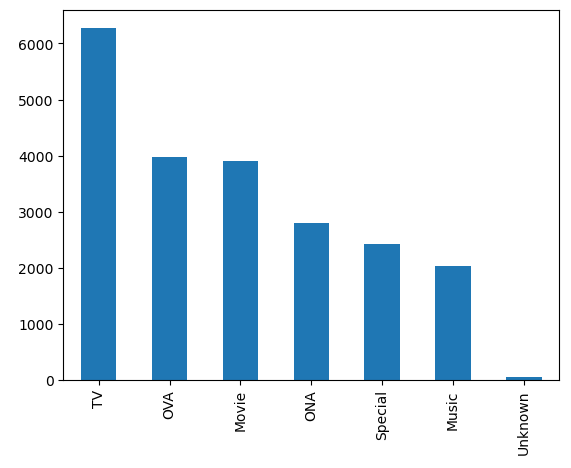
# Assignment 101

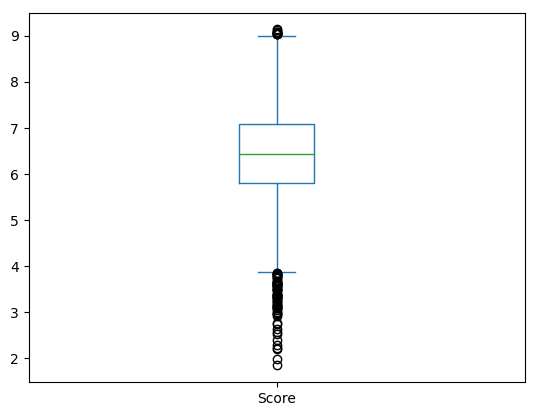
Data science is het verkrijgen van inzichten en waardevolle informatie door middel van verschillende vaardigheden, zoals statistieken, programmeren en sector-/organisatie kennis.

Deze inzichten en informatie kunnen op verschillende manieren worden aangepakt. Datasets bestaan namelijk uit meerdere kolommen van categoriale en numerieke gegevens. Om informatie hieruit te halen moet bepaald worden hoe er met deze data moet worden omgegaan. Voor categoriale gegevens kan je makkelijk een bar grafiek maken op basis van groeperingen (zie figuur 1)



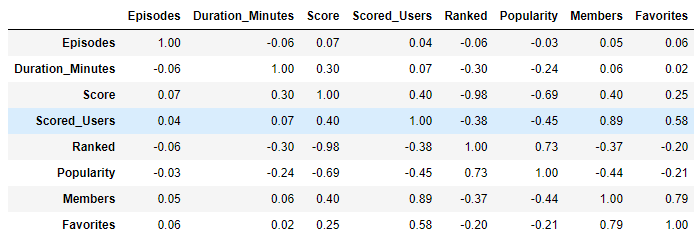
Figuur 1: Bar grafiek

Voor numerieke waardes kan er op verschillende manieren naar kijken. Zo kan je verschillende plots maken waar je informatie uit kan halen zoals gemiddelde waarde, de meest voorkomende waardes en de uitzonderingen (zie figuur 2)



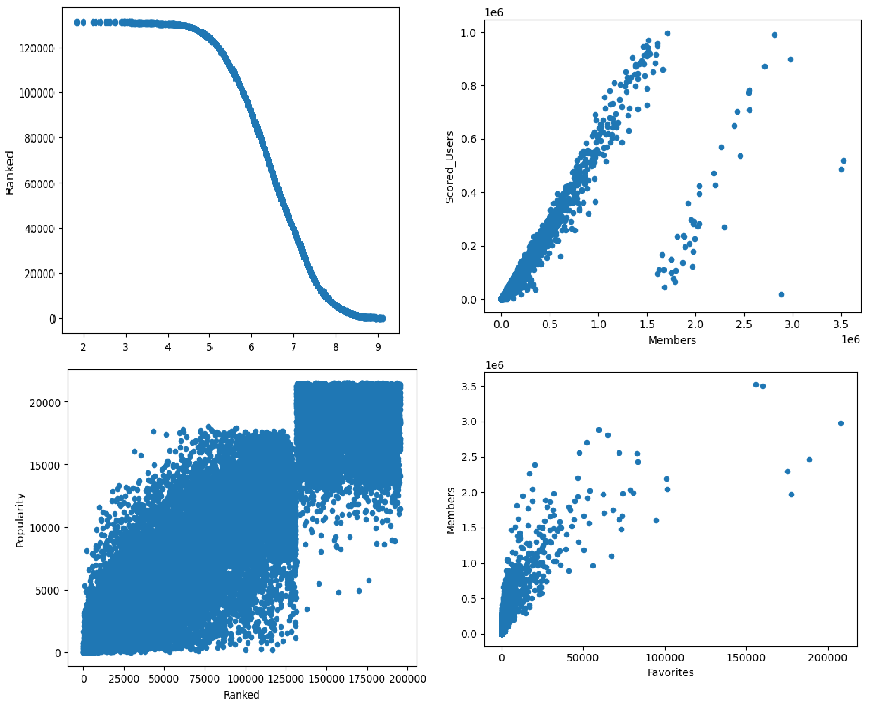
Figuur 2: Box plot

Tussen gegevens kunnen relaties liggen, ofwel correlaties. Deze correlaties kunnene worden geplot in een tabel (zie figuur 3) waarbij waardes dicht bij de -1/1 liggen een sterkere correlatie hebben dan waardes dicht bij de 0.



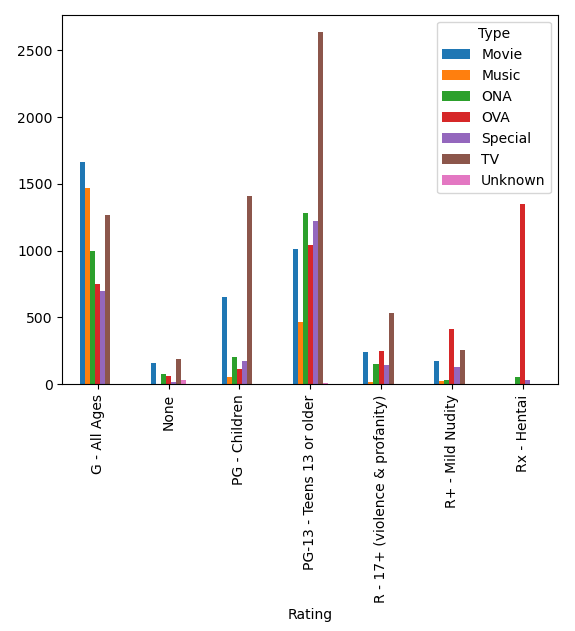
Figuur 3: Correlatie tabel

Deze correlaties tussen twee kolommen kunnen geplot worden in een scatter-plot. Met een hoge correlatie ga je in een scatter-plot een duidelijke trend zien en bij een lagere correlatie wordt het meer een warrige boel aan stippen (zie figuur 4). Binnen Scatterplots spreken we wel over numerieke waardes tegenover elkaar.



Figuur 4: Scatter plots

Voor categoriale waardes met elkaar te vergelijken bestaat er een uitgebreidere versie van de barplot waar op verdere groeperingen in kleuren worden verdeeld. Zo kan je bijvoorbeeld zien in figuur 5 binnen elke rating de meeste type anime’s zijn.



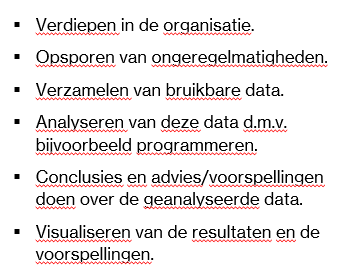
Figuur 5: Barplot correlatie categoriaal

We hebben nu een aantal verschillende voorbeelden gezien van grafieken en tabellen. Ook hebben we gezien dat er op verschillende manieren naar gegevens gekeken kunnen worden. Data Science is ook wel de werkwijze waarin de gegeven data inzichtelijk wordt weergegeven in tabellen en grafieken.

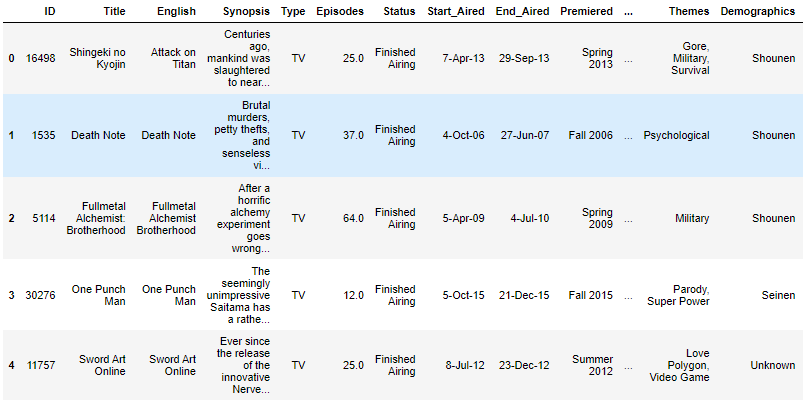
De relatie met BI hierin is dat deze grafieken en tabellen die uit Data Science voortkomen door de organisaties die met BI werkt hier gebruik van kunnen maken om belangrijke inzichten te krijgen. Zo kan je aan bepaalde trends of correlaties conclusies trekken die je normaal gesproken niet zouden zijn opgevallen.

# Assignment 102

Een Data Scientist moet de volgende zaken doen voor het proces:

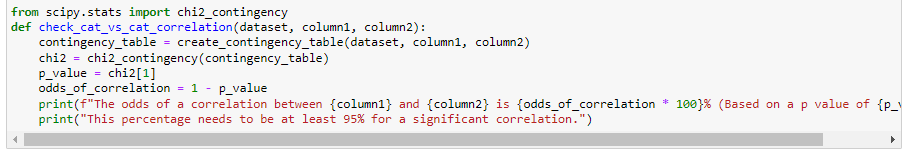


Als eerste moet een Data scientist zich gaan verdiepen in de organisatie en doet dit aan de hand van de verkregen dataset. In deze dataset gaat de scientist kijken met wat voor data hij/zij te maken heeft en of deze data geen onregelmatigheden heeft. Het kan namelijk zo zijn dat data niet handig is opgeslagen, zodat er niet meer over de jaren heen kan worden gekeken. Zo is het in de dataset hieronder dat bij de première van een anime het seizoen en het jaar in een kolom zijn gezet, waardoor heel veel verschillende groeperingen ontstaan. Een scientist zou hier bijvoorbeeld een oplossing voor moeten zoeken om de data wel bruikbaar te maken.



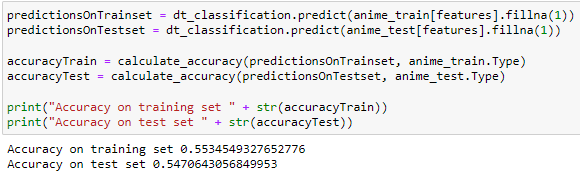
Figuur 6: Dataset

Om te kijken dat correlaties of zaken wel degelijk relevante conclusies uit getrokken mogen worden, zal een scientist ook moeten programmeren met formules om dit te controleren (zie figuur 6).



Figuur 7: Chi2 formule

Verder kan er door het analyseren van een dataset voorspellingen worden gedaan. Zo kan je op bepaalde waardes waar sterke correlaties zijn onderscheid maken, zodanig dat je kan zeggen dat een nieuwe record met een waarde bij die specifieke categorie zal horen. Hiervoor zal een scientist ook een stuk moeten kunnen programmeren. Dit doet hij door de dataset bijvoorbeeld op te splitsen in een test en trainingsgedeelte, dor machine learning van de trainingset te laten leren, na het leren de geleerde informatie op de testset toe te passen en dan te kijken hoe accuraat dit is gegaan.



Figuur 8: Accuracy train-/testset

Het visualiseren van gevonden inzichten doet de scientist op verschillende manieren. Dit kan door middel van grafieken, tabellen en resultaten uit berekende formules tonen.