Data science portfolio

Juliet van Bezooijen (2183903)

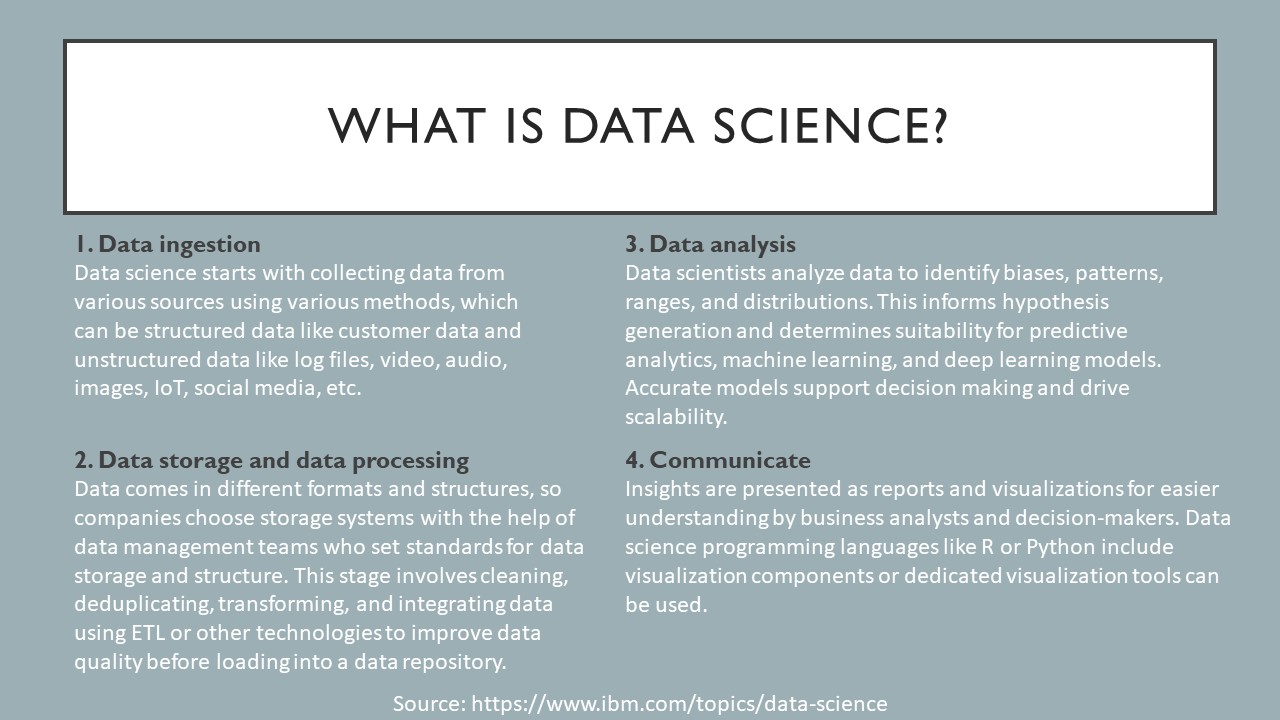
Avans Hogeschool Breda

Informatica | Introductie data science

# Portfolio assignment 101

## 101.1

In portfolio assignment 1 werd er onderzoek gedaan naar de inhoud van data science. Wat houdt data science nou eigenlijk in? Deze informatie is weergegeven in een powerpoint slide (zie Figuur 1).



Figuur 1: Portfolio assignment 1 – Wat is data science?

Als er teruggekeken word naar opdracht 1 werd er vooral gesproken over het hele proces van het verzamelen van data tot het presenteren. Echter werd er in de opdrachten vooral gefocust op het analyseren en presenteren van de data. De onderdelen van data science zijn dus wel goed weergegeven, alleen is er door het vak vooral ingezoomd op de taken Data analysis en Communicate.

## 101.2

#### Data ingestion

Data ingestion heeft in dit traject niet zo’n grote rol gespeeld. Alleen portfolio assignment 4 was gefocust op het verzamelen van data. Hierin werd de opdracht gegeven om een dataset te zoeken waarop verschillende analyses uitgevoerd worden. De eisen van deze data set waren dat de tabel minstens 100 rijen moest bevatten. De tabel moest ook minstens 2 categorische en numerieke kolommen bevatten. Tijdens deze opdracht heeft er dus data ingestion plaatsgevonden.

#### Data storage and data processing

Data processing is in verschillende portfolio assingments voorgekomen. Om bepaalde analyses te maken moeten er aanpassingen gedaan worden aan de dataset. Deze aanpassingen kunnen verschillen van het verwijderen van lege rijen, tot het beperken van de dataset tot de top 1- waardes van een kolom. In veel assignments was het niet nodig omdat de analysetechniek goed kon omgaan met lege of unieke velden.

In assignment 17 moest er echter een aanpassing gemaakt worden aan de database om de dataset op te kunnen splitsen in een train en test set. Dit werd na het inladen van de dataset (Figuur 2, regel 1) gedaan door de rijen met lege waardes te verwijderen uit de dataset (Figuur 2, regel 2).

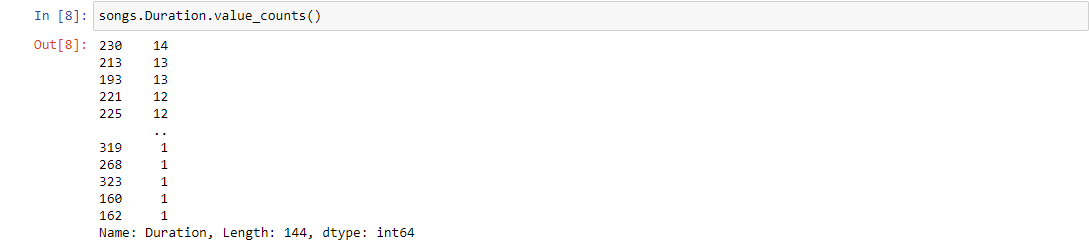


Figuur : Verwijderen lege rijen dataset

Zo zijn er ook in andere assignments verschillende filters toegepast om de dataset zo’n goed mogelijk te maken voor de specifieke analyse die uitgevoerd gaat worden.

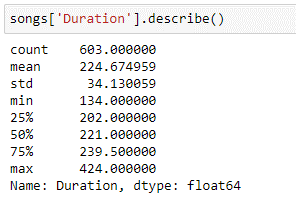
#### Data analysis & communicate

Vanaf assignment 6 heeft er in alle opdrachten data analysis plaatsgevonden. Er is per opdracht ingezoomd op 1 of meerdere kolommen. De data uit deze kolommen is vaak eerst weergegeven in een tabel of omgezet met behulp van een formule. Allereerst wordt er tijdens de analyse gezocht naar verbanden of patronen in de gegevens. Wanneer er een patroon of verband gevonden wordt, kan er op basis van deze informatie een hypothese opgesteld worden of een voorspelling gedaan worden over de toekomst. Om de analyseresultaten duidelijk over te brengen worden er grafieken en plots gemaakt. Deze zorgen ervoor dat alle bevindingen duidelijk gecommuniceerd kunnen worden met iemand die geen verstand heeft van de analysemethoden.



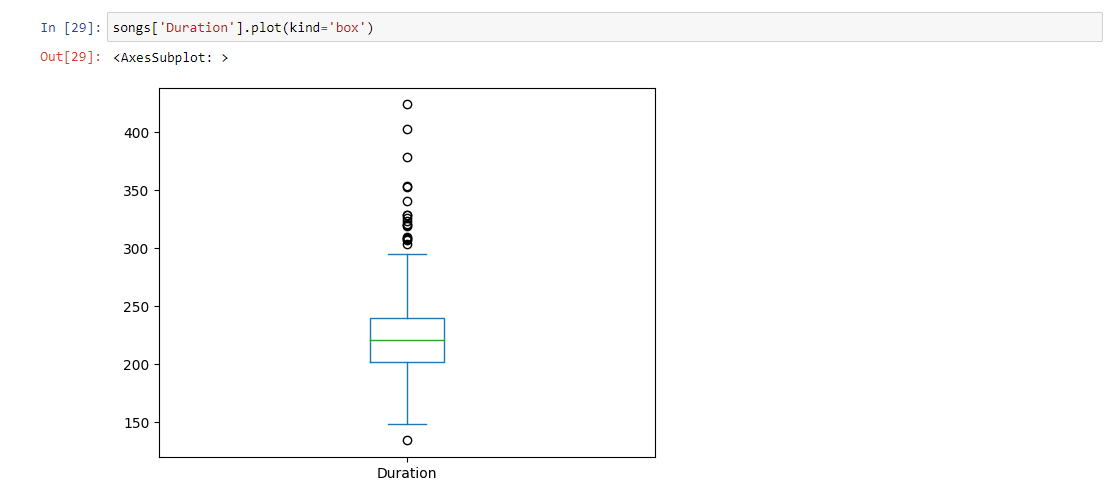
Figuur : Bekijken hoe vaak een waarde voorkomt in een dataset

In assignment 7 is er een voorbeeld getoond van hoe deze twee technieken (data analysis & communicate) in zijn werk gaan. In figuur 3 en figuur 4 wordt er uitgebreid bekeken welke waardes de kolom bevat en van welke types deze zijn. Van de data in figuur 4 kan vervolgens een hypothese gesteld worden.

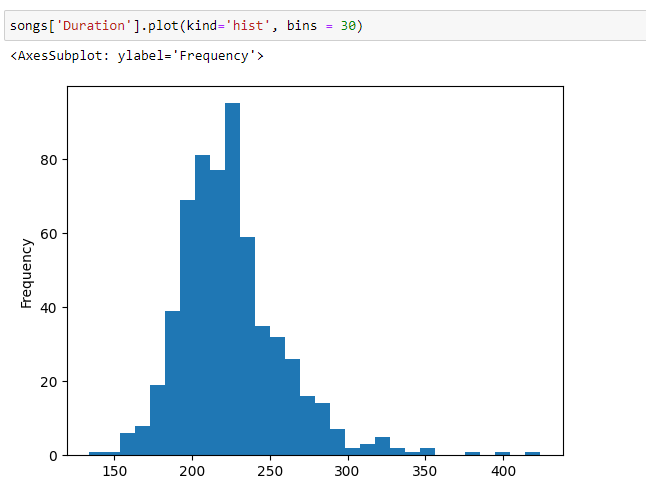


Figuur : Eigenschappen van een numerieke kolom

Voor mensen die de dataset niet kennen kan het lastig zijn om op basis van deze nummertjes een conclusie te trekken of verbanden te constateren. Daarom worden de resultaten van de analyse vaak weergegeven in een grafiek of plot (zie Figuur 5 en Figuur 6). Zo alle data uit de kolom makkelijk in 1 oogopslag te zien en zijn de verbanden en patronen duidelijk weergegeven. Dit valt onder de categorie communicate.



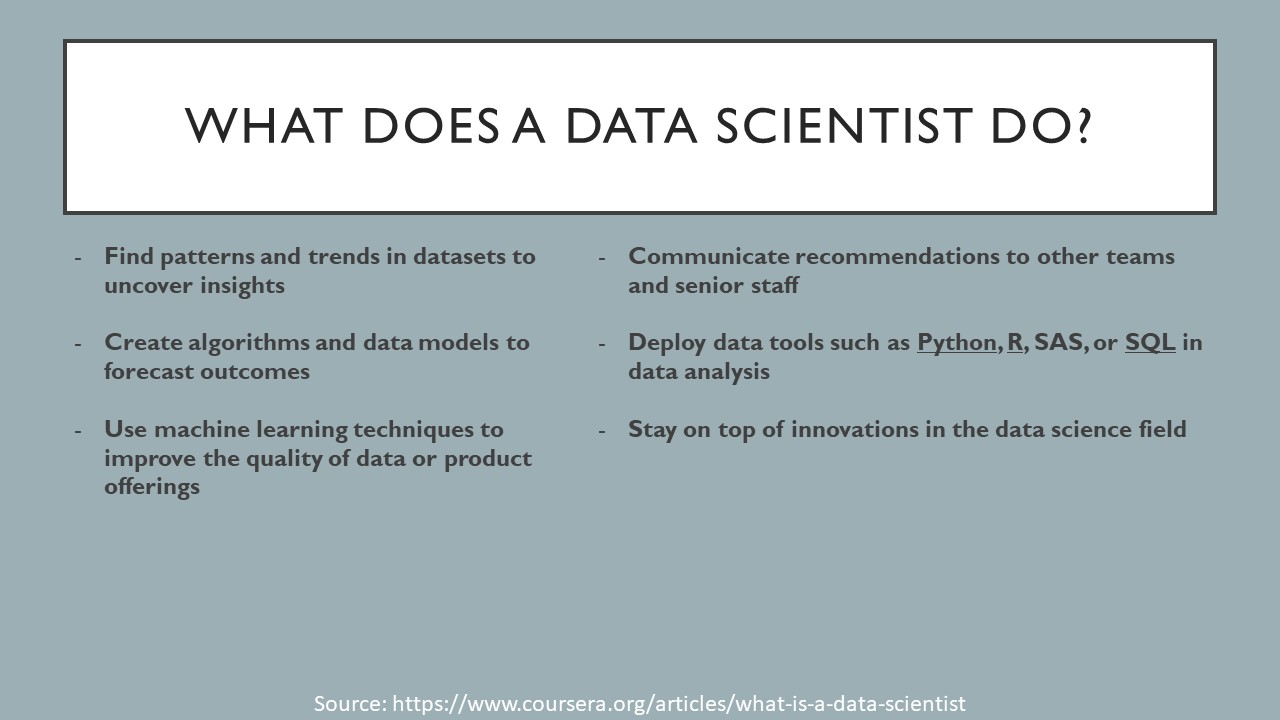
Figuur : Boxplot van een numerieke kolom



Figuur : Histogram van een numerieke kolom

# Portfolio assignment 102

In portfolio assignment 1 werd er onderzoek gedaan naar de inhoud van data science. Wat houdt data science nou eigenlijk in? Deze informatie is weergegeven in een powerpoint slide (zie Figuur 7).



Figuur : Portfolio assingment 1 - Wat doet een data scientist?

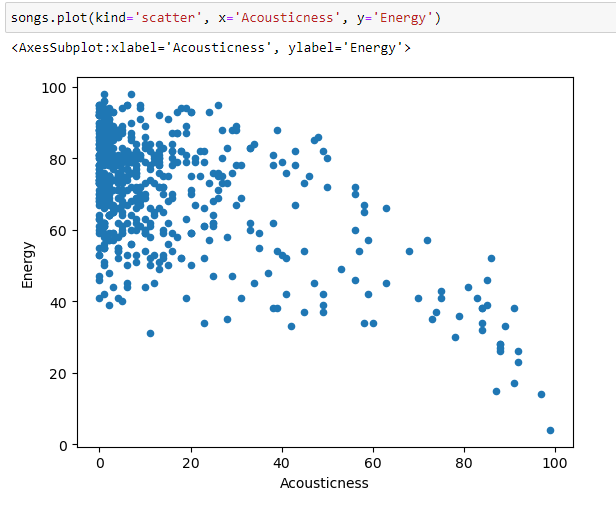
Tijdens het opbouwen van het portfolio zijn sommige van deze taken niet echt naar voren gekomen. De focus lag vooral op de linker kolom van taken. Daarom is een data scientist vooral verantwoordelijk voor:

* Het ontdekken van patronen, verbanden en trends in datasets
* Het creëren van algoritmes en modellen om voorspellingen te maken
* Het gebruiken van machine learning om zijn/haar bedrijf te optimaliseren.

Echter is er nog een punt naar voren gekomen die zeker belangrijk is als je als data scientist aan de slag gaat. Dit is namelijk het uitzoeken van de juiste bronnen, en daarmee zorgen voor een goede kwaliteit van de data.

#### Patronen, verbanden & trends

In assignment 10 is er een bivariate analyse gedaan op de kolom acousticness en energy. Dit zijn allebei eigenschappen van een liedje en deze zijn tegen elkaar uitgezet met behulp van een scatter plot (zie Figuur 8). De taak van een data scientist is om uit deze scatter plot patronen, verbanden of trends te herkennen en op basis daarvan een conclusie te trekken of een voorspelling te maken. Een conclusie die uit dit scatterplot getrokken zou kunnen worden is: Naarmate de acoustiek van een nummer toeneemt, daalt de energie. Dit komt omdat er een verband is tussen de twee eigenschappen.

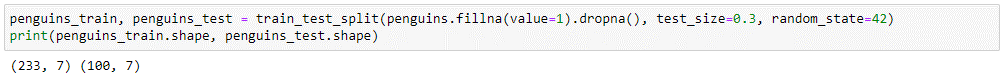


Figuur 8: Scatter plot Acousticness vs Energy

#### Machine learning & voorspellingen maken

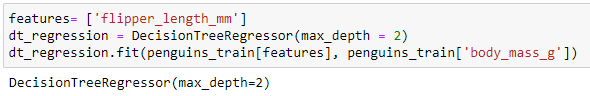
De tweede en derde taak van een data scientist vallen vaak samen. Door machine learning kan een algoritme voorspellingen maken over bepaalde data. Dit wordt bij supervised learning gedaan door middel van een train en een test set. In assignment 17 & 18 was het de taak door middel van de train set, voorspellingen te maken over de test set.

Allereerst is eer een algoritme geschreven waarmee de dataset opgesplitst wordt in 70% train set en 30% test set (zie Figuur 9). De data scientist bepaalt hoeveel van de dataset hij als training wilt gebruiken en hoeveel data als test.



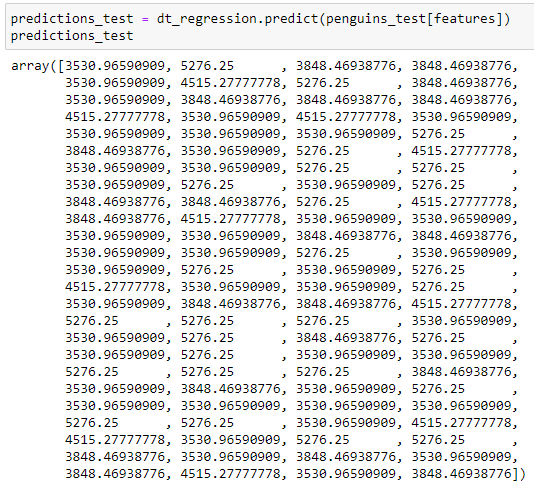
Figuur : Verdelen van een train en test set

Ook bepaald de data scientist welke eigenschap er voorspeld gaat worden. Nadat de dataset verdeeld is gaat de data scientist onderzoek doen naar welke eigenschappen invloed kunnen hebben op de voorspelde eigenschap. Dit kan door middel van een correlatietabel, of met behulp van voorkennis. De eigenschappen die invloed hebben en de eigenschap die voorspeld moet worden, worden meegegeven aan een decision tree (zie Figuur 10).

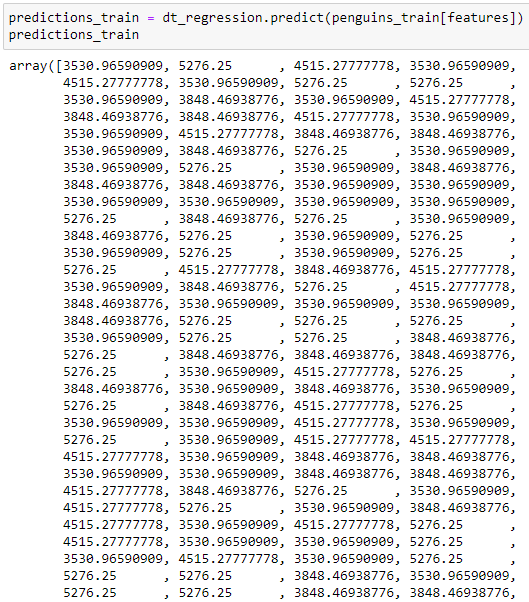


Figuur : Decision tree maken

Als de decision tree klaar is wordt het tijd om het algoritme te laten voorspellen. Hier hoeft de data scientist zelf niks voor te doen. De voorspellingen worden gedaan op zowel de train set als de test set (zie Figuur 11 & 12).



Figuur : Voorspellingen test set



Figuur : Voorspellingen train set

Tot slot kan de decision tree geplot worden om visueel weer te geven wanneer een eigenschap een bepalende factor is in een voorspelling.

Op basis van de voorspellingen van het algoritme kan de data scientist vooruit denken en beslissingen nemen om het bedrijf te optimaliseren. Er kan ingespeeld worden op eventuele veranderingen waardoor het bedrijf minder onverwachtse overhaaste beslissingen hoeft te maken.