Table of Contents

Data Quality Report – Initial Findings

1. Overview

This report will outline the initial findings based on the cleaned dataset (covid19-cdc-20200225.csv). It will summarise data, describe various data quality issues observed and how they will be addressed. Please see appendix for some background to this dataset. Appendix includes terminology, assumptions, explanations and summary of changes made to the original dataset. This also includes feature summaries, histograms and boxplots used to visualise the data.

On first indication the dataset appears relatively unclean. There are no duplicate columns. However, there were a large number of rows with no data. The main issues observed were there are significant number of columns that have a high percentage of missing values. 4 out of 11 columns have over 50% to 70% missing values. Also, several logical tests were carried out on the data a significant number of inconsistencies were found.

2. Summary

Several tests were carried out to check the logical integrity of the data. This brought about a significant number of failures of the data. In total 167 instances of irrational data was observed. For example, cdc_report_dt and pos_spec_dt appears to be earlier than cdc_case_earliest_dt, this is logically impossible. This irrational data will need to be dealt with and should be checked with the domain expert. See logical integrity section for further details.

In this dataset, there is one continuous features which is cdc_case_earliest_dt.

For the categorical values several changes are recommended. There are 11 main features including a wide variety of information about prediction of death risk of COVID-19 on data provide by CDC such as the earlier of the Clinical Date, Initial case report date to CDC, Date of first positive specimen collection, symptom onset date, case status, sex, age group, race ethnicity combined, Hospitalization status, ICU admission status, death status and Presence of underlying comorbidity or disease.

For columns of case status, sex, age group, race ethnicity combined, hospitalization status, ICU admission status, death status and presence of underlying comorbidity or disease, among all those have significant unknown value and missing value, which need to be replaced with unknown value as they delivered similar meaning. Except for sex column, missing value will be dropped as only a few affected rows. Columns namely sex, age group, race ethnicity combined, hospitalization status, ICU admission status, presence of underlying comorbidity or disease have 15, 13, 88, 2078, 7157, 7011 values with “Missing” category, which will be converted to “unknown” as missing data for data comparison and interpretation.

There was a significant number of missing values of across the feature set. However, on first induction these values appear to be plausible but should be investigated further.

3. Review Logical Integrity

tests were carried out. The failures are below;

- Test 1 – Check if any date of the earlier of the Clinical Date is later than date Initial case report date to CDC
    - 36 case found
- Test 2 – Check if any date of the earlier of the Clinical Date is later than date of first positive specimen collection
    - 48 cases found
- Test 3 – Check if any date of If symptomatic, onset date is later than date of first positive specimen collection
    - 83 case found

4. Review Categorical Feature

There are 7 categorical features in the dataset, 1 of which is the target and will not be evaluated here. The 6 remaining are sex, age group, race ethnicity combined, hospitalization status, ICU admission status and presence of underlying comorbidity or disease. The features will be summarised below;

4.1 Descriptive Statistics

- Current Status
    - Laboratory-confirmed case accounted for 93% of majority of case, which account for 8858 while probable case accounts for 7%.

- Sex
    - The majority of sex is female, which are roughly 700 cases as high as male. This indicate that female is more vulnerable to COVID-19.
    - Feature has 66 case is in the "unknown" category. Thus further investigation should be made and the issue will need to be addressed.
    - Feature has 0.1 % missing value as the affected rows are small, they will be dropped.

- Age group
    - Most of the age group from 10 to 69 years account for the majority cases, which is over 1000 cases across all the age groups, we can see an evenly distribution among age groups between 10 to 69 years. Among the groups, between 20 – 29 years dominated the largest cases, which is about 18%. The second and third goes to 30 – 39 years category and 40 – 49 years category. The fourth and fifth goes to 50 – 59 years and 60 – 69 years.

    - The second pattern found among the groups with cases reported around 500. It goes to categories of over 70 years and below 10 years.

    - Missing value of 0.8% will be merged with unknown value.

- Race ethnicity combined
    - Race ethnicity combined is dominated by "Unknown" category, which account for 3653 cases. Unknown is synonymous with missing data. There is a possibility that reported client is multiracial that the designed options cannot fit in or the data is simply missing. Missing value will be merged with unknown value

- While “White, Non-Hispanic” accounts for the second largest category, which is 3290 cases. The remaining categories accounted for less than 1000 cases. Hispanic/Latino accounted for 975 cases, Black/ non-Hispanic lasted for 698 cases. Multiple/ other accounted for 490 cases. Asian/ non-Hispanic accounted for 229 cases.

- Two of the other columns have less than 100 cases are “American Indian/Alaska Native, Non-Hispanic” and “Native Hawaiian/Other Pacific Islander, Non-Hispanic”
-

- Status of Hospitalized
  - Status of Hospitalized is dominated by “NO”, accounts for 5220 cases while “Yes” accounts for 669.
  - “Unknown” goes to third largest category. Further actions like merging cardinality are needed to handle the problem.

- ICU admission
  - While “Unknown” category accounts for 1297 cases, which is the largest. This column has to be addressed later like whether to merge cardinalities or not.

  - “No” accounts for 968 cases while “Yes” accounts for 82 cases.

- Status of death
  - The largest category goes to “No”, account for 9165 cases while “Yes” account for 339 cases
- Presence of underlying comorbidity or disease
  - It appears the majority cases reported “no” to presence of underlying comorbidity or disease, around 38%.
  - Unknown cases are around 720. Unknown is synonymous with missing data Missing value will be merged with unknown value

4.2 Histograms

All bar plots can be found on the appendix and in the accompanying notebook. The analysis of each bar plot will be discussed in this section:

- Current Status
  - Current Status is dominated by “Laboratory-confirmed” case, indicating that majority of reported case is laboratory-confirmed case.
  - The amount of probable case is hugely smaller than “Laboratory-confirmed” case

- Sex
  - Sex is dominated by female, indicating that the majority of reported case is female.
  - Unknown category account for the least for those did not fall into the group of female and male.
- Age group

- Age group is dominated by 20 – 29 years, which account for largest number. The second follows by 30 -39 years, the number decreases gradually across age groups.
- With three categories slightly less than others, 70 – 79 years and 80+ years and 0 – 9 years have around 500 cases, which is half less than the other.

- Race ethnicity combined
  - Unknown case account for the largest category.
  - White, non-Hispanic accounts for the second largest in race ethnicity combined category.
  - The number of other race ethnicity is hugely smaller. About one-third of that of white category.

- Status of Hospitalized
  - The majority of cases reported that they didn't hospitalized while only a slight amount of cases reported they did.
  - Missing issue needs to addressed like replacing to unknown value.

- ICU admission
  - Missing issue needs to addressed like replacing to unknown value
  - A Large majority case reported that they didn't admit to ICU while only a slight number of case reported they did.
- Status of death
  - A large (around 90%) majority case reported no to the status of death while only a slight number of cases reported yes.

- Presence of underlying comorbidity or disease
  - Missing issue needs to addressed like replacing to unknown value
  - Presence of underlying comorbidity or disease category is dominated by No while those case reported to Yes is slightly less than those reported No.

5. Action to take

Main actions will be taken summarised below;

- the earlier of the Clinical Date
  - date of first positive specimen collection and date Initial case report date to CDC earlier than the earlier of the Clinical Date is to be replaced
- "Missing" value found in column age group, status of hospitalized, ICU admission and medical condition
  - To be replaced with Unknown value
- Missing value found column – sex
  - Drop the affected row
- Column - date Initial case report, first positive specimen collection
  - Drop after replacement of the earlier of the Clinical Date

Reference
[1] https://www.nature.com/articles/d41586-020-02972-4

[2] https://covid.cdc.gov/covid-data-tracker/#datatracker-home

8. Appendix

8.1. Terminology

- First positive specimen collection
  Specimen collection is the process of obtaining tissue or fluids for laboratory analysis or near-patient testing.

- Symptom onset date
  When the virus does cause symptoms, common ones include fever, body ache, dry cough, fatigue, chills, headache, sore throat, loss of appetite, and loss of smell.

- Probable Case
  A possible case is usually a case with the clinical criteria as described in the case definition without epidemiological or laboratory evidence of the disease in question.

- Laboratory-confirmed Case
  Laboratory criteria include a list of laboratory methods that are used to confirm a case. Usually only one of the listed tests will be enough to confirm the case.

8.3. Categorical Features

Descriptive Statistics

	count	unique	top	freq
current_status	9504	2	Laboratory-confirmed case	8858
sex	9504	4	Female	5046
age_group	9504	10	20 - 29 Years	1778
race_ethnicity_combined	9504	9	Unknown	3653
hosp_yn	9504	5	No	5220
icu_yn	9504	4	Missing	7157
death_yn	9504	2	No	9165
medcond_yn	9504	4	Missing	7011

Box Plots