Data Quality Report - Initial Findings

1. Overview

The data and findings mentioned in this report are taken from the cleaned dataset (22205061_homework1_cleaned.pdf). It will summarise the data, describe the various data quality issues found and how they will be dealt with. If readers want to know more details about the data, you should see the appendix. The appendix includes referenced books and sources. This also includes features summaries, histograms and boxplots used to visualise the data.

The report will assess the data quality firstly. And then, continuous and categorical features will be discussed. Based on the above assessment, identifying any issues or problems with the data. Evaluate the impact of the data quality issues on the overall quality of the data and its potential impact on the business. This report will make some recommendations for these issues or problems based on the analytics. In the end of this report, there is a summary of the data analytics.

2. Summary of data

The initial data (covid19-cdc-22205061.csv) has 20000 rows and 19 columns. There are four columns whose values are float type (*state_fips_code, county_fips_code, case_positive_specimen_interval* and *case_onset_interval*)and the values of remaining columns are all object type. Number of duplicate (excluding first) rows in the table is 1078 rows. Number of duplicate rows (including first) in the table is 1915 rows. There is no constant column and no duplicated columns. Counting the number of empty value of each column, the empty value percentage of six columns are over fifty percent including *case_onset_interval* (55.95%)*, process* (91.13%), *exposure_yn* (89.59%), *symptom_status* (51.80%), *icu_yn* (91.32%) and *underlying_conditions_yn* (90.99%). Notice that Null, Missing, Unknown and empty values are considered as the same condition. Details about empty values are shown in the appendix.

The cleaned data (22205061_homework1_cleaned.pdf) has 18922 rows and 19 columns.There are two columns whose values are integer type (case_positive_specimen_interval, case_onset_interval) and the values of remaining columns are categorical type. All duplicate (excluding first rows) rows are dropped.

3. Review Continuous Features

3.1 Descriptive Statistics

There are 2 continuous features.

The values of *case_positive_specimen_interval* fall in the range (-56 - 53).

The values of *case_onset_interval* fall in the range (-81 - 65).

Their percentage of null values is around 50%. And almost all numbers are concentrated under 5. So the box can not be seen. Because the data has too many outliers in the boxplots of both continuous features.

More details will be shown in the appendix.

3.2 Histograms

All histograms can be found on the appendix as a summary sheet. Individual plots can be found in the accompanying pdf file (continuous_histogramplots.pdf).

3.3 Box plots

All boxplots can be found on the appendix as a summary sheet. Individual plots can be found in the accompanying pdf file (continuous_boxplots.pdf). However, outliers are not addressed immediately. They will be addressed in the future.

4. Review Categorical Features

4.1 Descriptive Statistics

There are 17 categorical features in the dataset.

There is no problem found in *res_state* and *state_fips_code* features temporarily.

The *res_county, county_fips_code, age_group* and *sex* features have fewer null values. Their percentage of null values is under 10%.

The *race, ethnicity* and *hosp_yn* features have the percentage of null values in the range from 20% to about 30%.

The *symptom_status* feature has the percentage of null values about 50%.

The *process, exposure_yn, icu_yn* and *underlying_conditions_yn* features have the percentage of null values approximately 90%.

The *state_fips_code* and *county_fips_code* features use a numerical value which maps to a specific meaning. Every numerical value represents a location.

4.2 Bar plots

The bar plots can be found in the accompanying pdf (categorical_barplots.pdf) and the appendix.

5. Action to take

1. Address all null values of each feature.
If the percentage of null values of a certain feature is too high, drop this feature.
If that is low, the rows which include null can be dropped.
In other cases, using other appropriate values to fill the null values.

2. Address the outliers in the two continuous features.

3. Save the re-cleaned data.

6. Summary

In this period, the data is initially processed. There are some issues found in the data. These issues are not addressed in this stage and will be addressed in the next stage.

7. Appendix

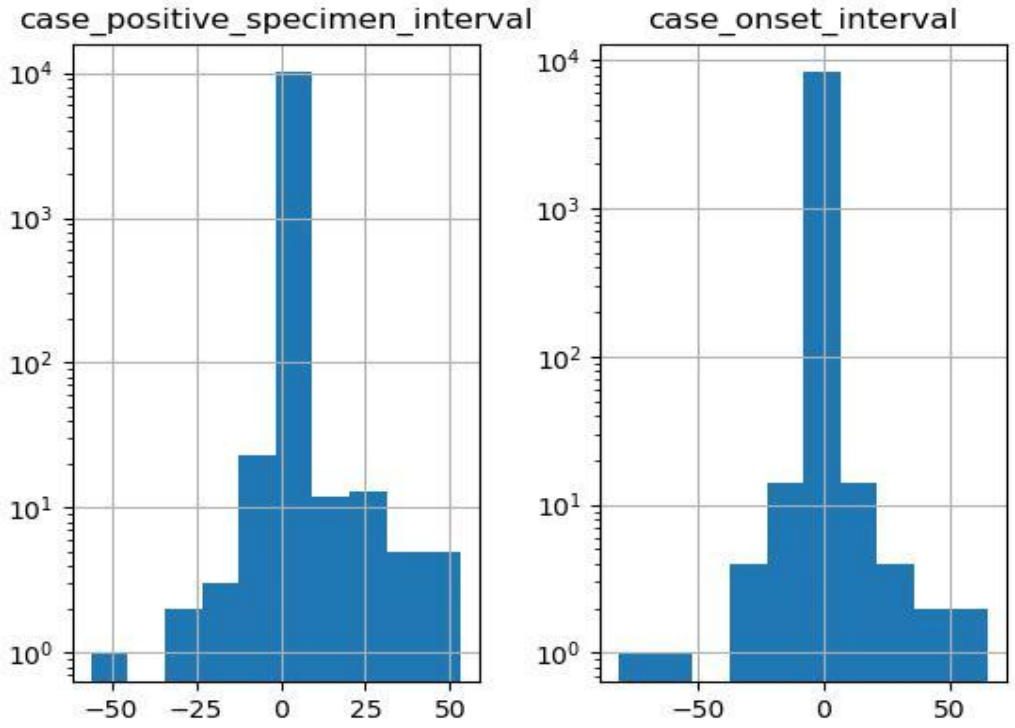
7.1 Continuous Features

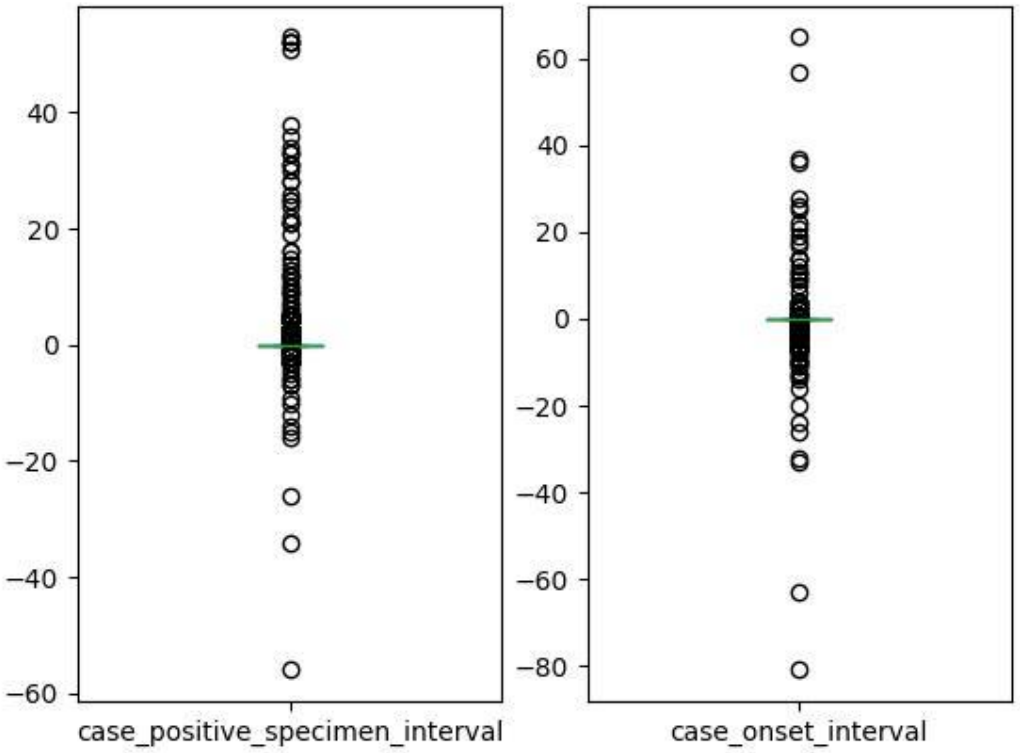
	count	mean	std	min	25%	50%	75%	max
case_positive_specimen_interval	10114.0	0.197548	1.961745	-56.0	0.0	0.0	0.0	53.0
case_onset_interval	8336.0	-0.053983	2.004306	-81.0	0.0	0.0	0.0	65.0

7.2 Categorical Features

	count	unique	top	freq
res_state	18922	49	NY	1946
state_fips_code	18922	49	36	1946
res_county	17780	868	MIAMI-DADE	370
county_fips_code	17780	1210	12086	370
age_group	18768	5	18 to 49 years	7226
sex	18530	4	Female	9521
race	16666	8	White	11797
ethnicity	16463	4	Non-Hispanic/Latino	11500
process	18922	9	Missing	17187
exposure_yn	18922	3	Missing	16185
current_status	18922	2	Laboratory-confirmed case	16006
symptom_status	18922	4	Symptomatic	8799
hosp_yn	18922	4	No	9501
icu_yn	18922	4	Missing	14668
death_yn	18922	2	No	14365
underlying_conditions_yn	1704	2	Yes	1684

7.3 Summary of Plots and Histograms





7.4 Missing value percentage

	Column Name	Count	Missing%
0	case_month	0	0.00
1	res_state	0	0.00
2	state_fips_code	0	0.00
3	res_county	1142	6.04
4	county_fips_code	1142	6.04
5	age_group	184	0.97
6	sex	493	2.61
7	race	4510	23.83
8	ethnicity	5831	30.82
9	case_positive_specimen_interval	8808	46.55
10	case_onset_interval	10586	55.95
11	process	17244	91.13
12	exposure_yn	16953	89.59
13	current_status	0	0.00
14	symptom_status	9801	51.80
15	hosp_yn	6259	33.08
16	icu_yn	17279	91.32
17	death_yn	0	0.00
18	underlying_conditions_yn	17218	90.99

7.5 Reference

[1] Fundamentals of Machine Learning for predictive Data Analytics 2015