Project Reflection

What obstacles did you meet when doing the project?

Before I started my project, I thought that Pandas was quite an inconsistently coded library, with many caveats and idiosyncrasies. However, I soon realized that open-source software in general is not that much more premium and flawless than the code that *I* write. For plotting the choropleth maps in question 1, I used the library Plotly instead of Folium, because I needed the maps to be animated and Folium did not have such a feature. Plotly's discrete choropleth maps, when animated, have some very interesting characteristics. It would take the first frame of the animation, see which categories are being plotted, and then use the *exact same* list of categories for all future frames. If a future frame had a category not present in the first frame, it would simply be discarded and the country would be colored gray, indicating no data. This would only change once one of the original categories was no longer present on the map, and a new category can be shifted into its place.

The only way I found to solve this problem (besides asking the developer on GitHub, to no avail) was to create placeholder data points such that all possible categories to be displayed were present on the first frame of the animation. Thankfully, Plotly's caveats worked in my favor this time, as when it encounters a country code that doesn't correspond to a country, it simply ignores it and carries on. This meant that I could put gibberish as the country code in the case of placeholders and it didn't matter. All that mattered was that all the categories were present in 1979-2016. In the end, my discrete choropleth plot is one of the most insightful plots for that question, as it showed a clear trend of countries' leading cause of death shifting from infectious to cardiovascular to neoplastic diseases. I think my efforts were quite worth it in the end.

How has this project changed your view of Data Analytics (or the Data Science Lifecycle)?

The only previous data analytics experience I had was during the year 1 Computational Thinking module, where I used Excel instead of a proper data analytics tool. Through this course, I really understood just how powerful Python is compared to Excel, especially when it comes to visualizing data. Because of the difficulty associated with plotting complex graphs in Excel, I always strived to use non-graphical methods of data visualization instead, such as correlation. This project made me realize the importance and insightfulness of graphical displays of data. For example, I only realized that Eastern Bloc countries had similar trends in terms of causes of death when I plotted my choropleth map. Before, I always thought that visualization was just for the lay people's sake, and that true data analysis was about numbers, but now I realize that perceiving something visually is crucial to developing a conclusion. We're after all humans, not machines that can analyze chunks of numbers on the fly.

How has this project help you gain insights on the topic you are investigating?

The choropleth maps made me realize the trend of leading cause of death throughout the years, which is probably the most interesting part of question 1. Question 2 was very surprising, as I always thought that when BMI increased, the number of people who die of heart-related diseases will also increase. However, the truth is that BMI can only be high when the level of development within a country is high, and in that case the healthcare infrastructure is probably also high, so in the end BMI and number of people who die of cardiovascular diseases is *negatively* correlated.

If given another chance, how would you approach the project differently?

I think I would have included some more indexes that correlate with a subset of the causes of deaths. For instance, the number of smokers per household which could correlate with respiratory illnesses. CMR and BMI are not enough to predict every cause of death, but I think it's a good start and if I had enough data, I could have built an index akin to Human Development Index, but for causes of death instead.

What content or skills have you learnt from the project?

In terms of skills that I could use in all aspects of life, probably perseverance. Data cleaning (especially when every organization uses a different naming standard for countries) is one of the most tedious things I've done for CS. Moreover, using external libraries also proved to be quite a challenge, and Jupyter notebook was also not cooperative at times. But in the end, I managed to produce some visualizations and results that I am proud of, so I would call it a win.

In terms of technical skill, this project really forced me to master Pandas, which could prove to be a really useful skill down the line when I need to do my ARP. Before the project, I only used variations of the examples in the notes, which restricted the number of ways there was to do something. However, afterwards I became almost as proficient in pandas as I am in Python. Proficiency in my opinion means that you develop an intuition on how to do something without needing to read documentation or think too deeply about it. That's something that only comes with hands-on experience. I wish I started my project earlier for this reason, as I could have used the experience in my tests and would have probably gotten a higher score.