**Individual Presentation Notes**

Phases of COVID-19 Model

As a group, it was critical to implement an overall project outline to ensure all deliverables of the culminating challenge were met. This included recognizing all the essential components of a successful machine learning model, such as exploratory analysis, data cleaning, feature engineering and so forth. It is critical to note that data cleaning and feature engineering are some of the highest contributors to a successful model, holding almost half its importance. As a result, the team noted this and ensured these components were adequately met. The first stage was geared toward data exploration, while the second focused on data analysis and model building. During both phases, data pre-preprocessing took place, although most of the heavy work took effect during our database development. Additionally, visual analytics were utilized during both parts to view overall data characteristics and patterns.

Data Exploration

Previous to building the database, it was necessary to employ a brief data exploration assessment in Pandas. Exploration allows for a deeper understanding of the dataset, making it significantly easier to navigate and utilize during database development and data analysis. This included determining overall column features and their application to the project scope, discovering mismatched or null data values, and statistical analysis. Transformation and concatenation of data was employed to ensure feature value (FIPS), a unique numerical code for each county, was present in all datasets for future merging.  Additional cleaning of the FIPS column was needed to certify that it only included US county data and not those outside the country, such as Puerto Rico, or any cities. Additionally, in both Tableau and Jupyter notebook, visuals were created to have a different view of the data feature characteristics at the county level. These evaluations aided in determining if the dataset was sufficient to be used further in the project. If the dataset or certain features were found to be insufficient, it was necessary to drop them from further research. After utilizing pandas and tableau for data exploration, further pre-processing took place in Postgres to merge and prepare data for analysis.

Recommendations for Future Analysis

Concerning future analysis, it would be beneficial to integrate a time series analysis to view quarterly, monthly, and daily differences. This can provide further insight if any features have significant changes in their importance depending on the time frame. Additionally, adding weather data in conjunction with the time series will be beneficial in evaluating weather's contribution to high case and death rates. Some experts have expressed that the COVID virus exhibits lower activity in hotter climates in the summer while climbing during winter. It would be fascinating to test this theory. Another possible avenue would be the inclusion of vaccination data. This can shed light if the county's complete, partial, or lack of vaccination significantly affects high death and case rates. Proceeding even further, it may be possible to view if a particular vaccine (Moderna, Pfizer, or J&J) was more prevalent in a specific county and if one was more effective than its counterparts based on death and case rates. Lastly, it would be interesting to implement the model for case and death rates during March 2021-2022 to view its performance.

Looking Back

Although the team is proud of what we have accomplished during our allotted project time frame, there are some areas that we would've executed differently. The first pertains to the primary steps of our project and establishing organization. It may have been beneficial to create a brief Gannt chart to delegate responsibilities, including major deliverables and their necessary subtasks. This chart will be helpful in seeing progress and viewing tasks ahead, behind, or on schedule. If a particular subtask is behind schedule, group members can quickly provide assistance.

 In terms of the machine learning model, it is best to start simple and then increase complexity each week. Initially, the group started with a neural network model, which then led to backtracking. Lastly, it would've been helpful to implement a thorough quality assessment of the data and their specific impact on the ML model earlier in the project. About 3/4 of the way into the project, it was discovered that one of our datasets was the contributor to Machine learning errors. This was discovered after implementing various ML models. Although a difficult lesson to learn, it is vital to know that a dataset can initially look promising but negatively impact performance. If the error had been found earlier due to a quality assessment, this would've negated unnecessary confusion and frustration during the machine learning building process