

Data verwerving en voorbereiding

In dit hoofdstuk leer je waar je data kunt vinden, hoe je deze ophaalt en voorbereidt voor visualisatie. Je maakt kennis met verschillende datatypen, bronnen en technieken voor het opschonen van data.

- Uitleg
- Opdracht

Uitleg

Datatypes en -structuren

- **Kwantitatieve data:** Numerieke waarden (bijv. aantallen, meetwaarden). Te visualiseren met histogrammen, boxplots.
- **Kwalitatieve data:** Categorieën of tekst (bijv. landen, productnamen). Te visualiseren met staafdiagrammen, woordwolken.
- **Gestructureerde data:** Tabellen in CSV, databases. Helpt bij kolom- en rijstructuur.
- **Ongestructureerde data:** Tekstbestanden, logs, afbeeldingen. Vereist extractie of transformatie.

Data bronnen

- **Open data portals:** CBS, Eurostat, World Bank.
- **API's:** RESTful endpoints, JSON- of XML-response.
- **Web scraping:** Automatisch gegevens ophalen van webpagina's met tools als BeautifulSoup.
- **Eigen data:** Sensoren, enquêtes of experimenten.

Data importeren

- In Python: gebruik `pandas.read_csv()`, `pandas.read_json()`, `sqlalchemy` voor databases.
- In Excel: data import wizard, Power Query voor complexere bewerkingen.

Data cleaning technieken

- **Ontbrekende waarden:** Identificeren (`isnull()`), verwijderen (`dropna()`) of imputeren (gemiddelde, mediaan).
- **Outliers:** Detectie via statistische methoden (IQR, z-score), visuele inspectie (boxplot).
- **Duplicaten:** Identificatie (`duplicated()`), verwijderen (`drop_duplicates()`).
- **Normalisatie/standaardisatie:** Schalen van kolommen (min-max, z-score) voor vergelijkbare eenheden.

Data profiling

Maak een eerste verkennende analyse:

- **Statistische samenvatting:** `describe()` toont mean, median, min-max.
- **Distributies bekijken:** Histogrammen, density plots.
- **Relaties onderzoeken:** Scatterplot matrix, correlatiematrix.

Opdracht

1. **Dataset kiezen:** Selecteer een dataset van een open data portal of via een API.
2. **Importeren:** Laad de dataset in Python (pandas) of Excel.
3. **Data cleaning:** Voer de volgende stappen uit en lever op:
 - Identificeer en behandel ontbrekende waarden.
 - Detecteer en behandel outliers.
 - Verwijder duplicaten.
 - Normaliseer één of twee kolommen.
4. **Data profiling:** Maak een overzicht met:
 - Statistische samenvatting (`describe()`).
 - Twee histogrammen voor numerieke kolommen.
 - Een correlatiematrix als heatmap.
5. **Documentatie:** Schrijf kort (300–500 woorden) over je reinigingsproces en uitdagingen.

Bundel alles in een verslag met de belangrijkste grafieken en een toelichting bij elke stap.