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.*

sdfbronnen!