COMP47350 Data Analytics

Name: Charlie Drumm

Student Number: 20478622

Title: Homework 1 COVID Data Analysis

Contents

1. Overview

This Data Quality Report is a outlines the initial findings of the dataset covid19-cdc-20478622.csv used in the COVID-19 project, designed to analyse the data's integrity across several critical areas. The dataset consists of a number of features, from case counts and mortality rates to demographic details and hospitalisations. This report summarises the critical features of the dataset and provides recommendations on how to clean the dataset.

The objective of this report is twofold: firstly, to assess the dataset's quality by examining its completeness, accuracy, consistency, and timeliness; and secondly, to outline the corrective measures undertaken to address any identified issues. Upon first look at the dataset there are a number of features with missing or null values. Also a series of logical tests were carried out on the data and a number of inconsistencies were found. These logical errors and missing data could be down to the fact of people making errors or skipping sections on the form which was used to create this data.

2. Summary

In this report we've conducted a thorough examination of a dataset sourced from the CDC that spans various aspects of the COVID-19 pandemic in the United States, including demographics, geography (county and state of residence), any exposure history, disease severity indicators and outcomes, and presence of any underlying medical conditions and risk behaviours. In the dataset each row represents a case. Our primary goal was to ensure that the data we rely on for our research and policy recommendations stands up to rigorous standards of quality. Identifying and addressing any shortcomings in the dataset is crucial.

Key Observations and Actions:

1. **Data Completeness and Accuracy**: We encountered several instances where data entries were missing or inaccurately reported. For example, significant gaps in the "Case Positive Specimen Interval" and "Case Onset Interval" were noted. These gaps potentially undermine our understanding of the virus's spread and its incubation period in the population. Also there were a number of negative numbers in "Case Onset Interval" 99% of which came from the state of Texas. Also three features “Process”, “ICU” and “Exposure” had a mode average of ‘Missing’, implying that people generally skipped that section of the form either when reporting it or reading it.

2. **Consistency Checks**: Our analysis revealed inconsistencies, such as symptomatic cases with no onset interval recorded and laboratory-confirmed cases with missing specimen intervals. These discrepancies suggest errors in data entry or collection methodologies. The majority of data within the dataset has a “Current Status” of ‘Laboratory-confirmed case’, however some cases have a “Current Status” of ‘Probable case’. These cases may not be COVID cases and perhalps they should be removed, this should be consulted with a domain expert.

Recommendations include employing statistical methods to impute missing values where appropriate, correcting inconsistent entries, and maintaining a rigorous standard for data collection and entry going forward. These steps are essential to improve the dataset's quality, ensuring that subsequent analyses and recommendations are based on solid, reliable data.

3. Review Logical Integrity

8 tests were conducted. The failures are below;

- ❖ Test 1 – Check if case_onset_interval is null when symptom_status is Symptomatic.
  - 1717 rows failed
- ❖ Test 2 – Check if case_onset_interval is not null when symptom_status is not Symptomatic
  - 493 rows failed
- ❖ Test 3 – Check if case_positive_specimen_interval is null when current_status is Laboratory-confirmed case
  - 20428 rows failed
- ❖ Test 4 – Check if case_positive_specimen_interval is not null when current_status is not Laboratory-confirmed case
  - 4548 rows failed
- ❖ Test 6 – Check if res_county is represented by 1 county_fips_code
  - 13 rows failed

4. Review Continuous Features

There are two continuous features in this dataset;

Case Positive Specimen Interval

Weeks between earliest date and date of first positive specimen collection. Most of the data points (25th to 75th percentile) are concentrated around 0 weeks, suggesting that for many cases, the specimen that tested positive was collected on same day of the case being reported. For each case where the current_status is Laboratory-confirmed case, case positive specimen interval of that row should not be null. However this is not always the case. There are 23,681 missing entries for this feature, 20428 of which are from cased which are current_status is Laboratory-confirmed case. This is a significant number of rows and if we were to change the values to the mode it would skew the results, if I was to use random sampling, missingness may not be at random and the imputed values would be biased. It is my belief that case positive specimen interval was not updated as the positive specimen may have been collected a number of days after the case was initially reported. Based on this I will leave the rows failing this test as there is no clear justification for the missing values and I cannot consult a domain expert to advise further, I also do not want to drop the feature as I believe it will be important in further analysis.

Case Onset Interval

Weeks between earliest date and date of symptom onset. The bulk of the data points are at 0 weeks, suggesting that for a significant number of cases, the symptoms occurred on the same day that the case was reported. There are 28,713 missing entries for this feature, however some values should be null if the person did not have symptoms. For each case where symptom_status is Symptomatic the case onset interval should not be null. There are 1717 rows where the case onset interval is null where the symptom_status is Symptomatic. I would recommend that these null values be replaced by the mode as it occurs 96.22% of the time.

Outliers as Possible Indicators of Re-infection

Negative Values: Negative case_positive_specimen_interval or case_onset_interval values could suggest data entry errors. I would recommend that the negative values are converted to positive values

Significantly Positive Values: Outliers with significantly positive intervals might represent cases where individuals were re-infected with COVID-19. If a person was tested and found positive or showing symptoms well beyond the typical incubation period from their case being initially reported, this might suggest a subsequent exposure and infection rather than a prolonged course of a single infection. The CDC incubation period is between 2 and 14 days [2]. Such outliers could also reflect delayed reporting or testing in relation to symptom onset. I would recommend removing these values as they do not accurately represent the typical onset timeline

5. Review Of Categorical Features

State and County of Residence and County and State FIPS codes

- **Description**: These features indicate the state and county of residence for each case.
- **Analysis**: Analysing these features helps understand the geographical distribution of cases. Variability in case counts across locations may reflect differences in testing rates, public health policies, or virus spread. There are a small number of res_county and county_fips_code null that I recommend imputing with Missing. There is no need to keep all four of these features as the codes correlate to the state/county. I would recommend dropping the state fips code and county fips code as they are more difficult to read and contains the same data as the state and county.
- **Implications**: Disparities in case distributions can inform targeted public health interventions and resource allocation.

Age Group

- **Description**: Categorises patients into predefined age ranges.
- **Analysis**: The age distribution can shed light on which groups are more affected or at higher risk. There are a small number of age_group values which are null. I recommend imputing the null values with 'Missing'. If the missingness is not completely random (for instance, certain age groups or genders might be less likely to report age or gender), proper imputation could reduce bias introduced by missing data.
- **Implications**: Identifying high-risk groups can guide vaccination and public health strategies.

Sex

- **Description**: The biological sex of the patient.
- **Analysis**: Differences in case counts or outcomes between sexes can indicate biological or behavioural risk factors. I recommend imputing the null values with 'Missing' as the missingness may not be completely random.
- **Implications**: Insights can inform sex-specific health advisories or interventions.

Race and Ethnicity

- **Description**: Provide demographic information about the race and ethnicity of the cases.
- **Analysis**: Analysis of these features can reveal disparities in COVID-19 impact among different racial and ethnic groups. I recommend imputing null values with ‘Missing’, these are sensitive demographic variables and it is important to consider ethical implications of imputing.
- **Implications**: Understanding these disparities is crucial for developing equitable public health responses and addressing social determinants of health.

Process and Exposure

- **Description**: **process** refers to the process that was used to first identify the case, while **exposure_yn** indicates known exposure to COVID-19 or certain environments.
- **Analysis**: These features can help understand the efficiency and findings of contact tracing efforts. The two of these features have a large number of missing values and I would consider dropping one or both of the features.
- **Implications**: Effective contact tracing is key to controlling spread, and insights here can help improve these processes.

Current Status

- **Description**: Indicates whether the case is Laboratory confirmed or probable.
- **Analysis**: This feature helps to understand the distribution of confirmed versus probable cases. However there are a large number of logical errors within this feature. There are 4548 rows where the case_positive_specimen_interval is not null but the current_status is ‘Probable case’. If there is a value for case_positive_specimen_interval that means that there has been a positive specimen collected and the current_status should be ‘Laboratory-confirmed case’. I would recommend changing the value of these rows failing to ‘Laboroatory-confirmed case’. The majority of values in this feature are ‘Laboratory-

confirmed case' and there are some logical issues with the feature regarding the case_positive_specimen_interval. I would also consider dropping this feature, however it is also useful to keep it as it regards to the testing status and confirms the presence of COVID. I feel that there could be some disparity in the testing methods in each state and this feature would be very useful to determine how cases were confirmed and how covid was tested for in each state.

- **Implications**: The number of confirmed to probable cases may reflect testing capacity and criteria.

Symptom Status

- **Description**: Indicates if the person was Symptomatic or Asymptomatic.
- **Analysis**: Symptom reporting can provide insights into the prevalence of asymptomatic cases and symptom patterns. There is a large number of rows where the symptom_status is unknown or missing. There are also 493 rows where the symptom_status is 'Missing' but the case_onset_interval is not null. I would recommend changing these values from 'Missing' to 'Symptomatic'.
- **Implications**: Understanding symptom profiles can inform public health guidelines and individual decision-making.

Hospitalization, ICU, and Death

- **Description**: These features indicate whether the case resulted in hospitalization, ICU admission, or death.
- **Analysis**: Analysis of these outcomes can highlight the severity of cases and identify which groups are most at risk of severe outcomes. Each of these features has a large number of missing or unknown, however due to the importance of each of them in indicating the severity of a case I would recommend keeping each feature.
- **Implications**: Insights into risk factors for severe disease can guide protective measures for vulnerable populations.

Underlying Conditions

- **Description**: Indicates whether the patient had underlying conditions known to increase COVID-19 risk.
- **Analysis**: This feature is vital for understanding how underlying conditions affect COVID-19 outcomes. Despite the large number of null values I recommend imputing the null values in underlying_conditions_yn with 'Missing'.
- **Implications**: Data on comorbidities can help prioritize individuals for vaccination and other preventive measures.

Key Considerations

- **Data Quality**: Categorical features may have missing or inconsistently reported values, affecting analyses.
- **Equity and Ethics**: Analysis of race, ethnicity, and underlying conditions requires sensitivity to avoid stigmatisation and ensure findings are used ethically to promote health equity.

6. Action to Take

9 main actions will be taken, summarised below;

1. **case_positive_specimen_interval (Negative Values):** Negative values indicate potential data entry errors and should be corrected. The recommended action is to convert all negative values to positive to reflect the correct time interval.
2. **case_positive_specimen_interval (Outliers):** Outliers may indicate errors or reinfection of COVID. It is advised to remove these outliers from the dataset to prevent them from skewing the analysis.
3. **case_onset_interval (57.426% null):** For the subset of these null values where symptom_status is Symptomatic (5.98%), the missing values are imputed with the mode.
4. **case_onset_interval (Negative values):** Similar to the specimen interval, negative onset intervals should be corrected by changing them to positive values.
5. **case_onset_interval (Outliers):** Outliers in this feature should be removed to avoid misrepresentation of the typical disease progression timeline.
6. **process (90.8% Missing):** Due to the high percentage of missing values, it is recommended to drop this feature entirely as it may not contribute valuable insights and could potentially introduce bias into the analysis.
7. **exposure_yn (86.14% Missing):** Similar to the process feature, the exposure_yn feature is largely incomplete. The recommendation is to also drop this feature to streamline the dataset.
8. **current_status :** For cases with a non-null positive specimen interval that are currently marked as 'Probable case', the status should be updated to 'Laboratory-confirmed case' to reflect the evidence of a positive test.
9. **symptom_status (2.32% of cases with not null case_onset_interval were Missing):** For cases where the case_onset_interval is not null but the symptom_status is missing, it is recommended to update the symptom_status to 'Symptomatic'. This change assumes that the presence of an onset interval implies symptoms were observed.

7. References

[1] *COVID-19 Case Surveillance Public Use Data with Geography | Data | Centers for Disease Control and Prevention* (2024). https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

[2] *Healthcare workers* (2020). https://www.cdc.gov/coronavirus/2019-ncov/hcp/non-us-settings/overview/index.html

8. Appendix

8.1 Continuous Features

Descriptive Statistics

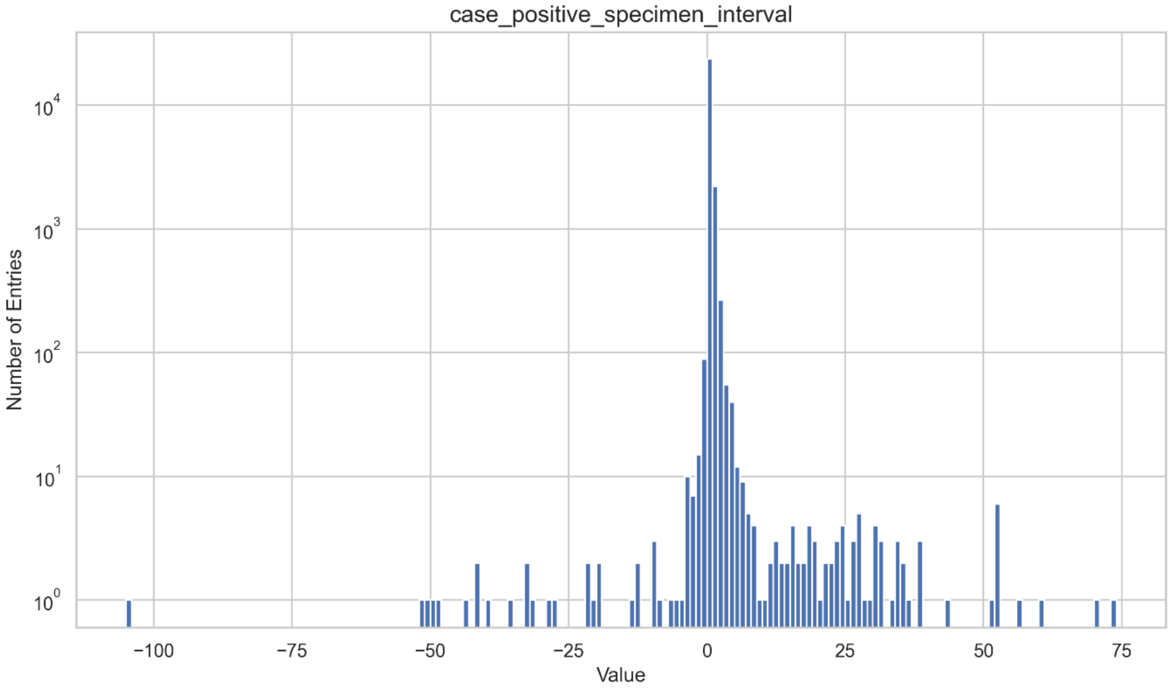
	count	mean	std	min	25%	50%	75%	max
case_positive_specimen_interval	26319	0.1663057	2.131756	-105	0	0	0	73
case_onset_interval	21287	-0.044581	1.807788	-107	0	0	0	59

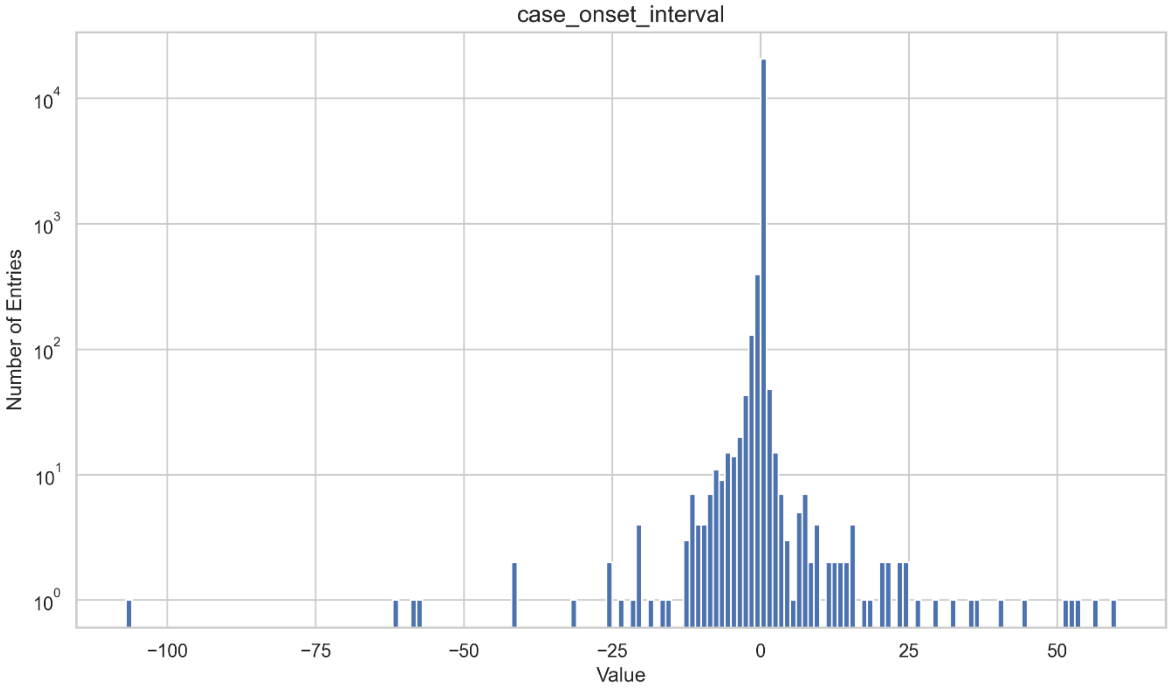
8.2 Categorical Features

Descriptive Statistics

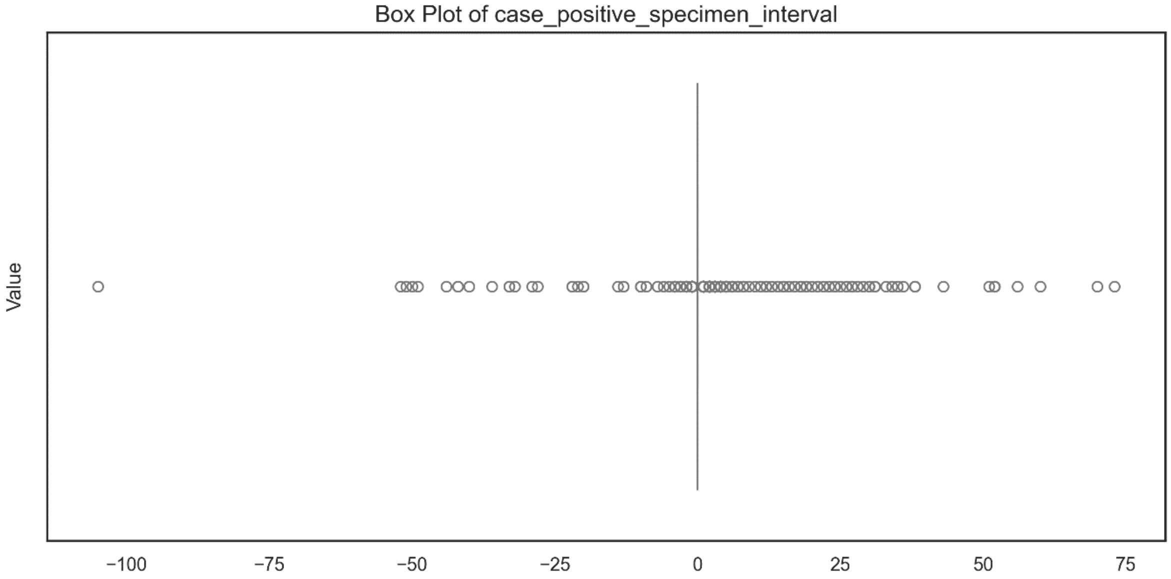
	count	unique	top	freq
res_state	50000	50	NY	5516
state_fips_code	50000	50	36	5516
res_county	47195	953	MIAMI-DADE	969
county_fips_code	47195	1358	12086	969
age_group	49637	5	18 to 49 years	20353
sex	48916	4	Female	25536
race	43810	8	White	30274
ethnicity	43257	4	Non-Hispanic/Latino	29602
process	50000	11	Missing	45399
exposure_yn	50000	3	Missing	43068
current_status	50000	2	Laboratory-confirmed case	42199
symptom_status	50000	4	Symptomatic	22511
hosp_yn	50000	4	No	25403
icu_yn	50000	4	Missing	39072
death_yn	50000	2	No	40000
underlying_conditions_yn	4114	2	Yes	4047

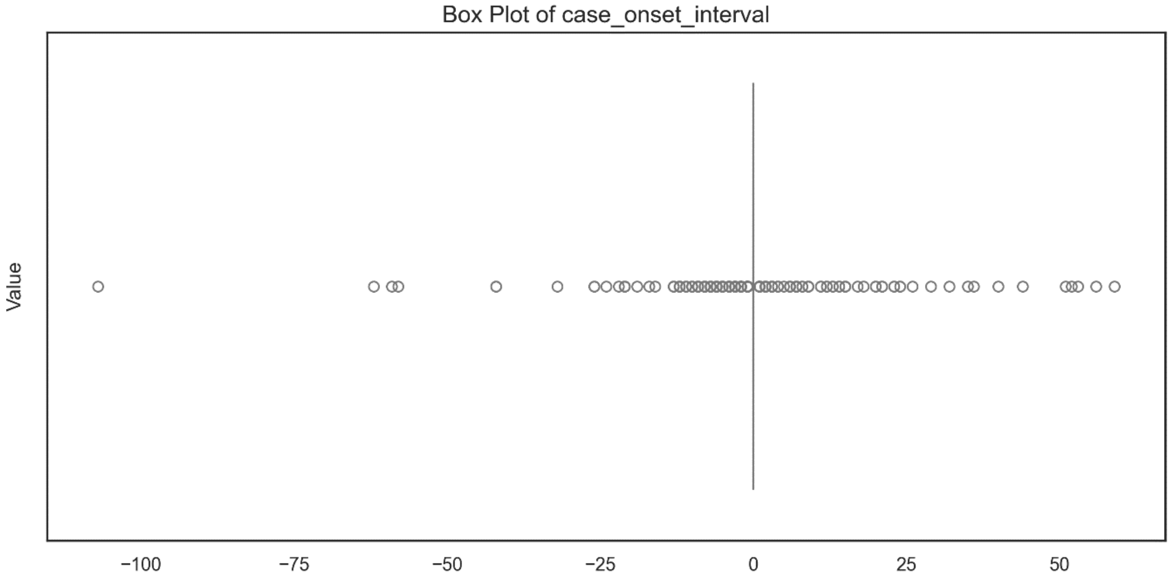
8.3 Continuous Histograms





8.4 Continuous Box Plots





8.5 Categorical Bar Plots

See accompanying pdf categorical_summary_sheet.pdf for categorical bar plots