# Opdracht 01 – Centrummaten en spreidingsmaten

## Doelstellingen

De doelstelling van deze opdracht is om voor verschillende types data een frequentietabel aan te maken en de verschillende centrummaten, spreidingsmaten van de gegevens te berekenen.

Er wordt ook verwacht uitschieters te kunnen opsporen en de scheefheid en kurtosis te bepalen en te interpreteren.

## Inhoud

* Opstellen van een frequentietabel met absolute, relatieve en cummulatieve waarden
* De frequentietabel omzetten naar een grafiek
* Berekenen van de centrummaten en spreidingsmaten via Excel
* Interpreteren van de centrummaten en spreidingsmaten
* Bepalen van de skew en kurtosis
* Zoeken van uitschieters

## Algemene toelichting

Deel 1 van de opdracht wordt uitgewerkt in Excel. Zie hiervoor het bestand ***‘Opdracht\_01\_Centrum\_Spreidingsmaten.xlsx’****.*

Benodigdheden: Een recente versie van Microsoft Excel.

Deel 2 van de opdracht wordt uitgewerkt in Python (Jupyter notebook). Zie hiervoor het bestand ***‘Opdracht\_01\_Centrum\_Spreidingsmaten.ipynb’****.*

De nodige software bestaat uit Python 3.6 (of hoger) en volgende Python libraries:

*jupyter, pandas, numpy, matplotlib, seaborn, scikit-learn*

Eenmaal Python geïnstalleerd is kunnen de libraries door middel van pip3 geïnstalleerd worden via de command line.

Voorbeeld: *pip3 install jupyter scikit-learn scikit-image pandas numpy matplotlib seaborn nltk*

Meer info over de installatie en gebruik van jupyter notebook kan je vinden op:

Installatie via pip3: [https://jupyter.readthedocs.io/en/latest/install.html#new-to-python-and-jupyter](https://jupyter.readthedocs.io/en/latest/install.html#_blank) (Koppelingen naar een externe site.)Koppelingen naar een externe site.

Opstarten van Jupyter notebook: [https://jupyter.readthedocs.io/en/latest/running.html (Koppelingen naar een externe site.)Koppelingen naar een externe site.](https://jupyter.readthedocs.io/en/latest/running.html)

Jupyter notebook interface: [https://jupyter-notebook.readthedocs.io/en/stable/notebook.html (Koppelingen naar een externe site.)Koppelingen naar een externe site.](https://jupyter-notebook.readthedocs.io/en/stable/notebook.html)

### Uitwerking en indienen van de opdracht

**Deel 1: Excel gedeelte**

Maak gebruik van de dataset te vinden in ***Opdracht\_01\_Centrum\_Spreidingsmaten.xlsx*** om de vragen die hier verder bij het opdracht gedeelte te vinden zijn uit te werken en te beantwoorden.

De uitwerkingen en de antwoorden komen **in het Excel document zelf** te staan.

Probeer ietwat structuur in het document aan te brengen. Je mag daarvoor gerust nieuwe werkbladen aanmaken in het Excel document zelf.

Zorg ervoor dat jouw naam in de bestandsnaam voorkomt:

Bijvoorbeeld: **Opdracht\_01\_Centrum\_Spreidingsmaten\_Wouter\_Gevaert.xlsx**

**Deel 2: Python gedeelte**

De volledige opdrachtsbeschrijving is te vinden in de Jupyter notebook ‘ **Opdracht\_01\_Centrum\_Spreidingsmaten.ipynb’.**

De uitwerkingen en de antwoorden vul je dan ook in die Jupyter notebook aan.

Zorg er ook hier voor dat jouw naam in de bestandsnaam voorkomt:

Bijvoorbeeld: **Opdracht\_01\_Centrum\_Spreidingsmaten\_Wouter\_Gevaert.ipynb**

**Indienen van de opgaves**

Maak 1 zip-bestand van het Excel-bestand en de Jupyter notebook en dien het in op Leho onder de opdracht Sessie\_01\_Centrummaten\_Spreidingsmaten.

## Opdracht DEEL 1 : Excel

De dataset is te vinden in Opdracht\_01\_Centrum\_Spreidingsmaten.xlsx.

Bekijk de video ‘Sessie\_01\_Toelichting\_Opdracht\_Excel\_Gedeelte’ .

### Vragen

1. Maak manueel een frequentietabel van de leeftijd van alle klanten. Verdeel daarbij in klasses.
2. Maak een histogram van de leeftijd van alle klanten.
3. Bepaal de scheefheid en kurtosis en geef conclusies. Komen de scheefheid en kurtosis overeen met wat je op het histogram ziet?
4. Maak een frequentietabel en histogram van het inkomen van alle klanten via de data analysis plugin. Indien nodig verdeel je ook hier in klasses. Bepaal via ‘descriptive statistics’ (data analysis plugin) alle centrum- en spreidingsmaten. Wat valt er te zeggen over het histogram van het inkomen?
5. Bepaal alle centrum- en spreidingsmaten volgens geslacht (dus van mannen en vrouwen apart).
6. Teken een boxplot van de leeftijden. Splits ook hier op volgens geslacht. Dus een aparte boxplot voor de leeftijd van de mannen en een aparte voor de leeftijd van de vrouwen. Kan je op basis van de boxplot verschillen in spreiding ontdekken tussen de leeftijden van de mannen versus die van de vrouwen?
7. Onderzoek op een gelijkaardige manier of er een verschil zit in uitgavescore tussen mannen en vrouwen. Welke conclusies kan je op basis van de boxplot trekken?
8. Onderzoek of er uitschieters te vinden zijn in de data. Gebruik enerzijds de methode via de interkwartielafstand (IQR). Waarden boven $Q3 +1,5 x IQR en waarden onder Q1 - 1,5 x IQR kan je als uitschieters beschouwen. Bepaal ook uitschieters door te kijken welke waarden meer dan drie standaardafwijkingen van het gemiddelde af gelegen zijn.

## Opdracht DEEL 2 : Python

De dataset en de opgave zijn te vinden in Opdracht\_01\_Centrum\_Spreidingsmaten.ipynb.

Bekijk de video ‘Sessie\_01\_Toelichting\_Opdracht\_Python\_Gedeelte’ .

Beantwoord de vragen in het Jupyter notebook document.