Portfolio Assignment 102

Een Data Scientist heeft verschillende taken bij een project. Dit noemen ze ook wel de Data Science Life Cycle. In figuur 1 is te zien hoe deze cycle eruit ziet.

Figuur 1 - Data Science Life Cycle

De meeste van deze stappen heb ik ook gedaan bij het maken van de Assignments. Om te laten zien wat een Data Scientist nou precies doet zal ik alle stappen nagaan met voorbeelden van de Assignments die ik heb gedaan.

Business Understanding

Als Data Scientist moet je natuurlijk precies weten wat het doel is van een project zodat je weet welke data je nodig hebt en wat je met die data moet gaan doen. Deze stap is natuurlijk niet echt relevant voor mij omdat ik alleen opdrachten moest maken, maar ik vond het wel interessant om bij die opdrachten na te denken welke data ik ook echt nodig had, wat ik ermee ging doen en wat dan uiteindelijk het doel was van de opdracht.

Data Mining

Het vinden van goede data in cruciaal voor een Data Scientist. Slechte data betekend natuurlijk slechte voorspellingen. Om zelf een dataset te vinden ben ik gaan zoeken op Kaggle.com waar veel interessante datasets te vinden zijn. Uiteindelijk is mijn oog op een dataset van Stockx.com gevallen waarin data te vinden was van verkochte sneakers in Amerika in een bepaalde tijdsperiode. Vervolgens is het natuurlijk de volgende stap om de dataset in je programma te laden zoals te zien is in figuur 2.

```
In [45]: stockx = pd.read_excel('StockX_Dataset.xlsx')

In [47]: stockx.head()
```

Out[47]:

	Order Date	Brand	Sneaker Name	Sale Price	Retail Price	Release Date	Shoe Size	Buyer Region
0	2017-09-01	Yeezy	Adidas-Yeezy-Boost-350-Low-V2-Beluga	1097.0	220	2016-09-24	11.0	California
1	2017-09-01	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Copper	685.0	220	2016-11-23	11.0	California
2	2017-09-01	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Green	690.0	220	2016-11-23	11.0	California
3	2017-09-01	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Red	1075.0	220	2016-11-23	11.5	Kentucky
4	2017-09-01	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017	828.0	220	2017-02-11	11.0	Rhode Island

Figuur 2 – Inladen van dataset

Data Cleaning

Een dataset kan inconsistente of verkeerde gegevens bevatten dus daarom is het ook belangrijk om deze aan te passen of weg te filteren. Een voorbeeld hiervan zijn kolommen die geen waarde hebben en dus null zijn. Sommige Machine Learning algoritmes kunnen niet goed om gaan met missende waardes dus daarom is het van belang om deze niet in je dataset te houden. Hoe ik dat heb gedaan bij de Assignments is te zien in figuur 3.

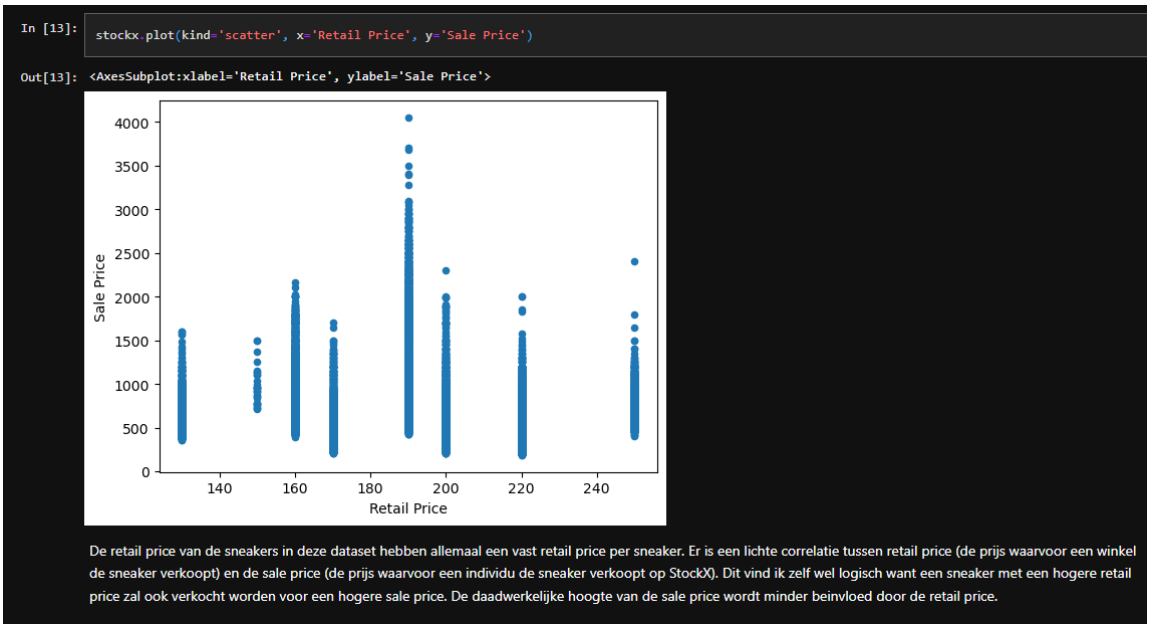
```
In [5]: stockx = pd.read_excel('StockX_Dataset.xlsx').dropna()
        stockx.head()
```

Figuur 3 – Weghalen null waardes

Op deze manier worden rijen met null waardes verwijderd. Een ander voorbeeld van Data Cleaning dat ik moest doen was het weghalen van dollar tekens. Bij de prijzen in mijn dataset waren deze toegevoegd maar daardoor kon ik hier geen Numerical Analasys op doen. Ook was er de Brand kolom waarbij het merk “Yeezy” een spatie ervoor stond “ Yeezy”. Deze moest ik dus ook weghalen.

Data Exploration

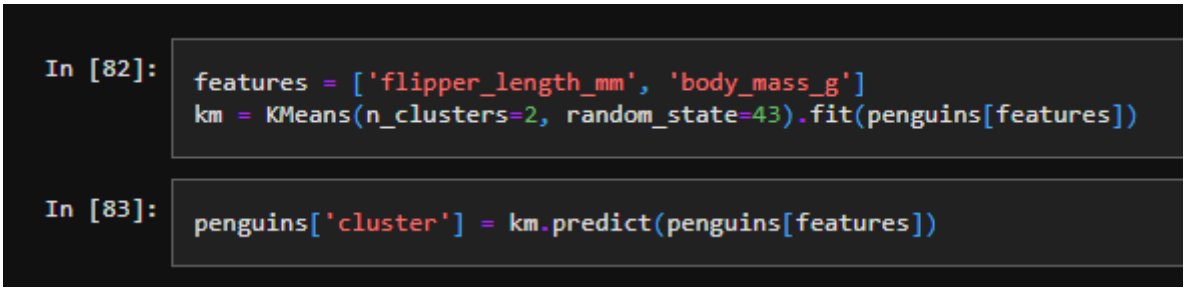
Het maken van hypotheses kan goed gedaan worden door je data visueel te tonen. Dit is ook iets van Data Scientists vaak doen. Het is namelijk veel makkelijker om visueel verbanden te vinden dan naar het staren van cijfers. Bij Assignment 10 heb ik dit gedaan om mogelijke correlaties te vinden tussen kolommen. Een voorbeeld hiervan is te zien in figuur 4.



Figuur 4 – Visualisatie correlatie

Feature Engineering

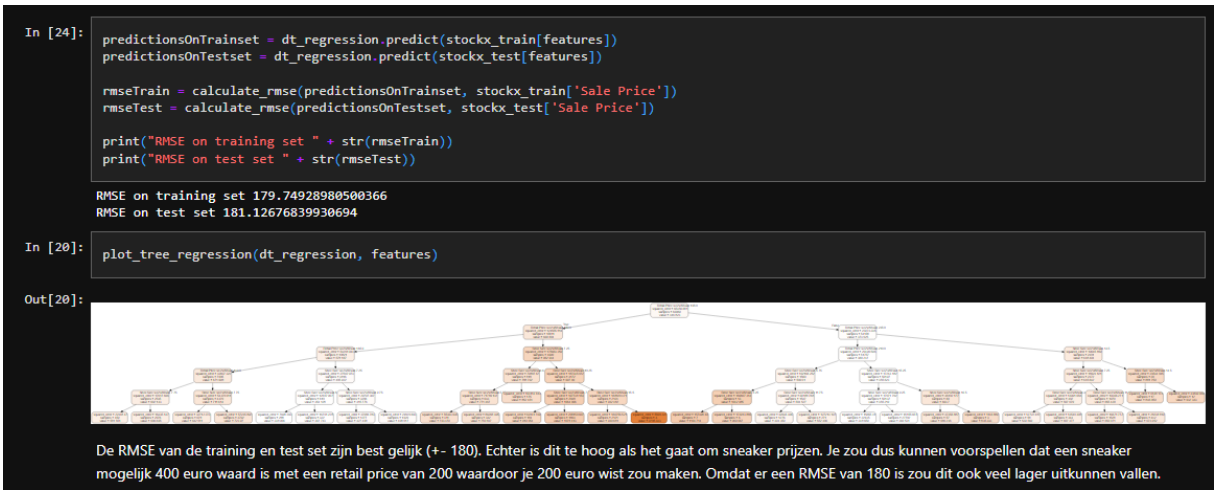
Feature Engineering is de stap die nodig is voordat je gebruik kan gaan maken van Artificial Intelligence en Machine Learning. Een voorbeeld hiervan is het weghalen van lege waardes zoals uitgelegd is bij Data Cleansing. Wat ook nodig is het aangeven van bepaalde features dat als invoergegevens gebruikt kan worden door het Machine Learning model. Bij Assignment 19 heb ik dit gedaan door clusters te vinden bij penguins.



Figuur 5 – Feature Engineering

Predictive Modeling

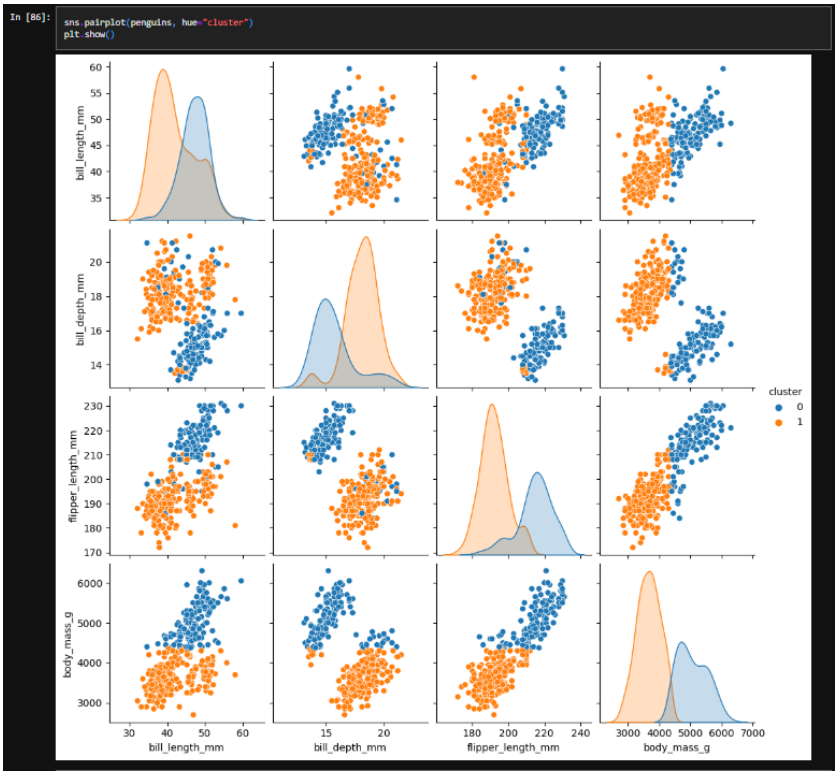
Het trainen van Machine Learning modellen is een onderdeel waar een Data Scientist erg bekend mee is, voor mij was dit helemaal nieuw. Data Scientist maken hiervan gebruik om gegevens te voorspellen op basis van andere gegevens. Ik heb dit bij Assigment 18 ook gedaan door te kijken of ik de prijzen van sneakers kan voorspellen op basis van een merk. Dit is te zien in het onderstaande figuur.



Figuur 6 – Voorspellen van Sale Price

Data Visualisatie

Tot slot is er de laatste stap om de gevonden gegevens te visualiseren voor de stakeholders. Vaak hebben stakeholders niet veel tijd om echt in data te duiken dus visualisatie is voor deze personen erg belangrijk. In mijn geval zijn er natuurlijk niet echt stakeholders van het visualiseren van data is wel iets wat ik heb gedaan bij Assignment 19 zoals te zien is in figuur 7.



Figuur 7 – Visualisatie van clusters