Data Scientist

Elco Mussert

2021

Inhoudsopgave

[Inleiding 2](#_Toc66292383)

[Business Understanding 2](#_Toc66292384)

[Data Acquisition & understanding 3](#_Toc66292385)

[Modeling 5](#_Toc66292386)

[Univariate analysis 5](#_Toc66292387)

[Bivariate analysis 6](#_Toc66292388)

[multivariate analysis 8](#_Toc66292389)

[Bibliografie 10](#_Toc66292390)

# Inleiding

In dit korte verslag wordt gesproken over het proces van een data scientist tot de development. Hiermee is de bedoeling dat de proces duidelijk te begrijpen valt. Er wordt vertelt over business understanding, data acquistition & understanding en modeling. Hierbij worden nog bijbehorende voorbeelden gegeven die gemaakt zijn gedurende de lessen introductie Data Science.

# Business Understanding

In business understanding zorg je ervoor dat een data scientist weet wat de bepaalde dataset of datasets die gebruikt gaan worden inhouden. Tijdens de opdrachten in de les stond er al een dataset vast en dat was pinguïns. Deze dataset bevat gegevens van verschillende pinguïns van eiland tot kenmerken van het dier zelf. De pinguïn dataset wordt geïmporteerd vanuit seaborn.



afbeelding Import van seaborn



afbeelding Ophalen van pinguïns dataset

Ook moest er een zelfgekozen dataset gekozen worden naar keuze. Hiervoor is gekozen voor een dataset van een league of legends toernooi uit 2018. Deze dataset bevat gegevens van spellers en teams die meededen gedurende het toernooi. De dataset bestaat uit een CSV bestand die wordt gelezen in jupyter notebook.

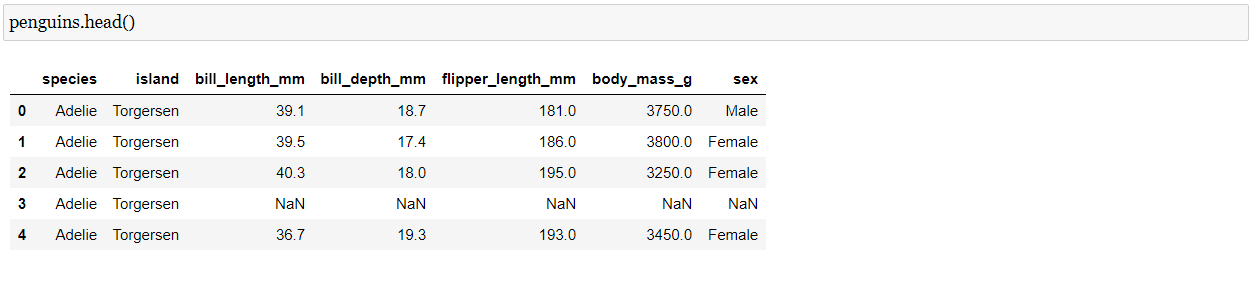


afbeelding Ophalen van LoL toernooi 2018 dataset

# Data Acquisition & understanding

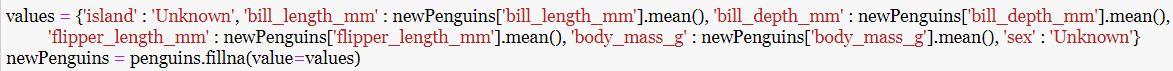
Met data acquisition en understanding kijken we naar de data in de datasets zelf. Zodat de data in de datasets ook begrepen worden. Zo kan er dieper gekeken worden wat voor data erin staat en of er eventuele aanpassingen gedaan moeten worden op de data zodat er goed op geanalyseerd kan worden.

Op de dataset van pinguïns is er in verschillende opdrachten gebruikt gemaakt van de beschikbare informatie. Hiermee is er vaak gebruikt gemaakt van alle data.

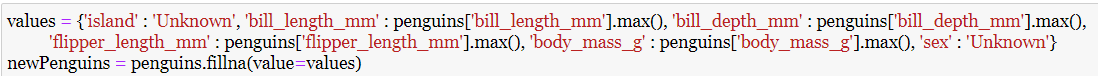


afbeelding Gegevens voorbeeld van pinguïns

Ook is een belangrijk deel van data acquisition en understanding om te kijken of alle data voldoet om op te kunnen analyseren. In de dataset over pinguïns kunnen verschillende NaN data gevonden worden (zie afbeelding 4). Dit betekent dat er geen data bestaat in die plekken. Er zijn verschillende oplossingen om dit te veranderen zodat sommige analyses gewoon uitgevoerd kunnen worden. Bij de opdrachten heb ik twee verschillende oplossingen gebruikt. Bij opdracht 15 heb ik alle NaN waardes vervangen door bij de numerieke gegevens de gemiddelden in te vullen en bij de rest is er ‘unknown’ toegevoegd (zie afbeelding 5), bij de andere opdracht namelijk opdracht 19 heb ik een ding anders gedaan en dat is niet het gemiddelde te nemen maar het hoogste getal (zie afbeelding 6).

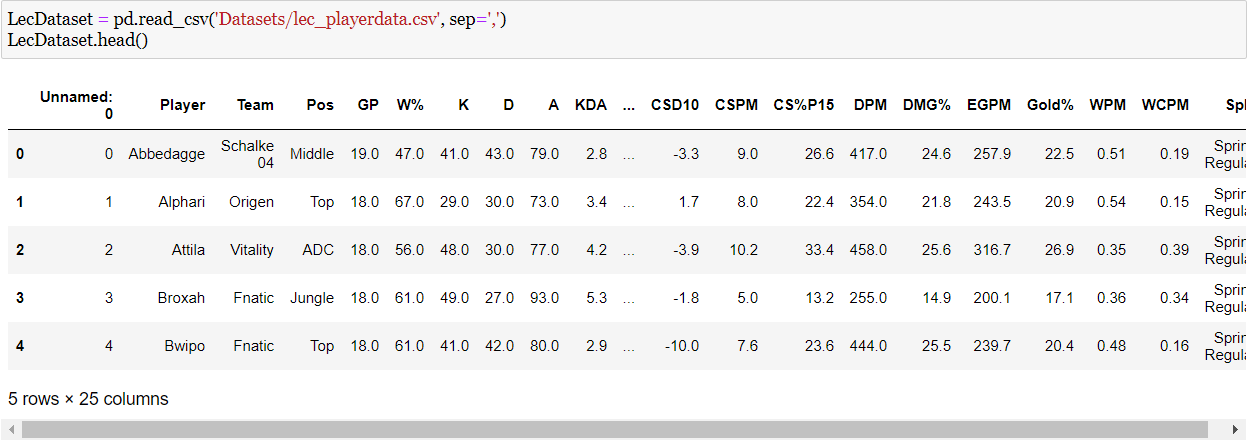


afbeelding NaN veranderen opdr. 15



afbeelding NaN veranderen opdr. 19

Bij de andere dataset die zelfgekozen is het league of legends toernooi van 2018 zijn niet alle gegevens gebruikt. Deze dataset is ook een stuk uitgebreider dan de pinguïn dataset. Ook hierin is gekeken of er eventuele data mist in de dataset, echter was dat niet het geval in deze dataset en was deze altijd klaar voor gebruik.



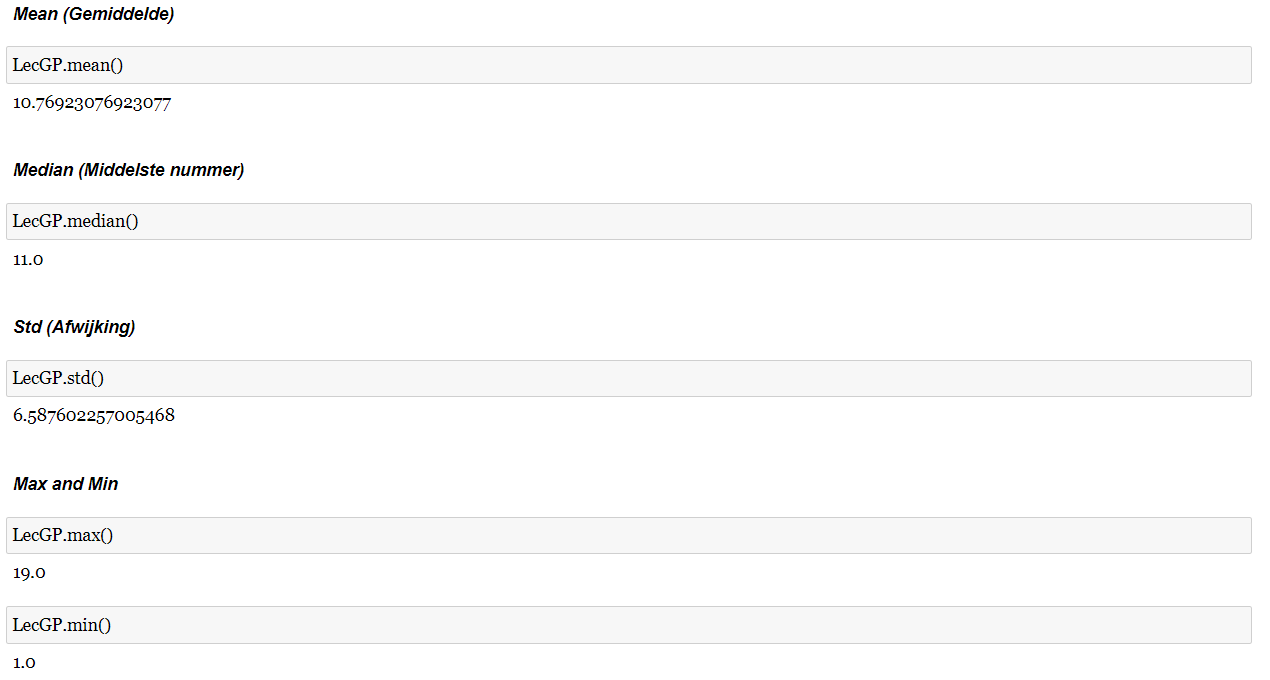
afbeelding Gegevens voorbeeld van dataset LoL toernooi 2018

# Modeling

Tijdens de modeling worden er verschillende analyses uitgevoerd op een bepaalde dataset. Er kunnen verschillende analyses worden uitgevoerd zoals univariate analysis, bivariate analysis en multivariate analysis. Door deze analyses kunnen we achter verschillende informatie komen en eventuele verbanden.

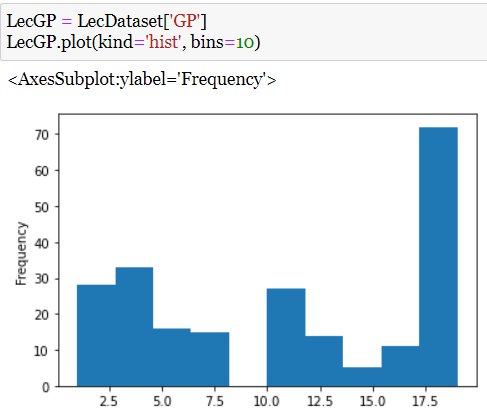
## Univariate analysis

Bij univariate analysis kijken we naar numerical data en categorical data. Bij numerical data kunnen we kijken naar statistics en visualisation. Statistics houdt in dat we een numerieke kolom nemen in bijvoorbeeld de league of legends toernooi 2018 en daar verschillende statistieken van opvragen dat is uitgevoerd in opdracht 6 (zie afbeelding 8).



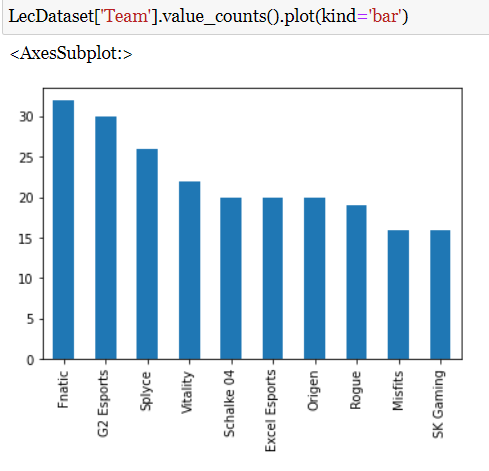
afbeelding Statistics over 'Gold percentage'

Ook kunnen visualisation opvragen van een numerieke kolom om te kijken hoeveel keer een bepaalde aantal voorkomt. In opdracht 6 is er zo’n visualisation gemaakt over hoeveel verschillende gold percentage te vinden is in de dataset.



afbeelding Visualisation over 'Gold percentage'

Bij categorial data heb je alleen een visualisation onderdeel en die werkt in principe bijna het zelfde als de visualisation van numerical data. Ook dit onderdeel is uitgevoerd in opdrachr 6 (zie afbeelding 10).

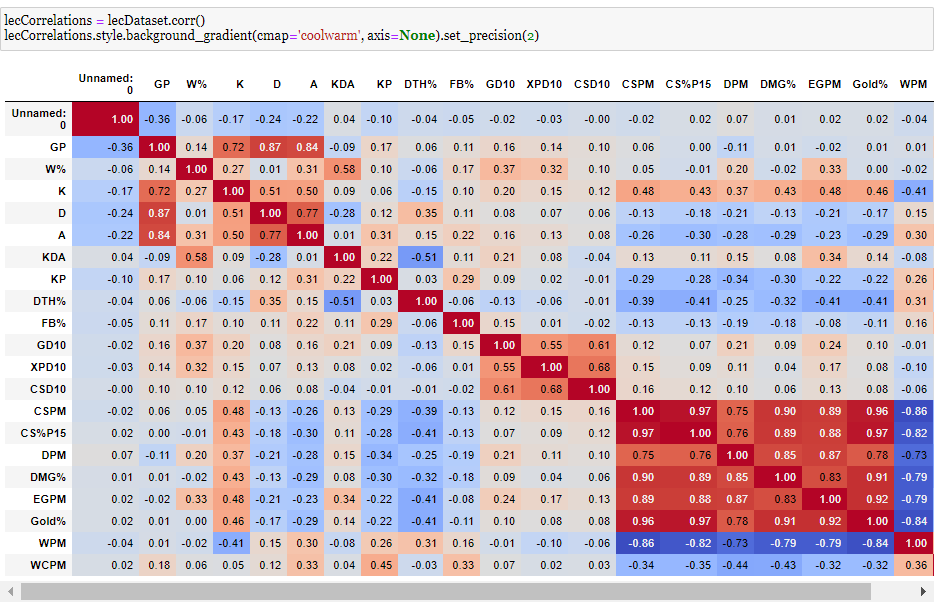


afbeelding 10 Visualisation over 'Teams'

## Bivariate analysis

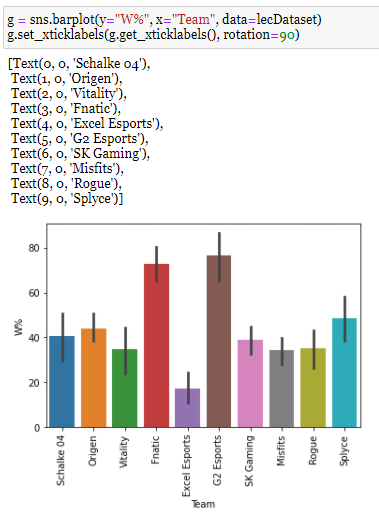
Bij bivariate analysis kunnen we verschillende data met elkaar vergelijken zoals numerical vs numerical, numerical vs categorical en categorical vs categorical. Ook hiervoor bestaan er weer statistics en visualisation.

In opdracht 10 zijn er correlations gemaakt over de eigen gekozen dataset. Hierin kan er gekeken worden of er verbanden bestaan tussen numerical data (zie afbeelding 11).



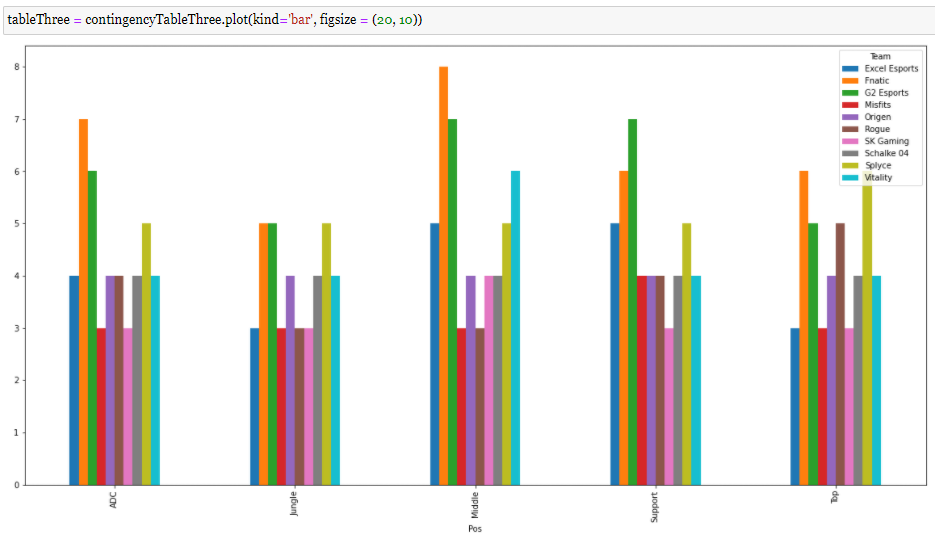
afbeelding Correlations op LoL Toernooi 2018 dataset

In opdracht 12 is er ook een voorbeeld van bivariate analysis met numerical vs categorical. Hierin is gebruikt gemaakt van barplots met confidence intervals (zie afbeelding 12).



afbeelding Barplot Teams LoL toernooi 2018 dataset

Om ook het laatste deel van bivariate analysis duidelijk te maken namelijk het categorical vs categorical er is een voorbeeld uit opdracht 14. Hierin wordt er gebruik gemaakt van een contingency tabel. In de league of legends toernooi 2018 dataset heb ik gekeken hoeveel spelers de verschillende rollen hebben aangeraakt per team (zie afbeelding 13).

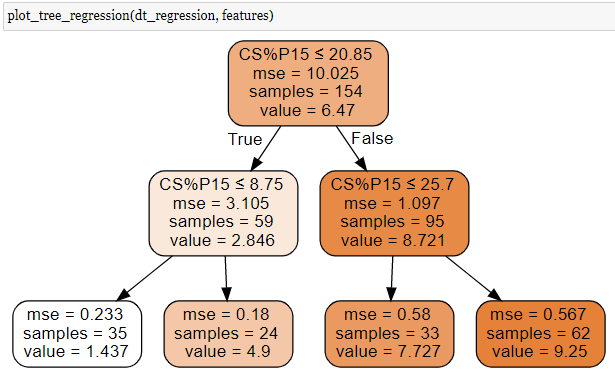


afbeelding Contingency Tabel LoL toernooi 2018 dataset

## Multivariate analysis

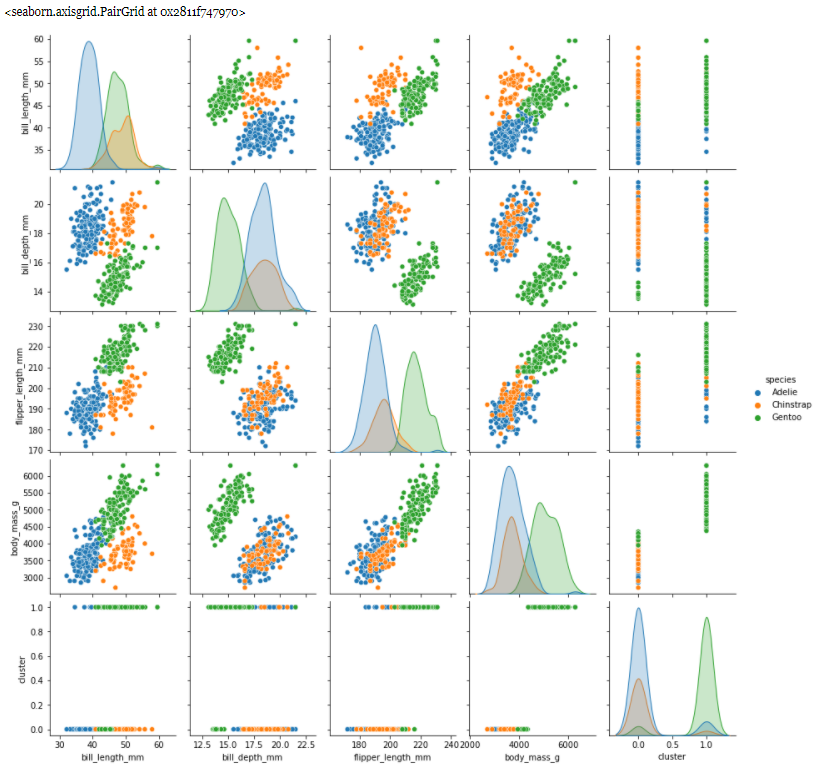
Als laatst is er nog multivariate analysis waar we op een dataset predicties en clusters op kunnen toepassen.

Een voorbeeld van de predicties kan zijn dat je gaat kijken in de league of legends toernooi 2018 dataset hoeveel CSPM een speler heeft gebaseerd op de CS%P15. Dit is uitgewerkt in opdracht 18 (zie afbeelding 14).



afbeelding Tree op LoL toernooi 2018 dataset

Als laatst is er nog een voorbeeld van clustering op de pinguïn dataset. Hiermee kan er gezien worden in welke groepen de species van pinguïns vallen. Deze analyse is uitgevoerd in opdracht 19 (zie afbeelding 15).



afbeelding Clustering op de dataset pinguïns