Data Quality Plan

* Mark down all the features where there are potential problems or data quality issues.
* Propose solutions to deal with the problems identified. Explain why you choose one solution over potentially many other. It is very important to provide justification for your thinking in this part and to list potential solutions, including the solution that will be implemented to clean the data.
* Apply your solutions to obtain a new CSV file where the identified data quality issues were addressed.
* Save the new CSV file with a self-explanatory name.
* Save the data quality plan to a single PDF file.

# Overview

The aim of the data quality report is to outline the condition of the raw data and how the data needed to be cleaned for in order for information to be extrapolated. The report will summarise the data, discuss where the issues with the data lie, detail how the data was cleaned and present the results of the data cleaning in the forms of graphs and tables. Additional information in regard to the terminology of the dataset and the additional information on the dataset can be found in the appendix.

On first look at the dataset, it was quite apparent that the data would not be easy to work with and would need a decent amount of work to clean it up. There were missing values scattered throughout the data. Due to the missing information, rows could appear as possible duplicates, due to the lack of identifying data and the missing data. The most difficult part of the dataset was the lack on numerical data in the dataset. Additionally, some columns had outlier information in regard to their dates, however, these dates appear at the end of the dataset and could be attributed to a lack of updated data.

Upon investigation it was determined that the dataset contained no duplicate columns, or columns with irregular cardinalities. A number of logic tests were carried out on the datasets which did not show any inaccuracies or inconsistencies in the data.

# Summary

Before any cleaning of the data the dataset had a size of 10,000 rows and 12 columns of data. The column names were *cdc\_case\_earliest\_dt, cdc\_report\_dt, pos\_spec\_dt, onset\_dt, current\_status, sex, age\_group, race\_ethnicity\_combined, hosp\_yn, icu\_yn, death\_yn,* and *medcond\_yn.* Descriptions of the columns can be found in the appendix.

Of the 12 columns, four had date/time data and the remaining eight were text data. Initially all the data was imported into the dataframe as objects. Hence, the four columns of dates were transferred into dattime64[ns] data types and the eight text-based data were recategorized as categorical data. Initially it was difficult to determine if the

# Review Logical Integrity

# Review Continuous Features

## Descriptive Statistics

## Histograms

## Box Plots

# Review Categorical Features

## Descriptive Statistics

## Bar Plots

# Actions to Take

# References

# Appendix

## Terminology & Assumptions