Tutorial and Laboratories

Week 11

The purpose of this week's tutorial and lab exercises is to understand the different ways to summarise the variation and the co-variations in categorical and numerical variables:

The *"Definition of Done"* for this tutorial and lab is to:

1- Understand the data summarisation methods of Categorical variables.
2- Understand the data summarisation methods of numerical variables.

Data Summarisation – Categorical variables:

The following methods can be used to summarise categorical variables:

1. **Frequencies**: Counting the number of elements per each group.
2. **Proportions**: Computing the probabilities of each group of elements.
3. **Marginal**: Computing the frequencies or the proportions per a group.

We also can use the visualisations to show the distributions of the variables as well as the relationships between the variables.

Exercise 1

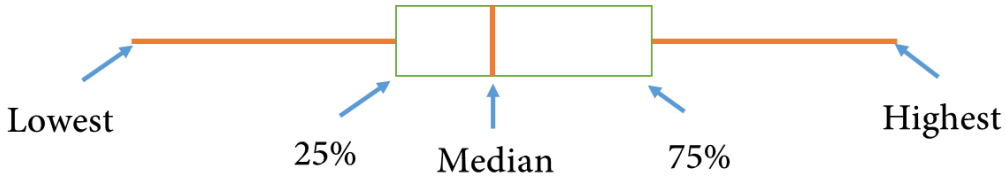
The following table simulates the data of a study that has been conducted on students of year 10, where they have been asked to define their favourite subject:

```
gender <- sample( c("Male", "Female", "Not-mentioned"), 100, replace=TRUE)
subject <- sample( c("Math", "Science", "Other"), 100, replace=TRUE)
data <- tibble(gender, subject)
```

1- Find the probability that the student chose math as his favourite subject, p(Math) = _____
2- Find the probability that the student was male, p(Male) = _____
3- Find the probability that the student was female, given the student chose Math as their favourite subject, p(Female | Math) = _____
4- Find the probability that the student chose Science as his favourite, given the student was female, p(Science | Female) = _____
5- Create a visualisation graph to show the counts and the proportions of the chosen favourite subjects per gender categories? Which geometrical shape would you use? Why?

Data Summarisation – Numerical variables:

The numerical data variables can be summarised by extracting the extracting the different measurements such as computing mean, median, minimum, maximum, range of the data values. One of the effective ways to summarise numerical variables is to compute the data quantiles. The following figure shows an illustration of one of the common quantiles (i.e. quartile) that presents a way to extract the data median, lower quartile, upper quartile and the Inter quartile range (IQR).



The inter quartile range can be computed as following:

$$IQR = lower\ quartile - upper\ quartile$$

Based on the IQR, we can specify the outlier values in the given data as any data points that fall before the lowest or after the highest. The lowest and the highest can be defined as following:

$$Lowest = lower\ quartile - 1.5 * IQR$$

$$Highest = upper\ quartile + 1.5 * IQR$$

Exercise 2

Given the following values, (199, 201, 236, 269,271,278,283,291, 301, 303,341):

1- compute the three quartiles and the inter-quartile range.
2- Use the box-plot to visualize the range of these values.
3- Identify the outlier values, if any.

Exercise 3

Given the data generated using the following code:

```
df <- data.frame(
gender=factor(rep(c("F", "M"), each=200)),
weight=round(c(rnorm(200,   mean=55,   sd=5),   rnorm(200,
mean=65, sd=5))))
```

1- Create a graph to visualize the distribution of the weight values.
2- Create a graph to visualize the distribution of the weight values per gender.
3- Add the average line to each group distribution.
4- Use the box plot to show the quartiles of each gender group.

Exercise 4 (Unsupervised activity)

1- Download and import the data from the following CSV file [Link].
2- Visualize the distribution of the species' weights grouped by their id.
(hint: let us use the boxplot to show this distribution)
3- Add the points of the weights of the species on top of the previous graph
4- Increase the transparency of the points.