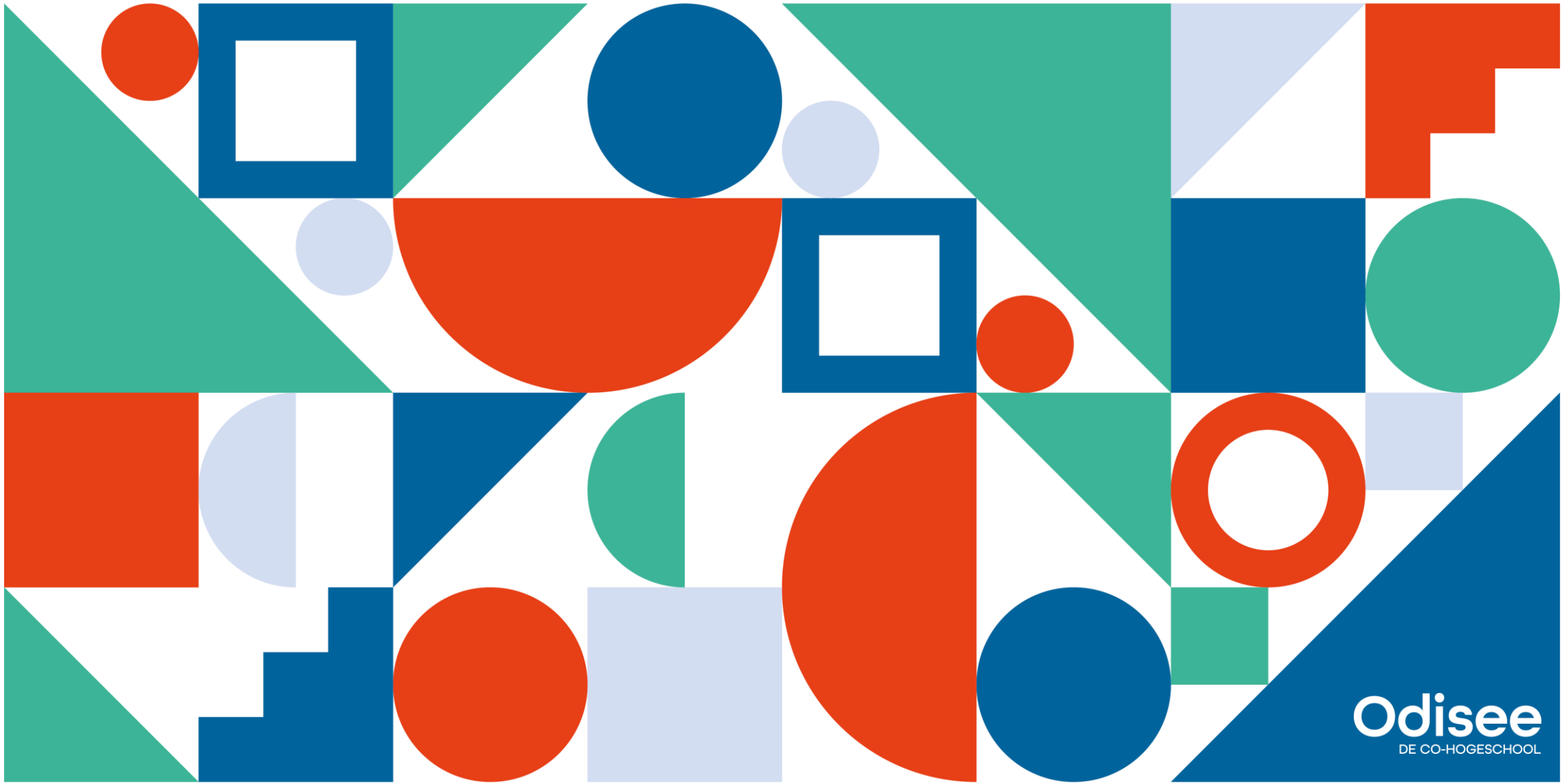

Odisee
DE CO-HOGESCHOOL

Data Science – week 4

Jens Baetens

How to participate?

Copy participation link

1. Go to wooclap.com
2. Enter the event code in the top banner

Event code
UROZPY

1. Send @UROZPY to 0460 200 711
2. You can participate

You cannot vote anymore

Welke functie kan je gebruiken om null waarden te selecteren

1. isna — 100% — 13
2. dropna — 0% — 0
3. fillna — 0% — 0

Go to **wooclap.com** and use the code **UROZPY**

Met de volgende functie kan je een aantal statistieken (mean, min, max, ...) berekenen voor de kolommen in een dataframe

1. info — 0% — 0
2. groupby — 0% — 0
3. select_dtypes — 0% — 0
4. describe — 0% — 0
5. drop — 0% — 0

00:59

Go to **wooclap.com** and use the code **UROZPY**

Met de dropna functie kan je

1. enkel rijen weglaten als er 1 of meerdere NaN waarden aanwezig zijn — 0% — 0
2. enkel kolommen weglaten als er 1 of meerdere NaN waarden aanwezig zijn — 0% — 0
3. rijen of kolommen weglaten als er 1 of meerdere NaN waarden aanwezig zijn — 0% — 0
4. rijen of kolommen weglaten als er exact 1 NaN waarde aanwezig is — 0% — 0
5. null-waarden in het dataframe vervangen — 0% — 0

00:50

You cannot vote anymore

Met de fillna functie kan je

1	willekeurige waarden vervangen door een andere waarde	9%	1
2	NaN waarden vervangen (alle NaN waarden door dezelfde waarde)	27%	3
3	NaN waarden invullen (alle NaN waarden in een kolom of rij met dezelfde waarde)	9%	1
4	NaN waarden invullen (ofwel met een vaste waarde of op basis van de vorige of volgende niet NaN waarde)	55%	6

You cannot vote anymore

Welke onderstaande technieken kunnen gebruikt worden om te voldoen aan privacy eisen over persoonsgegevens (meerdere zijn mogelijk)

Data analysis of EDA

Wat is het?

- EDA = Exploratory Data Analysis
- Proces waarbij de beschikbare data geanalyseerd wordt
  - Afwisselend met data cleaning
  - Rechtstreeks aansturen van bedrijfbeslissingen
  - Voorloper van ML/AI – modellen
    - Welk model is het best, welke parameters moeten we gebruiken, welke data is bruikbaar, …

Analyse op drie niveau's

- Algemene informatie over beschikbare data
- Informatie per kolom
- Informatie over het verband tussen verschillende kolommen

Niveau 1 - Algemene informatie

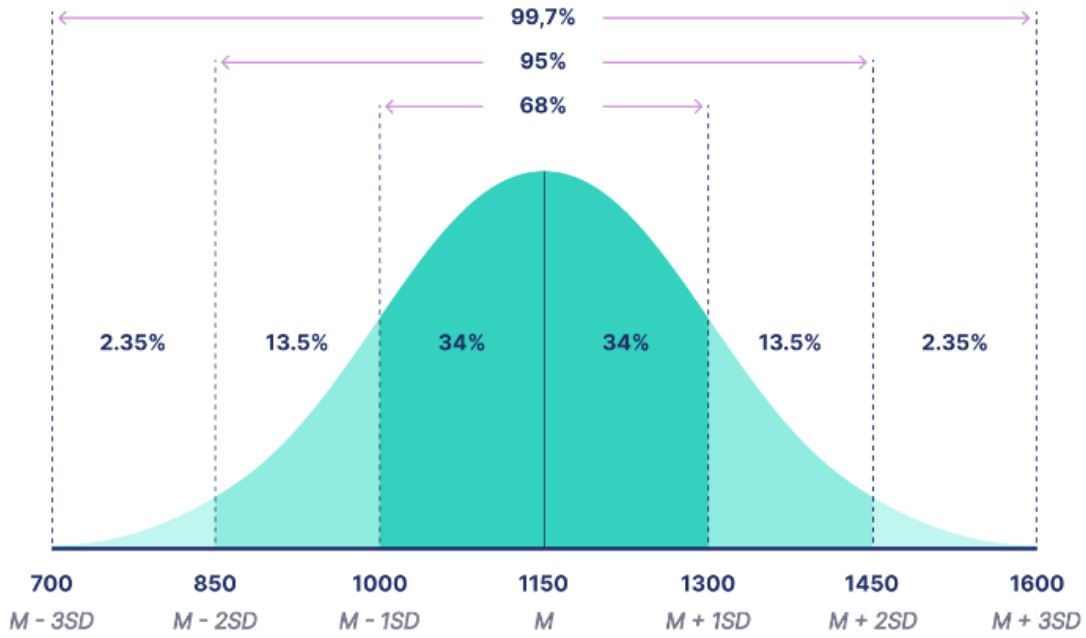
- Aantal rijen (observaties) en kolommen (features) zijn er
  - Welke data zit erin
  - Wat is het datatype
  - Categorieke vs numerieke data
  - Discrete vs continue data

Niveau 2 – Per kolom – unieke waarden

- Vooral voor categorieke data
  - Aantal unieke waarden
  - Aantal elementen per categorie
- Gebruikt om gebalanceerdheid te controleren
  - Ongebalanceerd kan nadelig zijn voor ML

Niveau 2 – Per kolom – statistische waarden

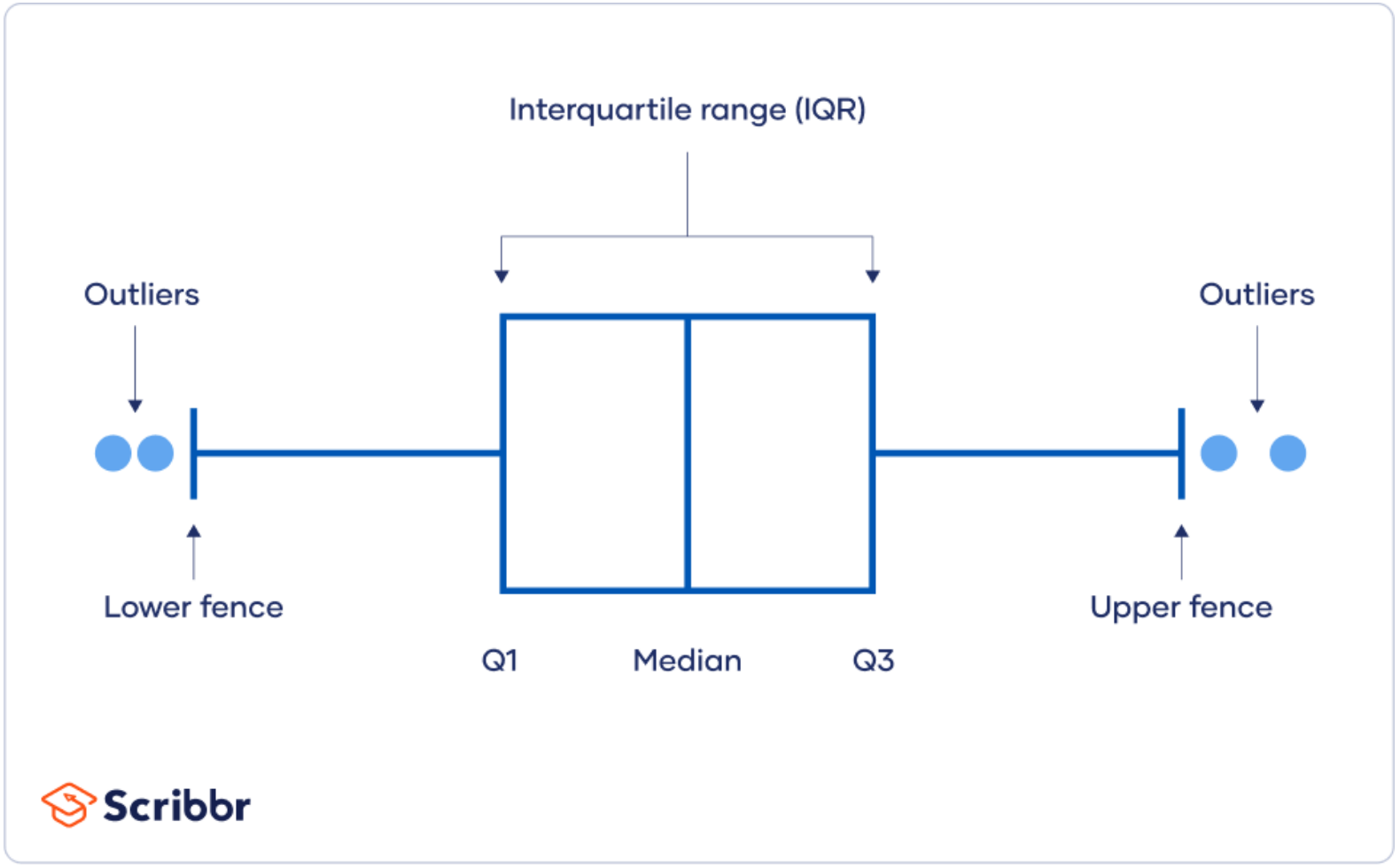
- Voor numerieke kolommen
  - Minimum, maximum, gemiddelde, standaardafwijking, percentielen, outliers



Niveau 2 – Per kolom – outliers

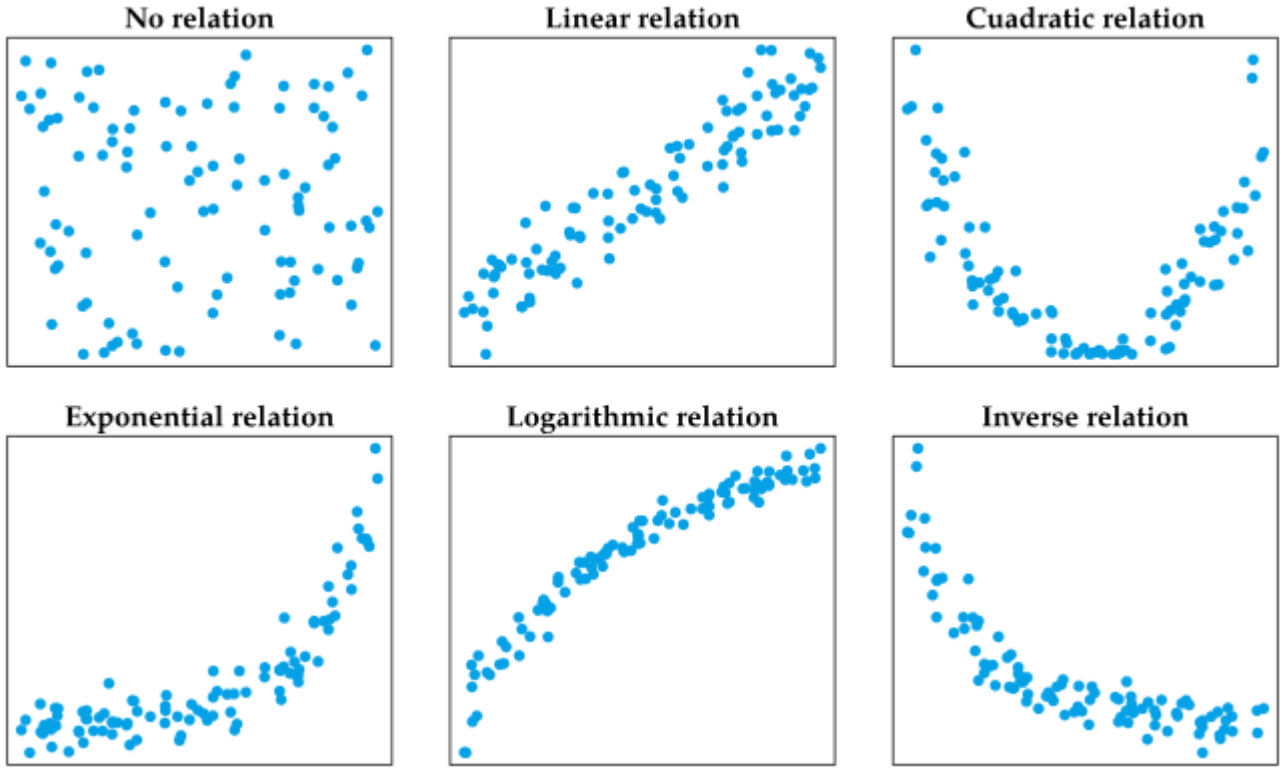
- Sorting method (manueel zoeken)
- Op basis van box plot (nieuwe data < minimum of > maximum)
- Statistische analyse
  - # std van het gemiddelde
  - Threshold manueel
- Interkwartiel methode

Niveau 2 – Per kolom – outliers – interkwartiel methode



Niveau 3 – tussen kolommen – scatter plot

- Zoek naar het verband tussen features
- Voor numerieke waarden
- Rekenintensief



Niveau 3 – tussen kolommen – correlation

- Correlatie coëfficiënt: indicatie van hoe sterk het verband is tussen twee variabelen
  - Waarde tussen -1 en 1
  - Teken of de verandering in dezelfde richting is
  - Grootte van het getal = hoe sterk het verband is

Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

Niveau 3 – tussen kolommen – correlation

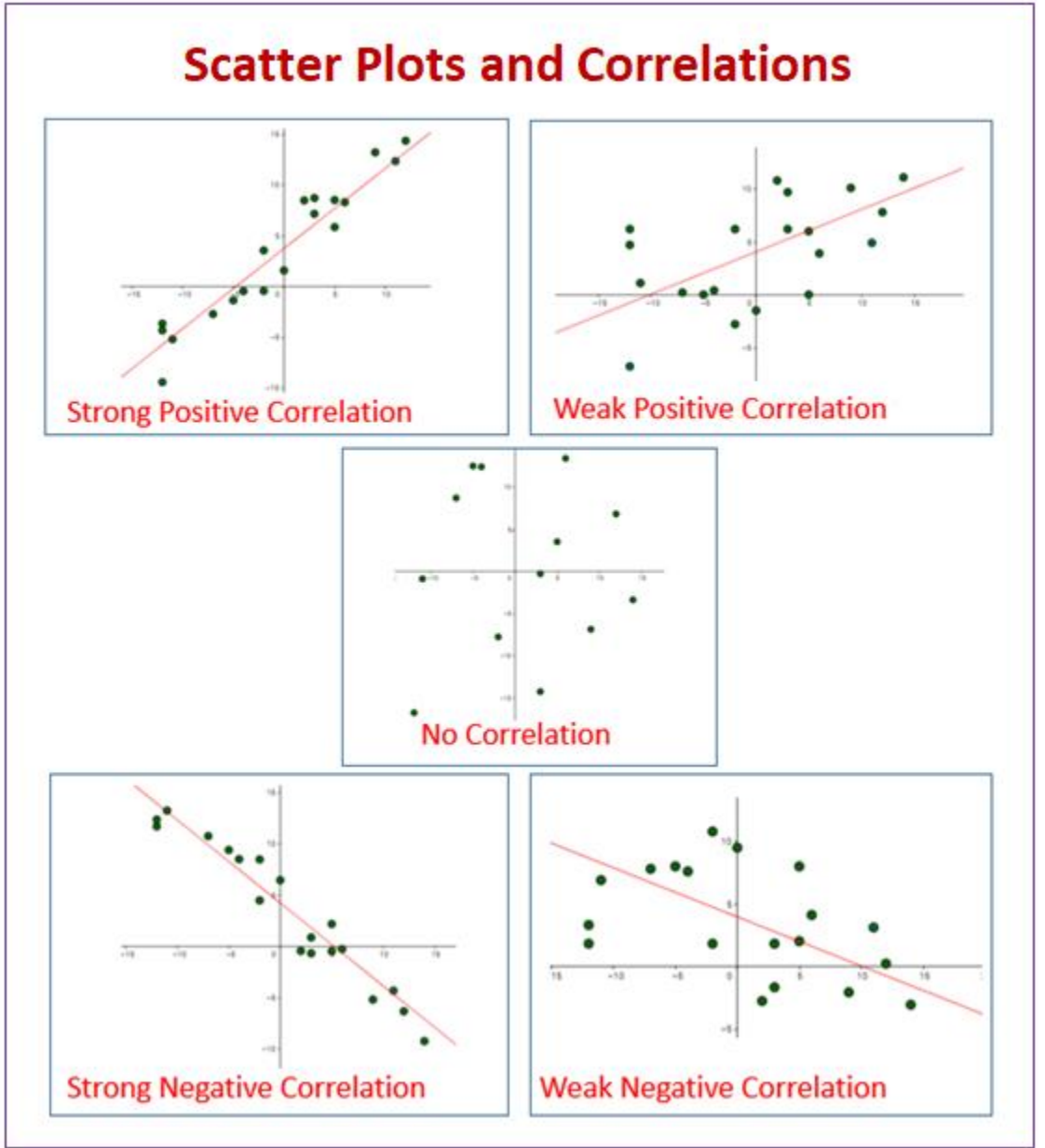
Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Ordinal: volgorde is belangrijk
Nominal: volgorde is niet belangrijk

Niveau 3 – tussen kolommen – Pearson's correlation

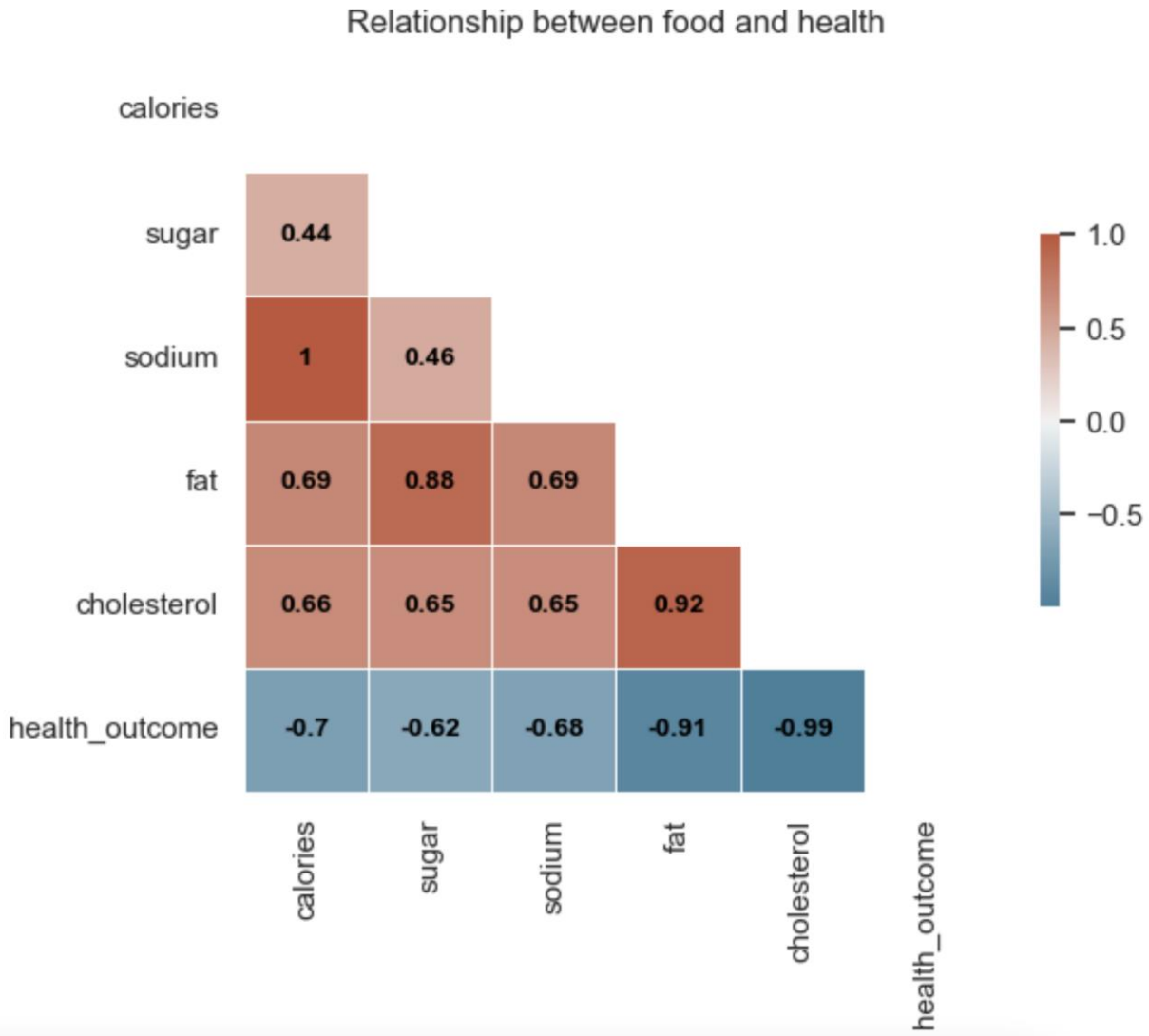
- Quantificieer het verband
  - Pearson correlatie

- Waarde tussen -1 en 1
  - -1 = sterk negatief
  - 0 = geen correlatie
  - 1 = sterk positief



Niveau 3 – tussen kolommen – Pearson's correlation

- Eenvoudig te bereken via .corr()
- Vaak voorgesteld als heatmap



Niveau 3 – tussen kolommen – Spearman correlation

- .corr(method="spearman")

- Te gebruiken wanneer
  - Minstens 1 ordinale variabele aanwezig
  - Minstens 1 variabele niet volgens standaardverdeling

Niveau 3 – tussen kolommen – Cramer's V correlation

- Tussen categorieke kolommen

$$\mathbf{V} = \sqrt{\frac{\chi^2}{N(k-1)}}$$

  - X^2 is het resultaat van de chi-square test
  - N aantal rijen
  - K = min(aantal rijen, aantal kolommen)

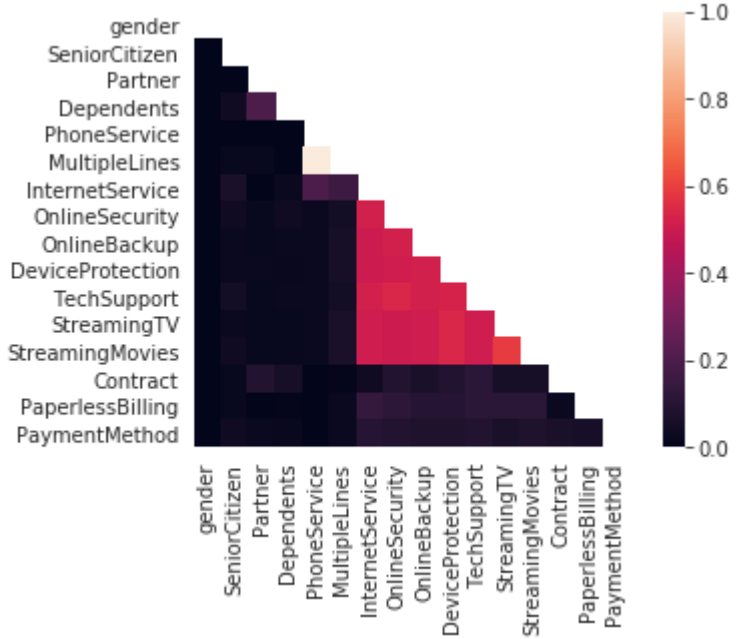
Niveau 3 – tussen kolommen – Cramer’s V correlation

```python
import pandas as pd
import scipy.stats as ss

# create the contingency table
contingency_table = pd.crosstab(df['Gender'], df['Marital Status'])

# calculate the chi-square test statistic
chi2, _, _, _ = ss.chi2_contingency(contingency_table)

# calculate the minimum of the number of categories in the two variables
min_categories = min(contingency_table.shape[0], contingency_table.shape[1])

# calculate Cramer's V coefficient
n = contingency_table.sum().sum()
V = np.sqrt(chi2 / (n * (min_categories - 1)))
```

$$\mathbf{V} = \sqrt{\frac{\chi^2}{N(k-1)}}$$

Niveau 3 – tussen kolommen – Cramer’s V correlation

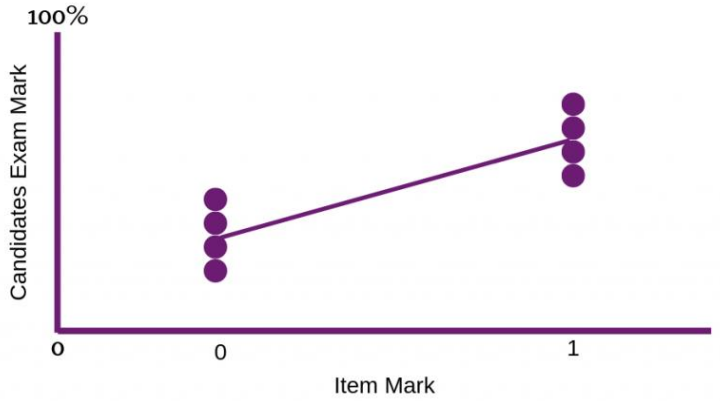
- V < 0.1 -> geen verband
- 0.1 < V < 0.3 -> zwak verband
- 0.3 < V < 0.5 -> gemiddeld verband
- V > 0.5 -> sterk verband



Niveau 3 – tussen kolommen – Point Biserial Correlation

- Tussen continue en categorieke kolom met twee mogelijkheden
  - Bvb bij one-hot encodings

```
>>> from scipy import stats
>>> a = np.array([0, 0, 0, 1, 1, 1, 1])
>>> b = np.arange(7)
>>> stats.pointbiserialr(a, b)
(0.8660254037844386, 0.011724811003954652)
```



Point Biserial

Hoe gebruiken we de correlation?

- Vooral kijken naar rij/kolom van wat we willen weten
- De sterkste correlaties bepalen welke waarden een grote impact erop hebben
- Deze waarden worden dan best gebruikt voor analyses / ML-modellen

- Let op: Verschillende correlatiecoefficienten kunnen niet zomaar vergeleken worden

Belangrijke termen

- Exploratory Data Analysis
- Correlation Coefficient
- Pearson’s correlation
- Cramer’s V correlation
- Interkwartielafstand
- Outliers

Huiswerk

EDA extra tutorial

- Ga naar:
  - https://www.kaggle.com/learn/intermediate-machine-learning
  - Volg de tutorial van hoofdstuk 1 tot en met 3.

  - De informatie in de tutorials is te kennen leerstof en helpt bij het maken van de oefeningen

EDA oefening

- Opgave: https://classroom.github.com/a/UAH3l74x
- Maak de oefening individueel tegen volgende week

- Deze oefening wordt geëvalueerd