Data Science Report

Ali Biyikli

2168022

2021

Contents

[1. Wat is Data Science? 2](#_Toc66308890)

[2. De taken van een Data Scientist 3](#_Toc66308891)

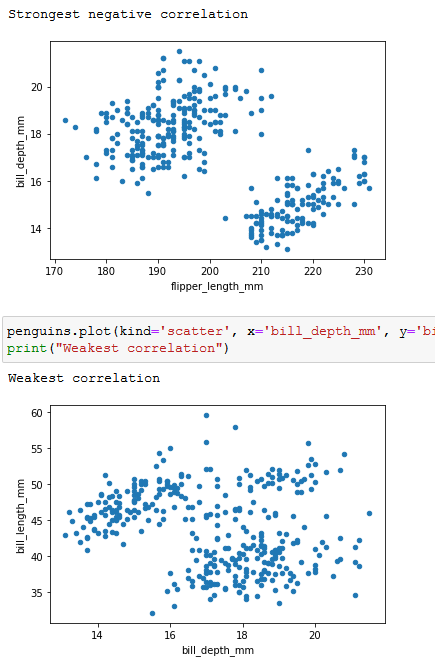
[2.2 Business Understanding 3](#_Toc66308892)

[2.3 Data Acquisition & Understanding 4](#_Toc66308893)

[2.4 Modeling 5](#_Toc66308894)

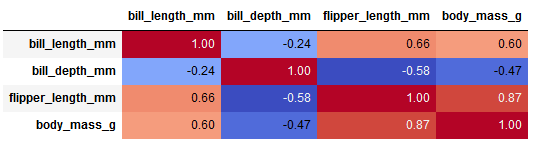
# Wat is Data Science?

Het hoofddoel van Data Science is om waarde en inzicht te krijgen uit data. Dit kan bereikt worden door bestaande data te analyseren en vervolgens te visualiseren in de vorm van grafieken en tekeningen. Een data scientist gebruikt tools om dit te kunnen bereiken. Er zijn veel tools beschikbaar, zoals DataRobot, PowerPivot, SAS, PowerBI, Jupyter Notebook etc.



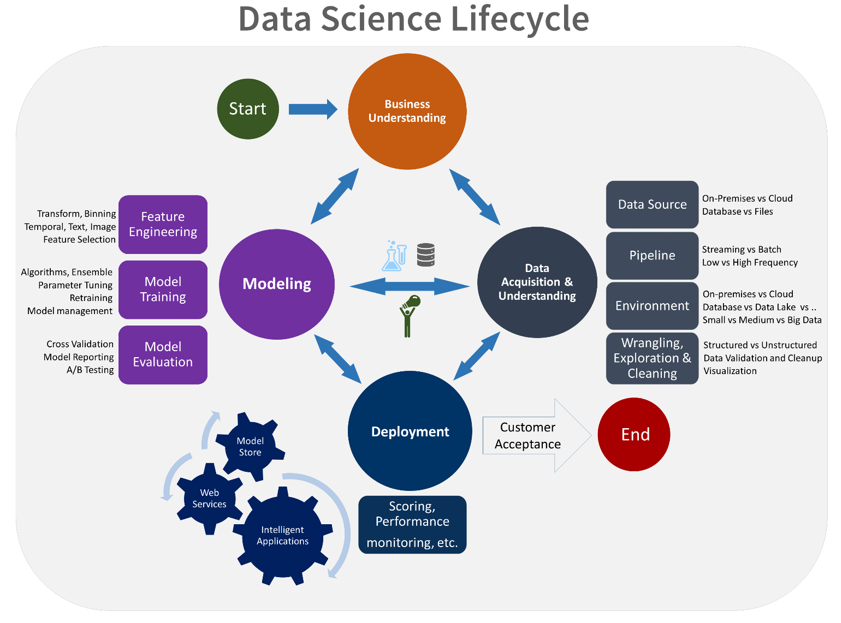
Een goede voorbeeld is Assignment 9 in mijn portfolio.   
Je kan hier zien dat er een dataset word ingeladen en   
vervolgens word gezocht naar correlaties tussen twee   
verschillende gegevens (bivariate analysis).

*correlaties tussen de colommen*



Met behulp van het analyseren en modelleren van de gegevens kan je dus mogelijk conclusies trekken met als wenselijk doel om de nieuw verkregen kennis en inzicht te kunnen gebruiken om een bedrijf te helpen met keuzes maken.

# De taken van een Data Scientist

Een data scientist heeft meerdere taken. Hieronder is een plaatje te zien van de lifecycle van een data scientist. Ik ga een aantal van de portfolio opdrachten gebruiken als voorbeelden om uit te leggen wat de taken zijn.

## 2.2 Business Understanding

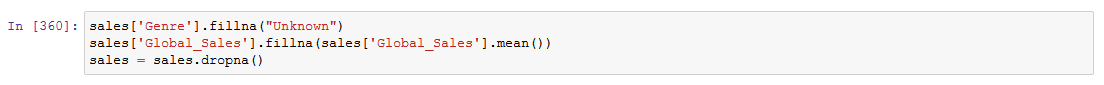
De business understanding houd in dat er een onderliggende begrip is voor de data die geanalyseerd worden, bijv. klant gegevens, of bloemen, of pinguïns of videogames. Deze context moet vanaf het begin worden begrepen, maar tijdens het proces van data analyseren word het begrip ook sterker.

## 2.3 Data Acquisition & Understanding

Voordat je data kunt gebruiken voor analyse moet natuurlijk is de data worden verzamelt. In elke portfolio opdracht word een dataset ingeladen:



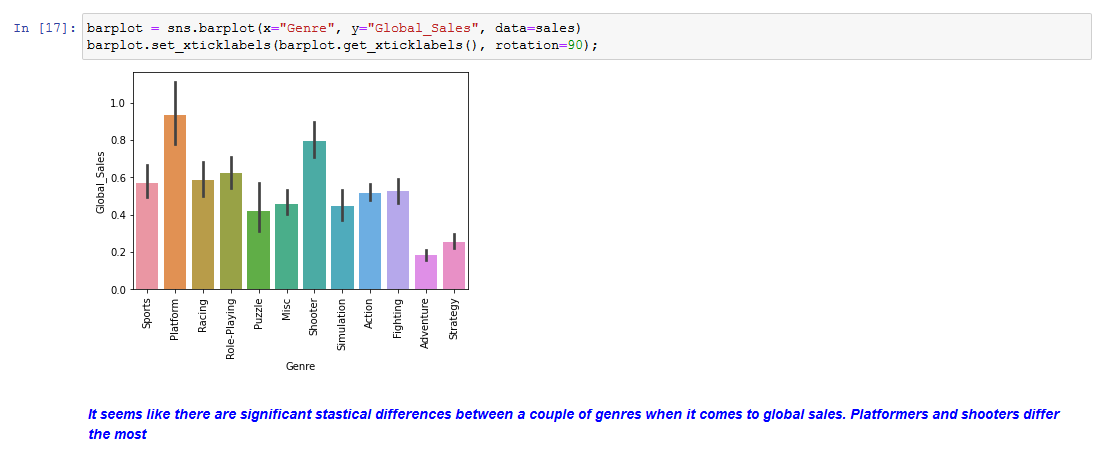
Vervolgens moet de data ook opgeschoond worden zodat hij bruikbaar is voor de gewenste analyse. Soms is het namelijk niet mogelijk om lege velden te hebben, zoals bijv. bij Decision Trees. Daar moet dus iets aan gedaan worden. In bijv. assignment 15 en 16 worden legen velden opgevuld maar ook verwijderd, zodat de data bruikbaar is voor de opkomende analyse:

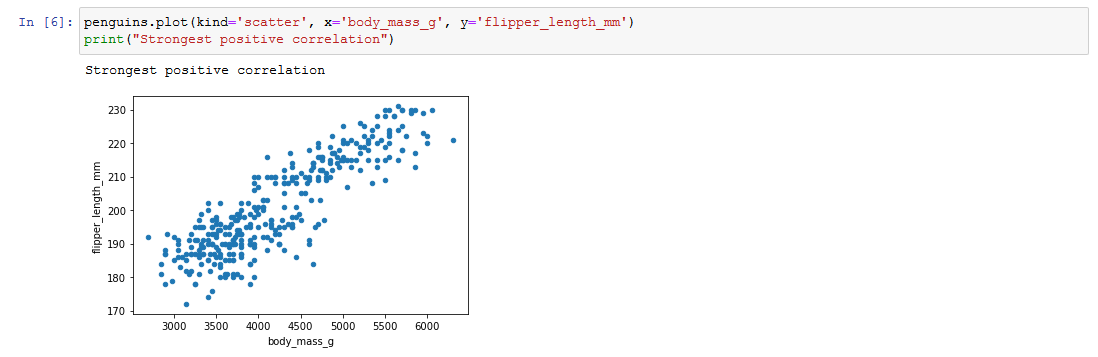


## 2.4 Modeling

Tijdens het modelleren van de data kan een data scientist inzicht krijgen op de data. Er kan bijvoorbeeld met behulp van barplots en scatterplots snel inzicht verkregen worden op grote hoeveelheden data.

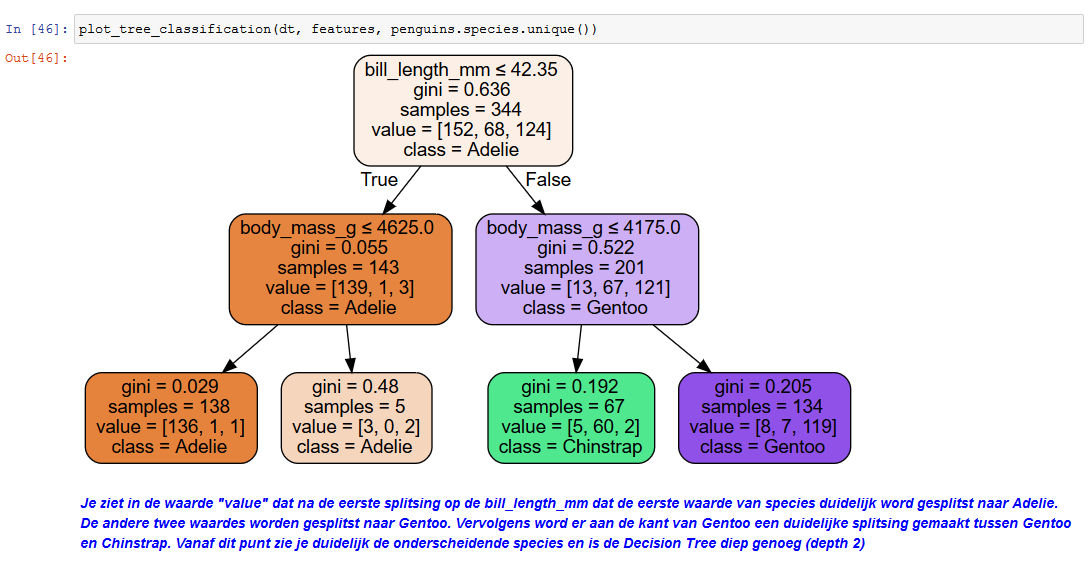
*Barplot in assignment 12*



*Scatter plot in assignment 9*

Ook kunnen er algoritmes en machine learning worden toegepast, met bijvoorbeeld Decision Trees:

*Decision Tree in assignment 15*



Vervolgens moet de gemaakte model getest worden of hij accuraat is en betrouwbaar is. Dit kan je doen door de model te testen op een test dataset en te kijken hoe goed hij presteert en of hij accuraat is.

*De getrainde model word uitgevoerd op een test data set om te kijken hoe accuraat hij is, assignment 16*

