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 showed inaccuracies with the *cdc\_case\_earliest\_dt* column.

# 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 data frame as objects. Hence, the four columns of dates were transformed into datetime64[ns] data types and the eight text-based data were recategorized as categorical data. Initially it was difficult to determine if the dates should be categorical or continuous, however, as the dates were continual time ranges, it was determined that the date columns should be treated as continuous data.

Once the data had been categorised into the relevant date types the data was checked for duplications, percentage missing, cardinality and constant columns. As discussed in the overview, no constant columns, duplicate columns or cardinality irregularities were discovered.

There were 969 duplicate rows in the table, which was reduced down to 540, when the first instance of the row was skipped. Upon looking at the data being presented the decision to keep the duplicate entries was made. The decision was made due to the type of data being discussed. The dataset had been anonymised before issue of the data and as such any identifying data had been removed. The personal data left in the dataset was very ambiguous and could very easily apply to more than one person as is the point of anonymising the data.

Next the data was checked for the missing value rate of the information. As the database owners fill in ‘Missing’ for some columns where information is missing the dataset needed to be reformatted so that the missing values were uniform. All cells which contained ‘Missing’ were replaced with ‘NaN’ and the missing dates were replaced with ‘NaT’. This meant the dataset could easily be searched to determine the number of missing values in the data set for each column.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Num of miss | total rows | %miss |
| cdc\_case\_earliest\_dt | 0 | 10000 | 0 |
| cdc\_report\_dt | 2311 | 10000 | 23.11 |
| pos\_spec\_dt | 7195 | 10000 | 71.95 |
| onset\_dt | 4966 | 10000 | 49.66 |
| current\_status | 0 | 10000 | 0 |
| sex | 10 | 10000 | 0.1 |
| age\_group | 15 | 10000 | 0.15 |
| race\_ethnicity\_combined | 89 | 10000 | 0.89 |
| hosp\_yn | 2366 | 10000 | 23.66 |
| icu\_yn | 7649 | 10000 | 76.49 |
| death\_yn | 0 | 10000 | 0 |
| medcond\_yn | 7448 | 10000 | 74.48 |

As indicated in the table above, the miss rate of *pos\_spec\_dt, icu\_yn,* and *medcond\_yn* was over 70%, and as such they were dropped from the table. The data could have been kept, however, with the level of missing data, any attempt at filling in the data could have skewed the results of a data analysis performed on the cleaned data at a later date.

Additionally, *onset\_dt* was looked at for potentially removing from the dataset, however as the data had an approximate 50% miss rate, it was decided to keep the column and perform some logical tests on it to see if it was worth keeping. It was determined that the *onset\_dt* could be beneficial for determining spikes in the transmission of the virus.

After the columns had been dropped a check was performed on the remaining data to determine how many rows contained data in all of the columns. Unfortunately, it was determined that of the 10000 rows of data, 54.46% of the rows were missing data from one or more column. At risk of losing over half of the data set, the discission was made to keep all the rows and work with the NaN values where they arose. As the NaN values occur in the categorical and continuous data, it was difficult to mask them, as such they were left as NaN due to the type of information we were dealing with. The NaN values can be useful with this data set in determining what type of data a person is willing to give.

Imputation, i.e. the mapping of the mean to a missing continuous number could not work in this instance as the data was in date format and could throw of analysis or cause inconsistencies in the data.

# Review Logical Integrity

A total of eight tests were performed on the data to determine if the data was logical. Most of the tests focused on the deaths of individuals. Originally more tests were scheduled, however due to the decision to drop columns due to the high levels, some tests couldn’t be run.

From the results there are inconsistencies with the cdc\_case\_earliest\_dt being the earliest date in the data set. As this is more of an indicator data, it was left in as one of the most complete columns in the data set. Additionaly, it is noted that there are quite a few instances of the cdc\_report\_dt being before the onset\_dt. This may be an inconsistency or additional data could show that the patient was a close contact and so was tested before the onset of symptoms. As such the data was kept.

Test 1a: Test if *cdc\_case\_earliest\_dt* contains dates before or equal to the *cdc\_report\_dt.*

26 instances were detected as being set later than the *cdc\_report\_dt*

Test 1b: Of the 26 instances check if *cdc\_case\_earliest\_dt* contains dates later than *onset\_dt*.

0 instances of *cdc\_case\_earliest\_dt* containing a date later than *onset\_dt.*

Test 1c: Of the 26 instances check that cdc\_report does not have a date later than *onset\_dt*

O instances of *cdc\_report\_dt* containing a date later than *onset\_dt*

Test 1d: tests if cdc\_case\_earliest\_dt occurs after onset\_dt

1 occurrence of this in the dataset

Test 1e: of the 1 instance check if cdc\_report\_dt is later than onset\_dt

1 occurrence of this failing

Test 1f: test if onset\_dt occurs always occurs before cdc\_report\_dt

3500 instances of this failing

Test 1g: Test if cdc\_report\_dt occurs before onset\_dt

39 instances of this failing

Test 2a: Determine if the deaths per age\_group and sex conform to the accepted rates.

A picture containing text, receipt

Description automatically generated

The counts detail the trend that older people are more susceptible to the virus than younger people

# Review Continuous Features

## Descriptive Statistics

There are three continuous features in the data set. All three continuous features are date information and as such can be difficult to work with as continuous data. In the figure below we can see the description of the continuous data. Note that the time part of the mean date is used to signify a fraction of a day.

Table

Description automatically generated

## Histograms

A set of histograms for the data is available in the appendix.

The histograms indicate a right-skewed dataset whereby the number of cases being reported per day are increasing every day. It is interesting to see that although the number of reports is still high in January 2021, there is a reduction in the number of people experiencing the onset of symptoms. Unfortunately, as onset\_dt had a 50% miss rate, this analysis might be out slightly considering onset tracks with cdc\_reports\_dt quite well for the previous bins.

## Box Plots

Not completed

# Review Categorical Features

## Descriptive Statistics

## Bar Plots

# Actions to Take

# References

# Appendix

## Terminology & Assumptions