DataFoundations

Portfolio

Kobe Vervoort
3ITSOF1

Inhoudsopgave

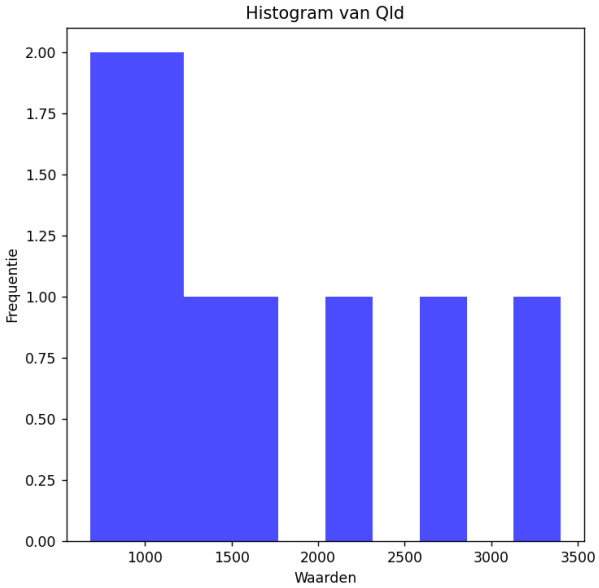
Labo 1 – Data-verwerking en -visualisatie met Python

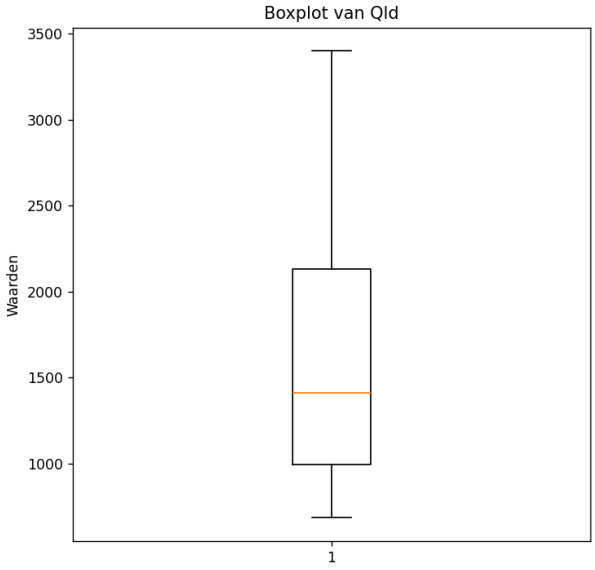
Inzichten en verbanden

In het eerste labo en de bijhorende theorieles kregen we uitleg over de verschillende vormen van data die gebruikt kunnen worden in het data-verwerkingsproces. Zo leerden we over de verschillen tussen Data Science & ML, Data Engineering en DevOps en hoe dit allemaal deel uitmaakt van MLOps. Ook kregen we een duidelijk inzicht in de bestaande datastructuren (Data Science Lifecycle) en de manieren waarop we data visueel kunnen voorstellen.

Werkproces

Tijdens het eerste labo ging ik aan de slag met de “pandas”-library in Python om zo data te verzamelen en te importeren. Hierdoor kon ik allerlei transformaties uitvoeren op de data en deze ook op allerhande manieren visualiseren. Zo lukte het me om o.a. de centrale tendens, de spreiding, een histogram en een boxplot van de verdeling te visualiseren.





	rownames	year	NSW	Vic	Qld	SA	WA	Tas	NT	ACT	Aust
0	1	1917	1904	1409	683	440	306	193	5	3	4941
1	2	1927	2402	1727	873	565	392	211	4	8	6182
2	3	1937	2693	1853	993	589	457	233	6	11	6836
3	4	1947	2985	2055	1106	646	502	257	11	17	7579
4	5	1957	3625	2656	1413	873	688	326	21	38	9640
5	6	1967	4295	3274	1700	1110	879	375	62	103	11799
6	7	1977	5002	3837	2130	1286	1204	415	104	214	14192
7	8	1987	5617	4210	2675	1393	1496	449	158	265	16264
8	9	1997	6274	4605	3401	1480	1798	474	187	310	18532

```
Gemiddelde: 1663.7777777777778
Mediaan: 1413.0
Modus: 683
Standaardafwijking: 913.125782378553
Eerste kwartiel (Q1): 993.0
Derde kwartiel (q3): 2130.0
Interkwartielafstand (IQR): 1137.0
```

Zoals je hieronder kunt zien, heb ik een zo goed mogelijke conclusie proberen trekken uit de verkregen data-visualisatie van vragen drie en vier:

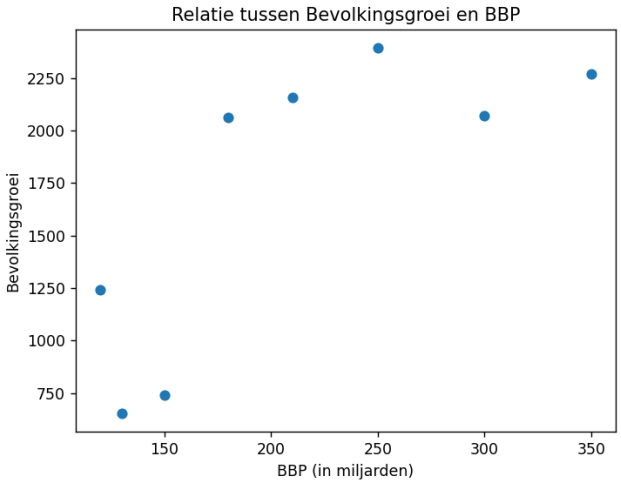
Stap 3: Interpretatie labo 1.1

Zoals we kunnen zien op de boxplot en het histogram is de data van variabele 'Vic' niet normaal verdeeld. We zien wel degelijk uitschieters. We zien in de boxplot dat de mediaan eerder rond de 2500 ligt. De spreiding van de data is groot, dit zien we aan de lengte van de boxplot. De standaardafwijking is 1177.

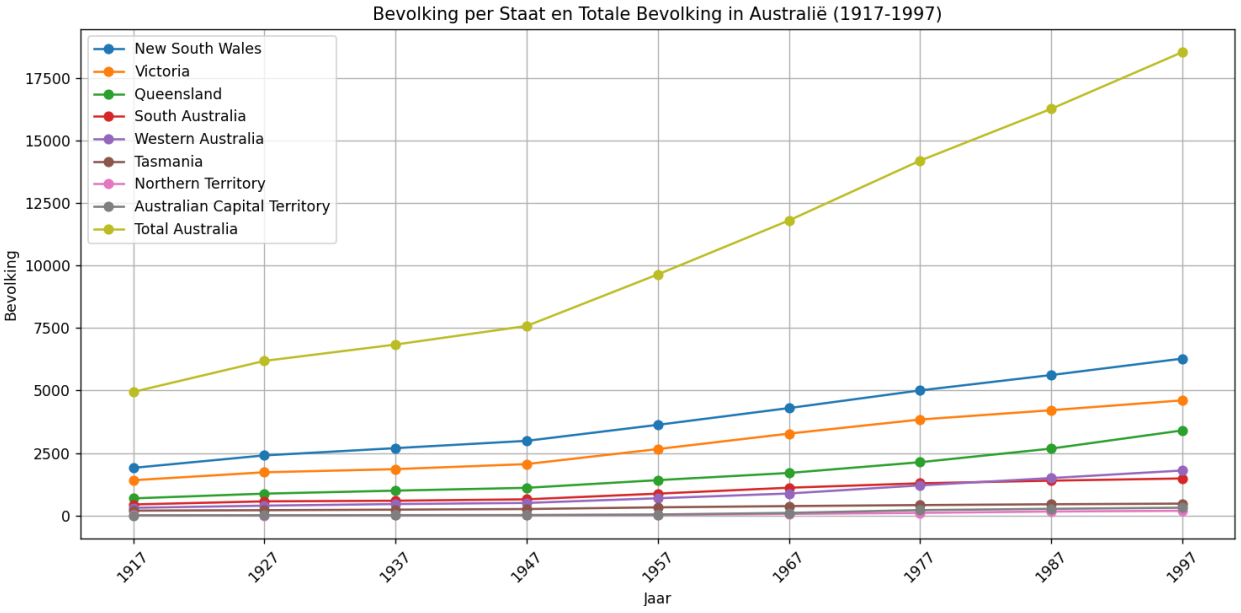
Stap 4: Rapportage labo 1.1

Bij het vergelijken van de variabelen 'Vic' en 'Qld' zien we dat het gemiddelde, de mediaan en de modus van 'Qld' hoger liggen dan die van 'Vic'. De interkwartielafstand (IQR) van 'Qld' is kleiner (1137 tegenover 1984), wat wijst op een meer geconcentreerde verdeling van de data rond de mediaan. Het histogram van 'Qld' toont een scheve verdeling met de meeste waarden onder 1000, wat op links-scheefheid wijst. 'Vic' heeft een bredere spreiding, wat duidt op meer variatie binnen de dataset. Deze verschillen kunnen wijzen op regionale of demografische verschillen in de dataset.

Ook maakte ik voor dit labo de extra opdracht. Daarin probeerde ik naast de basisanalyse ook enkele meer geavanceerde data-analysetechnieken uit. In plaats van het werken met één variabele, maakte ik nu een visuele analyse van meerdere variabelen en hun relaties.



	rownames	year	NSW	Vic	Qld	SA	WA	Tas	NT	ACT	Aust
0	1	1917	1904	1409	683	440	306	193	5	3	4941
1	2	1927	2402	1727	873	565	392	211	4	8	6182
2	3	1937	2693	1853	993	589	457	233	6	11	6836
3	4	1947	2985	2055	1106	646	502	257	11	17	7579
4	5	1957	3625	2656	1413	873	688	326	21	38	9640
5	6	1967	4295	3274	1700	1110	879	375	62	103	11799
6	7	1977	5002	3837	2130	1286	1204	415	104	214	14192
7	8	1987	5617	4210	2675	1393	1496	449	158	265	16264
8	9	1997	6274	4605	3401	1480	1798	474	187	310	18532



Conclusie

In dit labo leerde ik omgaan met data-manipulatie en -visualisatie in Python. De “pandas”-library voorziet handige functies en overzichtelijke grafieken om dit te doen. Door de data te visualiseren en dan te analyseren, kon ik gemakkelijker correlaties leggen en verkreeg ik nieuwe inzichten in de data.

Labo 2 – Power BI en data visualisatie

Inzichten en verbanden

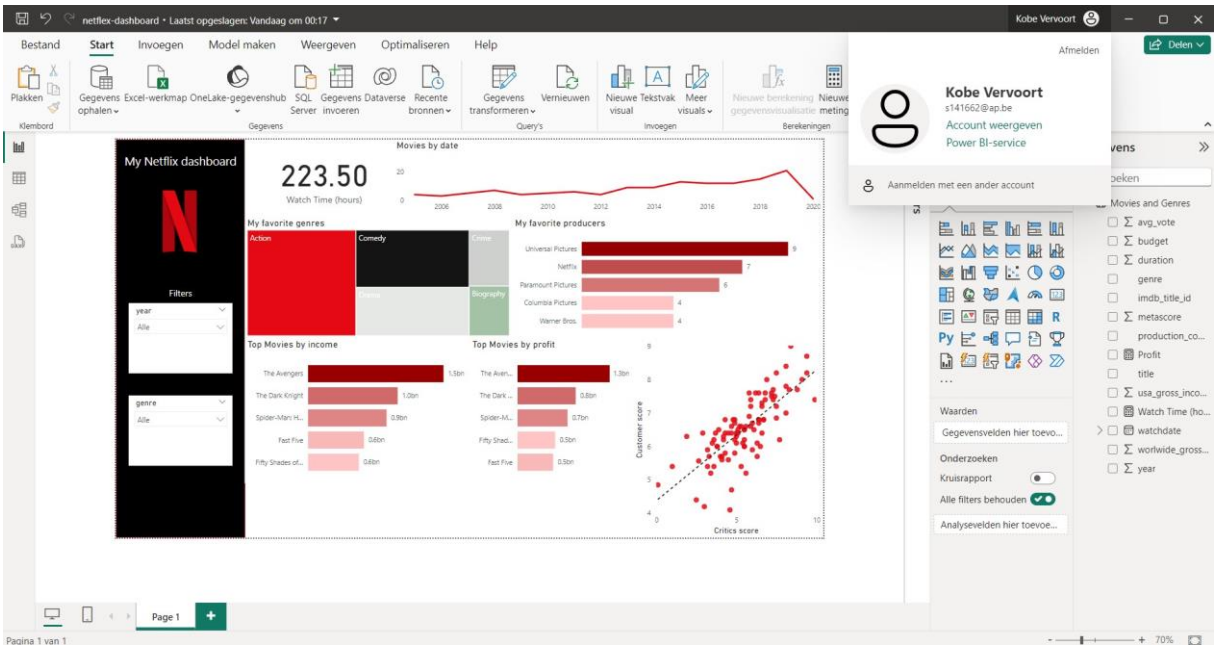
Tijdens de weken van labo 2 zijn we aan de slag gegaan met PowerBI. PowerBI is een tool van Microsoft om data te verwerken/transformeren en handig te visualiseren aan de hand van een aanpasbaar dashboard. Ook leerde ik bij over het nut van data-visualisatie en de mogelijke manieren waarop data gevisualiseerd kan worden, ieder met een andere focus, dus ieder geschikt voor een ander doel voor het weergeven van deze data.

Tijdens de theorieles leerden we dat data visualisatie erg belangrijk is. Data visualisatie zorgt er namelijk voor dat we data op een duidelijke en betere manier kunnen bestuderen en begrijpen, door dit in een visuele context te plaatsen. Door bepaalde statistieken in een andere weergave te zien, kunnen we er ook andere dingen van afleiden. Dit geeft ons andere inzichten!

Werkproces

Tijdens het eerste deel van het labo was het de bedoeling om verschillende oefeningen uit te proberen en wat wegwijs te geraken in de online versie van PowerBI. Ik volgde de gegeven tutorials van Microsoft zelf, maar raakte niet zo ver, aangezien er stappen ontbraken… Toch las ik alle tutorials door en begreep ik wel al een beetje van Power BI, omdat ik in een online virtuele omgeving dit zelf al kon verkennen. Ik leerde gegevens inladen en er dan transformaties op toepassen. Dit had ik nodig om het tweede deel van dit labo te kunnen voltooien.

Tijdens het tweede deel van het labo was het de bedoeling om het stappenplan van de docent (Dhr. Haddouchi) te volgen om zo tot een mooi Netflix-dashboard te komen. Dit is me wel goed gelukt, aangezien het gegeven stappenplan wel klopte en zeer duidelijk uitgelegd was! Zoals je op de screenshot hieronder kunt zien, kwam ik tot een mooi eindresultaat.



Conclusie

In de afgelopen lessen heb ik dus veel bijgeleerd over het gebruik van Power BI en het visualiseren van data. Ik onthoud dat ik altijd rekening moet houden met wat voor data ik wil visualiseren en de manier waarop ik dit best doe. Het blijft natuurlijk belangrijk om dit op zo'n manier te doen, dat men er de juiste inzichten uit kan halen.

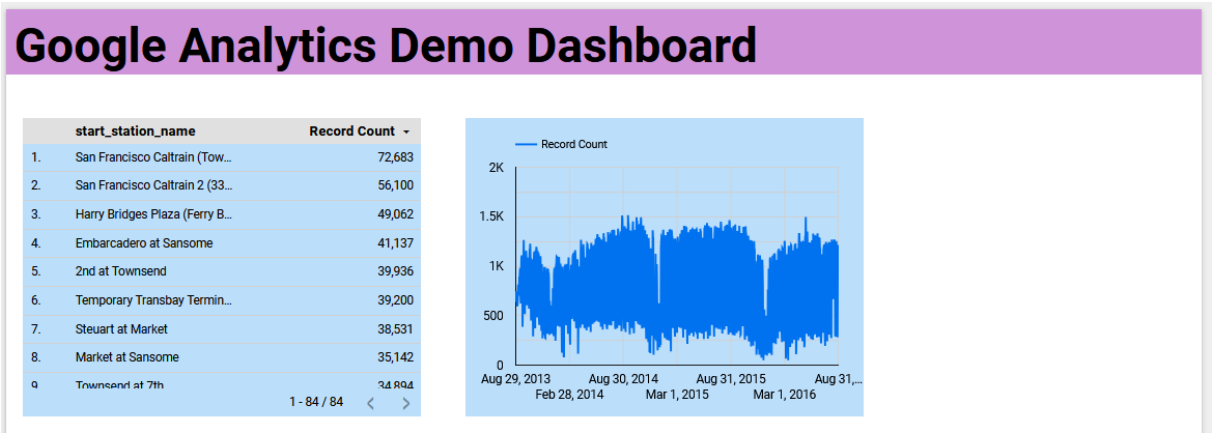
Labo 3 – Data Analytics in Google Cloud

Inzichten en verbanden (deel 1)

Tijdens de lessen van de afgelopen weken zijn we bezig geweest met het ontdekken van data analytics in Google Cloud. Google heeft namelijk ook een platform waarin men met data en analyse aan de slag kan gaan: Google Cloud. In Google Cloud vinden we enkele modules, zoals BigQuery, Looker Studio … Deze modules hebben telkens een eigen doel met een eigen functie. Zo is BigQuery bv. een enterprise data warehouse en dient het vooral voor het verwerken, transformeren en visualiseren van grote hoeveelheden data. Verder vind je ook een aantal interessante cursussen/labo's op Google CloudSkills.

Werkproces

Tijdens het eerste deel van het labo ging ik aan de slag met deze Google CloudSkills labo's. Ik maakte de introductiecursus over het "Google Analytics Demo Dashboard". Dit was niet zo moeilijk, aangezien de opgave goed uitgelegd werd. Het was interessant om hiermee aan de slag te gaan via Google Cloud. Hoewel ik eerder al wat over Google Cloud had gehoord, had ik er zelf nog niet mee gewerkt tot aan deze labo's.



Tijdens het tweede deel van het labo maakte ik een andere Google Cloud oefeningscursus, genaamd "Weather Data in BigQuery". In deze cursus ging ik aan de slag met BigQuery. Dat is een soort van framework voor het transformeren van data door het gebruik van SQL-queries.

Inzichten en verbanden (deel 2)

Tijdens de theorieles legde Dhr. Haddouchi uit dat data analytics een proces is waarbij data uit diverse bronnen en types wordt verzameld, verwerkt en geanalyseerd. In dit proces worden volgende onderdelen doorlopen: Ingest Collect, Store, Process Analyze en Consume. De 5 V's (volume, velocity, variety, veracity en value) zijn hierbij belangrijke uitdagingen. Ook werd het verschil tussen een data warehouse (relationeel, schema-on-write) en een data lake (flexibel, schema-on-read) in detail uitgelegd. Verder zagen we de batch- en streaming-dataverwerkingsmethoden en hebben we de verschillende types databases (relationele, niet-relationele en key-value) besproken.

Conclusie

Door deze verschillende soorten databronnen en opslagmethoden te begrijpen, kreeg ik een beter inzicht in hoe data verzameld, opgeslagen en geanalyseerd kan worden voor het nemen van effectieve besluitvorming en het verkrijgen van interessante inzichten. Deze lesinhoud is erg handig bij het kiezen van een geschikte database voor een bepaald project als toekomstige software-ontwikkelaar.

Labo 4 – Data Management

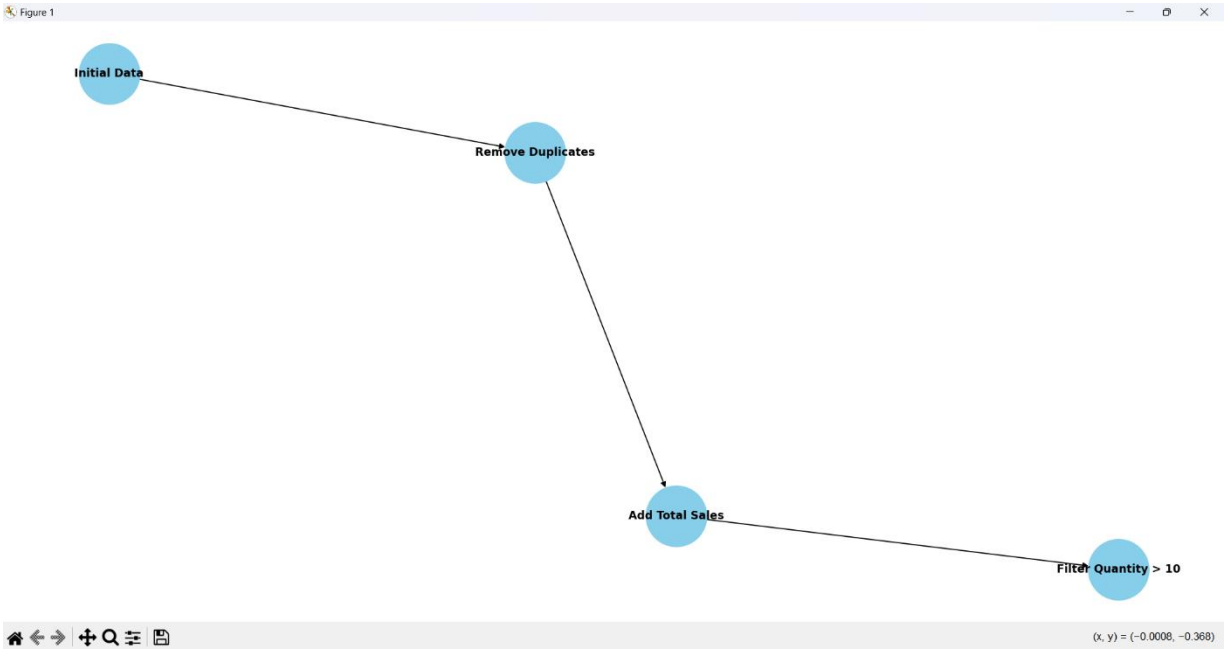
Inzichten en verbanden (deel 1)

Tijdens de theorieles en de labo's van afgelopen weken zijn we bezig geweest met data management en data lineage. Data management gaat over het behandelen, filteren en transformeren van data, zodat de datakwaliteit en privacy geoptimaliseerd kunnen worden. Data lineage gaat daarentegen over het traceren van elke stap in de transformatie van gegevens, waardoor de transparantie, controleerbaarheid en kwaliteit van de data verzekerd kan worden.

Werkproces

In het eerste deel van het labo ben ik aan de slag gegaan met Looker Studio. Looker Studio is een tool van Google waarmee we datavisualisaties en interactieve dashboards kunnen creëren vanuit verschillende databronnen. Ik maakte op Google CloudSkills het online labo over "Visualizing Billing Data with Looker Studio". Aangezien in de vorige labo's Looker Studio ook al even aangehaald werd, maar omdat we hier nog niet mee gewerkt hadden, vond ik het dus erg interessant om hiermee aan de slag te kunnen gaan.

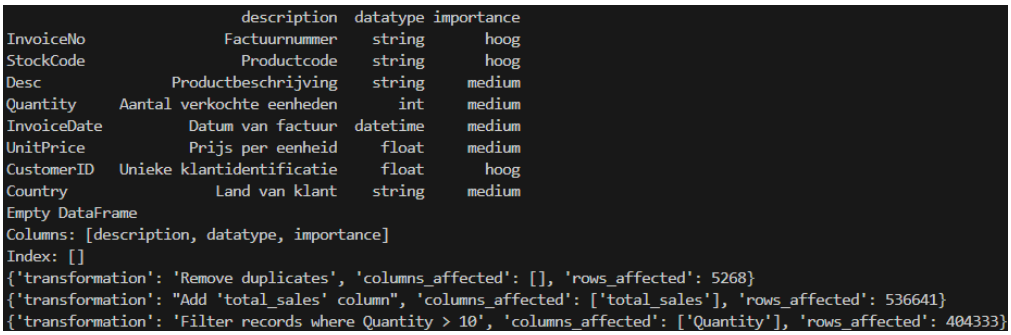
Het tweede deel van het labo ging vooral over data lineage en deze visualiseren door middel van de "pandas"-library van Python. Ook maakte ik een data catalogus en maakte ik metadata aan.



Samenvatting van de metadata en data catalogus (labo 4.2):

Op onderstaande afbeeldingen zien we de gegenereerde data catalogus, het overzicht van de metadata die bij elke kolom hoort en de data lineage.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France



```
                        description  datatype importance
InvoiceNo               Factuurnummer    string       hoog
StockCode                 Productcode    string       hoog
Desc              Productbeschrijving    string     medium
Quantity     Aantal verkochte eenheden       int     medium
InvoiceDate          Datum van factuur  datetime     medium
UnitPrice             Prijs per eenheid     float     medium
CustomerID   Unieke klantidentificatie     float       hoog
Country                 Land van klant    string     medium
Empty DataFrame
Columns: [description, datatype, importance]
Index: []
{'transformation': 'Remove duplicates', 'columns_affected': [], 'rows_affected': 5268}
{'transformation': "Add 'total_sales' column", 'columns_affected': ['total_sales'], 'rows_affected': 536641}
{'transformation': 'Filter records where Quantity > 10', 'columns_affected': ['Quantity'], 'rows_affected': 404333}
```

Hieronder vind je mijn eindrapportage en reflectie op de data lineage grafiek van labo 4.2:

> ***Reflectie labo 4.2:***
>
> *De transformaties die de data heeft ondergaan, worden weergegeven in de data lineage grafiek. Zoals we kunnen zien op de grafiek werden volgende transformaties uitgevoerd op de data:*
>
> 1. ***Verwijder duplicaten:** Deze transformatie verwijderde rijen die identieke waarden bevatten voor alle kolommen. Het doel was om de dataset schoon te maken en de overbodige records eruit te halen.*
> 2. ***Toevoegen van de "total_sales" kolom:** Deze kolom werd toegevoegd door de prijs per eenheid (UnitPrice) te vermenigvuldigen met het aantal verkochte eenheden (Quantity). Dit helpt bij het berekenen van de totale verkoopwaarde voor elke transactie.*
> 3. ***Filteren op Quantity > 10:** Deze transformatie filterde alle records uit waarin de hoeveelheid (Quantity) kleiner was dan 10. Dit werd gedaan om te focussen op de grotere sales, die relevanter zijn voor sales analyses.*
>
> *Deze transformaties hadden telkens een andere impact op de dataset. Door duplicaten te verwijderen werd de dataset schoner, wat de betrouwbaarheid verhoogde. De toevoeging van de "total_sales" kolom maakte het mogelijk om de omzet per transactie te berekenen. Het filteren van de dataset op grotere hoeveelheden zorgt voor een beter overzicht van de belangrijkste sales.*
>
> *Data lineage is belangrijk, aangezien het helpt bij het traceren van elke stap in de transformatie van gegevens, waardoor de transparantie, controleerbaarheid en kwaliteit van de data verzekerd kan worden. Dit is cruciaal voor het maken van betrouwbare analyses en handig bij het trekken van conclusies (uit de data) en het maken van besluiten.*

Inzichten en verbanden (deel 2)

In de theorieles leerde ik dat data management cruciaal is om bruikbare data te verkrijgen voor besluitvorming. Het omvat stappen zoals dataverzameling, doelbepaling, kwaliteitscontrole, delen en toegankelijkheid. Dit wordt gestructureerd door het DMBoK Wheel, met data governance als centrale component, en focus op data-architectuur, modellering, opslag, integratie, masterdata en kwaliteit. Metadata is hierbij belangrijk voor context. Data governance zorgt voor accurate en veilige databehandeling. Master Data Management (MDM) zorgt voor consistente kerngegevens. De data lineage grafiek toont alle transformaties die data ondergaat, zoals het verwijderen van duplicaten, het toevoegen van een nieuwe kolom en het filteren van rijen, wat cruciaal is voor de transparantie en betrouwbaarheid.

Conclusie

Tijdens dit labo heb ik geleerd dat data management draait om het optimaliseren van datakwaliteit en privacy door middel van verschillende processen en tools, waarbij data governance centraal staat. Met de visualisaties van Looker Studio kon ik data visualiseren, terwijl de focus op data lineage me het belang van transparantie en controle in datatransformaties heeft laten zien. Het creëren van een data catalogus en metadata in Python heeft mijn inzicht vergroot in hoe data getransformeerd kan worden. Kortom, ik heb geleerd hoe ik data kan beheren, analyseren en visualiseren om bruikbare inzichten te verkrijgen en deze om te zetten in een betrouwbaar besluit.

Labo 5 – Leren uit data

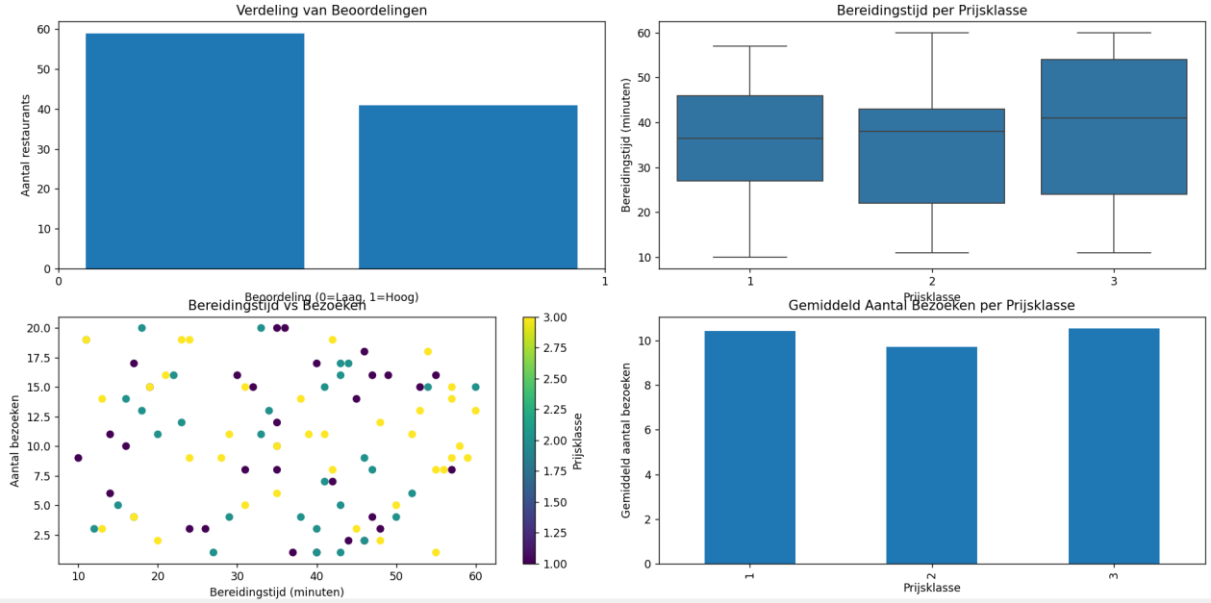
Inzichten en verbanden

Tijdens de lessen van afgelopen weken leerde ik omgaan met data om er zo conclusies uit te kunnen trekken. Ik leerde hoe Ik correlaties kon interpreteren en hoe er op deze manier aan fraudedetectie gedaan kan worden. Ik gebruikte hiervoor een zelfgetraind model.

Ook leerde ik over de mogelijkheid tot classificatie door middel van supervised en unsupervised learning. Verder werd er gesproken over reinforcement learning en werden classificatie en regressie met elkaar vergeleken. We zagen ook enkele veelvoorkomende data, waaronder clustering en anomalie detectie.

Werkproces

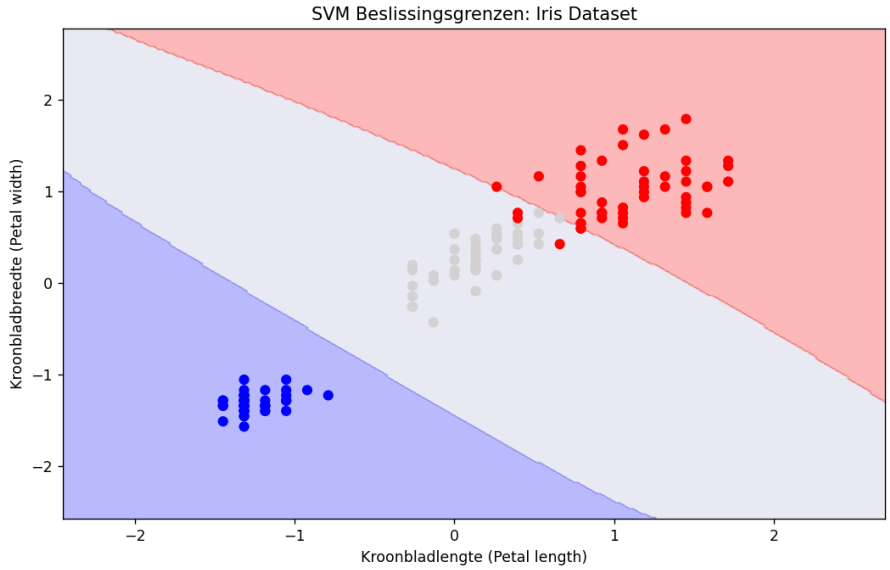
Tijdens het eerste deel van het labo was het de bedoeling om wat wegwijs te geraken met het trainen van een data model met de “pandas”-library in Python. Ik deed dit aan de hand van een zelf gegenereerde dataset. Daarna trainde ik mijn model (20-80 ratio) op deze dataset en vervolgens analyseerde ik deze. Ik kon enkele resultaten aflezen waaronder bereidingstijden en prijsklassen en hoe deze verhouding hielden tot elkaar.

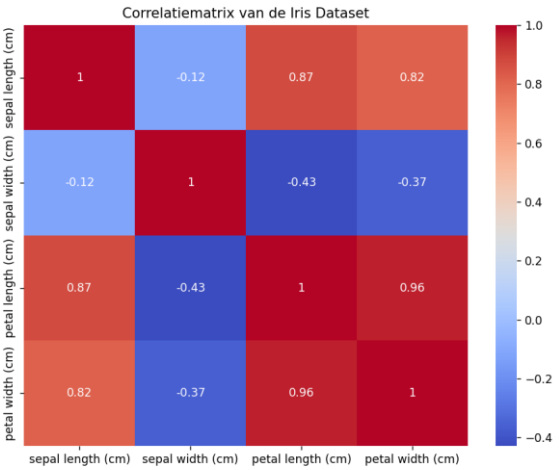


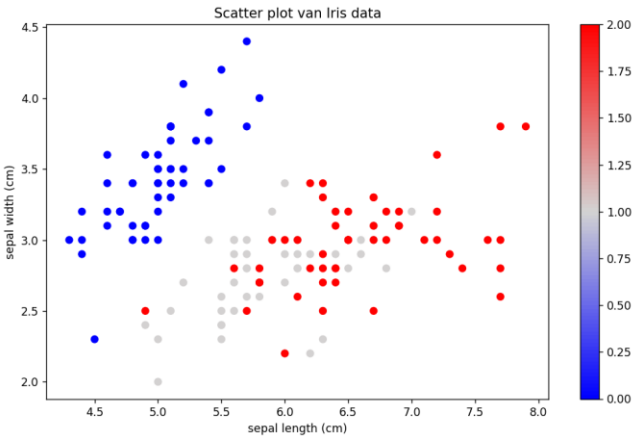
```
Gemiddelde waarden per kolom:
prijs              2.07
bereidingstijd    36.44
bezoeken          10.24
beoordeling        0.41
dtype: float64

Mediaan waarden per kolom:
prijs              2.0
bereidingstijd    38.5
bezoeken          10.5
beoordeling        0.0
dtype: float64
```

```
Gemiddelde bereidingstijd per prijsklasse:
prijs
1    35.666667
2    34.363636
3    38.918919
Name: bereidingstijd, dtype: float64

Gemiddelde beoordeling per prijsklasse:
prijs
1    0.666667
2    0.484848
3    0.135135
Name: beoordeling, dtype: float64
```

Tijdens het tweede deel van het labo maakte ik het Google CloudSkills labo over “Getting Started with BigQuery ML”. Dit gaf me een inkijk in hoe BigQuery ook ondersteuning heeft voor het werken met ML. Ook werd er verwacht dat ik een Python script maakte, waarmee ik een dataset over irissen kan analyseren. Het was de bedoeling dat er aan classificatie gedaan werd en dat deze resultaten ook gevisualiseerd werden. Ik koos twee variabele factoren om te vergelijken met elkaar. Ook stelde ik een correlatiematrix op, waarin ik duidelijk maakte hoe sterk de verhoudingen tussen alle factoren zijn.







Zoals je hieronder kunt zien, heb ik geprobeerd om de vragen uit het labo (5.2) zo goed mogelijk te beantwoorden:

> ***Welke kenmerken zijn het meest informatief voor de classificatie van bloemen?***
>
> *In de Iris-dataset zijn er vier kenmerken (features) die worden gebruikt voor de classificatie van bloemen. Van deze kenmerken zijn kroonbladlengte en kroonbladbreedte het meest informatief voor de classificatie van de verschillende Iris-soorten. Dit komt doordat deze kenmerken meestal een duidelijkere scheiding/beter verschil aantonen tussen de verschillende klassen (soorten) in de dataset.*
>
> ***Hoe kan de correlatiematrix helpen bij het vereenvoudigen of verbeteren van het model?***
>
> *Aangezien de correlatiematrix de correlatie tussen de verschillende kenmerken in de dataset aantoont, kan deze helpen bij het vereenvoudigen/verbeteren van het model op volgende manieren:*
>
> 1. *Identificatie van redundante kenmerken: Als twee kenmerken sterk gecorreleerd zijn, kan het nuttig zijn om één van hen te verwijderen. Dit vermindert de complexiteit van het model zonder al te veel informatie te verliezen.*
> 2. *Selectie van kenmerken: Door te kijken naar de correlatie tussen kenmerken en de doelvariabele (de klasse), kun je bepalen welke kenmerken het meest informatief zijn voor de classificatie. Dit kan helpen bij het selecteren van de beste subset van kenmerken voor het model.*
> 3. *Verbetering van modelprestaties: Door alleen de meest informatieve kenmerken te gebruiken, kan het model sneller trainen en betere prestaties leveren, omdat het minder ruis en irrelevante informatie bevat.*
> 4. *Visualisatie van relaties: De correlatiematrix kan helpen bij het visualiseren van de relaties tussen kenmerken, wat nuttig kan zijn voor het begrijpen van de data en het maken van beslissingen over feature engineering.*

Conclusie

Tijdens deze les leerde ik vooral de manieren waarop er gegroepeerd kan worden met data. Dit omvat o.a. regressie en classificatie. Ook leerde ik bij over wanneer ik deze het best kan gebruiken en waarom deze nuttig zijn. Om een goede conclusie te kunnen trekken uit een dataset is het classificeren van data en het visualiseren van correlatie erg van belang!

Labo 6 – Data ethics

Inzichten en verbanden

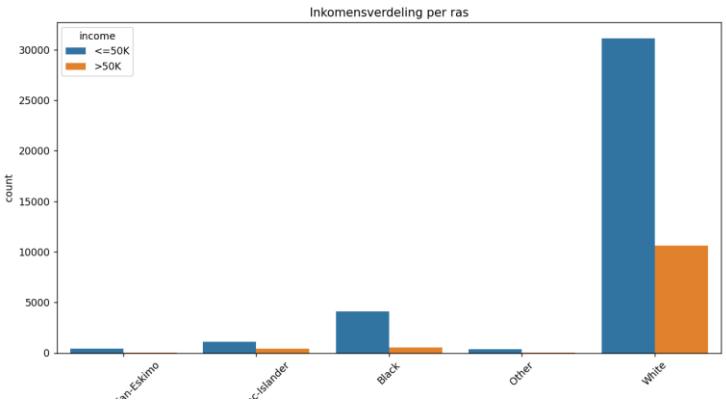
De lessen van afgelopen week gingen vooral over de morele kant van data verwerking en de evolutie van AI. Er werden verschillende kanttekeningen gemaakt over data privacy en de GDPR. Ook werd er voldoende aandacht geschonken aan het fenomeen bias. Bias is namelijk een vertekening in het data-resultaat doordat er te weinig aandacht wordt geschonken aan een onevenwichtige verdeling in de brongegevens. Hier moet altijd rekening mee gehouden worden, want anders krijg je vertekende resultaten, die een negatieve invloed kunnen hebben op het nemen van beslissingen waarvan deze data resultaten aan de basis liggen.

Verder werd er ingegaan op enkele ethische vraagstukken, waaronder het Trolley Problem. Hierbij werd er verduidelijkt dat ethiek, data en ML erg sterk verbonden zijn. Bij het maken van een AI willen we immers ook dat deze AI de ethische beslissingen neemt, die het meest logisch en ethisch verantwoord zijn.

Werkproces

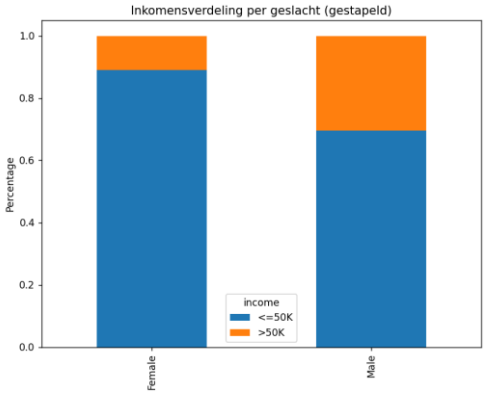
Bij het eerste deel van het labo was het de bedoeling om een verslag te schrijven over een gegeven case. Ik schreef dit verslag zo goed mogelijk uit met mijn visie op data-ethiek centraal. Ik denk dat dit me wel vrij goed gelukt is, aangezien ik rekening heb gehouden met alle mogelijke factoren zoals GDPR, morele aspecten, privacy en de AI-act.

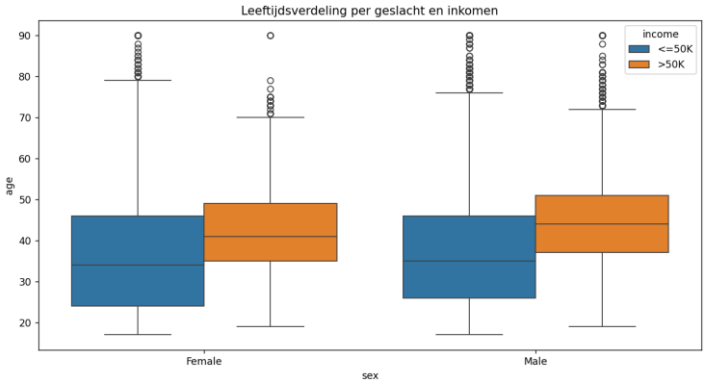
In het tweede deel van het labo ging ik weer aan de slag met de “pandas”-library in Python. Het doel van de opdracht was om een script te maken dat een gegeven dataset zou classificeren. Dit deed ik door de dataset in te laden, deze dan uit te lezen, transformaties er op uit te voeren, bewerkingen er mee te doen, een model te trainen voor het classificeren (20-80 ratio) en deze resultaten in een visuele voorstelling te gieten. Ook werd er gevraagd om de data te anonimiseren.




```
Percentage van inkomens per geslacht:
income      <=50K      >50K
sex
Female  0.890749  0.109251
Male    0.696233  0.303767

Verdeling van rassen:
race
White                 41762
Black                  4685
Asian-Pac-Islander     1519
Amer-Indian-Eskimo      470
Other                   406
Name: count, dtype: int64
```

Op bovenstaande resultaten zien we dat het witte ras over het algemeen veel meer verdient dan de andere rassen. Dit klopt echter niet, aangezien het witte ras echter veel meer vertegenwoordigd is in de dataset dan alle andere rassen. Hierdoor zou bias kunnen ontstaan als ik geen rekening zou houden met het relatief aantal. Daarom is het belangrijk dat de verhouding per ras wordt berekend en dan op relatieve basis vergeleken wordt met de verhoudingen van de andere rassen.





Anonymized Dataframe:

	workclass	fnlwgt	education	education-num	marital-status	occupation	...	native-country	class	income	age_group	sex_anon	race_anon
0	Private	226802	11th	7	Never-married	Machine-op-inspct	...	United-States	<=50K	0	26-35	M	A
1	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	...	United-States	<=50K	0	36-45	M	B
2	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	...	United-States	>50K	1	26-35	M	B
3	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	...	United-States	>50K	1	36-45	M	A
4	NaN	103497	Some-college	10	Never-married	NaN	...	United-States	<=50K	0	18-25	V	B

Hieronder vind je mijn reflectie op de gevonden bias, de prestatie-analyse van het model, de interpretatie van de Fairness-metrics en de impact van de anonimisatie op de data:

> ***Reflectie labo 6.2:***
>
> *De dataset vertoont een duidelijke bias, namelijk bias in inkomen tussen mannen en vrouwen en tussen de verschillende rassen. Het model presteert beter voor de meerderheidsgroepen (mannen, wit) en minder goed voor de minderheidsgroepen (vrouwen, andere rassen). Fairness metrics zoals Demographic Parity en Equal Opportunity laten zien dat er sprake is van ongelijkheid tussen de groepen. Anonimisatie door het groeperen van leeftijd en het vervangen van geslacht en ras door generieke labels verlaagt de mogelijkheid om direct informatie uit de dataset te herleiden. Het nadeel is een kleine achteruitgang in de modelprestaties. De keuze om wel of niet te anonimiseren en in welke mate is een afweging tussen privacy en de nauwkeurigheid van het model die we willen behalen.*

Conclusie

Tot slot gaf Dhr. Haddouchi ons ook nog enkele best practices voor ethische ontwikkeling in data mee. Het is namelijk belangrijk dat we rekening houden met zaken als eigenaarschap, transparantie in dataverzameling, privacy, intentie en uitkomst van data-anlyse bij het werken met data.

Uit deze les neem ik vooral het ethische en relatieve aspect mee. Hiermee bedoel ik dat het rekening houden met deze zaken cruciaal is en aan de basis ligt van het werken met data en AI. Het is van belang dat de verhouding van verkregen data-resultaten altijd mee verrekend wordt!

Bronnen

Dashboard. (2024, oktober 27). Retrieved from Google Cloud Skills Boost: https://www.cloudskillsboost.google/

Google Cloud Platform lets you build, deploy, and scale applications, websites, and services on the same infrastructure as Google. (2024, oktober 27). Retrieved from Google Cloud Platform: https://console.cloud.google.com/

pandas documentation - pandas 2.2.3 documentation. (2024, september 20). Retrieved from pydata: https://pandas.pydata.org/docs/

s.n. (2024, oktober 10). *Oefening: gegevens laden in Power BI Desktop*. Retrieved from Microsoft Learn: https://learn.microsoft.com/nl-nl/training/modules/clean-data-power-bi/8-lab