**Data Science Assignments**



Klas: INVT2A

Datum: 10-3-2021

Docent: Alexander Van den Bulck & Dion Koeze

Vak: Intro Data Science

Student: Job Haast – 2151057

Contents

[Inleiding 3](#_Toc65606938)

[1. Data Science 3](#_Toc65606939)

[1.1 Definitie 3](#_Toc65606940)

[1.2 Data Science Lifecycle 3](#_Toc65606941)

[1.3 Opdrachten 4](#_Toc65606942)

[1.3.1 Business Understanding 4](#_Toc65606943)

[1.3.2 Data Acquisition & (further) Understanding 5](#_Toc65606944)

[1.4 Eigen Definitie 5](#_Toc65606945)

[2. Taken Data Scientist 6](#_Toc65606946)

[3. Conclusie 6](#_Toc65606947)

[Literatuurlijst 7](#_Toc65606948)

# Inleiding

In dit rapport zullen de verschillende opdrachten die zijn gemaakt tijdens het vak Intro Data Science worden gebruikt om uit te leggen wat Data Science nou precies inhoudt. Dit wordt gedaan om voor de schrijver van dit rapport een koppeling te leggen tussen de theorie achter Data Science en de praktijk, om dus vervolgens een beter begrijpen van Data Science te krijgen. Het uitleggen aan de hand van de gemaakte opdrachten wordt gedaan in hoofdstuk 1. Daarnaast zullen we in dit rapport bepreken wat een Data Scientist nou concreet doet. Dit bespreken wordt gedaan in hoofdstuk 2.

# 1. Data Science

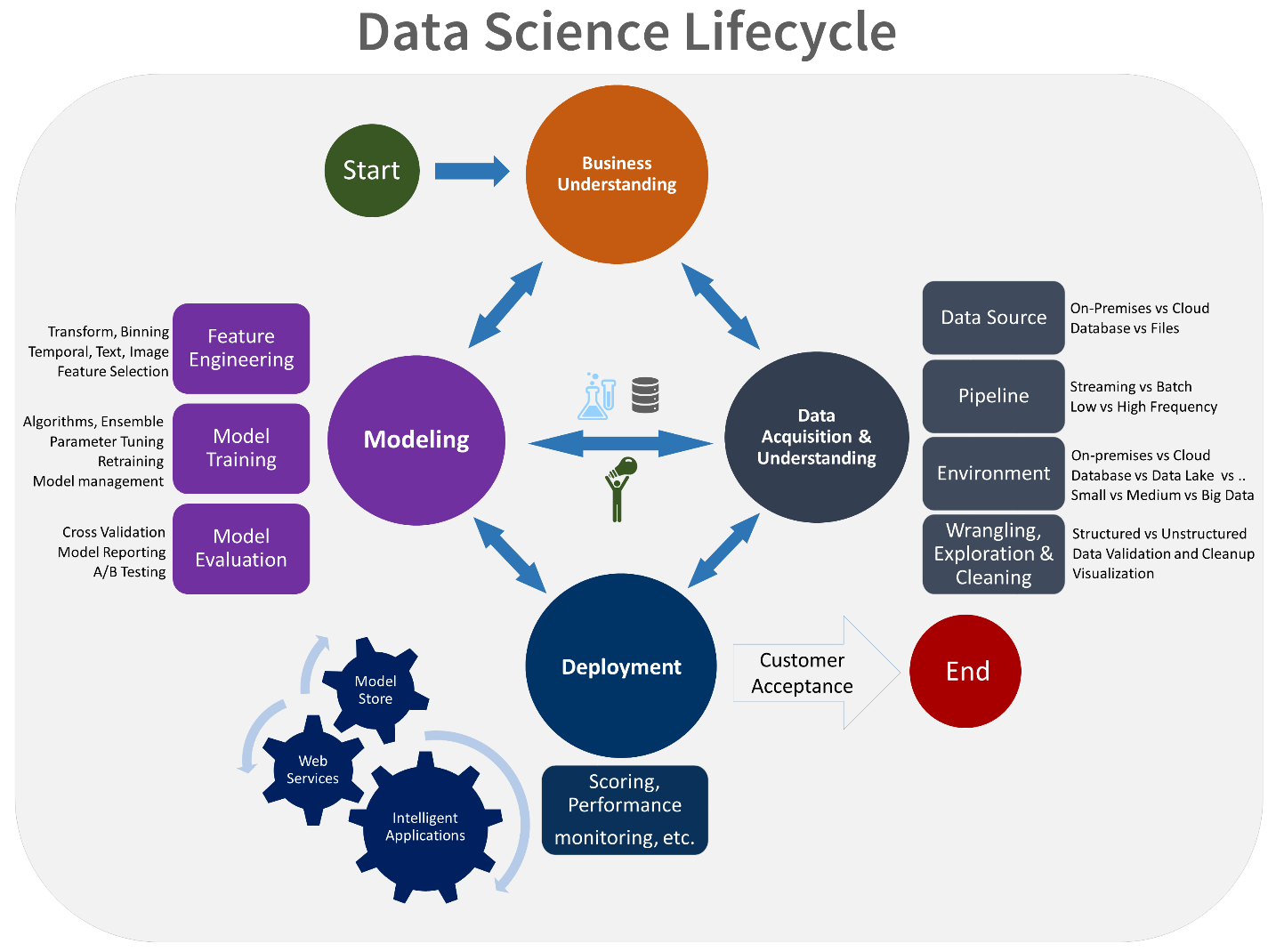
In dit hoofdstuk wordt aan de hand van gemaakte opdrachten tijdens het vak Intro Data Science uitgelegd wat Data Science is.

## Definitie

Om duidelijk te krijgen wat Data Science precies is begin ik met de definitie van Data Science. Deze definitie luidt ruw vertaald: "Data Science is een interdisciplinair vakgebied dat gebruik maakt van wetenschappelijke methoden, processen, algoritmen en systemen om kennis en inzichten te extraheren uit vele structurele en ongestructureerde gegevens... Data science is een "concept om statistiek, data-analyse en hun verwante methoden te verenigen" om "actuele verschijnselen te begrijpen en te analyseren" met gegevens. Het maakt gebruik van technieken en theorieën uit vele gebieden binnen de context van wiskunde, statistiek, computerwetenschap, domeinkennis en informatiewetenschap." - (Wikipedia contributors, 2021).

## Data Science Lifecycle

Nu dat we ongeveer weten wat Data Science is kunnen we kijken naar de onderdelen waar Data Science uit bestaat. De onderdelen waar Data Science uit bestaat is Business Understanding, Data Acuisition & Understanding, Deployment en Modeling. Deze onderdelen zijn dus de onderdelen waar Data Science uit bestaat. Dit kunnen we ook weergeven als de Data Science Lifecycle. In figuur 1 wordt aangegeven hoe deze lifecycle er dan uitziet.



Figuur 1 - (Microsoft, 2020)

## Opdrachten

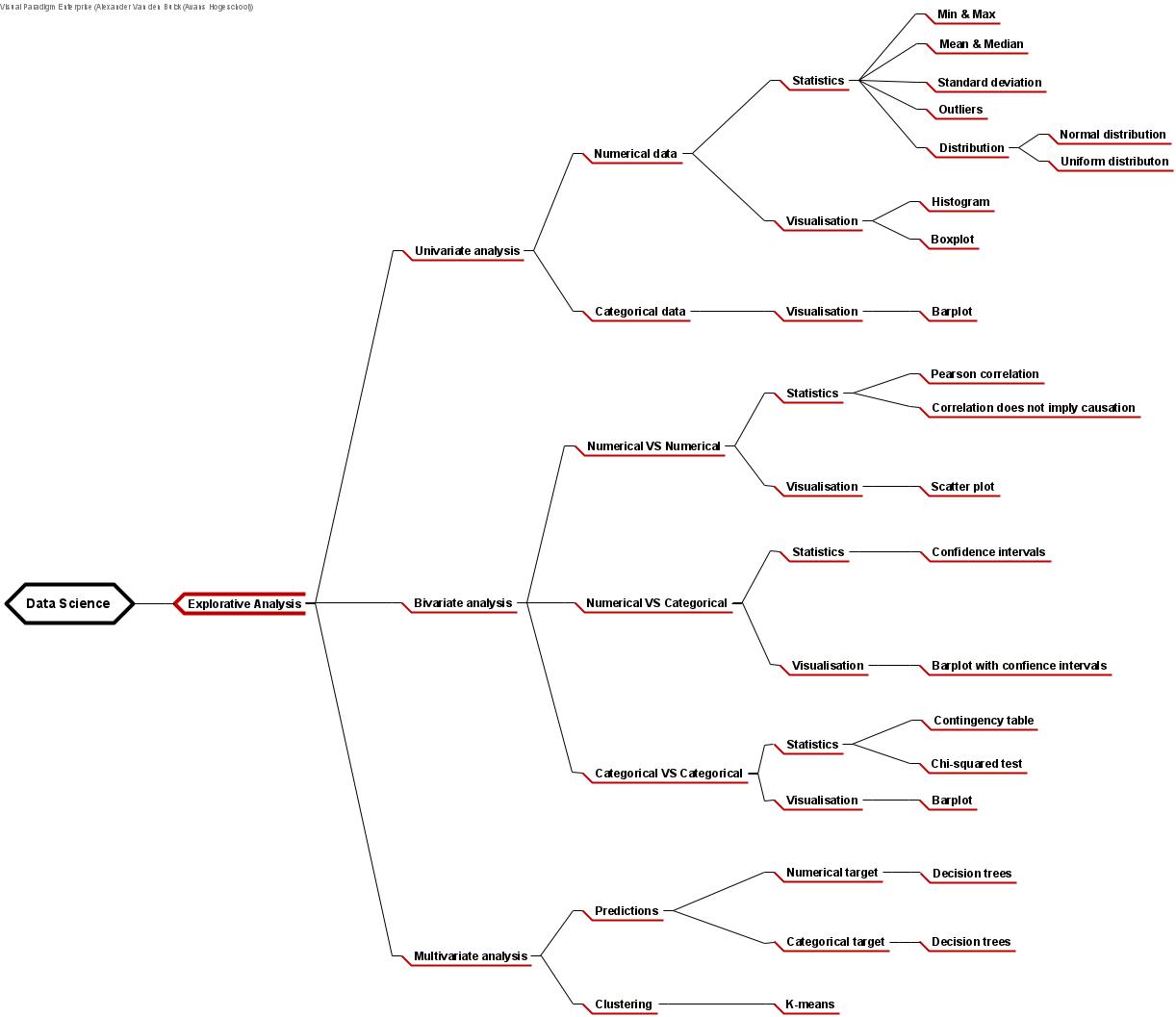
Nu dat de definitie en de lifecycle van Data Science is getoond, kunnen eindelijk de gemaakte opdrachten worden ingedeeld op de verschillende onderdelen in de Data Science Lifecycle. Dit indelen wordt hieronder gedaan.

### Business Understanding

Om een beter begrijpen te krijgen van een bedrijf moet een Data Scientist eerst begrijpen wat precies het bedrijf doet en welke gegevens beschikbaar zijn. De opdrachten die onder dit onderdeel vallen zijn er niet echt. Dit komt omdat er geen overkoepelende casus was voor de opdrachten. Daarom wordt voor dit onderdeel meer gekeken naar opdrachten die de begrippen en de algemene understanding van Data Science scheppen gekozen. De opdrachten die algemene understanding schepte waren de opdrachten 3, 5, 6, 7, 8. Opdracht 3 zorgt er voor dat er kennis wordt opgedaan over wat nou precies categorische data is door verschillende categorieën tegen elkaar uit te zetten in een bar-plot. Opdracht 5 zorgt er voor dat er kennis wordt opgedaan over wat nou precies numerieke data is door de numerieke data in een histogram te tonen. Opdracht 6 verbeterd de opgedane kennis in opdrachten 3 en 5 door op de besproken onderdelen in de opdrachten verder te bouwen. Opdracht 7 leert je kennis maken met Distribution en opdracht 8 met Confidence. Eigenlijk zorgt iedere opdracht voor een bredere kennis in Data Science. De genoemde opdrachten zijn de opdrachten die het beste de basis vormen.

### Data Acquisition & (further) Understanding

Data acquisition en understanding houdt in dat je gegevens gaat verzamelen en deze zo gaat analyseren dat je tot nieuwe conclusies/theorieën kunt komen. De opdrachten onder dit onderdeel vallen zijn de opdrachten 10, 11, 12, 13, 14, 15, 16, 17, 18, en 19. Iedere opdracht leert je meer kennis over Data Science door verschillende onderdelen van Data Science te doen. Deze onderdelen zijn Bivariate analysis, Numerical vs Categorical bivariate analysis, Categorical vs Categorical bivariate analysis, Categorical Multivariate analysis, Numerical Multivariate analysis en Clustering. In figuur 2 wordt weergegeven waaronder welke onderdeel valt.



Figuur 2 - (Van den Bluck, 2021)

## Eigen Definitie

Data Science is een lastig begrip om te verwoorden. Daarom wordt in deze paragraaf geprobeerd in eigen worden het begrip Data Science, aan de hand van de hiervoor gegeven informatie, te definiëren. Data Science is een competentie die een persoon kan bezitten, waarbij hij of zij aan de hand van data verschillende analysis doet, zoals Univariate Analysis (opdrachten 3, 5, 6), Bivariate analysis (opdrachten 9, 10, 11, 12, 13, 14) en Multivariate Analysis (opdrachten 15, 16, 17, 18) om nieuwe inzichten te krijgen.

# 2. Taken Data Scientist

Eén van de belangrijkste taken van een Data Scientist is het begrijpen van het bedrijf/data. Om dit te doen kan een Data Scientist bijvoorbeeld een Univariate Analysis uitvoeren om kennis te maken met de data, zoals is gedaan in opdrachten 3, 5 en 6. Wat een Data Scientist af een toe nog moet doen, om met de gegevens aan de slag te gaan, is het opschonen van de data. Dit kan op twee manieren gedaan worden. De eerste manier is om de rijen die gegevens missen niet mee te nemen in de analyses, en de tweede manier is het opvullen van gegevens met gemiddeldes of veel voorkomende gegevens. Dit opvullen werd gedaan in opdrachten 16, 17, 18 en 19. Als de Data Scientist klaar is met het opschonen van de data, dan kunnen de geavanceerdere analyses plaatsvinden. Dit kunnen bijvoorbeeld Bivariate Analysis (opdrachten 9, 10, 11, 12, 13, 14), Multivariate Analysis (opdrachten 15, 16, 17, 18) en Clustering (opdracht 19) zijn. Met deze analyses kan de Data Scientist inzicht krijgen in hoe de data tegenover elkaar correleert.

# 3. Conclusie

Data Science en Scientist zijn moeilijke begrippen om te begrijpen. In dit rapport is besproken wat Data Science is aan de hand van een definitie in combinatie met de Data Science Lifecyle en de gemaakte opdrachten. Daarnaast is geprobeerd uit te leggen wat de taken van een Data Scientist zijn. Dit is gedaan aan de hand van de gemaakte opdrachten.

Samengevat is Data Science Data een competentie die een persoon kan bezitten, waarbij hij of zij aan de hand van data verschillende analysis doet, zoals Univariate Analysis, Bivariate en Multivariate Analysis om nieuwe inzichten te krijgen. Een Data Scientist heeft verschillende taken. Deze taken zijn het begrijpen van het bedrijf/data, het opschonen van de data en het daadwerkelijk analyseren van de data.

# Literatuurlijst

Microsoft. (2020, 1 oktober). Data science lifecycle [Foto]. Microsoft. <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle>

University of London. (2018, 1 maart). Data science [Foto]. University of London. <https://london.ac.uk/sites/default/files/styles/max_1300x1300/public/2018-03/data-science.jpg?itok=bTPDs5nf>

Van den Bluck, A. V. D. B. (2021, 14 januari). Intro data science mind map V3 [Foto]. Blackboard. <https://bb.avans.nl/webapps/blackboard/content/listContent.jsp?course_id=_17731_1&content_id=_1790526_1>

Wikipedia contributors. (2021, 28 februari). Data science. Wikipedia. <https://en.wikipedia.org/wiki/Data_science>