**Samenvatting**

HoB blijkt vanaf haar 3e jaar een competatief salaris te bieden waarin het meer dan 50% van de markt verslaat. In de eerste 2 jaren verdienen werknemers boven het modale loon. Het blijkt echter dat House of Bèta nog ruimte heeft om de hard skills van haar personeel te ontwikkeling. Het belangrijkste zijn skills m.b.t de cloud, big data, en geavanceerde data science skills. Op het moment van schrijven staan de volgende potentiële opdrachtgevers boven aan de lijst met openstaande vacatures:

**Inleiding**

Dit onderzoek richt zich op de vraag hoe competitief House of Bèta is op de markt voor data professionals. Hierin wordt gekeken naar zowel het salaris (dus wat House of Bèta biedt aan het personeel) als de mogelijkheden voor House of Bèta om een groter deel van de markt te bedienen (gevraagde vaardigheden en potentiële opdrachtgevers).

House of Bèta heeft al langer de wens om inzichten te krijgen in de reden van uitstroom van haar consultants, gezien dit een belangrijke asset is van de organisatie. Wegens juridische en technische beperkingen is het echter niet mogelijk om hierop diepgaand onderzoek op te doen.

Een tweede wens is om de markt beter te kunnen bedienen en te groeien als de leverancier van IT personeel. Na overleg is gebleken dat beide wensen gekoppeld kunnen worden in dit onderzoek. Het onderzoek helpt House of Bèta inzichten te krijgen in een ontwikkelende markt, haar eigen arbeidsvoorwaarden op waarden te schatten en biedt een basis voor verdere analyse aangaande de churn van haar personeel.

**House of Beta**

House of Bèta (HoB) is een detacheerder op het gebied van (business) IT. HoB probeert voornamelijk starters op de arbeidsmarkt aan te trekken en probeert deze te matchen aan passende opdrachtgevers. De consultants hebben de mogelijkheid zich na verloop van tijd te specialiseren waarin HoB de trainingen en opleiding faciliteert of betaald.

HoB is een relatief nieuwe naam op de markt maar voert al sinds 2015 haar werkzaamheden uit onder het moederbedrijf Talent&Pro. De uitstroom van consultants die T&P heeft gezien over2021 is de aanleiding geweest om meer grip op deze uitstroom te willen krijgen (de eerste wens).

Tijdens de een recente informatieavond van HoB zijn nieuwe doelstellingen voor 2022 gepresenteerd waar ook de ontwikkeling en groei van data professionals naar voren komt. Dit is dan ook de aanleiding voor de tweede wens van HoB.

**Businessvraag**

Dit onderzoek en de wensen van HoB kunnen worden vastgelegd in een centrale vraag: “Hoe positioneert HoB zich ten opzichten van de Nederlandse arbeidsmarkt op het gebied van data professionals kijkend naar loon, vaardigheden en potentiële opdrachtgevers?”

Deze vraag kan verder worden opgedeeld in 3 deelvragen:

1. Hoe verhoudt het salaris van HoB zich tot de nederlandse markt?

2. In hoeverre sluit het “curriculum” van HoB zich aan bij de ontwikkelende markt?

3. Welke opdrachtgevers actief op de markt worden nog niet bediend door HoB?

**data gebruik**

Binnen HoB vinden nog geen datamining activiteiten op dit gebied plaats. Buiten de resultaten van dit onderzoek wordt er dan ook een een script aangeleverd die het mogelijk maakt om de datamining voor dit onderzoek voort te zetten. Dit script is beperkt tot het zoekwoord “data” op de vacaturesite Indeed, maar kan uitgebreid worden om op andere woorden te zoeken.

**Business Success criteria**

De volgende kwalitatieve eisen worden gesteld: er moet er een analyse komen van HoB versus markt inkomen, een analyse over de vraag verdeling over bedrijven en een over skills. Daarnaast moet er een bruikbare scraper voor vervolg onderzoek worden opgeleverd

**Methoden**

Dit hoofdstuk behandelt de data en het gebruik hiervan. In “data understanding” worden basale kenmerken van de dataset besproken, “data preperation en modelling” gaat in op het schonen een verrijken van de data. Als laatste gaat “reproduceren van het onderzoek” over de mogelijkheid om dit onderzoek na te bootsten.

**Data** **understanding**

Hieronder worden de scope, de gebruikte datasets en de basale kenmerken beschreven.

**Gekozen datasets en scope**

Om te beginnen is hieronder een versimpelde datalog weergegeven. Dit bevat informatie over de datasets in de ruwe vorm (te vinden in ".\Data\_raw")".

D001 is na vooronderzoek gekozen als basis dataset. Hierin staan onder andere (hard)skills, functie type (data analist, -engineer, -scientist). Dit legt een goede basis voor het onderzoek naar data specialisaties binnen HoB, hier worden namelijk dezelfde specialisaties ook aangeboden. Er zijn andere datasets overwogen echter bevatte deze te weinig relevantie functie type (< 200 rijen) of ontbraken hier relevante skills.

D002 is gescrapte data van Indeed. Het rationaal achter het zelfstandig scrapen van Indeed data komt voort uit het tekortschieten van bestaande API’s en afwezigheid van relevante datasets.

**Beschrijving van de ruwe data**

Beginnend bij de beschrijving van D001. Deze dataset is al erg opgeschoond. Een aantal basale kenmerken zijn hieronder programmatische weergegeven. Allereerst is duidelijk dat er 43 kolommen aanwezig zijn. Echter voor het doel van dit onderzoek zijn maar twee kolommen echt interessant. Dit zijn “Skill” en “Job\_Type”. Hierin staan respectievelijk een lijst van (hard)skills en de functie soort per observatie.

Zoals hierboven getoond komt duidelijk naar voren dat het aantal unieke “Skills” bijna gelijk is aan het aantal unieke rijen. Dit betekent echter niet dat er zoveel verschil zit tussen de vacatures. Het is namelijk zo dat [SQL, Python] en [Python, SQL] hetzelfde betekenen maar wordt gezien als twee unieke rijen. Deze informatie wordt gebruikt in latere ETL stappen. Daarnaast is opvallend dat de verdeling tussen data analist (31%), data engineer (24%) en data scientist (45%) niet volledig evenredig is.

Dataset D002 bevat datavan Indeed. Hiervoor is er gezocht naar alle functies waar het woord "data" in voorkomt. D002 bestaat in de ruwe vorm uit meerdere losse bestanden (een voor elke dag dat er een run is geweest van de scraper). Om een indruk te krijgen van wat er in een dergelijk bestand staat kan er gekeken worden naar ".\Data\_raw\D002\2022-03-08 scraped indeed data.csv". Hieronder worden weer de voornaamste kenmerken.

Uit bovenstaande informatie vallen een aantal dingen op te maken. Allereerst gaat de data niet verder terug dan begin februari 2022. Dit komt omdat de webscraper pas op 6 maart 2022 is afgerond. Dit zal ook effect hebben op de analyses gezien de data niet ver genoeg teruggaat om echte trends te ontdekken. Ten tweede valt op dat het aantal rijen en het aantal unieke URL's niet overeenkomen. Dit geeft het vermoeden dat er dubbele rijen zijn, of in iedergeval een functie meermaals is gescraped. Ook valt op dat er een fout in de salaris kolom staat. Hierin is het euro-symbool (€) niet goed vertaald is (ï¿½). Dit en nadere opschoning vindt plaats in de volgende paragraaf "Data preperation en modeling".

**Data preperation en modeling**

Voor het binnenhalen, schoonmaken en verijken van de data zijn diverse stappen genomen. In de onderstaande processflow is te zien hoe het process hoogover doorlopen wordt om tot de uiteindelijke analyses te komen.

Het process is in 3 segmenten op te delen. Het bovenste segment heeft betrekking op de data vanuit Kaggle (D001), het tweede segement, in het midden, heeft betrekking op de Indeed data die gescraped wordt (D002) vervolgens is de onderste laag waar alles samenkomt. Na het schonen verijken en wegschrijven van de data kan er een dataframe worden opgezet die goed interpeteerbaar is.

Er zijn een aantal belangrijke kanttekenen die moeten worden gemaakt. Wegens het gebrek aan tijd zijn een aantal aannamens gedaan en is data versimpeld weergegeven. De meest noemenswaardigen zijn:

Process stap

Betreffende kolom

Versimpelde weergaven

Scraping

Salary

Er is bij weergaven salarissen altijd voor het laagste salaris gegaan

ETL (midden-rechts)

job\_type

Er zijn maar 3 types gedefinieëerd (data analist, scientist, en engineer). Ongeacht de functietitel zal 1 van deze types worden toegewezen. Zo kan data architect of teamleider data analytics worden toegewezen aan het type data analist.

ETL (midden-rechts)

salary

er is aangenomen dat alles boven de €10.000,- jaarsalaris betreft. dit wordt omgerekend naar maand. Alles onder de €100,- is als uur tarief aangenomen en wordt omgerekend naar maan (obv. 168 uur per maand). alles tussen de €100,- en €1000,- wordt als scraping fout gezien en omgezet naar NA.

Aannamens tijdens process stappen,

De scripts voor deze stappen zijn te vinden in "./Supporting\_code", specifiek:

Process stap

Script naam

ETL (linksboven)

Generate\_ref\_skill\_list.R

Scraping

indeed\_scraper.R

ETL (midden-rechts)

Scraped\_indeed\_data\_wrangler.R

process stappen en bijbehorende scripts.

Het draaien van deze scripts zorgt uitindelijk voor een dataframe dat gebruikt kan worden voor het analyseren en beantwoorden van de businessvraag. Voor een duidelijker beeld wat de code doet zijn de .R scripts voorzien van comments.

Reproduceren van het onderzoek

Om dit project te reproduceren is het nodig om de mappen: "Data\_cleaned", "Data\_raw", en "Supporting\_code" te gebruiken. Hierin staan bron data en hulp .R scripts. In deze Markdown worden hulp scripts aangeroepen om de code complexiteit in deze Markdown te beperken. Wat betreft externe libraries moeten de volgende worden geïnstalleerd:

"tidyverse", "stringr", "lubridate", "tibble", "textTinyR". Indien er nieuwe data gescraped moet worden (niet nodig voor reproductie): "XML" en "rvest".

Er zijn verder geen afhankelijkheden buiten het bovengenoemde om deze markdown te draaien.

**Resultaten**

In dit hoofdstukworden de deelvragen beantwoord aan de hand van de analyse resultaten. Ook wordt uiteengezet welke vragen niet beantwoord kunnen worden en welke artifacten opgeleverd worden. Als eerste wordt gekeken naar de verhouding van het HoB salaris t.o.v. het salaris op Indeed, vervolgens wordt gekeken

**Hoe verhoudt het loon van HoB zich tot de Nederlandse banenmarkt op het gebied van data specialisten?**

Binnen HoB is het mogelijk om op verschillende manieren salarisverhoging te behalen. Zo wordt een consultant beloont voor het behalen van cursussen en opleidingen en hangt de verhoging samen met de beoordeling. Afhankelijk van het niveau van de consultant is dit een absolute stijging of een relatieve stijging ten opzichten van het huidige loon. Ook worden consultants beloond voor het behalen van cursussen. Om deze reden is het lasting om een accuraat salaris te bepalen. Gelukkig heeft HoB op haar website de referentie salarissen voor 1ste, 2de en 3de jaars consultants staan (€2.763,- , €3.087,- , €3.591,- respectievelijk).

De vraag van HoB is om dit geboden salaris aanbod op waarden te schatten ten opzichten van de markt. In onderstaande grafiek is te zien dat het HoB salaris de eerste twee jaren lager ligt dan het mediaan en modale salaris aangeboden op Indeed.

In de bovenstaande analyse is het het aantal jaren werkervaring van de indeed vacatures niet meegeneomen. De dataset is wel vereikt met een kolom die het aantal jaren ervaring aangeeft. Echter, doordat er 'regex' is gebruikt en de data erg ongestructureerd is zal deze kolom niet helemaal betrouwbaar zijn. Desondanks kan deze data gebruikt worden om toch een indruk te krijgen van de salarisverdeling over de verschillende eisen qua jaren werkervaring. In onderstaande grafiek komt duidelijk naar voren dat het salaris van HoB personeel in jaar 3 hoger ligt dan het modale salaris dat op indeed geboden wordt tot en met jaar 3. Ook is opvallend dat bij 8 jaar werkervaring de curve naar links verschuifd

Het salaris bij HoB wordt niet alleen bepaald door werkervaring. Het is namelijk ook zo dat er veel in de consultants wordt geïnvesteerd doormiddel van trainingen. HoB richt zich zowel op hard-, als softskills. Binnen dit onderzoek zijn de hardskills het meest accurate uit de tekst te halen en daardoor het beste te onderzoeken. Bij HoB zal een consultant data gemiddeld 3 tot 5 skills overkoepelende skills machtig zijn, In kader van dit project is niet direct meetbaar welke skills het meest voorkomen binnen HoB data consultants, echter op ervaring afgaande zal dit de volgende skills betreffen: SQL, Python, R, Excel en Data Modeleren. In onderstaande analyse is de salaris verdeling voor 3, 4, en 5 skills te zien.

Hieronder zijn drie grafieken weergegeven. Iedere grafiek laat het mediaan salaris per aantal skills zien. De balken zijn groen gekleurd zodra het referentie salaris van HoB hoger ligt dan het mediaan salaris. Er vallen hier twee belangrijke dingen op, namelijk:

Er is een klein positief lineair verband tussen het aantal skills en het mediaan gebode salaris

Het referentie salaris van HoB is pas vanaf het 3e jaar 'competitief' te noemen.

Een verdieping zou nog bereikt kunnen worden door ook te kijken naar het aantal jaren werkervaring als extra dimensie. Helaas zijn er op dit moment te weinig observaties om deze extra dimensie toe te voegen in deze analyse.

**Vraag naar Skills**

De volgende deelvraag voor dit onderzoek gaat om de vraag ontwikkeling naar specifieke skills. Hoe ontwikkelt de vraag naar specifieke skills zich over de tijd? Deze deelvraag is lastig te beantwoorden op binnen dit onderzoek. De data loopt niet verder terug dan begin februari 2022. Voor een duidelijke trend analyse is er data nodig van ~3 jaar. Dit maakt dat onderstaande analyses een versimpelde weergaven zullen zijn. De artefacten die aangeleverd worden (waaronder de scraper) kunnen gebruikt worden om deze dataset verder op te bouwen totdat er genoeg data is om de onderstaande analyse opnieuw te doen.

Als eerst kan er gekeken worden naar de totale vraag naar skills. In de twee onderstaande grafieken zijn zowel de 20 meest gevraagde skills als de 20 minst gevraagde skills terug te vinden. de top drie is hier weinig verrassend en behelst de toch wel meest bekende talen binnen het data vakgebied. Verassender is om Tableau in de top 20 te zien maar geen PowerBI. Verder wordt nu duidelijk wanneer gekeken wordt naar zowel de top als de bottom 20 dat er een mix van echte hard skills en bepaalde conventies (zoals "https" en "rest") worden opgenomen in de skill referentie lijst uit D001. Hier is met namen opvallend dat "marketing" en "network" hoog scoren. Hierin kan er aangenomen worden dat "network" het cloud aspect zijn van het data vak, maar, het zou ook kunnen duiden op neurale netwerken. Helaas is dit een fout in het scrapen en binnen dit project niet direct meer te herstellen. Wat betreft "marketing" kan verwacht worden dat dit gaat om marketing analytics en marketing automation. Twee velden die erg groot zijn binnen data science en data analytics.

Het opmerkelijkste is dat PowerBI niet hoger in de lijst staat. Een van de platformen waarop HoB haar energie richt wat betreft trainingen. In tegenstelling is het Tableau waar dagelijks meer vacatures voor verschijnen. Dit om te overeen met de initiële analyse van D001 waarin de aannamen was dat de Amerikaanse markt andere voorkeuren heeft dan de Nederlandse markt, wat aanleiding zou kunnen zijn voor verder onderzoek.

De bottom 20 geeft weinig informatie. Het zijn in meer en mindere maten bekende talen/tools. het signaleert een niche binnen de vacatures maar buiten het feit dat HoB haar activiteiten hier niet op moet focussen levert dit ons geen extra informatie op.

**Instroom vacatures**

Een volgende stap is patroon analyse. Omdat de data niet vergenoeg terug gaat kan er geen accurate voorspelling worden gedaan over de ontwikkeling naar de vraag. Naast dat het inspelen op deze vraag ontwikkeling ook erg lastig is wegens de doorlooptijd van het opzetten van opleidingen en het opleiden van kandidaten. Wel kan er gestart worden met een analyse of er een bepaald patroon is m.b.t. de instroom.

De eerste grafiek hieronder maakt duidelijk dat er een grote instroom van vacatures is op het eerste moment van scrapen. Dit heeft een logische verklaring: Indeed kent drie soorten waarden m.b.t plaatsings moment "Vandaag", "1-30 dagen geleden" "30+ dagen geleden". De scraper maakt geen onderscheid tussen "30" en "30+" Dit betekent dat het eerste moment van scrapen alle "30" en "30+" kenmerken om worden gezet naar "30". Dit verklaart een spike wanneer deze twee waarden voorkomen, doordat tijdens het data schonen dubbele rijen worden verwijderd (de eerste wordt behouden) zien we dit effect alleen terug in het begin van de totale dataset. Wanneer er dus gekeken wordt naar de vraag over tijd zal het binnen deze dataset noodzakelijk zijn om vanaf 5 februari 2022 te kijken. In beide grafieken is lastig vast te stellen of er een correlatie of een bepaald patroon aanwezig is. Wel valt op te maken dat er over het algemeen een dalende trend lijkt te zijn. Om hier echter meer over te zeggen zal minstens 3 jaar aan data nodig zijn om trends en patronen te kunnen vaststellen.

Beide grafieken lijken pieken op bepaalde intervallen te tonen. In de derde grafiek wordt daarom de autocorrelatie weerggeven hieruit blijkt dat er geen duidelijk patroon aanwezig is, hierbij zou 1 (of -1) een mogelijke correlatie of patroon aanduiden. Deze vindt alleen plaats op lag 0. Dit betekent dat er alleen een verband is op de dag zelf. Kortom lijkt er geen patroon zichtbaar te zijn wat betreft de instroom.

**Ontwikkeling skill behoeften**

Hoewel er met de huidige data nog geen overkoepelend patroon te signaleren is kan er wel gekeken worden naar de ontwikkeling van de skill behoeften. Ook hier mist er een hoeveelheid data waarmee een goede tijdanalyse gedaan kan worden. Daarom dienen onderstaande grafieken en bijbehorende code ook als artefact waar op een later moment gebruik van gemaakt kan worden. In onderstaande grafieken zijn in iedergeval de drie meest gebruikte programmeertalen rondom data specialisaties te vinden. Daarnaast is het analyse platform Tableau weergegeven en de twee cloud oplossingen AWS en Azure.

Ook is hier weer een dalende trend te zien in nagenoeg alle skill aanvragen. Zowel in onderstaande grafieken als in het overzicht van totale aanvragen kan dit te verklaren zijn door de rijkwijdte van het scraping script (max 300 vacatures) en de willekeurige sortering op de Indeed. Daarom zal ook hier de tijd een belangrijke factor spelen en een goede analyse na 3 jaar data verzamelen gedaan kunnen worden.

Wat we uit deze en de voorgaande analyse in iedergeval kunnen opmaken is het feit dat er een daling is in vacatures na een piek aan het begin van februari.

**Ontwikkeling opdrachtgevers**

De laatste deelvraag voor dit onderzoek betreft de ontwikkeling en verhouding van de vraag vanuit mogelijke opdrachtgevers. Omdat het gebleken is dat de ontwikkeling van de vraag te weinig zegt wordt hier alleen nog gekeken naar de totale vraag vanuit opdrachtgevers. Hierop kunnen we binnen een aantal elementen segmenteren, namelijk: functietype en gevraagde skills.

Allereerst wordt gekeken naar de 25 bedrijven met het grootste aanbod . in de eerste grafiek is te zien dat er overwegend gevraagd wordt naar data scientisten, vervolgens data analisten, en in veel mindere maten naar data engineers. Wanneer andere detacheerders / uitzendbureaus / consultancies worden weggelaten blijft de volgende top 5 bedrijven over (ten tijden schrijven):

Bertelsmann

NN

Rabobank

Belastingdienst

Rijkswaterstaat

Vanuit bovenstaande groep werkt HoB nog niet samen met Bertelsmann (een mediaconglomeraat) wat zou kunnen duiden op mogelijkheden tot een nieuwe samenwerking.

In de volgende twee grafieken zijn de top 10 bedrijven (van de top 25) gekozen die zich niet richten op detachering, consultancy of traineeships. Hierin is de verdeling over functie type en de 10 meest gevraagde skills te zien. Het is duidelijk te zien dat zowel de functie type en de gevraagde vaardigheden in zekere zinnen overeenkomen met de eerder vastgestelde verdeling over de gehele gescrapte populatie (een paar afwijkingen achterwege gelaten). Zo lijkt de data analist het meest gezocht te worden, gevolgd door data scientist en als laatste de data engineers. Wat betreft vaardigheden komen de top 3 vaardigheden overeen met de algehele populatie. Het meest opvallende binnen de vaardigheden zijn "marketing" en "Tableau". Deze zijn niet veel vertegenwoordigd in de vacatures geplaatst door de meest frequente opdrachtgevers.

Wanneer we echter naar de vierde grafiek kijken zien we de meest gevraagde skills door de top 10 opdrachtgevers. Het beeld lijkt hier vertekend door Bertelsmann omdat deze een groot deel van de aanvragen uit heeft staan. Des al niet te min blijft de top 4 (sql, python, R, (neural) network) ongewijzigd. Hier verder op in gaande is ook te zien data Bertelsmann in zet op big data technologie (Hive, scala en spark). Dit zijn interessante signalene komende van de grootst potentiele opdrachtgever op Indeed.

Bovenstaande geeft handvatten om te bepalen welke skills momenteel belangrijk zijn om te ontwikkelen onder consultants en welke opdrachtgevers benaderd kunnen worden wegens het aantal vacatures. De code en artefacten die dit document opleveren geven de mogelijkheid om naar voorkeur de analyse aan te passen door bijvoorbeeld de focus van de bedrijven te verleggen (indien grootste aanbieders van functies over de tijd wijzigt).

**Ontwikkeling functie type en vraag opdrachtgevers**

Ondanks dat het niet mogelijk is om nu nog een uitspraak te doen over de ontwikkeling van vraag bij opdrachtgevers wordt er wel vast een artefact aangeleverd in de vorm van een tijdserie op basis van functietype. Deze grafiek en de bijbehorende code zijn hieronder en in de ondersteunende scripts terug te vinden voor toekomstig gebruik. Omdat er nu geen uitspraken worden gedaan i.v.m de statistische validiteit zal er geen verdere toelichting bij de grafiek te vinden zijn.

Er is hier ook overweogen een tweede artefact mee te leveren, namelijk een stuk code om het verloop van de vraag door individuele opdracht gevers weer te geven. Echter zegt de grafiek door de geringe hoeveelheid data te weinig. Om verkeerde interpetaties van de data te voorkomen wordt deze code dan ook niet meegeleverd. Zodra er genoeg data verzamelt is kan aan de hand van deze file en bijbehorende code een nieuw onderdeel worden toegevoegd om deze verandering over de tijd accuraat weer te geven.

In dit hoofdstuk koment in iedergeval de kritische verbeterpunten en tekortkoming aanbod. Ook wordt er aandacht besteed aan de deployement en wat er daadwerkelijk van waarde wordt opgeleverd.

**Data omvang**

Het eerste verbeterpunt zit in de omvang van de dataset. Het grootste deel van dit onderzoek kon goed beantwoord worden binnen de tijd en met de aanwezige middelen. Echter blijft een interessante deelvraag: de ontwikkeling van de vraag. Grotendeels onbeantwoord. Inherent aan tijdseries analyses is de lange termijn data die hiervoor nodig is. Bij aanvang van dit project was dit al een duidelijke beperking. Echter was de hoop er om met de hoeveelheid tijd die beschikbaar was toch een goede tijdsreeks op te bouwen die inzichten kon verschaffen. Dit is achteraf toch niet gelukt. Hoewel er in de basis genoeg data punten zijn om een tijdsreeks analyse te kunnen doen, maken de gewenste groeperingen (bedrijf, functie, skill) dat de dataset te ver versnipperd raakt. Dit is vooral voor toekomstige project goed om rekening mee te houden

**Data kwaliteit**

In de methoden is de datakwaliteit al kort aanbod gekomen. Toch wordt dit punt hier nog aangestipt. Door het ontbreken van een betrouwbare API, de tijdspannen van dit onderzoek en de bewerkelijkheid van de data zijn een aantal keuzes gemaakt die maken dat het eindproduct niet volledig accuraat is. De belangrijkste is de analyse omtrent functietypen. Deze is verdeeld in drie specialisaties en elke observatie in de dataset is zo goed mogelijk in een van deze drie specialisaties onderverdeeld. Echter, er zijn observaties waar eigenlijk geen van de drie specialisaties bij passen (denk aan "data entry" functies). Ook heeft er wegens tijdsgebrek maar een beperkte check plaatsgevonden op deze categorisering en andere bewerkingen zoals het omrekening naar maandsalaris of het extraheren het aantal jaren werkervaring. Dit zijn kritische verbeterpunten. Bij het door ontwikkelen van de functionaliteiten van dit project zou er daarom gekeken kunnen worden naar:

Het uitbreiden van de functietypen

betere extractie van werkervaring

betere extractie van vaardigheden

validatie van de formule en de bepaling van salarissoorten (uur, maand, en jaar)

**Data scraping**

Als laatste zijn er twee kantekeningen bij het data scrapen. Allereerst is alleen Indeed als bron gebruikt. Qua website omvang is dit misschien een goede representatie maar het zou kunnen dat hier nog een beter alternatief voor is. Het gevaar van dubbele vacatures binnenhalen bij het gebruik van meerdere scrapers is dan natuurlijk een gevaar. Wegens tijdsgebrek is hier geen verder onderzoek naar gedaan en staat dit nog open voor verbetering.

De tweede kantekening betreft de zoek titel. De scraper en data is gericht op een zo breed mogelijk profiel. Om deze reden is als zoekopdracht "data" gebruikt. Echter een verbetering doe wordt voorgesteld is de scraper om te bouwen naar een meer generieke scraper. Hierbij kan de gebruikte zoek woorden vrij worden ingevuld en opgeslagen in een variabele die ook wordt opgeslagen in de dataset. Voor dit onderzoek zou dit hebben geholpen om accurater te zoeken naar analisten, engineers, en scientisten. Dit sluit aan op de verdere kritiek onder "data kwaliteit".

**Deployment**

Als laatste onderwerp het stuk deployment. Vanuit dit onderzoek worden er een aantal artefacten opgeleverd. Deze zullen niet standaard in een productie omgeving draaien maar kunnen wel voor ad-hoc analyses of door ontwikkeling worden gebruikt.

De scripts zijn lost van dit notebook te draaien en geven House of Bèta de mogelijkheid niet alleen het onderzoek te reproduceren maar ook verder uit te breiden. In onderstaande tabel de artefacten die deel uitmaken van dit onderzoek.

Artefact

Gebruik

rmd\_Capstone\_DMAE\_MaelvDijk.Rmd

Markdown document met de analyses om de vraag "Hoe positioneert HoB wat betreft het loon ten opzichten van de ontwikkeling van de vraag naar 'data specialisten'in Nederland zich kijkende naar de gevraagde vaardigheden, het geboden loon en mogelijke opdrachtgevers?" te kunnen beantwoorden

Data\_cleaned / Data\_raw

diverse .CSV bestanden die als brondata dienen voor het hier uitgevoerde onderzoek.

indeed\_scraper.R (Supporting\_code)

webscraper voor de Indeed vacature website. Te gebruiken om nieuwe data binnen te halen. de mogelijkheid is er om de functionaliteit uit te breiden met onder andere andere zoekwoorden.

Scraped\_indeed\_data\_wrangler.R (Supporting\_code)

script om de gescrapte data (opgeslagen in .csv format) op te schonen en verijken met nieuwe datapunten. De mogelijkheid is er om de functionaliteit uit te breiden naar eigen inzien.

Scraped\_indeed\_plotter.R (Supporting\_code)

script om de gescrapte data na schonen en verrijken op een uniforme wijze te plotten. Ook dit bestand kan naar eigen inzien uitgebreid worden.

# Conclusie

Het laatste onderdeel van dit onderzoek tracht de hoofdvraag: "Hoe positioneert HoB wat betreft het loon ten opzichten van de ontwikkeling van de vraag naar 'data specialisten'in Nederland zich kijkende naar de gevraagde vaardigheden, het geboden loon en mogelijke opdrachtgevers?" te beantwoorden.

Uit het onderzoek is gebleken dat HoB haar werknemers vanaf het 3e jaar een competatief salaris lijkt te bieden. Hiermee zou de werknemer in het 3e jaar meer verdienen dan hij/zij in meer dan 50% van de gevallen had kunnen verdienen afgaande op de vacatures op Indeed. In het eerste en tweede jaar ligt het salaris boven het modale salaris op Indeed.

Wat betreft de vaardigheden die binnen HoB worden aangemoedigd om aan te leren zit HoB in het segment waar de meeste vraag naar is (sql, R, python, excel). Echter duikt hier ook een tekort op. Een groot deel van dataprofessionals bezitten deze gevraagde skills of zijn deze aan het ontwikkelen. Hoewel de vraag naar deze skills klaarblijkelijk nog steeds aanwezig is zijn dit niet de plekken waar HoB zich op kan onderscheiden. De vaardigheden waar HoB zich daadwerkelijk in zou kunnen onderscheiden zijn: neural networks, ai, machine learning, azure en aws, sas, en big data frameworks zoals spark en scala. Deze aanbeveling berust enerzijds op het resultaat uit dit onderzoek zoals de meest gevraagde skills bij de meest frequente opdrachtgevers, anderzijds komt dit vanuit persoonlijke ervaring en gesprekken tijdens het uitvoeren van mijn eigen opdrachten als data analist / data scientist.

het laatste onderdeel m.b.t de opdrachtgevers is voor nu niet te beantwoorden door het tekort aan data. Wel kan HoB haar inspanningen potentieel richten op de top 10 bedrijven (uitgezonderd andere consultant bureaus) voor haar business development, op het moment van schrijven zijn dat:

Bertelsmann SE & Co

NN Group

Rabobank

Belastingdienst

Rijkswaterstaat

Alliander

MN

ING

Amarant

ABN AMRO

In het kort, HoB investeert in de ontwikkeling van haar personeel waarbij het salaris in het 3e jaar een groot deel van de markt verslaat. Er zijn voornamelijk kansen voor HoB om met haar specialisten verder te richten op kennis van de cloud, big data oplossingen en diepgaandere kennis over ai en machine learning