|  |  |
| --- | --- |
|  | Opdracht 101 & 102 |
|  |  |
|  | Stan Barkmeijer  Introductie Data Science  3/10/21 |

Contents

[Wat is data science 2](#_Toc66267468)

[Terugkijk naar assignment 1 2](#_Toc66267469)

[Portofolio assignments 2](#_Toc66267470)

[Univariate Analysis 2](#_Toc66267471)

[Bivariate Analysis 2](#_Toc66267472)

[Multivariate analysis 2](#_Toc66267473)

# Wat is data science

## Terugkijk naar assignment 1

In assessment 1 heb ik uitgeschreven dat Data Science het veld dat wetenschappelijke methoden, algoritmen en systemen gebruikt om kennis en inzichten te verkrijgen uit data is.  
Deze data kan vervolgens gebruikt worden als ondersteuning van keuzes door Data Analisten.

Na het volgen van de lessen, en het maken van assignments ben ik tevreden met mijn aanname over wat Data Science is. Ik zou hier niets over willen veranderen.

## Portofolio assignments

Binnen data science kunnen we onderscheid maken tussen verschillende analyse categoriën.  
Deze zijn Univariate Analysis, Bivariate Analysis en Multivariate Analysis.

### Univariate Analysis

Univariate analyse is het analyseren van numerieke of categorische data. De numerieke data kan statistisch of visueel geanalyseerd worden. De categorische data kan gevisualiseerd worden.   
Door visualisatie krijgen we een simpel in te zien overzicht over de verdeling van de data. De statistische informatie kan moeilijk te begrijpen zijn voor mensen die geen kennis hebben over wat de nummers betekenen.  
Visueel kunnen we histogrammen, boxplots en barplots maken.

### Bivariate Analysis

Bivariate analyse is het analyseren van Numerical VS Numerical, Numerical VS Categorical en Categorical VS Categorical data. Al deze soorten data kan statistisch en visueel ingekeken worden.   
Statistisch kunnen we bij N VS N de correlatie bekijken tussen kolommen. Zo kunnen we zien of er een relatie hangt tussen de kolommen. Visueel kunnen we een scatterplot maken, om zo de verdeling van de correlatie te bekijken.  
Bij N VS C kunnen we met statistieken zekerheidsintervallen berekenen, en daarmee visueel barplots maken met die intervals.  
Bij C VS C kunnen we met statistieken contingency tabellen en chi2 tests uitvoeren. Met de contingency tabellen kunnen we twee variabelen/kolommen tegen elkaar opzetten om te zien wie welke eigenschap heeft. Met de chi2 test kunnen we de significance berekenen tussen de twee verhoudingen. Hier kunnen we vervolgens een barplot mee maken.

### Multivariate analysis

Binnen multivariate analysis kunnen we spreken over predictions door middel van machine learning, AI en andere slimme algoritmen, maar ook clustering.  
Bij predictions kunnen we een numeriek of categorisch doelwit onderzoeken door middel van decision trees. Hiermee kunnen we inschatten hoe een numerieke of categorische waarde tot stand kan komen in verschillende stappen (diepte).   
Hier maken we gebruik van train en test sets, om de zekerheid van de prediction vast te stellen.  
Bij clustering maken we gebruik van een Kmeans algoritme, hiermee wordt de data ingedeeld in clusters. Deze kunnen gebruikt worden om een duidelijke pairplot te maken.