deze twee datasets bestonden uit informatie over de huizenmarkt en hadden dezelfde identifiers zodat deze ook aan elkaar gekoppeld/gelinkt konden worden. Deze opgave werd aangeleverd in een R markdown bestand waarin reeds een lege *chunk* (*runnable code-block*) was aangemaakt.

DE DATA
De voorziene datasets bevatten data van een huizenmarkt met verschillende parameters over de huizen alsook de verkoopprijs. Een gedetailleerde uitleg van de datasets kan gevonden worden in de description file.

 # Moest je schermopname nog niet gestart zijn, start hem dan zeker nu!

BESCHRIJVING OPDRACHT
 # Opdracht 1
Visualiseer de distributie van de verkoopprijs.

 # Opdracht 2
Visualiseer de relatie tussen het aantal slaapkamers en de verkoopprijs van de huizen. Toon enkel de categorieën (elk aantal slaapkamers is een categorie, dus alle huizen die vier slaapkamers hebben zitten in dezelfde categorie) met meer dan 10 datapunten.

 # Opdracht 3
Bereken de gemiddelde verkoopprijs van de subset van huizen die een overall condition en(!) een overall quality hebben van 8 of hoger.

BRONNEN
Voor het oplossen van deze opdrachten is er volledige vrijheid. Er mag gebruik gemaakt worden van alle bronnen zoals het internet en dergelijke. Het onderzoek probeert het proces van data-analyse in kaart te brengen en dus zeker niet het resultaat. Probeer uiteraard wel een zinvolle analyse te bekomen die informatie uit de data haalt.

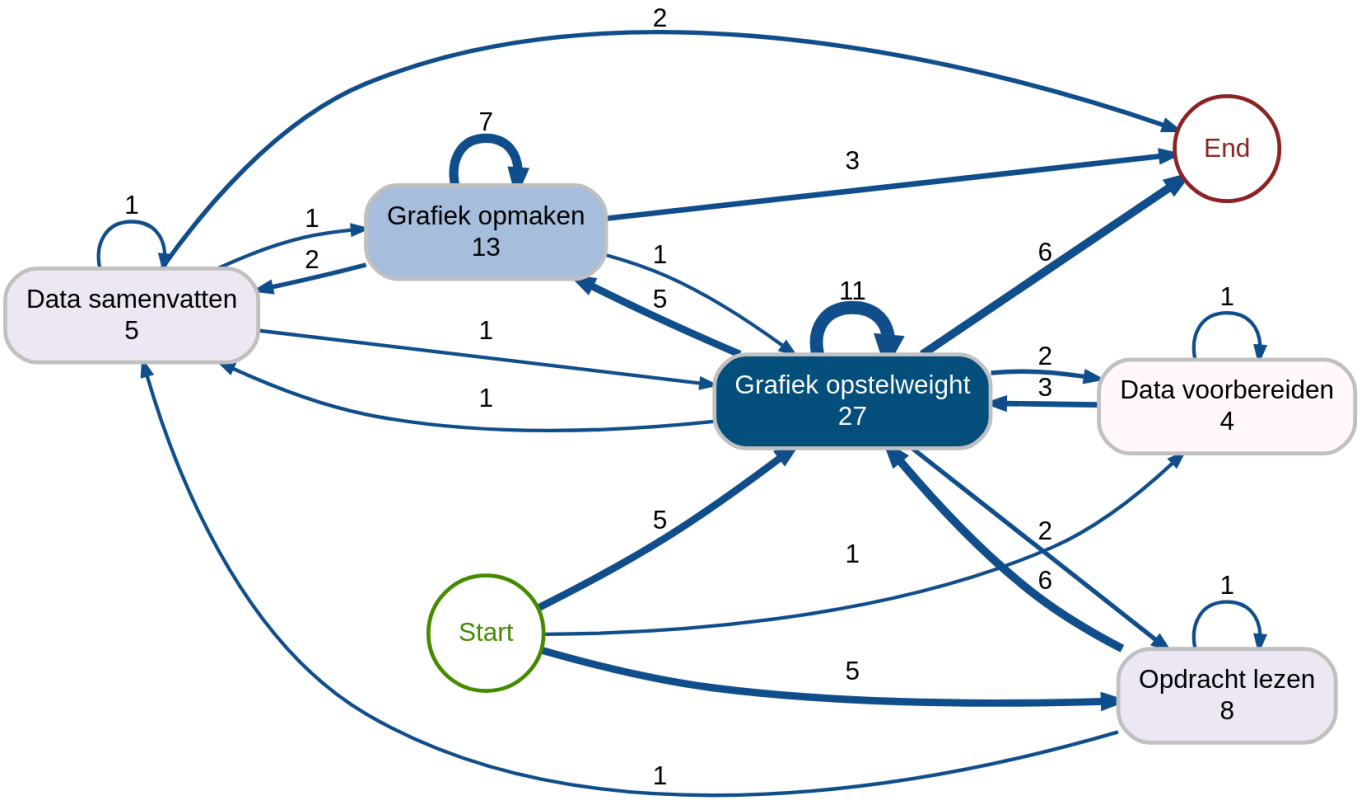
Veel succes!

7.3 Bijlage 3: proces visualisaties per vraag

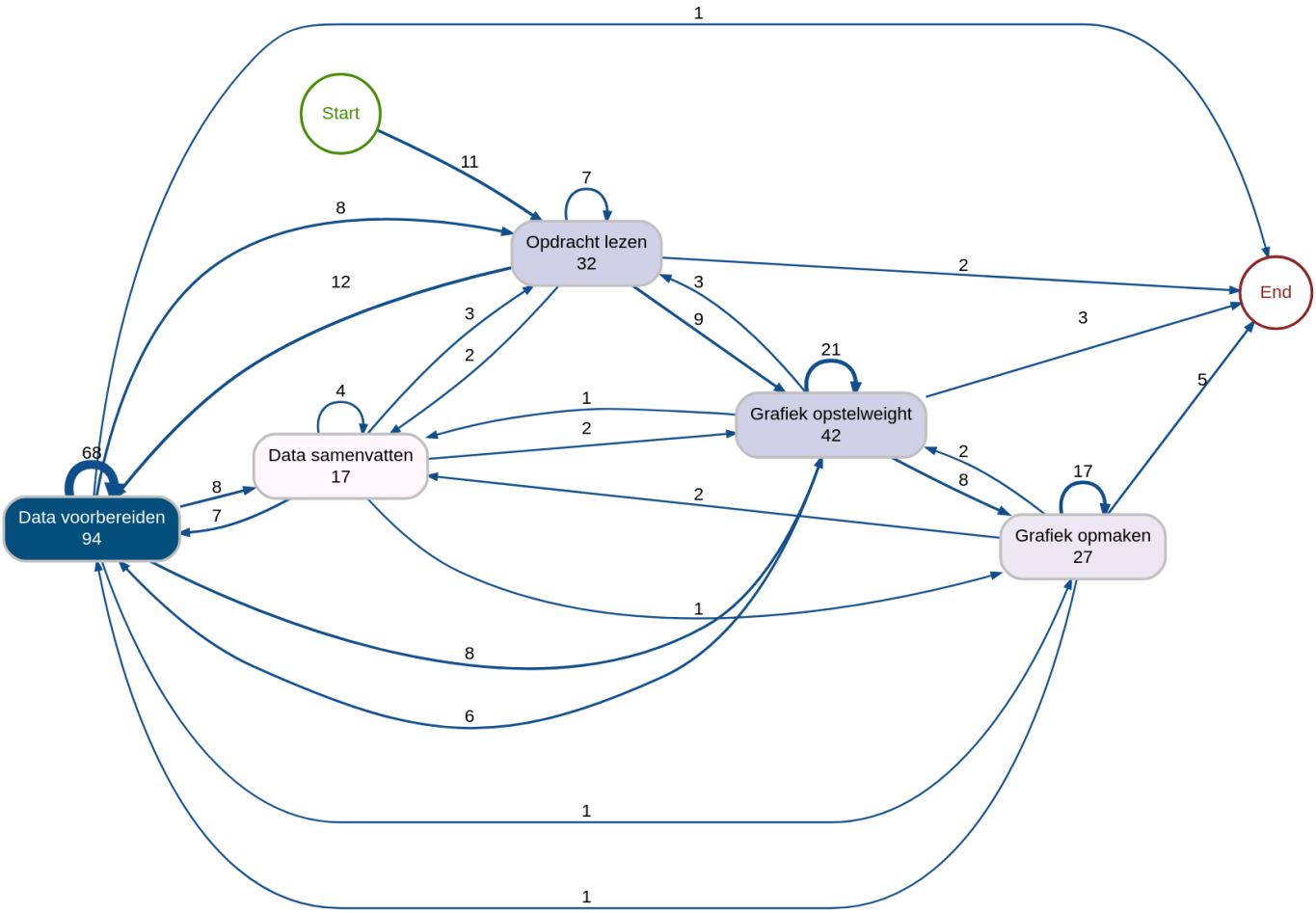
In Figuur 4, 5 en 6 kan het procesverloop per vraag visueel teruggevonden worden. Deze figuren waren de basis voor de volgorde die gedefinieerd werd in Figuur 3.

Vraag	Antwoordmogelijkheden
Ik vind mezelf een data analist	Likertschaal 1-5
Ik kan een data-analyse uitvoeren	Likertschaal 1-5
Ik kan een data-analyse in R uitvoeren	Likertschaal 1-5
Hoe ervaren ben je met het analyseren van data	Likertschaal 1-5
Welke momenten kwam je reeds in contact met data-analyse?	Keuzemogelijkheden – Ik volgde een data-analyse vak zoals EDDA – Ik heb mezelf verdiept in het analyseren van data – Ik gebruikte data-analyse bij één of meer projecten voor school – Ik gebruikte data-analyse reeds in mijn vrije tijd – *Extra mogelijkheid, zelf in te vullen*

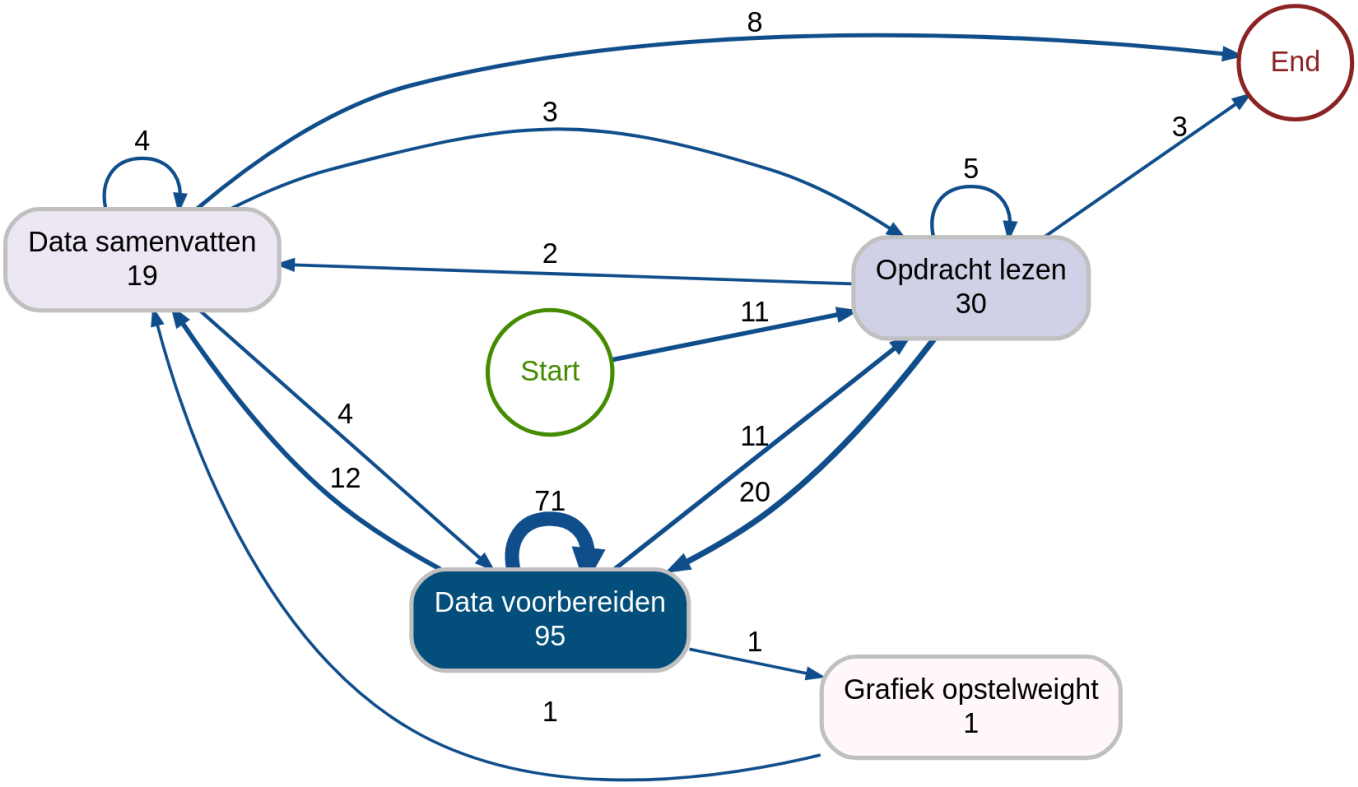
Tabel 14: Vragen inleidende enquête



Figuur 4: Proces van de eerste vraag



Figuur 5: Proces van de tweede vraag



Figuur 6: Proces van de derde vraag