Data Quality Report - Initial Findings

1. Overview

Introduction to Data Quality Report for CDC COVID-19 Dataset

The dataset obtained from the Centers for Disease Control and Prevention (CDC) serves as a critical foundation for developing a data analytics solution aimed at predicting death risk in individuals affected by COVID-19 within the United States. As a prominent health protection agency, the CDC diligently collects and disseminates individual-case data, providing a comprehensive snapshot of the pandemic. The purpose of this data quality report is to assess the reliability, completeness, and consistency of the CDC COVID-19 dataset. Ensuring data quality is crucial for generating accurate and trustworthy predictions related to death risk. This report will scrutinise various aspects of the dataset, highlighting potential challenges and areas for improvement, to facilitate robust analyses and meaningful insights.

This report will outline the initial findings based on the cleaned dataset (CDC_COVID_1_1.csv) It will summarise the data, describe the various data quality issues observed and how they will be addressed. Please see the 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, plots used to visualise the data.

On first indication the dataset appears relatively rough. There are many null values in many different features. The main issue observed was regarding these null values.

2. Summary

For the continuous features there was an extremely high number of null values.

case_positive_specimen_interval (float64)	Null Values (47%), 0 values (47%), High and Low Outliers	Too many Null Values, Drop Feature
case_onset_interval (float64)	Null values (57%), 0 value (41%), High and Low Outliers	Too many Null Values, Drop Feature

For ‘case_positive_specimen_interval’, 47% of all values were ‘Nan’ and a further 47% were the value 0. Similarly for ‘case_onset_interval’, 57% of all values were ‘Nan’ and a further 41% of all values were 0. These features were dropped.
For the categorical values several changes are recommended. There were several features with suppressed values due to privacy protection limitations. These values were initially

displayed as null values. They were changed to ‘Missing’ or ‘Suppressed’ instead. One feature displayed irregular cardinality but was solved by combining two features.

Logical tests were performed on the cleaned data but nothing irregular was found. Therefore, these tests were removed to keep the report concise.

3. Review Continuous Features

3.1. Descriptive Statistics

There are 2 continuous features. Categories which will be summarised below;

case_positive_specimen_interval (float64)	Null Values (47%), 0 values (47%), High and Low Outliers	Too many Null Values, Drop Feature
case_onset_interval (float64)	Null values (57%), 0 value (41%), High and Low Outliers	Too many Null Values, Drop Feature

3.2. Histograms

All histograms can be found on the appendix. Individual plots can also be found in the accompanying notebook. Overall the features showed irregular distributions. The majority of data of both features lie in the value 0 as seen from the plots.

3.3. Box plots

All boxplots can be found on the appendix and can be found in the accompanying notebook. Again, the majority of data of both features lie in the value 0. Outliers can evidently be seen in the boxplots also. However, the points are only seen as ‘outliers’ because of the huge percentage of data with value 0.

4. Review Categorical Features

4.1. Descriptive Statistics

There are 17 categorical features in the dataset, 1 of which is the target (death_yn) and will not be evaluated here. They can essentially be divided into a few groups.

Time

- case_month (category) - The earlier month of the Clinical Date (date related to the illness or specimen collection) or the Date Received by CDC

Problems

No problems with data quality. Feature kept as is.

Geographical Factors

- res_state (category) - State of residence
- state_fips_code (category) - State of residence code
- res_county (category) - County of residence
- county_fips_code (category) - County of residence code

Problems

Some States and Counties are protected with suppressed values due to case numbers. These nulls must be filled. Res_county and county_fips_code do not have matching cardinalities.

Patient Demographics

- age_group (category) - Age group of patient
- sex (category) - Sex of patient
- race (category) - Race of patient
- ethnicity (category) - Ethnicity of patient

Problems

All features have suppressed values due to privacy protection.

Process/Status

- process (category) - Under what process was the case first identified?
- exposure_yn (category) - In the 14 days prior to illness onset, did the patient have any of the following known exposures: domestic travel, international travel, cruise ship or vessel travel as a passenger or crew member, workplace, airport/aeroplane, adult congregate living facility (nursing, assisted living, or long-term care facility), school/university/childcare centre, correctional facility, community event/mass gathering, animal with confirmed or suspected COVID-19, other exposure, contact with a known COVID-19 case?
- current_status (category) - What is the current status of this person?
- Symptom_status (category) - What is the symptom status of this person?

Problems
Many categories have large Missing/Unknown values in either their first or second modes. Categories such as exposure_yn, process have large majorities of their data points as Missing or Unknown making them extremely poor data quality columns.
When combining the Unknown and Missing values in symptom_status column, these basically null values take up 54% of the total values.

Hospital/Patient Care

- hosp_yn (category) - Was the patient hospitalised?
- icu_yn (category) - Was the patient admitted to an intensive care unit (ICU)?
- death_yn (category) - Did the patient die as a result of this illness?
- underlying_conditions_yn (category) - Did the patient have one or more of the underlying medical conditions and risk behaviours: diabetes mellitus, hypertension, severe obesity (Body Mass Index ≥40 kg/m2), cardiovascular disease, chronic renal disease, chronic liver disease, chronic lung disease, other chronic diseases, immunosuppressive condition, autoimmune condition, current smoker, former smoker, substance abuse or misuse, disability, psychological/psychiatric, pregnancy, other?

Problems
Many categories have large Missing/Unknown values in either their first or second modes. Categories have large majorities of their data points as Missing or Unknown making them extremely poor data quality columns.

4.2 Plots

All plots can be found in the accompanying notebook.

5. Action to take

Two main problems to address;

Null Values
- The main issue with the majority of the features is the large proportion of null or missing values. Many of the features had large percentages of null values as well as Missing and Unknown values.

Irregular Cardinality
- Cardinality did not match up for res_state and county_fips code where they should have.

Propose solutions to deal with the problems identified.

Null Values

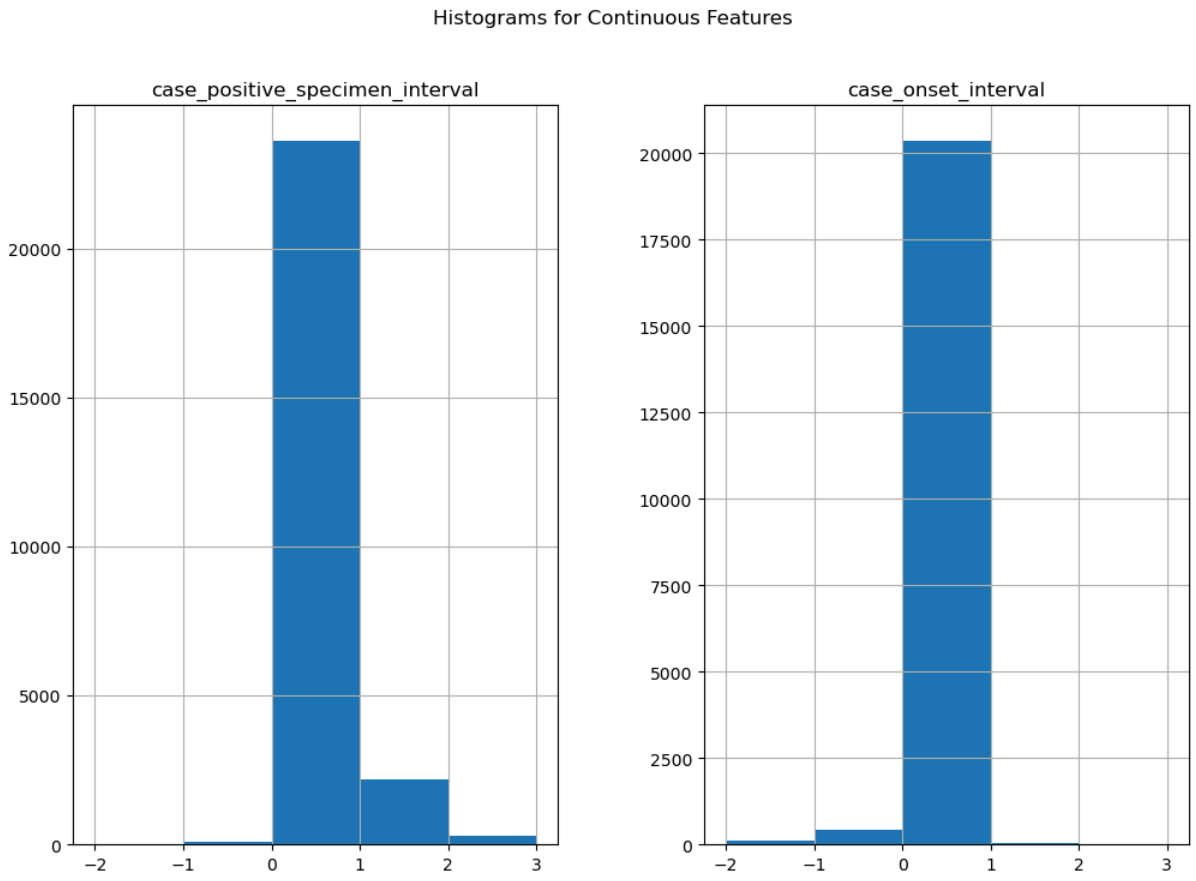
- Appears in 12 features.
  res_state & *state_fips_code* only contained 2 rows with null values. After examining these two rows it was clear that these rows were made up of mostly empty values.
- Many features such as *res_county* contained null values due to privacy protection and not just through missing values. In the case of *res_county* and *county_fips_code*, I replace these nulls with 'Suppressed' values. However, in other features that already had a 'Missing' field, I simply combined the null values with the 'Missing' values.
- Also some feature's rows were filled with majority null or missing values. In many cases, over 50% of the rows were null or missing values. These features were dropped due to poor data quality.

Irregular Cardinality

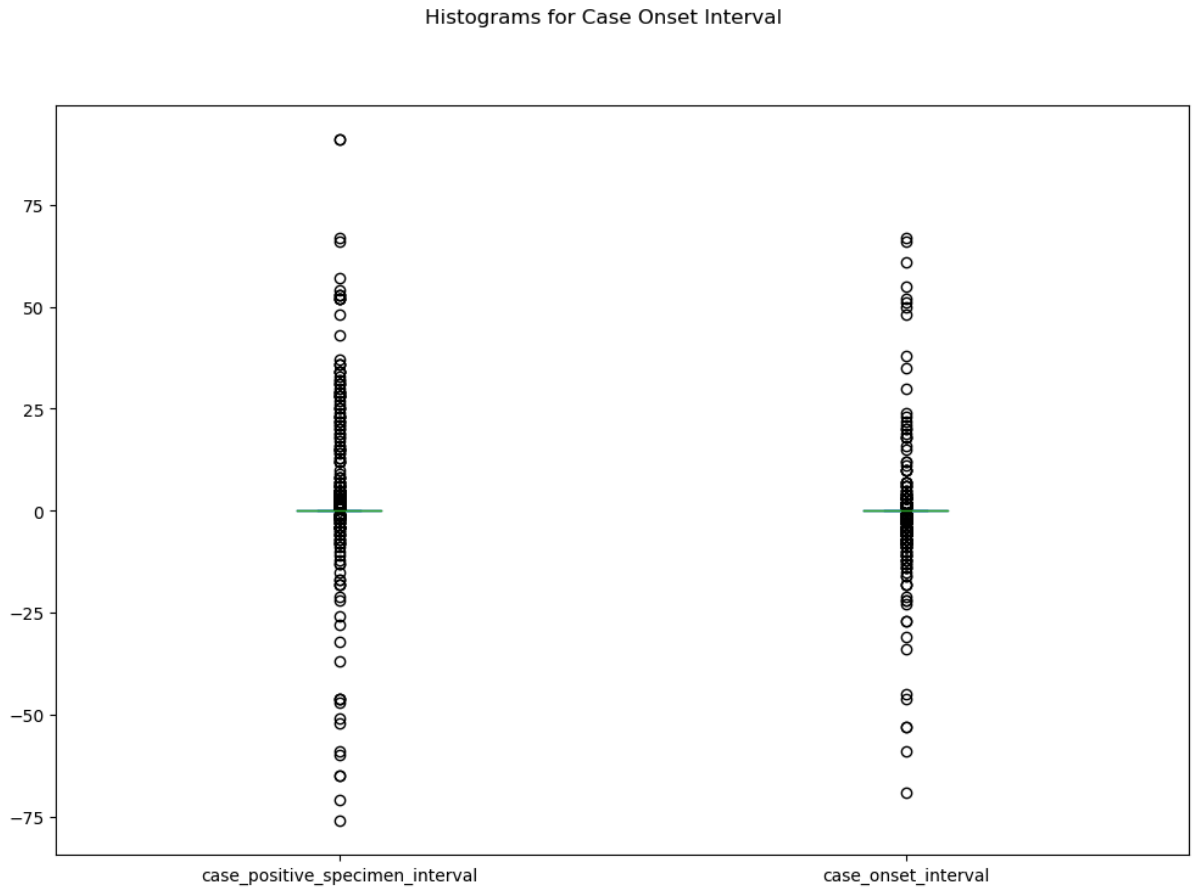
- Cardinality did not match up for *res_state* and *county_fips* code where they should have. This was evident from the bar plot as they had a slightly different shape. This was due to the fact that some counties shared their names but resided in different states. This resulted in a lower cardinality in the *res_state* column than the *county_fips_code* column. I combined the two features to fix this problem.

8. Appendix

8.1. Histograms



8.2 Boxplots


Histograms for Case Onset Interval
75
50
25
0
−25
−50
−75
case_positive_specimen_interval
case_onset_interval