Data Quality Report

1. Overview

This report will discuss the initial findings based on the cleaned dataset (covid19-cdc-20204844_1-1_cleaned.csv). It will summarise the data, review any issues observed and layout the actions to be taken. The appendix contains the tables of descriptive statistics for both the continuous and categorical features.

2. Summary

A few tests were carried out to check the logical integrity of the data. This highlighted a small number of failures within the dataset. Overall, 52 instances of irrational data were observed.

For both the continuous and categorical features there are a significant number of missing values. For many of the categorical features there are also a significant number of entries with the value 'Unknown' which makes analysis ineffective.

There main problem in this dataset is that there are a significant number of missing values for features in this data set.

3. Review Logical Integrity

There were 4 tests were carried out. The failures are below:

- Test 1 – Check that *cdc_case_earliest_dt* < *onset_dt*
    - 1 case found

- Test 2 – Check if any entries were admitted to ICU but not Hospitalised
    - 0 cases found

- Test 3 – Check if any entries were probable cases that resulted in death
    - 26 cases found

- Test 4 – Check if any entries showed symptoms more than 14 days after *cdc_case_earliest_dt*
    - 25 cases found

4. Review Continuous Features

4.1 Descriptive Statistics

There are 4 continuous features which are datetime features.

- cdc_case_earliest_dt
    - There are no null values for this feature, and it has a range of 381 days.

- cdc_report_dt
    - There are 17.9% of the values missing for this feature. According to the CDC (2021) this feature is now deprecated, and they recommend researchers to use cdc_case_earliest_dt instead.

- pos_spec_dt
    - There is 70% of the values missing for this feature. This feature represents the 'Date of first positive specimen collection'.

- onset_dt
    - Nearly half of the values are missing for this feature however, this represents 'symptom onset date, *if symptomatic*'. As this feature is conditional, I do not believe the missing values are of concern.

Most of these features have missing values however in future analysis I would only be concerned in using *cdc_case_earliest_dt* and *onset_dt*.

4.2 Histograms

The histograms are included in the accompanying pdf. From these histograms we can see that a majority of the cases occurred in the final months of 2020.

4.3 Box Plots

As the only continuous features in this dataset are time series, I do not feel it is necessary to create box plots for these features as I am not concerned about the central tendency of these features

5. Review Categorical Features

5.1 Descriptive Statistics

There are 8 categorical features in this dataset. One of these, *death_yn*, is the target. Four of these features are Yes/No variables.

- current_status
  - 7.4% of the entries are probable cases. It may be more accurate to build a model using the 'Laboratory-confirmed case' entries and therefore remove the 7.4% of entries that are probable cases.

- sex
  - There is a slight majority of female entries however the split between Male and Female is relatively even. There is <1% of entries with the value 'Unknown' and 0.1% of entries are missing values.

- age_group
  - The greatest number of cases are between the ages of 20-29 with the majority being spread between 20-60 years. The are 0.1% of entries missing a value for age_group.

- race_ethnicity_combined
  - There is a significant number of entries with 'Unknown' value (35%). The most common value for this feature is 'White, Non-Hispanic' (36.6%).

- hosp_yn
  - Over half of the entries have the value 'No' for this feature. 21% are 'Unknown' and 16% are missing.

- icu_yn
  - Nearly three quarters of the data is missing for this feature. Only 11% of the entries are either Yes/No.

- medcond_yn
  - Over 70% of the values are missing for this feature

5.2 Histograms

The histograms can be found in the accompanying pdf.

6. Action to Take

The following 6 actions will be taken:

For Logical Integrity Test 1 replace *cdc_case_earliest_dt* with *cdc_report_dt.*

For Logical Integrity Test 4 drop rows where time from *cdc_case_earliest_dt* to *onset_dt* was greater than 14 days (covid-19 incubation period).

Drop the features *cdc_report_dt* and *pos_spec_dt* as they have a high number of missing values and I believe only *cdc_case_earliest_dt* and *onset_dt* will be needed in future analysis.

Drop rows that are 'Probable case' as any future modelling should be done on data from confirmed cases.

Drop entries where *age_group* is missing.

For *sex*, *race_ethnicity_combined*, *hosp_yn*, *icu_yn*, and *medcond_yn* replace missing values with 'Unknown'.

7. References

CDC, 2021. Covid-19 Case Surveillance Public Use Data.

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

8. Appendix

8.1 Continuous Features

Table 1. Continuous Features Descriptive Statistics

	count	mean	min	25%	50%	75%	max
cdc_case_earliest_dt	9223	30/09/2020	01/01/2020	22/07/2020	01/11/2020	11/12/2020	16/01/2021
cdc_report_dt	7573	15/10/2020	01/01/2020	14/08/2020	10/11/2020	20/12/2020	29/01/2021
pos_spec_dt	2749	19/09/2020	12/03/2020	06/07/2020	20/10/2020	04/12/2020	26/01/2021
onset_dt	4995	20/09/2020	01/01/2020	14/07/2020	18/10/2020	01/12/2020	29/01/2021

8.2 Categorical Features

Table 2. Categorical Features Descriptive Statistics

	count	unique	top	freq
current_status	9223	2	Laboratory-confirmed case	8540
sex	9223	4	Female	4754
age_group	9223	10	20 - 29 Years	1705
race_ethnicity_combined	9223	9	White, Non-Hispanic	3379
hosp_yn	9223	4	No	5166
icu_yn	9223	4	Missing	6872
death_yn	9223	2	No	8897
medcond_yn	9223	4	Missing	6694

8.3 Bar Plots & Histograms

The plots can be found in the accompanying pdf files.