PROJECT 4: LUNG CANCER ANALYSIS

Description of the project:

For this third project we were asked to work with the data of our liking. I chose to have an analysis on cancer and specifically on lung cancer across the globe.

The project itself:

Lung Cancer is now one of the deadliest types of cancer. Therefore, I focused my analysis on finding a match between tobacco consumption and lung cancer.

During my analysis I tried to answer, as precisely as I could with the data I had in my hands, to the following questions:

- Impact of Cigarettes on cancer?

- Share of Cancer deaths related to tobacco?

- Impact of Wealth on the survival rate?

I used different tables of data from the same source to achieve my analysis:
https://ourworldindata.org/cancer

What I wanted to achieve with this analysis:

With the different data sets that I we had I tried to answer to very basic questions:

- Does tobacco have an influence on lung cancer and where is it the most present?
- What are the regions of the world where the share of cancer deaths attributed to tobacco (%) is more significant?
- Does wealth influences survival rate?

The problems that I encountered during the process:

- To find the data it took me quite a long time.
- The data I have used to build my analysis was useful, but the quantity of information provided was not enough to use a probabilistic approach.
- I did not get the expected results.

The Workflow and Organisation:

I tried to look at health organisations to find data about cancer. Eventually, I got the link to the website I used to retrieve information. The datasets used were very precise, thus not very furnished with data, however they were still very interesting to use. I tried to clean several separate datasets and combine them to find correlations. This process was not always very successful. Regarding the questions that I asked myself to do the analysis I thought about using the most appealing topics related to lung cancer within the website

:https://ourworldindata.org/cancer.

Conclusions:

Access to relevant data is key to make a good analysis. Therefore, even though my analysis is in accordance with my initial data, the result that I got do not reflect completely the reality of the subject due to the fact of the data I have used. For instance, the correlation ratios do not give much evidence that the level of cigarettes sold influences lung cancer level. I believe it does not reflect reality as it should.
Although the analysis is not as precise as I hoped, the overall picture makes sense and give some insights about lung cancer.