EI Assignment 07

Vithusan Ramalingam (21-105-515)

Jan Ellenberger (21-103-643)

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**
  - We can visualize this with a x-y-table where the x-axis is the time expressed by dates of the songs and the y-axis is the number of streamed songs
- **Number of visitors on a webpage and location of visitors**
  - With a global map with points placed in different locations. Those points vary in size according to the number of visitors of the webpage.
- **Numbers of student applications for two different universities that are received by day of the week**
  - With a x-y-table where the x-axis represents the different days of the week. And the y-axis the numbers of student applications. We could add the values with two different colors to represent each university.

2. Variables

Which of the following variables are discrete and which continuous?

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

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?**
  - $(2 + 10 + 4 + 15 + 7 + 10 + 5 + 9 + 4)/(9) = 7.333$
- **What is the standard deviation?**
  - $\sqrt{\frac{\left(2-\frac{22}{3}\right)^2+\left(10-\frac{22}{3}\right)^2+\left(4-\frac{22}{3}\right)^2+\left(15-\frac{22}{3}\right)^2+\left(7-\frac{22}{3}\right)^2+\left(10-\frac{22}{3}\right)^2+\left(5-\frac{22}{3}\right)^2+\left(9-\frac{22}{3}\right)^2+\left(4-\frac{22}{3}\right)^2}{8}} = 4.062$

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

- predictors

(b) Sales of the museum restaurant per day

- response

➔ We can explain the number of sales of the restaurant by the number of visitors of the museum.

(a) Number of steps that a user of a wearable device makes per day

- response

(b) Precipitation levels in mm

- predictor

(c) Temperature in degrees Celsius

- predictor

➔ We can explain the steps of wearable users with the help of the temperature variable and the precipitation level

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.

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

(a) Describe a linear regression model that can be used to investigate the relation between time of the day and number of applications.

- Predictor: (x) = time(h)
- Response: (y) = number of applications

(b) Compute the slope and intercept of this linear model

- Slope:

$\bar{x} = \frac{(\sum_{i=1}^{n} xi)}{n} = 9.778$

$\bar{y} = \frac{(\sum_{i=1}^{n} yi)}{n} = 7.333$

$s_{xx} = \sum_{i=1}^{n}(xi - \bar{x})^2 = \frac{176}{9}$

$s_{xy} = \sum_{i=1}^{n}(yi - \bar{y})(xi - \bar{x}) = \frac{11}{3}$

$\widehat{\beta 1} = \frac{sxy}{sxx} = 0.1875$

- Intercept:
-

$\widehat{\beta 0} = \bar{y} - \widehat{\beta 1}\bar{x} = 5.5$

(c) How many student applications do you expect to receive at 13 h?

Y = mx + b

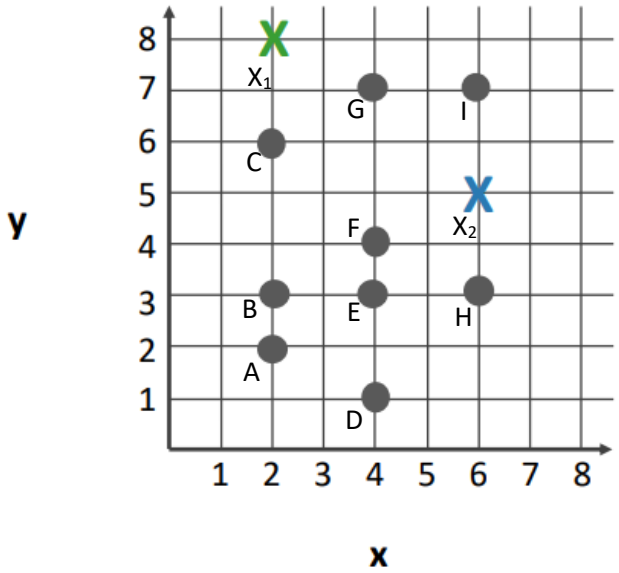
m = $\widehat{\beta 1}$
b = $\widehat{\beta 0}$
x = 13

y = 7.9375

We can expect that at 13h there we'll be around 11 applications.

4. Clustering

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



Step 1: choose total number of clusters k, in this case $k=2$

Step 2: select two random point as centroids, in this case $X_1=(2,8)$, $X_2=(6,5)$

Step 3: assign all remaining points to one centroid, the closest one

	X_1	X_2	A	B	C	D	E	F	G	H	I
x	2	6	2	2	2	4	4	4	4	6	6
y	8	5	2	3	6	1	3	4	7	3	7

$$d(x_i, x_j) = \left[\sum_{d=1}^{n} (x_{id} - x_{jd})^2\right]^{1/2}$$

Iteration 0:

	X_1	X_2	A	B	C	D	E	F	G	H	I	
	2	6	6	5	2	7,28	5,39	4,47	2,24	6,4	4,12	K_1
	8	5	5	4,47	4,12	4,47	2,83	2,24	2,83	3	2	K_2

Cluster k_1: $X_1(2,8)$: C(2,6), G (4,7)

Cluster k_2: $X_2(6,5)$: A(2,2), B(2,3), D(4,1), E(4,3), F(4,4), H(6,3), I(6,7)

Step 4: compute the new centroid of each cluster

Cluster k_1: $\left(\frac{2+4}{2}, \frac{6+7}{2}\right) = \left(\frac{6}{2}, \frac{13}{2}\right) = (3, 6{,}5)$

Cluster k_2: $\left(\frac{2+2+4+4+4+6+6}{7}, \frac{1+2+3+3+3+4+7}{7}\right) = (4, 3.29)$

Step 5: re-assign the points, given the new centroids

$$\left[\sum_{d=1}^{n} (x_{id} - x_{jd})^2\right]^{1/2}$$

Iteration 1:

X_1	X_2	A	B	C	D	E	F	G	H	I	
3	4	4,61	3,64	1,12	5,59	3,64	2,69	1,12	4,61	3,04	k_{1_2}
6,5	3,29	2,38	2,02	3,37	2,29	0,29	0,71	3,71	2,02	4,21	k_{2_2}

Cluster k_{1_2} : X_1 (3/6,5): C(2,6), G(4,7), I(6,7)

Cluster k_{2_2} : X_2 (4/3.29): A(2,2), B(2,3), D(4,1), E(4,3), F(4,4), H(6,3)