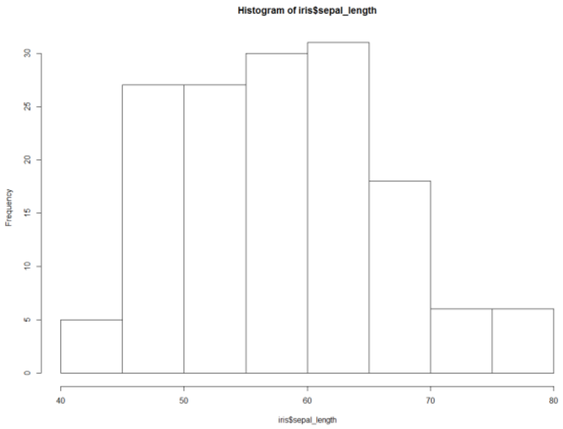
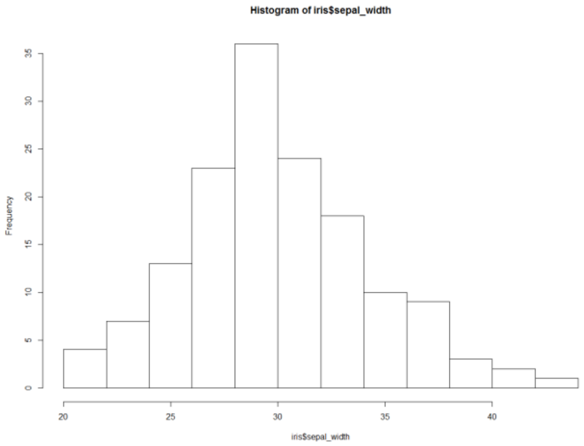
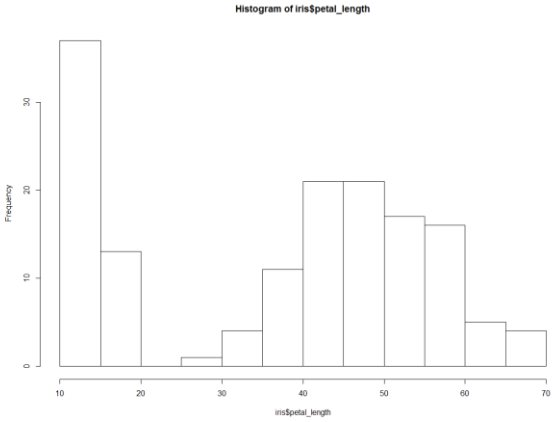
Isep Lab2, elements of answers

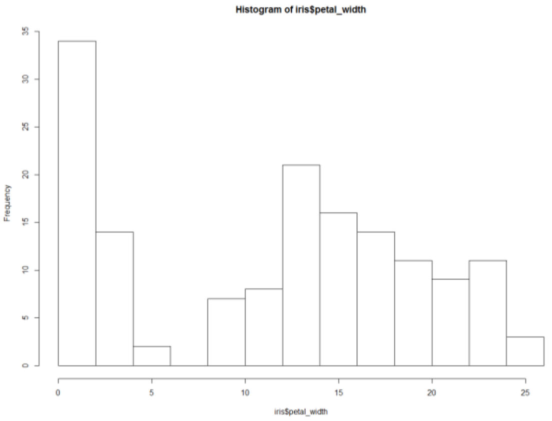
A Multivariate data set: Fisher Iris

A)1)

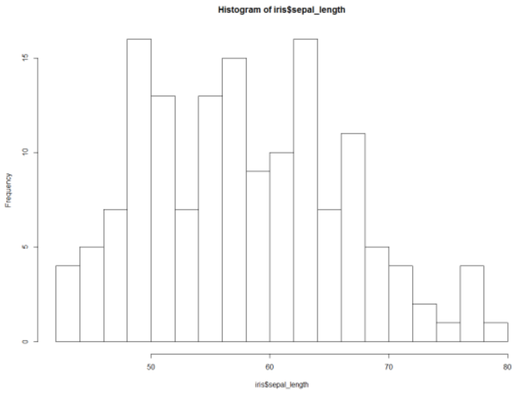


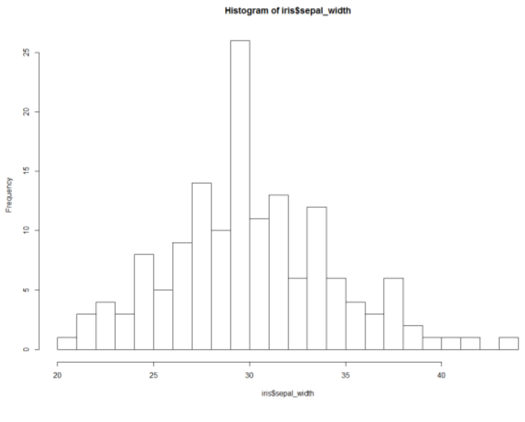


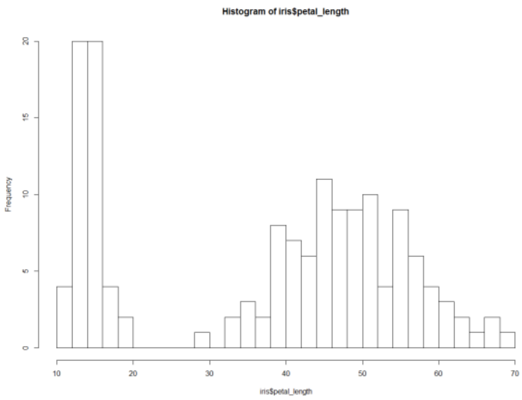


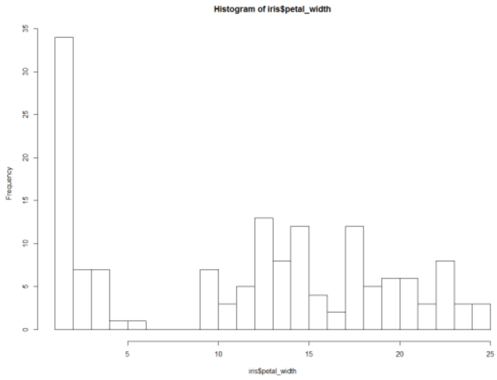


The sepal width attribute seems to be following a somewhat Gaussian distribution. The 2 petal attributes look bimodal, as if we have at least 2 different species. We need to zoom in a bit to be sure, so we need more breaks.

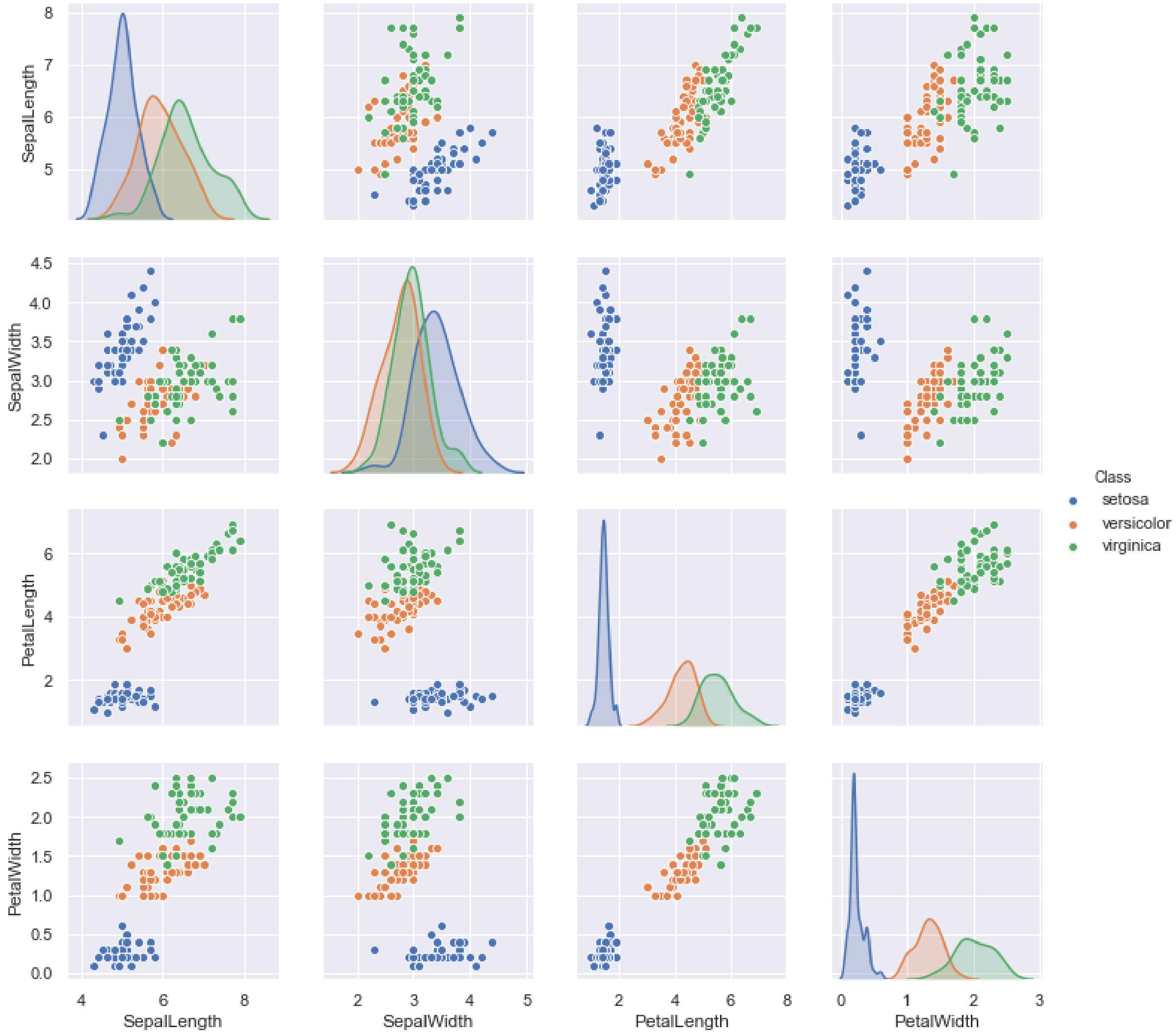


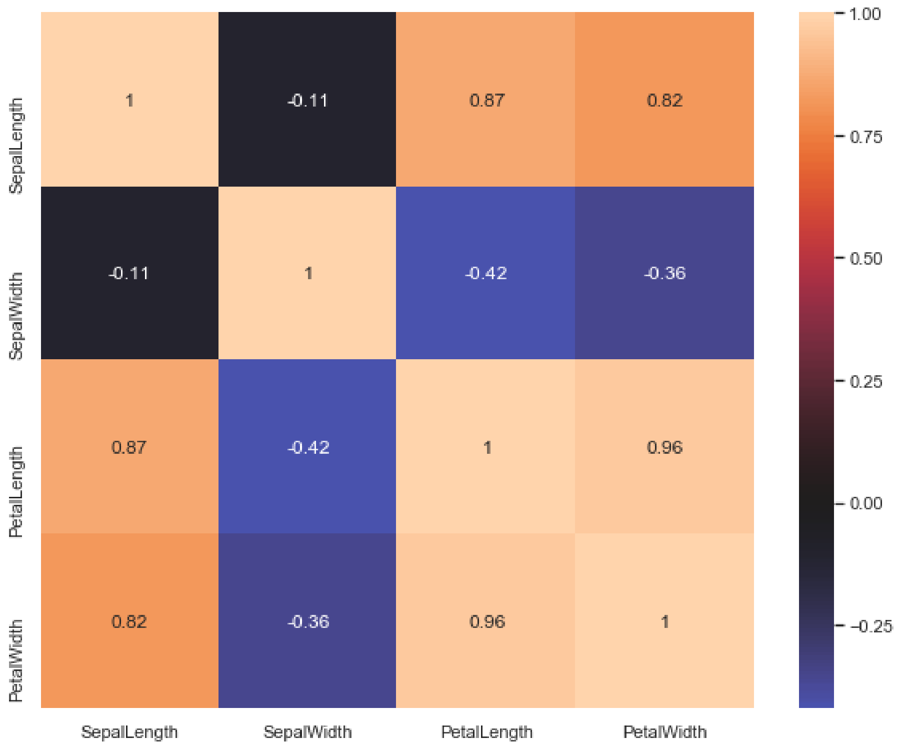






We can now confirm that both sepal attributes are following a somewhat gaussian distribution. We also clearly see 2 groups in the petal attributes, and we have 2 gaussian-like distributions in the petal_length attribute.

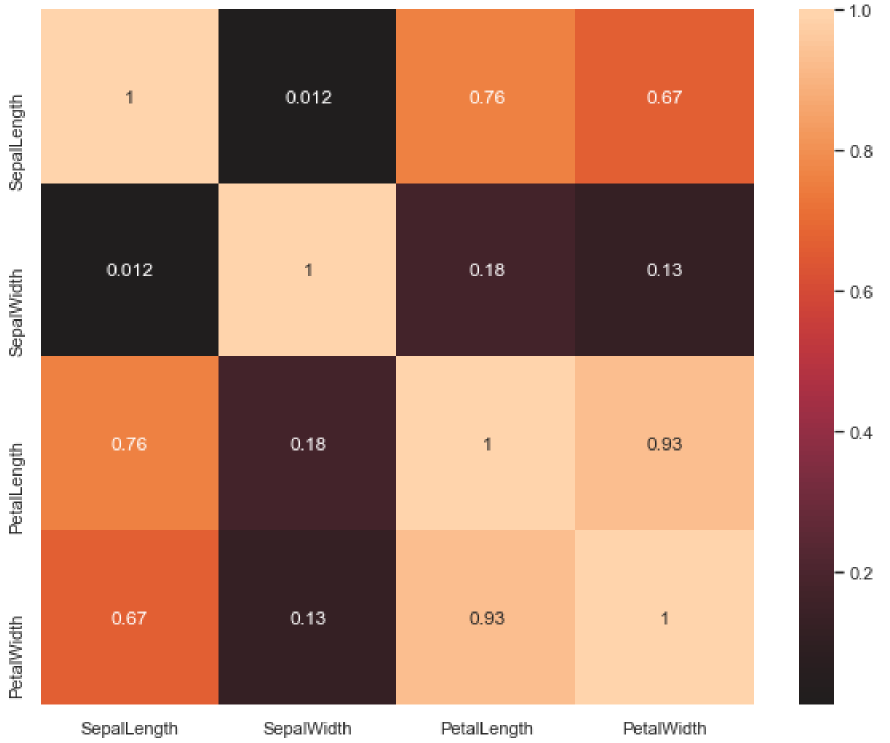




From the heatmap, we can see that we have 3 strong positive correlations :

- Petal Width and Sepal Length
- Petal Width and Petal Length
- Petal Length and Septal Length

All other correlations seem to be too low to say anything.



The determination heatmap confirms the Petal Width and Petal Length to be the mostly strongly linked attributes.

5) Using the course formulae, we can confirm the previous results. The strong correlation we found all have relatively narrow confidence intervals.

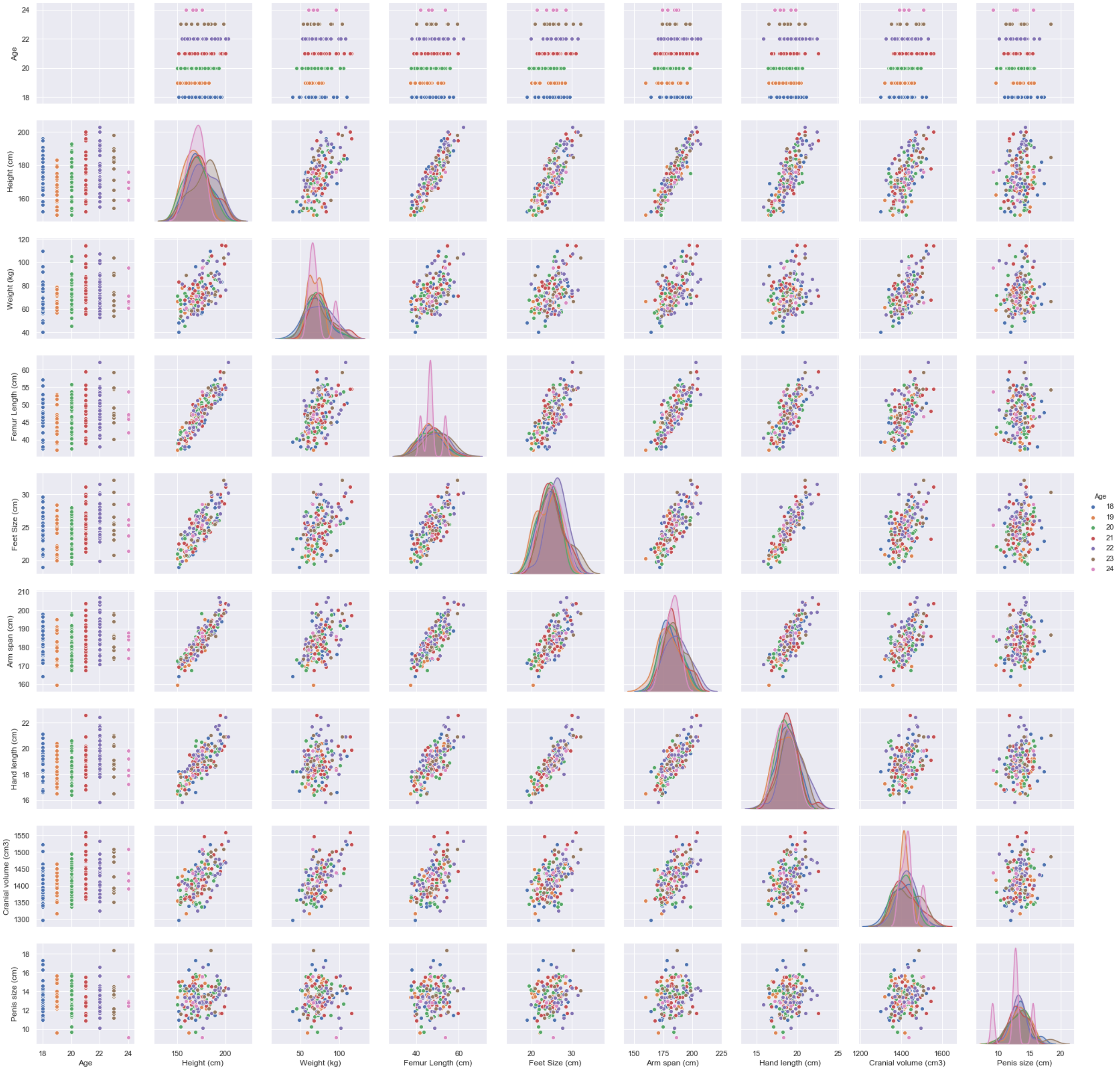
B Anthropometric data

1)

The function describe can be used to display interesting univariate information on all attributes.

2) The histograms of the attributes show distributions that looks Gaussian for most attributes except the age.

3)



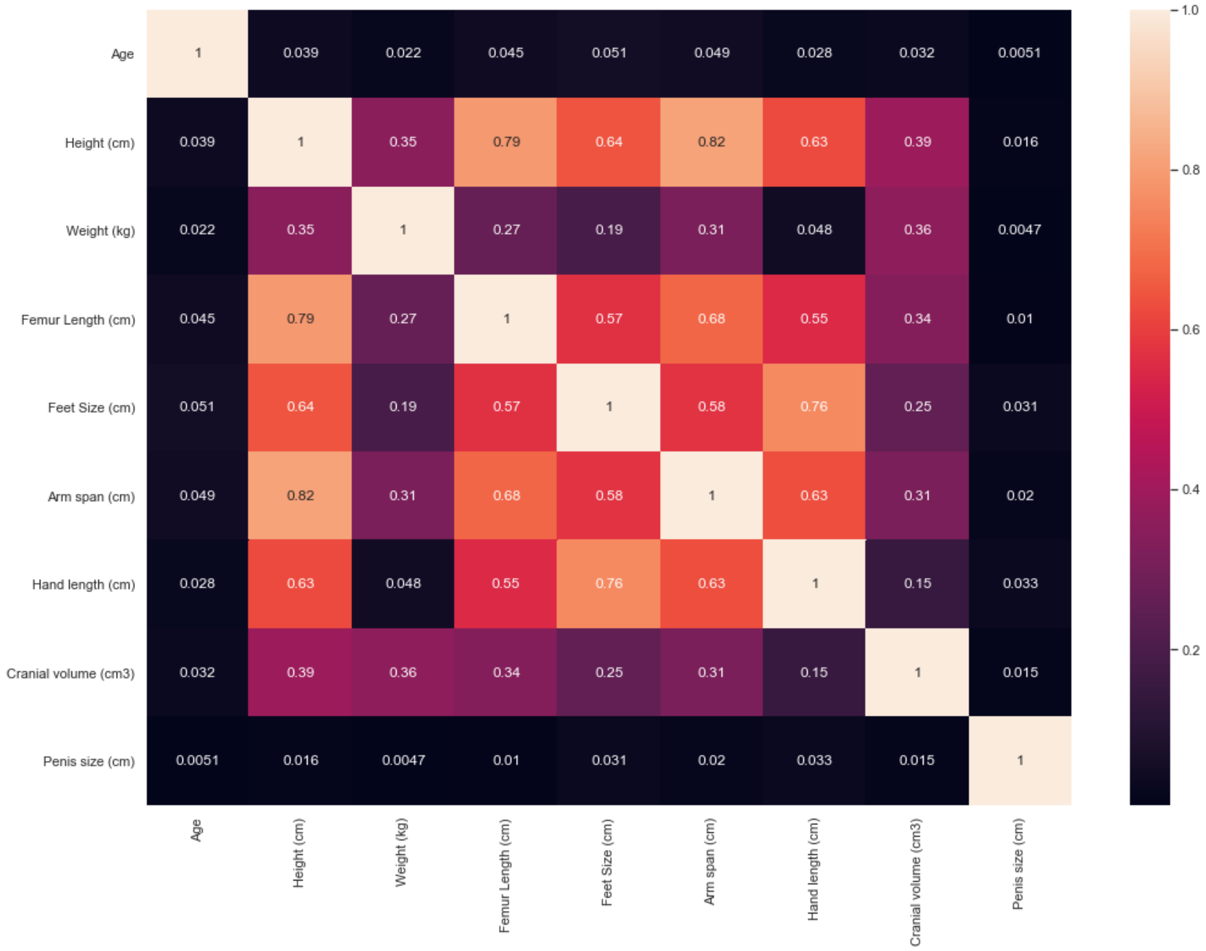


Figure 1: Determination heatmap

Using the cloud points and the heatmap, we can see that Age and Penis size don't have any correlations with the other variables. All other variables seem to have more or less strong linear connections.

With a correlation of 89%, the femur size is pertinent to predict the size of an individual in archeology. Note that the armspan has an even higher correlation, however femur bones are more likely to survive time than all bones from hand to hand. It is therefore a good indicator of human size in archeology.

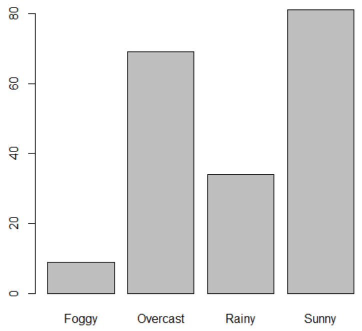
5)See course for the confidence intervals
Comments can be made on whether or not the intervals are large for both correlation and determination. Most interval are rather large, and centered around a low determination coefficient.

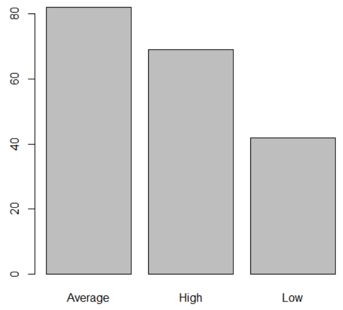
C Chi square on wheather data

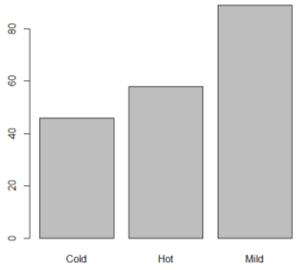
We have 3 categorial variables :

- Weather which can be foggy, overcast, rainy or sunny
- Humidity : average, low, or high
- Temperature : cold, hot, or mild

We can see that foggy and cold weather are less frequent in this dataset than the other cases. Especially in the case of foggy weather which is severely under-represented, this could leas to biases in the analysis.







Chi2 test

H0: The variables are independent

H1: The variables are not independent

Temperature/Weather

	Cold	Hot	Mild
Foggy	4	3	2
Overcast	19	14	36
Rainy	6	11	17
Sunny	17	30	34

For weather and temperature, we have : chi^{2}=8.4933 and p-value of 0.2041. The chi^{2} value cannot be directly interpreted. However, the p-value is way above 0.05, which means that we cannot reject H0. Here there is 20% chance that weather and temperature are independent. We can’t therefore say that there is a significant dependency between these 2 variables.

Weather/Humidity

	Average	High	Low
Foggy	3	6	0
Overcast	34	30	5
Rainy	10	24	0
Sunny	35	9	37

chi^{2}=68.4897 and p-value=8.34e-13

Here, we have a very low p-value, we have therefore a very high probability that there is a significant link between weather and humidity (99.999+% chance). If we look at the contingency table, we can see that obviously humidity is high when it is foggy or rainy. Therefore, this result makes sense.

Using Cramer coefficient, we find a 42% dependency strength between the 2 variables.

Temperature/Humidity

	Average	High	Low
Cold	20	22	4
Hot	24	15	19
Mild	38	32	19

Temperature and humidity : chi^2=10.3307 and p-value=0.03521

The independency hypothesis can be rejected with a 3.5% chance of error. The 2 variables are most likely linked. This can probably be explained by the low occurrence of cold days with a low temperature in the dataset. However, we should remember that both cold temperature and low humidity are under represented in the data, which may lead to a bias in the results.

Cramer and Chuprov coefficient return a strenght of 16.35% between the 2 variables, which is quite low. Therefore, we can conclude that while there is probably a link between humidity and temperature, the relationship between the 2 is weak.