DataFoundations

Portfolio

Kobe Vervoort   
3ITSOF1

Inhoudsopgave

[Labo 1 – Data-verwerking en -visualisatie met Python 3](#_Toc182746828)

[Labo 2 – Power BI en data visualisatie 5](#_Toc182746829)

[Labo 3 – Data Analytics in Google Cloud 7](#_Toc182746830)

[Labo 4 – Data Management 10](#_Toc182746831)

[Bronnen 11](#_Toc182746832)

# **Labo 1 – Data-verwerking en -visualisatie met Python**

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.

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.

Afbeelding met tekst, diagram, schermopname, Perceel

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijving

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*

*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*

*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 hun relaties.

Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijvingAfbeelding met tekst, schermopname, scherm, lijn

Automatisch gegenereerde beschrijvingAfbeelding met tekst, lijn, Perceel, schermopname

Automatisch gegenereerde beschrijving

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**

Deze week zijn we tijdens het labo 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 van 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!

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.

Afbeelding met tekst, schermopname, software, Computerpictogram

Automatisch gegenereerde beschrijving

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**

Tijdens de labo’s van 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. 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.

Tijdens het eerste deel van het labo ging ik aan de slag met deze 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.

Afbeelding met tekst, schermopname, software, Lettertype

Automatisch gegenereerde beschrijving

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.

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.

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**

Tijdens de theorieles en de labo’s van afgelopen weken zijn we bezig geweest met data management en datalineage. 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.

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.

Afbeelding met schermopname, lijn, tekst, diagram

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermopname, Lettertype, zwart-wit

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermopname, menu, Lettertype

Automatisch gegenereerde beschrijving

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.

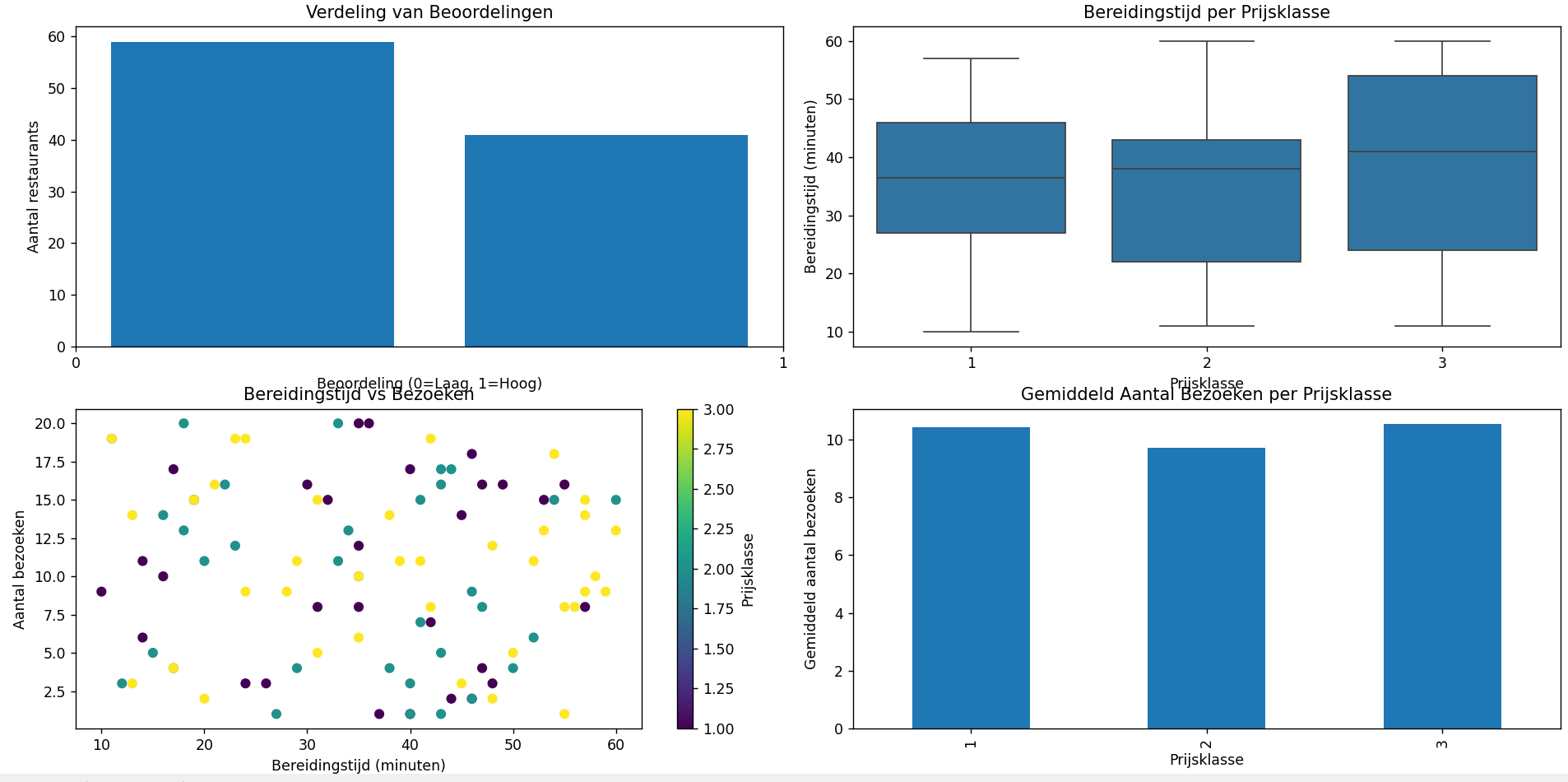
Tijdens dit labo heb ik geleerd dat data management draait om het optimaliseren van datakwaliteit en privacy door middel van verschillende processen en tools, waarbij data governance centraal staat. Met de visualisaties van Looker Studio kon ik data visualiseren, terwijl de focus op data lineage me het 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**

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 tegenover elkaar afgetoetst. We zagen ook enkele veelvoorkomende data, waaronder clustering, anomalie detectie.

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) op deze dataset en vervolgens analyseerde ik deze. Ik kon enkele resultaten aflezen waaronder bereidingstijden en prijsklasses en hoe deze verhouding hielden tot elkaar.



Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijvingAfbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijving

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 hen te vergelijken met elkaar. Ook stelde ik een correlatiematrix op, waarin ik duidelijk maakte hoe sterk de verhoudingen tussen alle factoren zijn.

Afbeelding met tekst, schermopname, Lettertype, diagram

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermopname, plein, Rechthoek

Automatisch gegenereerde beschrijvingAfbeelding met tekst, schermopname, diagram, Perceel

Automatisch gegenereerde beschrijving

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**

De lessen van afgelopen week gingen vooral over de morele kant achter 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 beslissing met deze data resultaten als argument.

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.

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, bewerken er mee te doen, een model te trainen voor het classificeren (20-80) en deze resultaten in een visuele voorstelling te gieten. Ook werd er gevraagd om de data te anonimiseren.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijving

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.

Afbeelding met tekst, schermopname, diagram, Rechthoek

Automatisch gegenereerde beschrijving Afbeelding met diagram, tekst, schermopname, Rechthoek

Automatisch gegenereerde beschrijvingAfbeelding met schermopname, Lettertype

Automatisch gegenereerde beschrijving

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