DS 2002: Reflection

Upon working on this final project with our team, we found it very interesting, insightful, and informative. Although we had a fair share of challenges, our team was to work together to share resources to overcome them. Beginning with data selection and exploration, we had difficulty narrowing down which dataset to use. We found a plethora of different datasets that could've been used on Kaggle; however, there was a lot of confusion about how they could be intersected to make meaningful insights. Ultimately, we settled on global data trends because merging the datasets based on their country would be relatively straightforward. Particularly, the use of ISO codes (USA, AFG, etc.) made it incredibly easy to merge our datasets and remove all non-country observations from the data.

From there, our team had to tackle the challenge of how to clean up the dataset to make it easier to merge. One of the main issues we encountered was we were unsure of what components of the data to keep, and which to discard. Our original datasets not only had data about GDP and Life expectancy, but they also included energy output and consumption. Ultimately, we decided as a team that it would still be beneficial to include this data, at least for the analysis portion of the project, to see if there are other correlations we may find. From there, we would see which data variables correlated with one another, and which ones didn't. The ones that did not have any merit in our data would then be discarded in a "trimmed" dataset.

After that, we attempted to extract meaningful insights and analysis from the data. A big issue with this step was that many of the variables had incomplete data, with many countries only tracking the explanatory variable in the last 20 years—or never at all. To overcome this, we decided to analyze only the data post-1950, despite the fact that some of our data went back several hundred years. This allowed for more consistent data visualizations that allowed for greater insights.

Furthermore, our team did not have too much trouble with converting our datasets to an SQL database. The bulk of our SQL code was used from our First Data Project, and upon conversion, no major bugs were found in our tests.

Lastly, a big challenge we faced was adding our database metrics to Google Cloud. The main issue we faced was that we had to figure out how to download the SQL database as a CSV and then upload that CSV to the cloud. Unfortunately, the tutorial lab we did for Google Cloud was not as helpful as we thought, as we had several errors with our "Schema" mismatching our data values and columns. After a lot of tweaking by trial and error, we were able to correct the data mismatch, and have the values stored in Google Cloud.

An important lesson we learned was project management. We ensured everyone was on the same page about the direction of the project through check-in Zoom meetings and continuous communication. To maximize efficiency we divided the work into sections—data analysis, clean-up/storage, and presentation—allowing each team member to focus on their specific tasks without overlapping efforts. However, we still had everyone review sections they hadn't worked on to stay updated on the project's progress and provide feedback. This streamlined process allowed for the seamless production of our final product while maintaining quality and accountability. We not only gained technical skills, learning how to successfully create an ETL pipeline and analyze our transformed data, but also soft skills having to tackle a major project as a group. If we were to expand upon this project in the future, we could try to see other avenues we could use to visualize the data, perhaps utilizing an API or Excel dashboard to selectively comb through our visualizations easier, while maintaining real-time dynamic data.

In conclusion, this project served as a valuable learning experience that enhanced both our data programming and collaborative abilities. Through overcoming our several challenges with data selection, cleaning, analysis, and cloud integration, we managed to deepen our understanding in developing pipelines and the more complex considerations and complications that follow it. Our ability to be adaptable and meet several times as a team to work cohesively played a crucial role in developing our pipeline. These experiences have equipped us with a stronger foundation for tackling future projects and collaborating effectively in professional environments.