DataFoundations

Portfolio

Kobe Vervoort   
3ITSOF1

Inhoudsopgave

[Labo 1 – Data-verwerking en -visualisatie met Python 3](#_Toc178458133)

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

Deze week hebben we het in de les gehad over het gebruik van data in ons dagelijkse leven. We zagen de verschillen tussen Data Science & ML, Data Engineering en DevOps en hoe dit allemaal deel uitmaakt van MLOps. Ik leerde hier veel over bij en stond versteld van de hoeveelheid Deep Learning die eigenlijk aanwezig is in ons leven (of de hoeveelheid van de data die we uit dagelijkse dingen kunnen halen, die gebruikt kan worden voor ML en DL).

Verder zagen we enkele types van data: gestructureerd, semi-gestructureerd en ongestructureerd. Ook leerde ik dat er een massa aan tools is voor het visualiseren van data, waaronder de “pandas”-library in Python. Ook zagen we de Data Science Lifecycle en hoe deze in zijn werk gaat.

De Data Science Lifecycle is een handig model waardoor je aan de hand van zes stappen gemakkelijk conclusies kunt trekken uit een bepaalde dataset (en zijn visualisatie). Dit heeft mij erg geholpen bij het oplossen van labo 1 en vooral bij het beschrijven en het nemen van conclusies bij vragen drie en vier. Door middel van de Data Science Lifecycle is het eenvoudiger om aan Data Storytelling te doen. Dit heeft me bewuster gemaakt van het gebruik van data in de hedendaagse wereld. Data wordt namelijk overal gebruikt: als informatiebron, als besluitvorming, maar ook als AI.

Verder vond ik het leuk en interessant om in de labo’s (labo 1 en labo 1 extra) zelf aan de slag te kunnen gaan met enkele datasets door middel van Python. Ik heb dan ook veel geleerd uit de labo-oefeningen, namelijk hoe ik op een nuttige en overzichtelijke manier data kan visualiseren en hier conclusies uit kan trekken.

Dit vond ik in het begin echter eerder moeilijk, aangezien ik niet goed wist wat er juist van mij verwacht werd dat ik neerschreef. Gelukkig kon ik hier hulp en feedback voor vragen aan Dhr. Haddouchi. Hij legde de theorie, maar ook de opdrachten erg grondig en op een goed tempo uit, zodat we de lesinhoud goed konden begrijpen. Uiteindelijk slaagde ik er wel in om de labo’s te voltooien.

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.*

# **Labo 2 – Power BI en datavisualisatie**

Deze week zijn we tijdens de lessen 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 leerden we bij over het nut van datavisualisatie en de mogelijke manieren waarop data gevisualiseerd kan worden, elks met een andere focus, dus elks geschikt voor een ander doel van het weergeven van deze data.

Tijdens de theorieles zagen we dat datavisualisatie erg belangrijk is. Datavisualisatie zorgt er namelijk voor dat we data op een duidelijke en betere manier kunnen bestuderen en begrijpen, doordat het in een visuele context wordt geplaatst. Door bepaalde statistieken in een andere weergave te zien, kunnen we er ook andere dingen van afleiden. Dit geeft ons andere inzichten!

Echter blijft het wel belangrijk om de keuze van de manier waarop je de data gaat visualiseren steeds af te stemmen op (1) het publiek, de personen voor wie je deze visualisatie realiseert, (2) de boodschap, datgene dat je graag wil overbrengen, maar ook (3) de impact die je wil bereiken aan de hand van je data-visualisatie. Voor elke “case”/context is er een andere, beter passende manier van visualiseren. Aan de hand van elk verschillende weergave of ander diagram, zul je je boodschap anders overbrengen.

In de Business Intelligence is datavisualisatie handig voor het analyseren van data en het bekomen van nuttige inzichten; op deze manier kan men die business ondersteunen. Daarom zijn er ook verschillende types van rapporten: statisch, interactief en dashboards. Ook deze rapporten zijn verschillende soorten van visualisaties die in verschillende contexten gebruikt worden en die in bepaalde gevallen nuttiger en duidelijker zullen zijn dan de andere types.

Verder maakten we kennis met enkele datavisualisatie tools, zoals Microsoft Power BI (dit gebruikten we in de labo’s), Tableau en QlikView. Ook hebben we enkele mogelijke visualisaties en diagrammen overlopen, zoals een tabel/matrix, staafdiagram, lijn- en oppervlaktediagram, spreidingsdiagram (scatter plot), kaarten, taart- en donutdiagram, tree map, combo diagram …

Ook leerden we over de “Grammar of Graphics”. Dit wil zeggen dat we bij het maken van een visualisatie steeds rekening moeten houden met zes componenten! Deze componenten helpen bij het begrijpen en het verduidelijken van de data die we willen weergeven. Aan de hand van deze componenten kunnen we betere inzichten geven en kunnen we bepaalde gegevens er doen uitspringen/opvallen!

1. Aan de hand van welke **data** we willen tonen, zullen we een keuze maken voor het diagram/de visualisatie waarop deze data het beste tot uiting zal komen.
2. Aan de hand van de getoonde, **esthetische** componenten (assen, relatieve posities, kleur, grootte …), kunnen we bepaalde inzichten bekomen.
3. Aan de hand van hoe elk onderdeel van de data wordt weergegeven (**geometrische** vormen), kunnen we gemakkelijk bepaalde verschillen zien.
4. De gebruikte **facetten** helpen bij het opmerken van verschillen tussen relatieve diagrammen.
5. Door het verschil tussen de **statistische** gegevens die getoond worden, zullen we andere, duidelijkere manieren van visualisatie gebruiken.
6. Ook de **coördinaten** en assen hebben invloed op het gegeven inzicht van een bepaald visualisatie/diagram.

Ten slotte kregen we nog enkele tips voor het maken van een datavisualisatie, zoals hou het functioneel, vermijd taartdiagrammen en 3D-grafieken, probeer niet te veel data tegelijk weer te geven (data overload), toon steeds de volledige as …

Daarna gingen we zelf aan de slag met Power BI tijdens de labo’s. Tijdens labo 1 was het de bedoeling om wat wegwijs te geraken in deze nieuw applicatie. 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.

Tijdens labo 2 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! Ik kwam tot een mooi eindresultaat.

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.

# **Bronnen**

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