Tutorial and Laboratories

Week 12

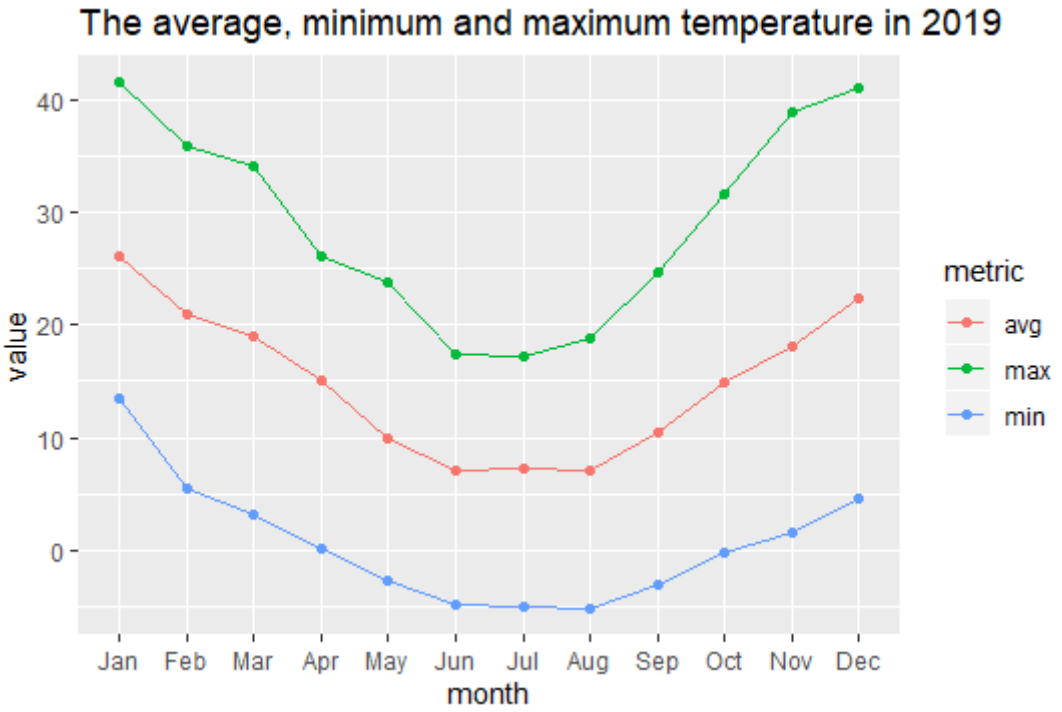
The purpose of this week's tutorial and lab exercises is to understand the Correlation coefficient and how we to implement it in R to check the linear relationship between the variables.

The objectives of this tutorial and lab are to:

1- Understand the concept of using the correlation to check the linear relationship between the variables.
2- Practice on using the R functions that are related to computing the correlation.

Before we start practising on the correlation, let us recap on how to use the visualisation as a tool to explore and analyse the data. Let us practise with the following exercise.

Exercise 1:
Please download the data on this link (or download from week 12) to be used in this exercise. Given these data, can you generate the following plot? What are



the steps would you follow?

Please note that, this graph is for 2019 only, as mentioned in the title of plot.

Correlation:

It is a statistical measure that can be used to measure the dependency relationship between the variables. In other words, it is a method to test the impact of changing predictor (i.e. explanatory) variables on a continuous, predicted (i.e. response) variable.

The correlation coefficient has the following properties:

1. The magnitude (absolute value) of the correlation coefficient represents the strength of the linear association between the explanatory and the response variable and it is always between -1 and 1.
2. The sign of the correlation coefficient indicates the direction of the relationship between the variables.
3. The correlation coefficient is unitless, i.e. the value will stay very similar even with scaling or normalising any of the variables.

Prerequisites

Please download the "Advertising.csv" file from this link (or download it from week 12) and put it in your working directory.

Exercise 2

1- Read the data from the advertising.csv file and get familiar with the included variables.
2- Remove the un-named data variable(s), if any
3- Check the summary of the "sales" variable

Exercise 3

Continue the previous steps as following:

1- Plot scatter plots between the sales and each of the other variables in the data.
2- Add a regression line using stat_smooth(method = lm) to the previous plots? What does lm mean?
3- Compute the correlation between sales variable and each of the other variables independently, using cor(var1, var2) function.
4- Generate the correlation matrix for all the variables together.
5- Visualise the correlation matrix using ggcorr(…) function, by replacing … with the data frame.

Exercise 4 (unsupervised activity)

We also can use the correlation matrix as a test the linear relationship between variables, using cor.test(var1, var2, method), where method could be “Pearson”, “Spearman”, etc.

1- Use the cor.test function to test the estimated correlation between sales and TV variables.
2- How do you interpret the ***p-value*** and ***estimate***?
3- Can you extract the ***p-value*** and ***estimate*** coefficients from the output of the correlation test?