Exercises 3 — Data analysis

Important notice! Always save the solutions for the exercises. They will be submitted and evaluated (*Avaliação Contínua*). Confirm the exact date in *Inforestudante*. The exercises should be solved with the use of D3 functions, unless it is explicitly stated to not to use. The documentation of the functions can be found here: https://github.com/d3/d3-array/blob/master/README.md#statistics

Exercise A

1. Given the table below identify the following:
2. Type of each variable;
3. Which attributes are independent and which are dependent.

ID	Date	Price	Weight	Model	Origin	GPS Location
1	30-11-2021	10.000,00	3400	BMW	Europe	40.21, -8.40
2	30-4-2021	5.600,00	2500	Audi	Europe	40.33, -8.45
3	31-11-2021	18.000,00	1400	Mercedes	Europe	40.52, -8.51
4	1-1-2021	25.000,00	3090	VW	Europe	41.62, -9.41
5	13-6-2021	2.300,00	4142	Opel	Europe	42.53, -8.51
6	26-8-2021	7.000,00	2700	BMW	Europe	41.51, -9.72

Exercise B (Frequency tables)

Given the dataset about cereal brands—which can be accessed here cereal.csv—load the data and execute the following tasks using D3:

1. Construct a frequency table for "MFR" (manufacturer) attribute:
   a. Calculate the relative frequencies and percentages, and print to the console;
   b. What is the most frequent manufacturer?

       c.   What is the proportion of cereals of the predominant manufacturer?

2. Construct a frequency table for "Calories" attribute and:
   a. Create a dot plot representation;
   b. Create a histogram representation;
   c. Observe the result and describe the shape of distribution;

Exercise C (Central tendency)

Given the same dataset about cereals, load the data and execute the following tasks using D3:

1. Compute the mean rating of the cereals without using method d3.mean()
2. Compute the mean sugar of the cereals
3. Compute the median for sugar
4. Compare it with the mean
5. Compute the mode for the "proteins" attribute.
6. Find the quartiles for the rating score
7. Determine the minimum and maximum values for rating
8. Advanced (optional):
   a. Find outliers in terms of rating score
   b. Represent it using the box plot model