Bachelor Toegepaste Informatica

Portfolio

Roan Heylen

3ITSOF1

Vak: Data Foundations

Lector: Haddouchi Hassan

Academiejaar: 2024-2025

1. Data-verwerking en visualisatie met Python

A. Intro

In deze labo en theorieles leerde we de basis van data. Welke soorten datastructuren er zijn, en hoe we deze soorten data visueel kunnen weergeven en interpreteren.

B. Interpretatie – Labo 01

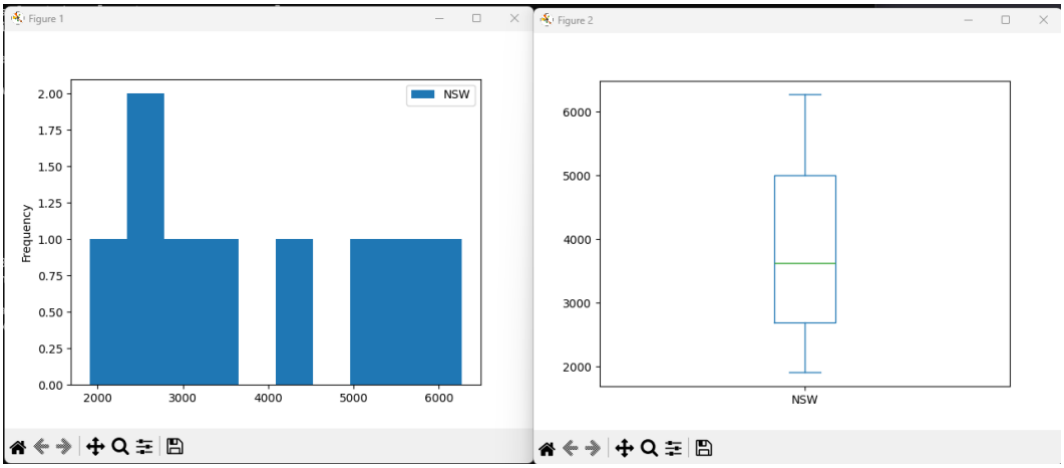
Populatie van: Western Australia population counts
We kunnen beide aan de mediaan zien dat doorheen de jaren de populatie zeer lang rond de 600-700 personen hing,

maar aan de som kunnen we zien dat dat niet doorheen de jaren altijd zo was. Som = 858 personen

Ook kunnen we zien aan de boxplot dat de mediaan lager hangt dan het gemiddelde, dit betekend dat de populatie toename of afname heeft getoond.

(Dat het niet volledig stabiel over de jaren was)

Er zijn geen uitschieters, er is geen normaal verdeling, omdat de modus en het gemiddelde niet hetzelfde zijn.



	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

```
Data: NSW
Sum: 3866.3333333333335
Median: 3625.0
Mode: 1904
Standard deviation: 1522.4122306392576
Interkwartielafstand 3625.0
```

C. Rapportage – Labo 01

Vergeleken met Western Australia heeft New South Wales een grotere populatie.

Aan de mediaan en de som kunnen we zien dat de populatie niet stabiel was.

Aan de boxplot kunnen we zien dat het verschil tussen het minimum en het maximum zeer groot is, dit betekend dat de populatie grote veranderingen heeft gezien.

Bij de frequentie is er een uitschieter: namelijk rond de 2000 populatie.

Dit betekend dat de populatie voor een tijd stabiel was rond de 2000 personen, waarbij hij dan afgenomen of toegenomen was.

D. Conclusie

Data kan soms gestructureerd zijn, maar er kunnen ook soms uitschieters liggen in data. Door het te kunnen visualiseren van alle data kan men gemakkelijker beslissingen trekken uit deze data.

Met tools zoals Pandas, is dit vrij gemakkelijk te doen in Python met enkel een paar lijnen code.

2. Data visualisatie met Power BI

A. Intro

In deze labo/theorieles kregen we een beeld hoe we data kunnen visualiseren met Power BI.

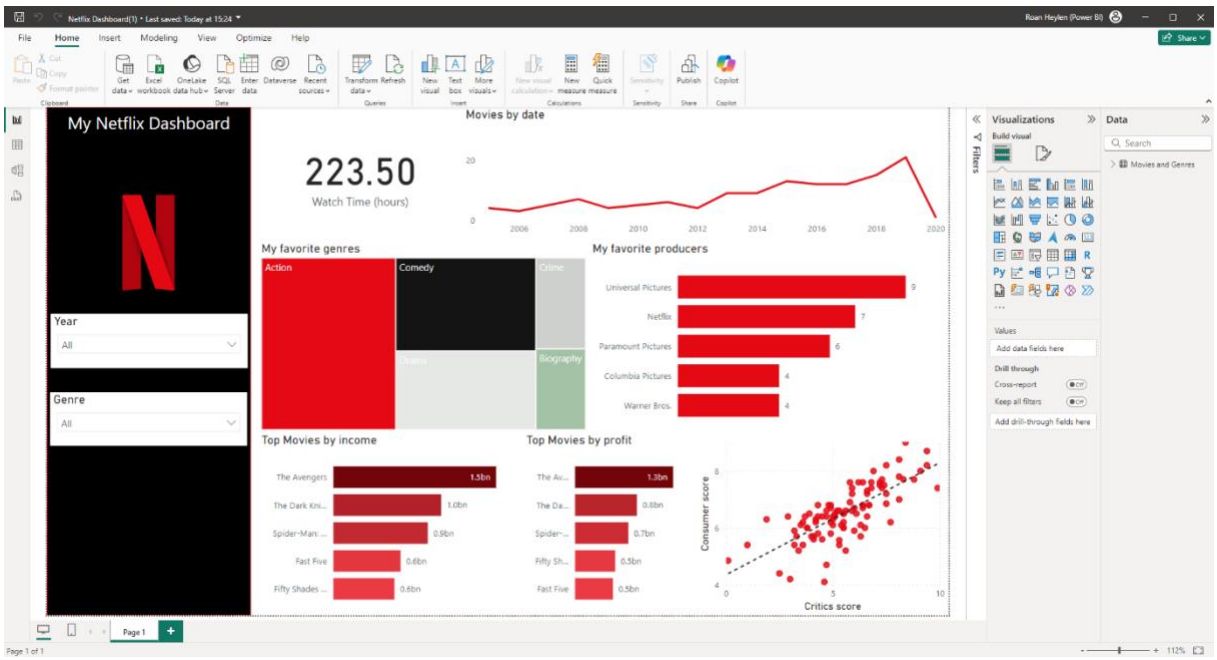
We leerde de verschillende soorten rapporten en visualisaties, en maakte hierbij zelf een dashboard van in Power BI.

We leerde de componenten van een grafiek, en kregen tips over hoe een grafiek best opgesteld wordt.

B. Labo

We hebben in de labo een dashboard gemaakt voor het te kunnen visualiseren van Netflix kijkersdata.

Met deze labo leerde we de ins-en-outs van PowerBI, hierdoor leerde ik over de verschillende soorten diagrammen die PowerBI in zijn arsenaal heeft, maar ook hoe we nu juist een grafiek moeten opstellen.



C. Conclusie

Diagrammen en dashboards maken met PowerBI was een zeer simpel proces. Met dit programma konden we echtelijke data omzetten naar een overzichtelijk dashboard.

In plaats dat we beslissingen uit ruwe data moesten maken, konden we dit doen uit een visuele representatie.

3. Data analytics on Google Cloud / Looker studio

A. Intro

In deze labo leerde we hoe we beslissingen kunnen maken uit data. Waar we over de verschillende componenten van data analytics leerde.

Ook, leerde we over de 5 V's, of terwijl de uitdagingen bij data. Met deze 5 V's kunnen we gemakkelijk beslissingen trekken uit data door de bepaalde eigenschappen van de data te bekijken.

We leerde over Volume, namelijk wat data warehouses, data marts en data lakes zijn.

Waarbij we ook leerde over de snelheid van data, en welke uitdagingen dat daar bijkomen. Zoals batch data of streaming data.

Daar bovenop leerde we over de verschillende soorten data: gestructureerd, semi-gestructureerd en ongestructureerd.

Flat files waren een voorbeeld van nog uitdagingen bij data, want bij een flat file hebben we geen relaties tussen objecten, duplicate waarden, en zo veel meer.

B. Labo

Bij de labo werkte we met Google Cloud, we kregen informatieve video's over het gebruik van BigQuery en Data Studio.

Ik leerde ik bij over het querieën van data met SQL queries, en het kunnen beslissingen maken met data door deze queries.

Met Looker Studio leerde ik hoe ik data uit BigQuery moest nemen en hiermee, net zoals PowerBI, data kan visualiseren door middel van een dashboard.

C. Conclusie

Bij deze labo en theorieles leerde ik meer over hoe ik beslissingen kan maken uit data en welke soorten data er en uitdagingen er nu juist zijn.

Hiermee kan ik betere beslissingen maken, en weet ik waar voor soort data ik met bezig ben, om gemakkelijker een beslissing te trekken uit data.

Ook, heb ik bijgeleerd hoe ik met BigQuery en Looker/Looker studio kan werken.

4. Data lineage

A. Intro

In deze les leerde we over hoe we data moeten onderhouden en bruikbaar kunnen blijven houden. En hoe we dit visueel konden weergeven met data lineage.

We leerde over het Body of Knowledge Wheel, een schema bedoeld voor de verschillende focusgebieden van datamanagement.

Ook leerde we over data governance waarbij een focus werdt gelegd over GDPR met een voorbeeld van ING.

We leerde ook over Master Data Management: het onderhouden van één enkele bron van waarheid bij kritieke gegevens.

B. Labo

In de labo leerde we over data visualisatie met Looker Studio, en data lineage visualisatie in Python.

Met data lineage krijgen we een mooi overzicht over hoe data eigenlijk veranderd doorheen het traject van deze data.

C. Samenvatting metadata en data catalogus

Overzicht van metadata:

```
[5 rows x 8 columns]
Data Catalogus:
    ColumnName              Description  Datatype Importance
0    InvoiceNo            Factuurnummer    string       hoog
1    StockCode              Productcode    string       hoog
2  Description    Productbeschrijving    string     medium
3     Quantity  Aantal verkochte eenheden       int     medium
4  InvoiceDate      Datum van de factuur  datetime       hoog
5    UnitPrice         Prijs per eenheid     float     medium
6   CustomerID           Unieke klant-ID     float       hoog
7      Country         Land van de klant    string     medium
```

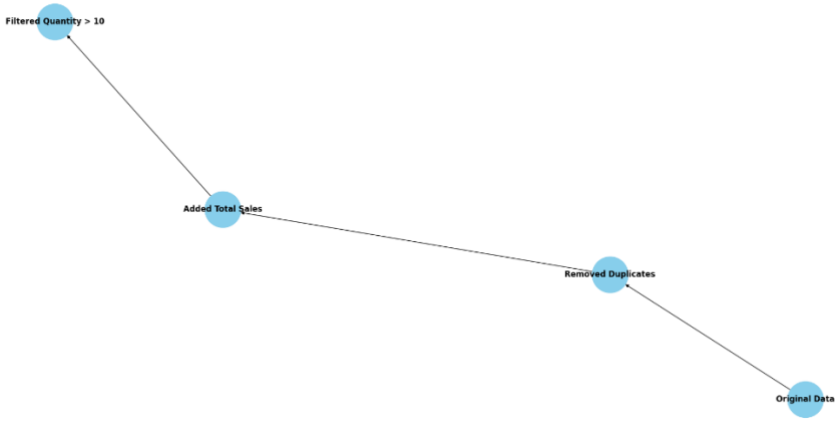
Gegenereerde data catalogus:

```
Gegenereerde data catalogus:
   InvoiceNo StockCode                          Description  ...  CustomerID         Country  total_sales
0     536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER  ...     17850.0  United Kingdom        15.30
1     536365     71053                  WHITE METAL LANTERN  ...     17850.0  United Kingdom        20.34
2     536365    84406B       CREAM CUPID HEARTS COAT HANGER  ...     17850.0  United Kingdom        22.00
3     536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE  ...     17850.0  United Kingdom        20.34
4     536365    84029E       RED WOOLLY HOTTIE WHITE HEART.  ...     17850.0  United Kingdom        20.34
```

D. Data lineage

De data heeft de volgende veranderingen ondergaan:

- Duplicaten zijn verwijdert uit de originele data
- Er is een kolom “total_sales” toegevoegd, dat aantoont hoeveel omzet er was per product.
- We hebben gefilterd enkel op data dat meer dan 10 stuks heeft verkocht.



Data lineage is belangrijk, want dan kunnen we duidelijk weergeven welke stappen de data is ondergaan voordat het uiteindelijk opgekuist is.

Het geeft ons een mooie visuele weergave, maar het kan ook gebruikt worden om fouten op te sporen.

E. Conclusie

Data lineage geeft een goed overzicht van hoe datatransformatie er aan toe gaat. Het zorgt ervoor dat we data bruikbaar houden, en dit gemakkelijker kunnen gebruiken bij het visualiseren van data.

Zelf, heb ik veel bijgeleerd over het transformeren van data met Pandas.

5. Leren uit data

A. Intro

We leerde over hoe fraudedetectie wordt gedaan door middel van ML/voorspellen van data.

Ook leerde we supervised vs unsupervised leren, waarbij je een model traint met labels of zonder labels.

We leerde wat reinforcement learning is: het trainen van een model met trail and error. Waarbij we leerde wat classificatie, regressie en clustering is.

We leerde wat voor machine learning algoritmes er bestaan voor models te trainen, waarbij: lineaire regressie, logitstische regressie en een beslissingsboom.

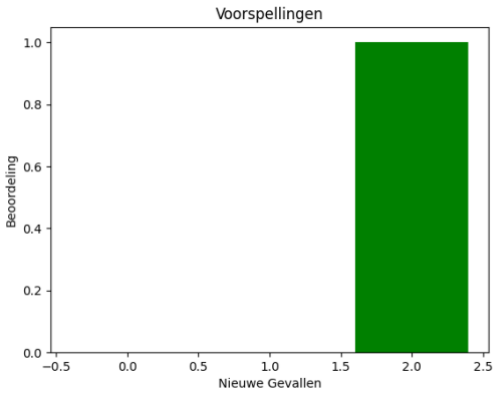
Ook, leerde we over het fenomeen “overfitten”, waarbij je een model te goed trained waardoor het alleen goed werkt op de exacte trainingsdata dat je de model geeft.

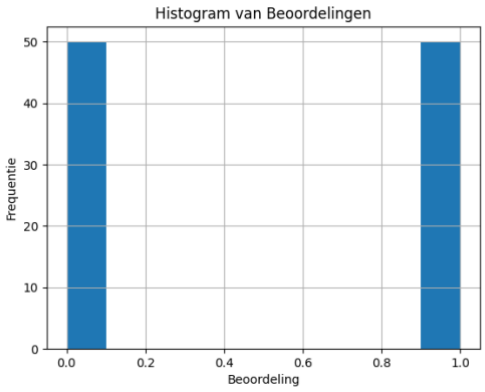
B. Labo

In deze labo leerde we een voorspellingsmodel te trainen op beoordelingsdata van een restaurant (willekeurig gegenereerd.)

	prijs	bereidingstijd	bezoeken	beoordeling
count	100.000000	100.0000	100.000000	100.000000
mean	1.980000	31.9700	9.130000	0.500000
std	0.803779	15.4994	6.089658	0.502519
min	1.000000	5.0000	1.000000	0.000000
25%	1.000000	17.7500	3.000000	0.000000
50%	2.000000	32.5000	8.000000	0.500000
75%	3.000000	43.2500	16.000000	1.000000
max	3.000000	59.0000	19.000000	1.000000

Nauwkeurigheid: 0.35





We kunnen zien dat het model niet zo nauwkeurig is met het voorspellen van beoordelingen, dit komt doordat we willekeurige data gebruiken.

Wel kunnen we zien dat het model voorspelt dat voor een verhoogde prijsverhoging of bereidingstijd en meer bezoeken, er hogere beoordelingen zullen zijn.

C. Conclusie

Bij deze labo heb ik bijgeleerd hoe ik een voorspellingsmodel kan maken, welke voordelen dat voorspellen van data heeft en hoe ML/voorspelling wordt gebruikt in verschillende applicaties.

6. Data ethics & Bias

D. Intro

We leerde over het juridische kader bij data.

Termen zoals: GDPR, CCPA. Waarbij een druk werdt gelegd op GDPR, een Europese wetgeving voor het verwerken van persoonsgegevens.

We leerde ook over morele problemen, zoals: welke beslissing is de correcte beslissing voor ML/AI? Het trolleyprobleem is hier een voorbeeld van.

Ook, leerde we dat er bias kan onstaan wanneer we een model met een grote dataset trainen. Het kan er voor zorgen dat de ML/AI onrechtvaardige beslissingen nemen.

Hierbij moest ik ook een verslag schrijven over een case, waarbij ik over de 4 framework componenten bij bias moest praten: vertrouwen, transparantie, eerlijkheid en privacy.

E. Labo

F. Welke bias heb ik gevonden?

G. Prestaties van het model

H. Fairness-metrics

I. Impact van anonimisatie op data

J. Conclusie

Ik heb bijgeleerd dat we ook stil moeten staan bij deze componenten wanneer we modellen trainen/data verzamelen en verwerken.

Ook, heb ik bijgeleerd dat er een groot juridisch kader zit achter data.