EI: Data handling and visualization

1. Data visualization

In which way can you visualize the following datasets?

- Number of songs in a streaming platform, time and dates when they are listened
- Number of visitors on a webpage and location of visitors
- Numbers of student applications for two different universities that are received by day of the week

2. Variables

Which of the following variables are discrete and which continuous?

- Nationality of students enrolled in a study program
- Daily temperature of a city

You are given the following dataset that contains the number of people that visited a website at different time points:

Number of visitors
2
10
4
15
7
10
5
9
4

- What is the mean number of visitors that the website has received?
- What is the standard deviation?

3. Relating variables

You are asked to describe the relation between two variables using linear regression. In each of the following cases, which variable will be the response, and which one the predictor?

(a) Number of visitors that museum receives per day
(b) Sales of the museum restaurant per day

(a) Number of steps that a user of a wearable device makes per day
(b) Precipitation levels in mm
(c) Temperature in degrees Celsius

4. Linear regression

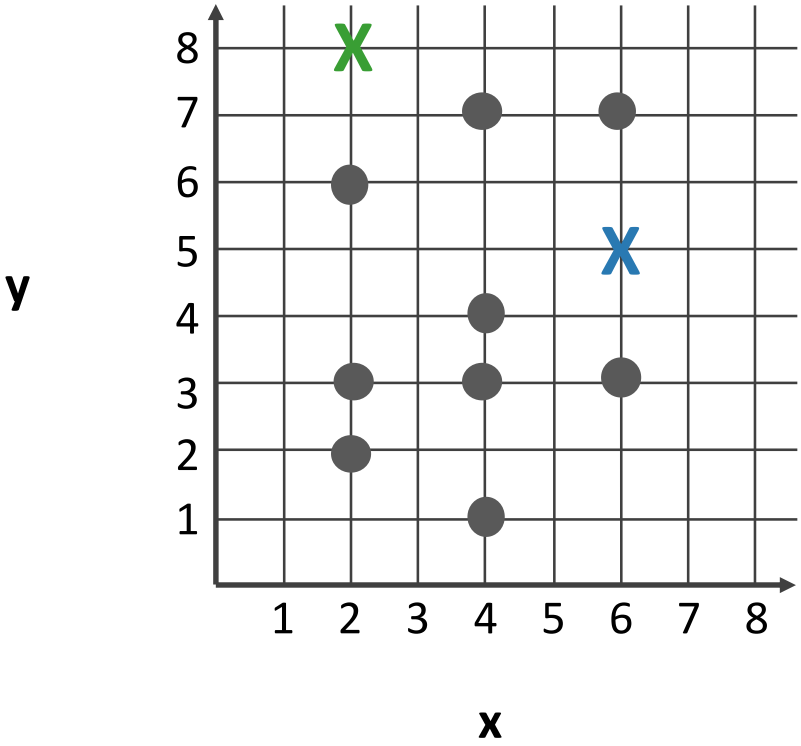
In the dataset that you received on the number of students applying at a study program, you are now additionally given the time of the day of these applications in hours.

(a) Describe a linear regression model that can be used to investigate the relation between time of the day and number of applications
(b) Compute the slope and intercept of this linear model
(c) How many student applications do you expect to receive at 13 h?

Number of applications	Time (hours)
2	8
10	9
4	12
15	9
7	12
10	11
5	8
9	10
4	9

4. Clustering

Perform one iteration of k-means clustering in the following set of points. The crosses indicate the initial centroids.


y
8
7
6
5
4
3
2
1
X
X
1
2
3
4
5
6
7
8
x