Final Project Reflection

Throughout this project, we faced several challenges that posed setbacks to our progress. However, these obstacles ultimately became opportunities to push us to develop a deeper understanding of the data and its analysis. The data selection process proved to be the most challenging aspect of the project, particularly in finding reliable datasets organized by county. After identifying datasets with information on education levels, unemployment rates, and median income, we encountered inconsistencies in the number of counties across the two sources, which required additional cleaning to ensure alignment. Also, we faced the challenge of managing multiple columns with overlapping information. For example, the education dataset included both the total number of people with a high school diploma and the percentage of the population. To streamline our analysis, we focused on percentages, as they offer a more proportional comparison across states. Another issue was the inconsistency in timeframes – the education data provided averages for 2018-2022, while the employment data was specific to 2021, which complicates direct comparisons. Missing data, particularly significant gaps in Puerto Rico's 2020 data, also posed a challenge. Next, the ETL process was relatively easy for us, but handling large datasets with a few thousand rows made initial data loading and processing time-consuming. Additionally, some records had missing NaN values, which we realized might skew the analysis. This came in the form of rows that only contained the state names with no valid information relevant to our data analysis. Utilizing the pandas library, we were able to conduct thorough modifications and cleaning of the dataset in preparations for storage where we used tools to convert to JSON and to drop NAs. In terms of visualization, selecting the most effective representation of our insights was tricky due to the abundance of countries. However, we solved this by looking at state-level analysis which presented a more comprehensive visual of our

findings. Lastly, cloud storage presented the issues of column name inconsistencies – many were large, complicated, and had spaces that made it difficult to execute SQL queries in Google Cloud. So, we had to clean these names further to adhere to the Big Query studio format by replacing certain characters within the column names. Additionally, navigating Google Cloud and adding API permissions was difficult with multiple users as owners. Only the original owner of the Google Cloud project was able to edit these permissions, which then allowed us to create python notebooks within the project.

Overall, we were satisfied with our group collaboration throughout the project. All parts were completed on time, and we felt like everyone contributed fairly. At the start, assigning roles was a bit challenging, especially since we were trying to coordinate over text. This led to some confusion about who would take on what role and responsibilities. However, once we met in person during class to work on the project, we were able to align our goals better, set a clear timeline, and determine what we wanted to accomplish by the end of Thanksgiving break. From this experience, we learned the importance of meeting face-to-face, especially at the beginning to ensure everyone is on the same page before diving into individual goals.

After completing this project, we gained valuable skills in data cleaning, analysis, and cloud storage management. We developed a deeper understanding of how to handle large datasets, address missing data, and align data to solve inconsistencies. Some potential improvements include using more advanced visualization tools like Tableau paired with Google Cloud to create more interactive visualizations for the users. Furthermore, storing the data in MongoDB could allow for more scalability as the datasets are constantly updated with the most current year's data where the schemas are flexible in nature as opposed to the MySQL database that was used. Additionally, we improved our ability to select appropriate visualizations that

effectively communicate insights, especially when working with complex data sets. These visualizations can also enable us to conduct statistical analysis on potential independent variables that have a relationship with the different rates that we see throughout the different datasets.