LECTURER: TAI LE QUY

DATA SCIENCE

TOPIC OUTLINE

UNIT 3

DATA PREPROCESSING

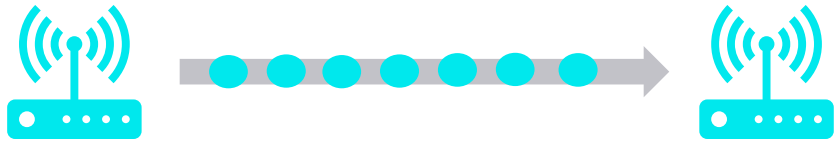
On completion of this unit, you will have learned …

- data transmission methods and techniques.
- how to handle missing values and outliers in a dataset.
- how to apply correlation analysis.
- data transformation approaches.
- data visualization tools.

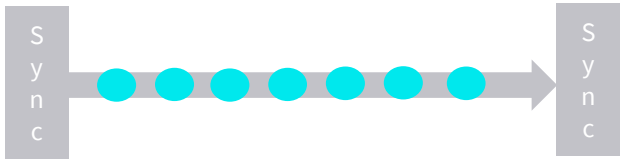
EXPLAIN SIMPLY

1. Explain the difference between asynchronous and synchronous transmission.
2. Describe correlation coefficients using your own words and why they are important in data preprocessing.
3. Identify three different data visualization forms and give an example of their respective usage.

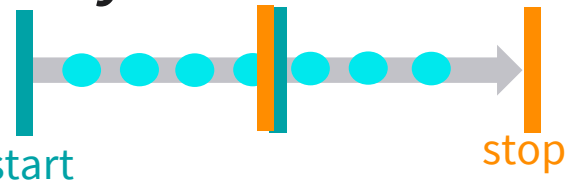
Serial data transmission



- reliable, lower cost
- synchronous

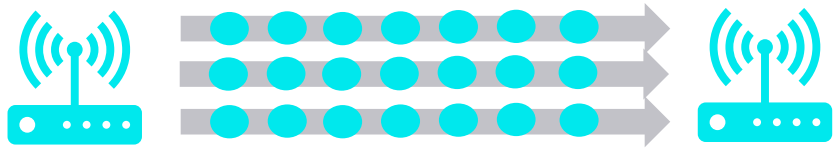


- asynchronous



Example: laptop ⟷ laptop

Parallel data transmission



- fast, short transmission distance

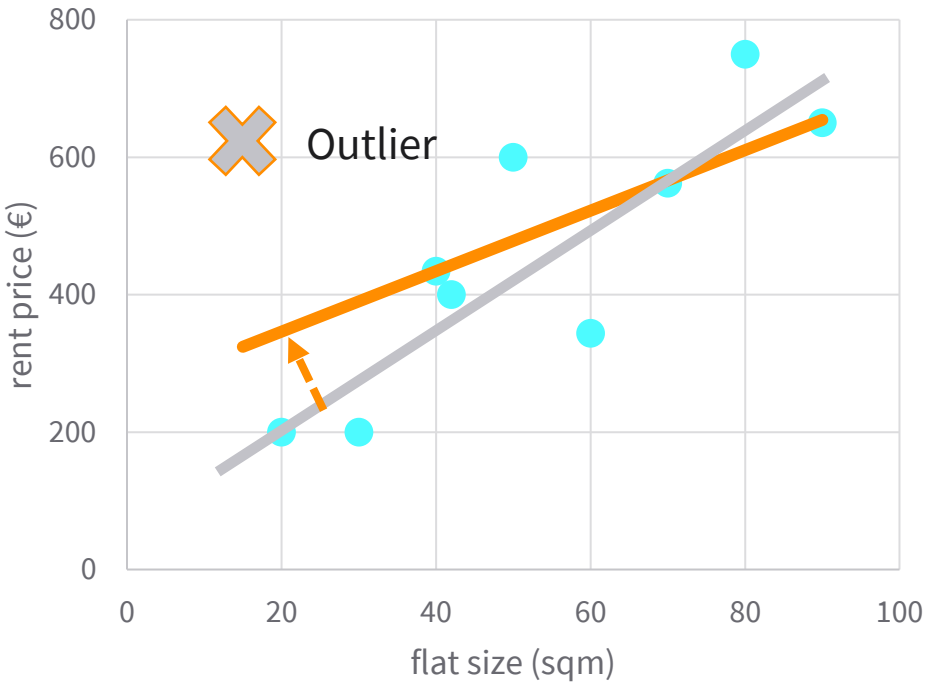
Example: laptop ⟷ printer

Source of the graphics: Drawing by Antonia Schulze, 2021.

DATA CLEANSING

MISSING VALUES

OUTLIERS



REMOVE OBSERVATIONS

rare condition/ anomalies

IMPUTE WITH INTERPOLATED VALUE

missing temperature in time series

IMPUTE WITH AVERAGE

numerical data: salary

IMPUTE WITH MODE

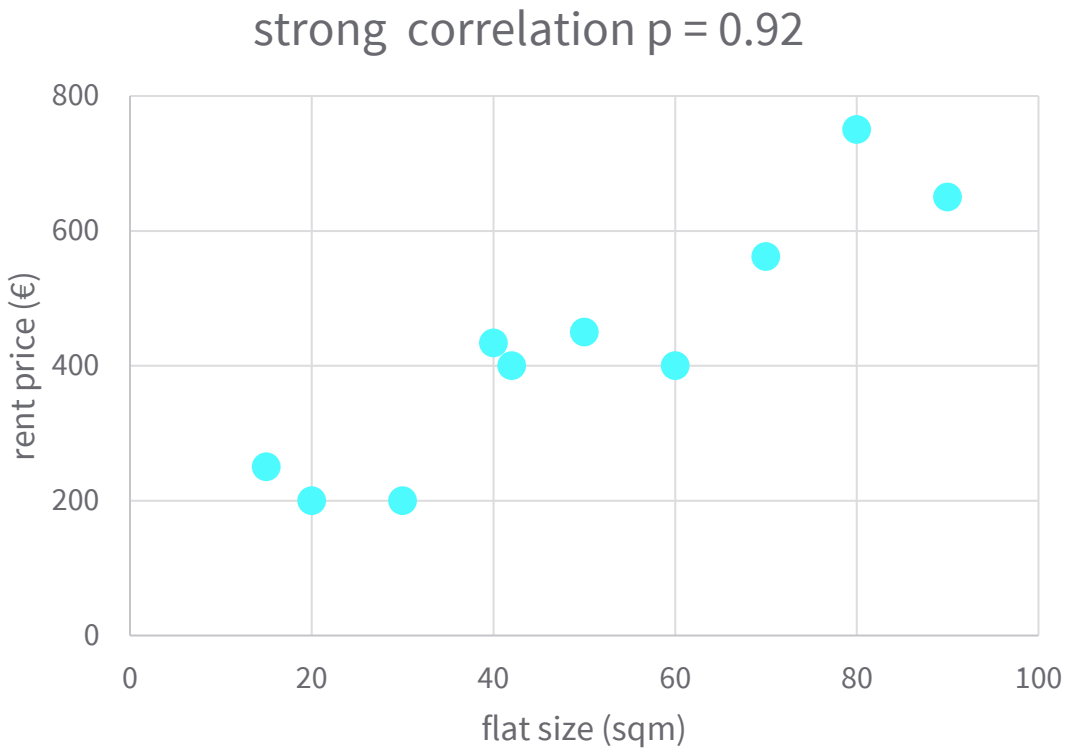
categorical data: gender

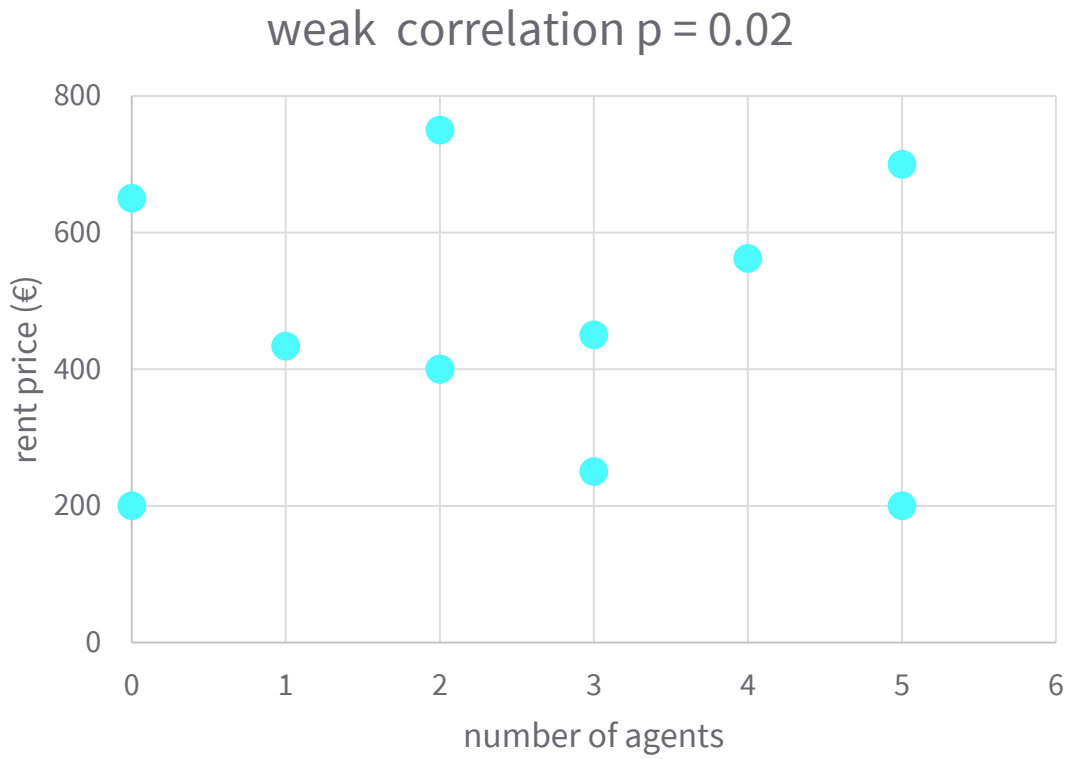
CONSIDER POTENTIAL BIASES BEFORE APPLYING ANY METHOD

Source of the graphics: Drawing by Antonia Schulze, 2021.

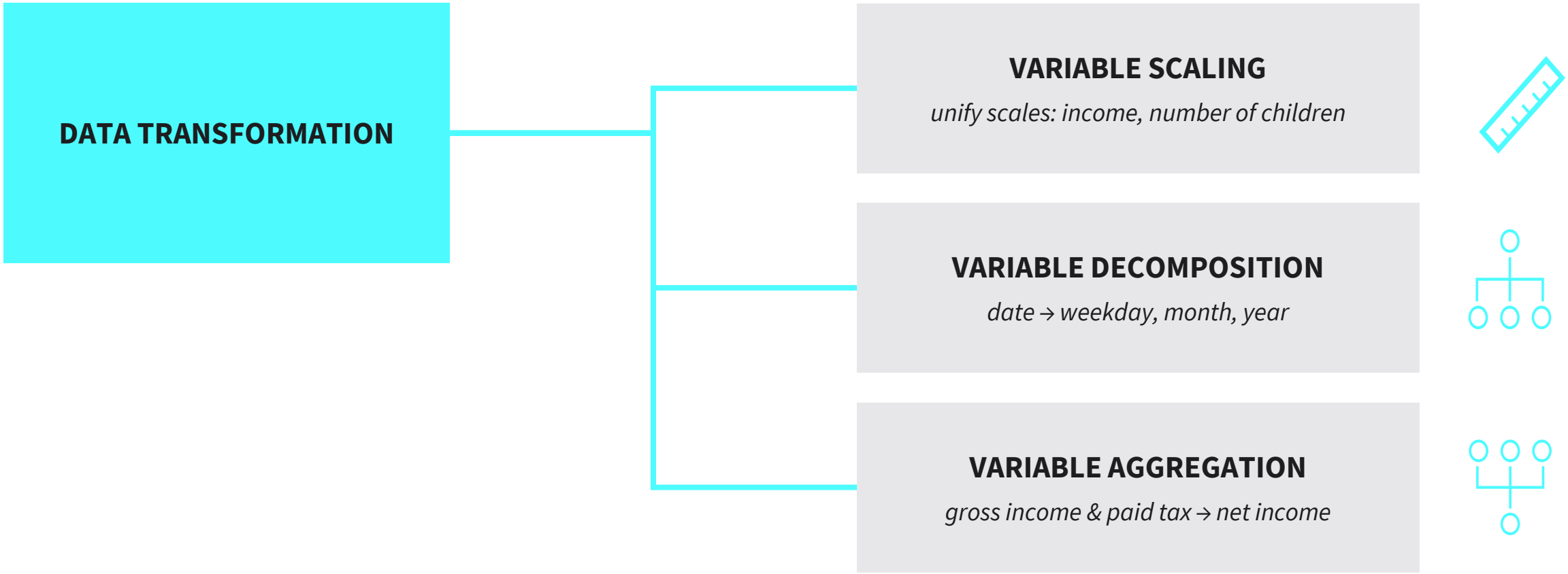
REDUNDANCY

Correlation = measure of direction and strength of the relationship between two numerical variables.

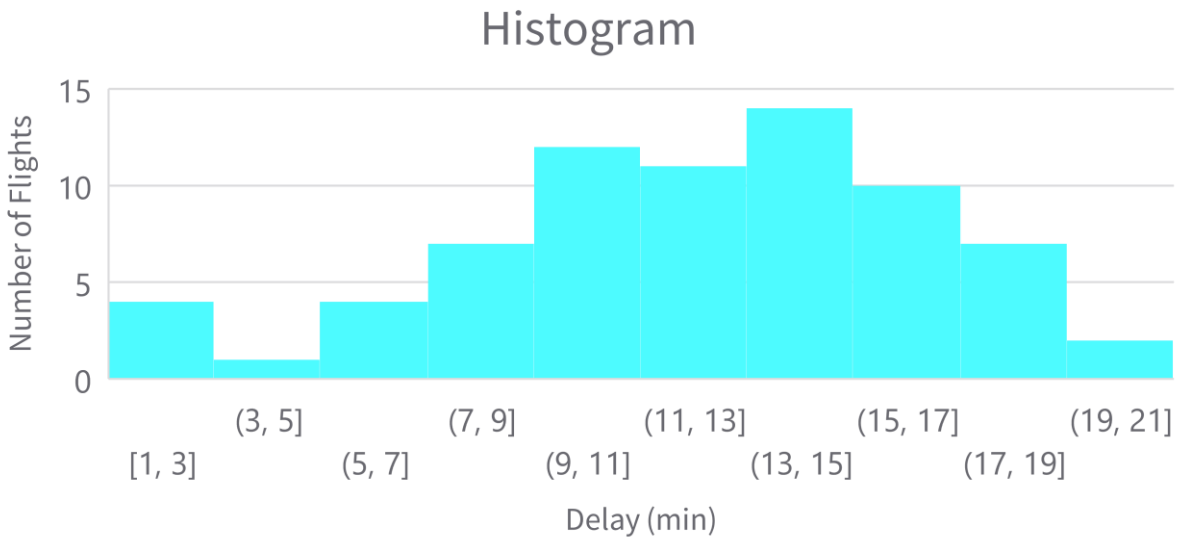


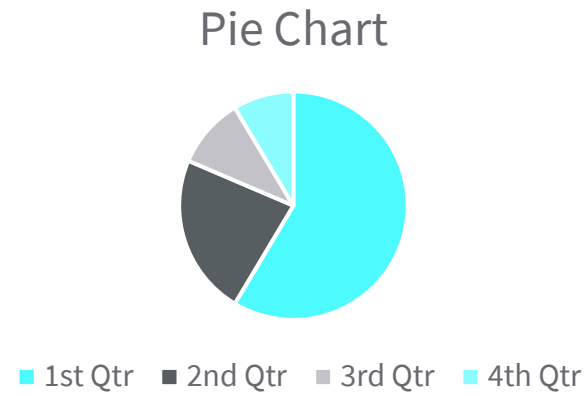


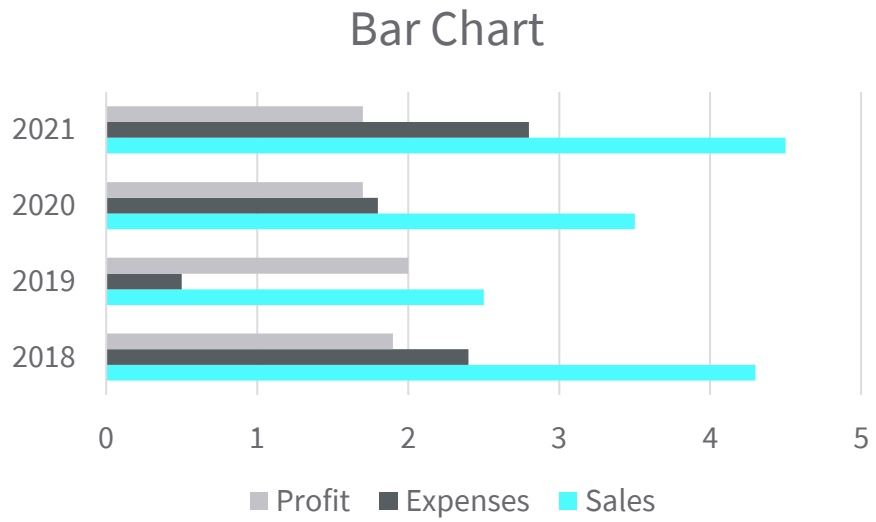
Source of the graphics: Drawing by Antonia Schulze, 2021.

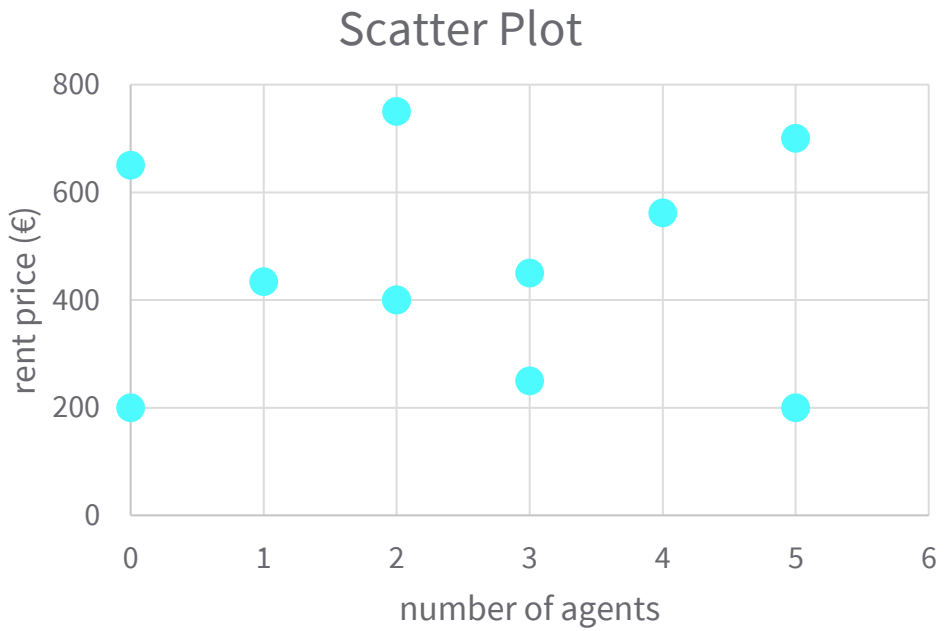

DATA TRANSFORMATION
VARIABLE SCALING
unify scales: income, number of children
VARIABLE DECOMPOSITION
date → weekday, month, year
VARIABLE AGGREGATION
gross income & paid tax → net income

DATA VISUALIZATION



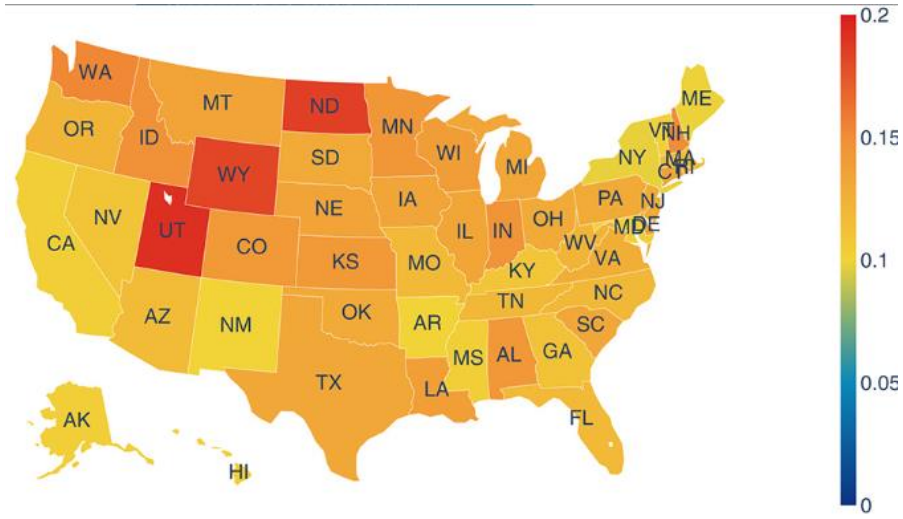




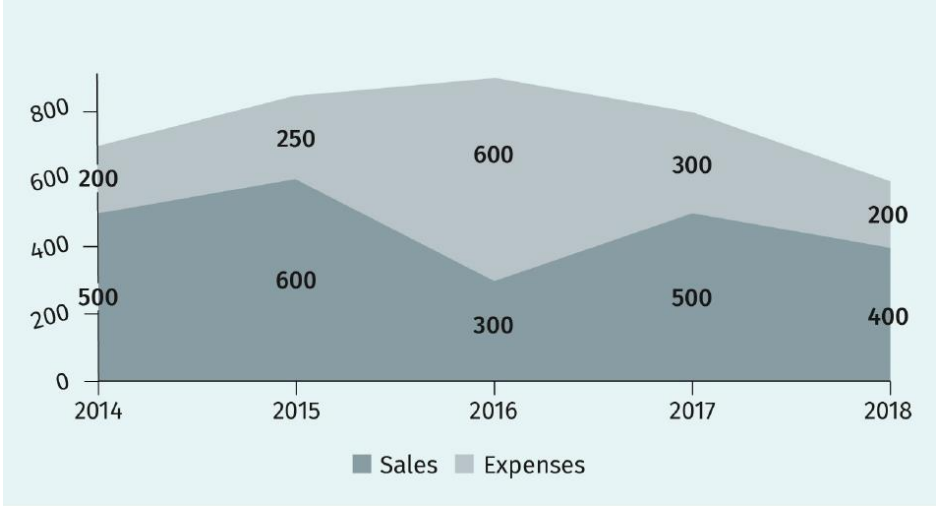


Source of the graphics: Drawing by Antonia Schulze, 2021.

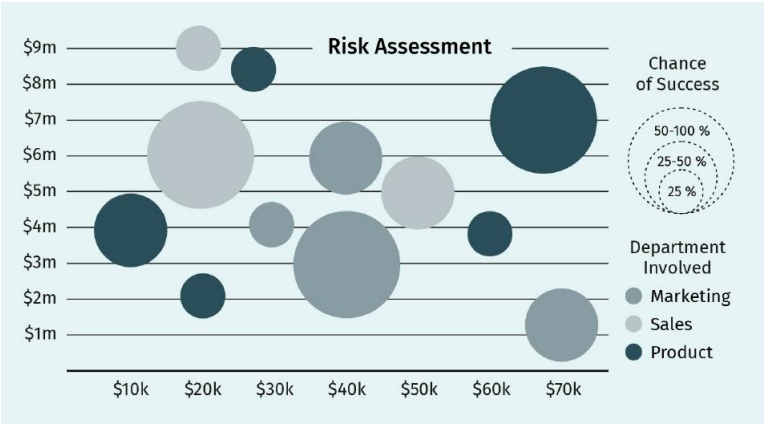
DATA VISUALIZATION



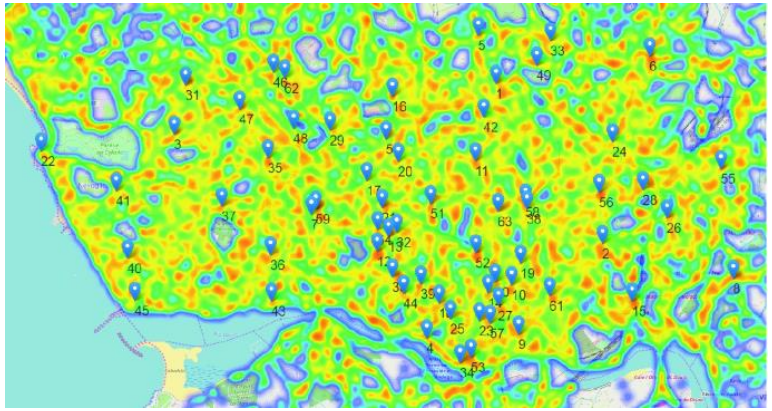
Geomap Visualization



Area Chart



Bubble Chart



Heat Map

You have learned …

- data transmission methods and techniques.
- how to handle missing values and outliers in a dataset.
- how to apply correlation analysis.
- data transformation approaches.
- data visualization tools.

SESSION 3

TRANSFER TASK

TRANSFER TASK

You work at a telecommunications company and your manager gives you the following dataset and would like you to present the data using data visualization techniques. Pay attention to potential steps of preprocessing and start drafting on a paper.

customerID	gender	age	location	tenure	PhoneService	InternetService	MovieSubscription	Contract	MonthlyCharges
234	Female	44	France	1	No	DSL	No	Month-to-month	29.85
784	Male	56	Germany	34	Yes	DSL	No	One year	56.95
893	Male	23	UK		Yes	No	No	Month-to-month	53.85
345	Male	80	USA	45	No	DSL	No	One year	42.3
831	Female	28	Germany	2	Yes	Fiber optic	No	Month-to-month	70.7
934	Female	174	Poland	8	Yes	Fiber optic	Yes	Month-to-month	99.65

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. The process of removing the variable's average and dividing by the variable's standard deviation is called a high false negative rate.

   a) variable decomposition
   b) variable scaling
   c) variable aggregation
   d) variable correlation analysis

2. Correlation analysis is applied to handle…

   a) outliers.
   b) missing values.
   c) duplicate records.
   d) redundant variables.

3. The data visualization tool which shows proportions of a whole, where the value of the variables is 100% is the …

   a) combo chart.
   b) area chart.
   c) pie chart.
   d) bubble chart.