Data Quality Report – Initial Findings

1. Overview

The following data quality report summarizes the steps to analyze and clean sample data set extracted from a public data set released by the Centers for Disease Contorl and Prevention (CDC). Over the course of the pandemic, the CDC colected data about Covid-19, analyzing cases, deaths, and trends throughout the United States. The dataset will be used to inform a linear regression model, where the target feature is ‘death_yn’, indicative of whether or not the patient died as a result of the virus, and coded as ‘yes’ or ‘no’.

The report outlines the various data quality issues observed in the data and addresses potential solutions. At a glance, there are MANY issues with the data. Most likely because the source of the data is a questionnaire filled out by a human on the specifics of the patient’s case, there is a lot of missing information. There are many null values, duplicate rows, and strange values for data that will have to be evaluated by a domain expert to determine its validity. Descriptive statistics for the set were gatherred and logical tests were carried out to both identify sources of error and explore options for cleaning it up.

2. Summary

To assess the quality of the data in the sample, several tests were conducted, including simple characterizations of the data set shape, statistical analyses to evaluate distribution, and logical integrity tests to verify consistency. However, we encountered a major issue with both the categorical and continuous data sets: a significant amount of missing data, which was encoded using various terms such as "NaN," "Missing," "Unknown," and "NA." To address this issue, the decision was made to consolidate these values into a single "Missing" value ("NA" for integer data types). This approach was chosen because the model cannot extract risk factor or probability of death information from these fields. While the other option is to impute the missing data, consolidating missing values into a single category is more practical. Doing this will facilitate data analysis by both human analysts and machine learning algorithms.

Additionally, it is recommended to drop the 'res_state', 'res_county', and 'state_fips_code' features, as 'res_state' and 'res_county' are redundant, while 'res_county' creates confusion for the model due to the presence of multiple counties with the same names across the United States. An analysis of the data indicates that the 'state_fips_code' is embedded in the 'county_fips_code', and 'county_fips_code' is available for all but two states (Puerto Rico and the US Virgin Islands). Therefore, a new 'county_fips_code' should be generated that is equal to the original 'county_fips_code' when available, but equal to the original 'state_fips_code' multiplied by 1000 when the 'county_fips_code' is null.

Duplicate values are not considered in the analysis. There are 1849 instances of duplicate information, out of 20000 rows, representing nearly 10% of the data. It is recommended that the data be removed from the data set. This is not because the data is believed to be invalid; based on knowledge of Covid-19, it would make sense that more than one person in a county of the same age, gender, race, and ethnicity, could contract Covid at the same time, representing 9 of the features already. However, in a linear regression model, duplicate values will indicate a strong correlation between the feature pairs, which can lead to skewed or misleading results. Therefore the duplicate columns are dropped.

Finally, it is recommended to take the absolute value of all 'case_positive_specimen_interval' and 'case_onset_interval' features, as negative intervals are not possible according to the formula derivation. For further details on the rationale behind these recommendations, please consult the remaining sections of the report and the accompanying Jupyter Notebook.

After cleaning the data to address the issues identified during the initial findings, missing values will be addressed using data imputation. Mean and mode imputation can be used for continuous data, as approximately 50% of the ‘case_positive_specimen_interval’ and ‘case_onset_intervals’ values consist of zeros. However, for categorical values, where mean or median imputation is not suitable, alternative methods have been breifly researched. One such method is the K-nearest neighbor clustering algorithm, which estimates data points based on similar rows. This method provides better estimates than simply filling in the mode, although it may introduce bias. Given its simplicity and potential to improve model accuracy, this method will be used. Various imputation techniques will be applied to address missing values, ensuring that accurate insights are obtained from our analysis.

3. Review Logical Integrity

A total of 15 tests were carried out. The failed tests are listed below. For information on the other tests carried out, refer to ‘22202398_Homework1-Notebook.ipynb’.

- Test 1: Check if any patient has a case_positive_specimen_interval < 0 (impossible)
  - 44 patients found
  - This entry is the difference between two values entered the sheet. They may have been subtracted wrong. For the remained of the tests, we will use the absolute value of this number.
- Test 2: Check if any patient has a case_onset_interval < 0 (impossible)
  - 249 patients found
  - Like the case_positive_specimen_interval, this entry is the difference between two values entered the sheet. They may have been subtracted wrong. For the remained of the tests, we will use the absolute value of this number.
- Test 5: Check if any patient has a ‘case_positive_specimen_interval’ not equal to null and ‘current_status’ not equal to ‘Laboratory confirmed case’ (impossible)
  - 1647 patients found
- Test 6: Check if any patient has a ‘case_onset_interval’ not equal to null and a ‘symptom_status’ not equal to ‘Symptomatic’ (impossible)
  - 221 patients
- Test 8: Check if the number of unique ‘res_county’ values is equal to the number of unique ‘county_fips_code’.
  - 853 unique ‘res_county’ values and ‘1203’ unique county_fips_code values.
  - Consider dropping this column to eliminate redundanct, ambiguous information.
- Test 12: Check that all entries where ‘hosp_yn’ is ‘Yes’ should also list ‘symptom_status’ as ‘Symptomatic’
  - 1410 patients found with ‘hosp_yn’ as ‘Yes’ and ‘symptom_status’ as ‘Asymptomatic’
- Test 14: Check that all entries where ‘icu_yn’ is ‘Yes’ should also list ‘hosp_yn’ as ‘Yes’
  - 2 patients found with ‘icu_yn’ as ‘Yes’ and ‘hosp_yn‘ as ‘No’

- Test 15: Check that all entries where ‘icu_yn’ is ‘Yes’ should also list ‘symptom_status’ as ‘Symptomatic’
    - 4563 patients found with with ‘icu_yn’ as ‘Yes’ and ‘symptom_status’ as ‘Asymptomatic’

4. Review Continuous Features

4.1 Descriptive statistics

There are 3 continuous features. They are listed and summarized below:

- ‘Case_month’ – Values range between January 2020 and November 2022. There are no 0% missing values and there is a cardinality of 35, which is equal to the number of months between the minimum and maximum.
- ‘Case_positive_specimen_interval’ – From the CDC data dictionary, “Calculated by dividing the days between pos_spec_dt and cdc_case_earliest_dt, rounding up. Blank when pos_spec_dt not provided.” The average value is 0.244, and ranges between -61 and 104. We can assume all negative values should have been entered as positive, which reduces the range to just 0 to 104. Even so, 104 weeks is exactly 2 years, which seems like a long time between the CDC earliest date and the first time a patient tested positive for the virus. These are likely outliers, but a domain expert would have to weigh in to discuss whether or not this is possible.
- ‘Case_onset_interval’ – From the CDC data dictionary, “Calculated by dividing the days between onset_dt and cdc_case_earliest_dt, rounding up. Since cdc_case_earliest_dt is calculated using the earliest of multiple dates, for some records this will mean the cdc_case_earliest_dt and onset_dt are the same and so the interval is 0 and have the same value as cases with 0 week intervals. Blank when onset_dt not provided.” The average value is -0.0428, and ranges between -56 and 72. We can assume all negative values should have been entered as positive, which reduces the range to just 0 to 72. Even so, 104 weeks is a 1.5 years, which seems like a long time between the CDC earliest date and the first time a patient experienced symptoms of the virus. These are likely outliers, but a domain expert would have to weigh in to discuss whether or not this is possible.

All of the values in the ‘case_month’ feature seem plausible, and the case distribution per month seems to align with the pandemid peaks experienced throughout the world. The others were harder to get a sense of using just the statistics, and required the visual plots for analysis. As already discussed, there seem to be some outliers, but a domain expert should be consulted to verify whether these values are possible. The full table of descriptive statistics can be found in the appendix.

4.2 Histograms

All histograms can be found in the appendix. The histograms for both ‘case_positive_specimen_interval’ and ‘case_onset_interval’ indicate a very large concentration of values around 0. This aligns with my personal expectations, since the majority of patients would test positive within days of showing symptoms or being exposed to the virus, or vice versa, showing symptoms within days of testing positive or being exposed. The histogram for ‘case_month’ again shows the expected case distrubution per the timeline of the pandemic.

4.3 Boxplots

All boxplots can be found in the appendix. Again, the boxplots for both ‘case_positive_specimen_interval’ and ‘case_onset_interval’ indicate a very large concentration of values around 0, but with outliers that are far

greater than (and less than, though these could be considered the same once the absolute value is taken during clean up). The distribution of data for 'case_month' is reasonable and makes sense.

5. Review Categorical Features

There are 16 categorical features in the dataset, 1 of which being the target feature. They can be organized into X groups. They are listed and summarized below:

- Demographic Information
    - Res_state –The name of the state the questionnaire was submitted in
    - State_fips_code – The unique fips code of the state the questionnaire was submitted
    - Res_county – The name of the county the questionnaire was submitted in
    - County_fips_code – The unique fips code of the county the questionnaire was submitted
    - Age_group – The age of patient, grouped broadly into one of 5 categories.
    - Sex – Sex of the patient, male, female, or 'Missing'
        - Note that the danger in grouping all 'Missing' and 'Unknown' values into one 'Missing' category is the risk of underrepresenting the LGBTQ+, however there are only 115 patients under 'Missing' even after this grouping
    - Race – The race of the patient
    - Ethnicity – The ethnicity of the patient
- Covid Detection and Status Information
    - Process – The process by which the case was discovered. This feature has 91% missing entries and the recommendation will be to drop it.
    - Exposure_yn – Whether or not the patient was exposed to the virus before contracting it. This feature has 90% missing entries and the recommendation will be to drop it.
    - Current_status_yn – The current status of the patient's Covid-19 diagnosis, whether it is a 'Laboratory confirmed case' or a 'Probable case'. See Logical Integrity section for potentially invalid representation of this feature.
    - Symptom_status – Whether the patient is 'Symptomatic' or 'Asymptomatic'. See Logical Integrity section for potentially invalid representation of this feature.
    - Hosp_yn – Whether the patient has been hospitalized. See Logical Integrity section for potentially invalid representation of this feature.
    - Icu_yn – Whether the patient has been admitted to an ICU. This feature has 90% missing entries, and the recommendation will be to drop it.
    - Underlying_conditions_yn – Whether the patient presented any underlying conditions to exacerbate their illness. This feature has 91% missing entries, and the recommendation will be to drop it.

5.1 Histograms

All histograms can be found in the appendix. For a lot of features, the historgrams give a good indication of the amount of possible values for a feature, whether there are very many or very few. They also make it clear when the 'Missing' data dominates the feature.

6. Action to Take

7 main actions will be taken in an initial round of data cleaning.

- “Missing”, “Unkown” and “NaN” values
    - Replace these Special values with one flag that encapsulates them all, “Missing.
- Duplicates
    - Drop all duplicate values
- Logical Integrity
    - Replace all negative ‘case_positive_specimen_interval’ and ‘case_onset_interval’ values with their absolute values.
    - Replace all ‘hospital_yn’ values with ‘Yes’ if ‘icu_yn’ is ‘Yes’
    - Replace all ‘symptom_status’ values with ‘Symptomatic’ if ‘hospital_yn’, ‘icu_yn’, or ‘death_yn’ values are ‘Yes’
    - Replace all ‘symptom_status’ values with ‘Symptomatic’ if ‘case_onset_interval’ is not Null
    - Replace all ‘currnet_status’ values with ‘Laboratory_confirmed_case’ if ‘case_positive_specimen_interval’ is not Null.
- Categorical Features
    - Drop ‘res_county’, ‘res_state’, and ‘state_fips_code’, and replace with a modified ‘county_fips_code’
    - Drop ‘process’, ‘exposure_yn’, ‘icu_yn’, and ‘underlying_conditions_yn’ features, with over 90% missing data.
- Subsequent Rounds of Data Cleaning will involve imputation for filling in the missing data
    -
- Outliers
    - Consult a domain expert to evaluate what range of outliers is possible. Drop features far beyond valid range.

7. References

Data Dictionaries and CDC Covid Information:

- https://www.cdc.gov/coronavirus/2019-ncov/downloads/php/covid-19-case-reporting-data-dictionary.pdf
- https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

8. Appendix

8.1 Categorical Features Descriptive Statistics

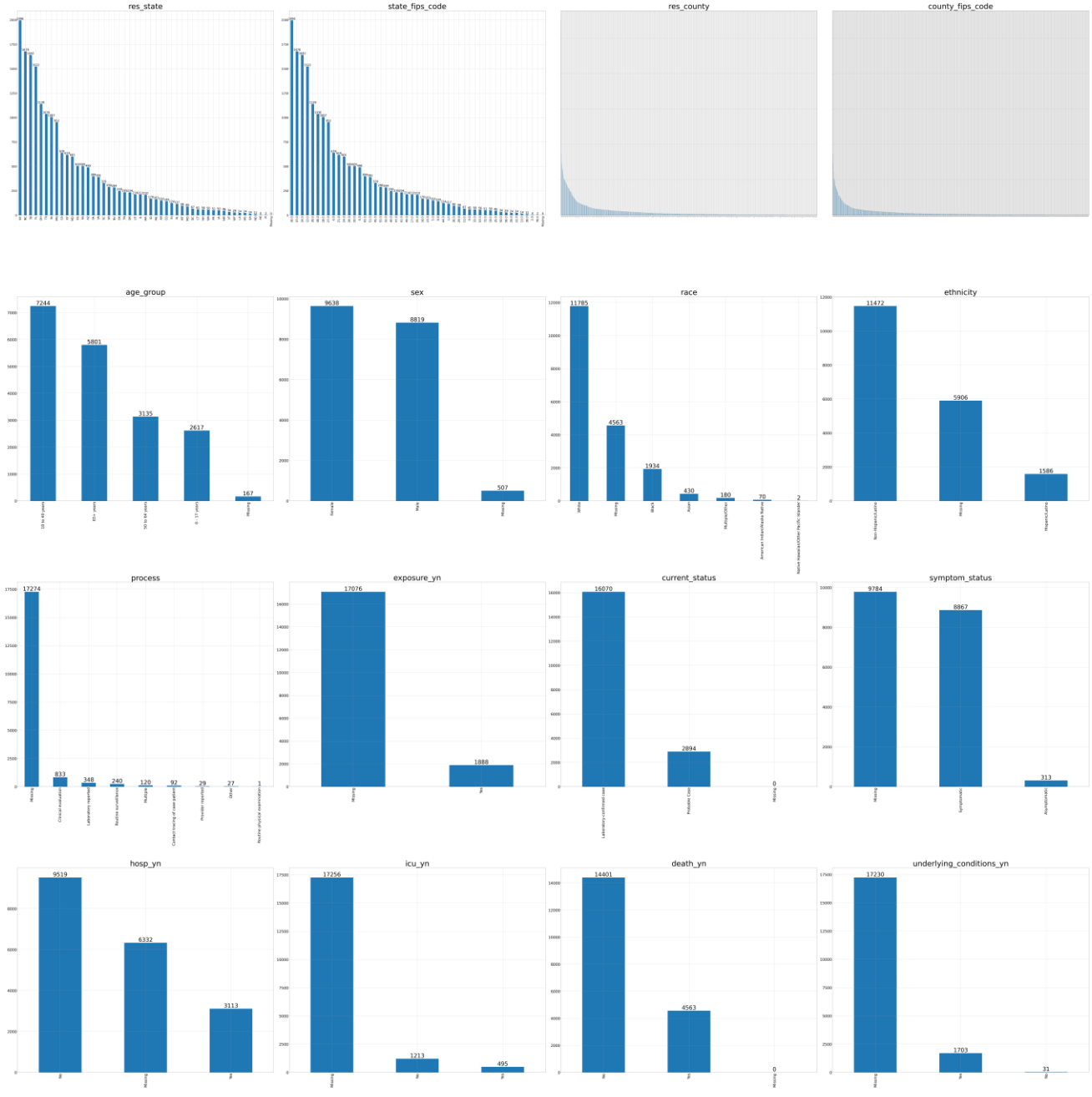
Feature	Count	Cardinality	1st Mode	1st Mode Fre	2nd Mode	2nd mode fre	# Missing	% Missing
process	18964	9	Missing	17274	Clinical evaluation	833	17274	91.1
icu_yn	18964	3	Missing	17256	No	1213	17256	91.0
underlying_conditions_y	18964	3	Missing	17230	Yes	1703	17230	90.9
exposure_yn	18964	2	Missing	17076	Yes	1888	17076	90.0
symptom_status	18964	3	Missing	9784	Symptomatic	8867	9784	51.6
hosp_yn	18964	3	No	9519	Missing	6332	6332	33.4
ethnicity	18964	3	Non-Hispanic/Latino	11472	Missing	5906	5906	31.1
race	18964	7	White	11785	Missing	4563	4563	24.1
res_county	18964	853	Missing	1103	MIAMI-DADE	348	1103	5.8
county_fips_code	18964	1203	Missing	1103	12086	348	1103	5.8
sex	18964	3	Female	9638	Male	8819	507	2.7
age_group	18964	5	18 to 49 years	7244	65+ years	5801	167	0.9
res_state	18964	48	NY	1996	NC	1678	0	0.0
state_fips_code	18964	48	36	1996	37	1678	0	0.0
current_status	18964	2	Laboratory-confirmed cas	16070	Probable Case	2894	0	0.0
death_yn	18964	2	No	14401	Yes	4563	0	0.0

8.2 Continuous Features Descriptive Statistics

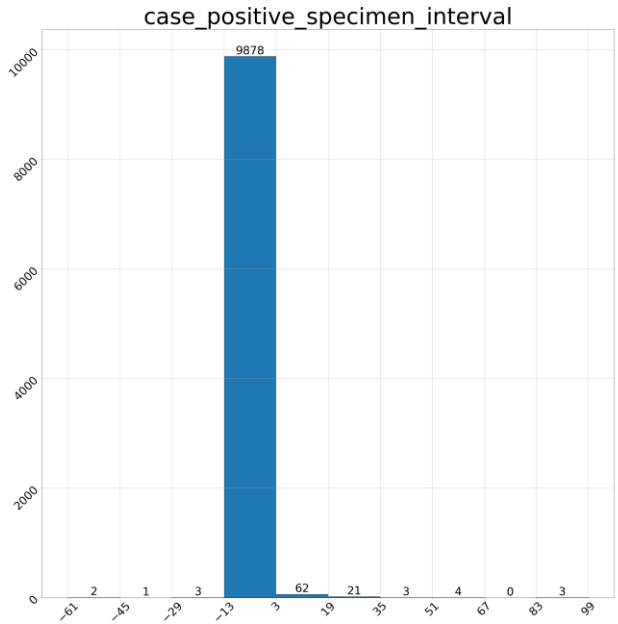
Feature	Count	Mean	Minimum	0.25	0.5	0.75	Max	StdDev
case_positive_specimen_interval	9978	0.244237322	-61	0	0	0	104	2.8
case_onset_interval	8428	-0.04283341	-56	0	0	0	72	1.7
case_month	18964	6/6/2021	1/1/2020	12/1/2020	7/1/2021	1/1/2022	11/1/2022	

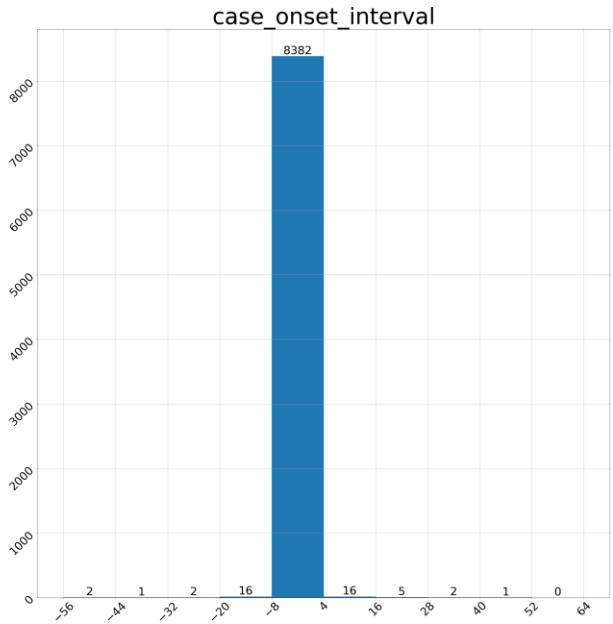
Feature	Cardinality	# Missing	% Missing	1st Mode	1st Mode Fre	2nd Mode	2nd Mode Fr
case_positive_specimen_interval	49	8986	47	0	8828	1	900.0
case_onset_interval	37	10536	56	0	8131	-1	142.0
case_month	35	0	0	1/1/2022	8/20/1906	12/1/2020	1513.0

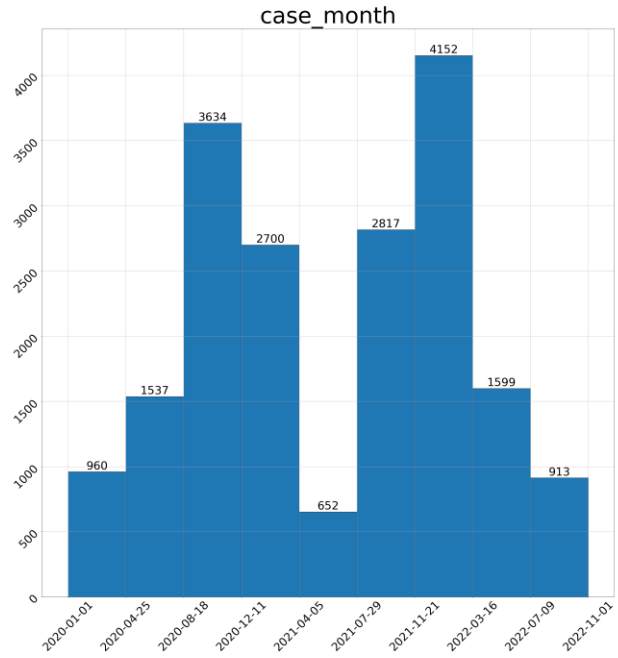
8.3 Categorical Features Histogram



8.4 CONTINUOUS FEATURES BAR PLOTS







8.5 Continuous Features Box-And-Whisker Plots