DS 2002 Final Project Reflection

Our final project for DS 2002 focused on analyzing the relationship between COVID-19 death rates and air quality using median AQI across various U.S. counties in 2020. We constructed an ETL pipeline that ingested raw datasets, transformed them to extracted insights, visualized trends using plots to understand potential health and environmental factors, and stored the transformed data in Google Cloud. Key features for data ingestion and transformation included merging AQI, COVID-19 deaths, and population datasets based on ‘State’ and ‘County’ columns and calculating metrics such as "Deaths per 100,000 people.” The analysis consisted of creating visualizations like scatter plots, box plots, and bar graphs, to explore the correlation between the variables.

Throughout the project, our group faced numerous challenges during data selection, ETL setup and implementation, analysis, and cloud storage that we had to overcome. One of the primary challenges encountered at the start during data selection was finding datasets that were relevant to our topic. While there were several datasets available on platforms like kaggle, many were outdated or incomplete. Ultimately, we were able to find three datasets that were relevant to our chosen topic, which included county and state-level information, enabling us to merge the data. The ETL setup and implementation also posed technical issues. During the transformation phase of the project, we had to meticulously clean the data by handling missing values, normalizing data types, and ensuring data integrity. Also, one major challenge was addressing inconsistencies in naming conventions and formatting. For example, the COVID-19 dataset included the word ‘county’ in the county names, while the AQI dataset did not. This mismatch initially caused issues in the merging process resulting in a combined dataset of 10 rows. However, after fixing this problem and standardizing the formatting of ‘County’ names, the

combined dataset was able to have 646 rows, ensuring compatibility and completeness for analysis. During analysis, there were some minor challenges in generating meaningful insights from the data. We were able to address outliers by creating boxplot visualizations. However, analyzing the relationship between COVID-19 death rate and median AQI ended up revealing no strong linear correlation, which was confirmed by the Pearson correlation coefficient. Our further analysis indicated that additional factors influencing the death rate were not considered, suggesting the need for a more targeted exploration in the future.

This project taught us many valuable lessons, including overcoming technical challenges, fostering efficient team coordination, and identifying ways to improve future projects. For instance, technical challenges could be prevented in future projects by ensuring the data sets are cleaned properly and inconsistencies in naming conventions and formatting are addressed beforehand. In addition, effective communication was essential for team coordination, especially in addressing challenges we encountered throughout the project. Further, using Google Colab to share and write the ETL code allowed for smooth collaboration and improved productivity. Lastly, a future improvement would be to allocate more time to dataset cleaning and transformation so inconsistencies could be caught early, minimizing issues during analysis.

In addition to lessons, we also gained skills that will aid us in future projects. Some of these are knowledge in designing and implementing ETL pipelines using python and advanced skills in data cleaning, transformation, and visualization using libraries such as pandas and matplotlib. Another new skill that we implemented in the project was learning how to store transformed data in Google Cloud and document the cloud storage setup. During data cleaning, we gained expertise in resolving dataset inconsistencies, such as standardizing county names for successful dataset merging. Apart from technical skills, some soft skills we acquired were

efficient team coordination, communication, and conflict resolution along with enhanced problem-solving abilities by combatting unforeseen challenges during dataset transformation and analysis. Areas for future development could include analyzing additional factors apart from air quality which would have affected the COVID-19 death rate such as age, demographic, etc. Overall, this project not only provided us with valuable lessons and skills for data ingestion, analysis, and storage, but also enhanced our ability to tackle future projects with efficiency, accuracy, and collaborative teamwork.