Data Quality Report - Initial Findings

1. Overview.

This report will outline the initial findings based on the cleaned dataset (cleaned_data.csv), which has been generated from the CSV file ‘covid19-cdc-23222213.csv’. This report serves as a summary of the data and the issues contained in aforementioned data, as well as how they will be addressed. An appendix is provided at the end of this report, which includes terminology and assumptions, as well as graphs generated from ‘cleaned_data.csv’.

From a cursory visual inspection, the data appeared to be straightforward and relatively easy to interpret. However, initial data analysis tasks revealed that there are underlying complexities and discrepancies that require detailed attention. The main issues observed were a high rate of data entries classified as ‘missing’, ‘nan’ or ‘unknown’. In addition, the dataset contained multiple duplicate entries, a substantial amount of outliers and high ca

2. Introduction

This dataset is a subset of the public data released by the Center for Disease Control and Prevention (CDC). This is de-identified individual case data, gathered via standardised forms. From this anonymised data, it is hoped that the CDC data can be used to build a data analytics solution for death risk prediction.

3. Summary.

This is a relatively large dataset, containing 5000 rows and 19 columns. Each patient entry contains a wide range of elements, including their demographics, the geographical location of the case, disease severity indicators (such as hospitalisation, ICU stay), the presence of underlying conditions and their death status. Further information on the columns of the dataset is available in appendix 9.3.

As part of the data review, CDC documentation was reviewed (available at the link below). The CDC note some unique issues within their dataset and explain the justification for these issues. Before moving on to discuss the results of the logical tests carried out in the initial analysis, it is important to briefly note the CDC points below:

- In order to maintain the privacy of patients, the CDC has redacted certain data elements. Low-frequency combinations of elements such as case month, geographical location and demographic elements could place the patient at risk of identification. Therefore, the CDC has suppressed certain data elements and re-coded them to NA.
- As per the CDC (https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/about_data), an initial logic check has already been completed for the date data. If an illogical date was provided, the CDC reviewed it with the reporting jurisdiction and temporarily sets the value to ‘null’.

Several tests were performed to check the logical integrity of the data. Multiple failures of the data were observed, as well as incidents of the data returning unexpected and/or illogical results. While on first assessment, some elements appear plausible or explainable, they may require further analysis.

4. Review Logical Integrity.

The tests carried out to analyse the logical integrity of the data are detailed below.

Test 1: Iterate through the columns to count the unique values.

- This test was performed to check whether the columns should be converted to an alternate form (categorical to continuous and vice versa).
- This test ‘passed’, in that the values displayed were in line with what should be expected. Some additional reading to understand the cultural context of the dataset (the specific meaning of FIPS codes and their reuse) was required. As this is not specifically of use to this report, it is not included - however, that learning is detailed in the Jupyter notebook, should it be of interest.

Test 2: Checking the total count of values in each column, then summing the missing data of each feature.

- This test revealed significant issues in the dataset. Multiple columns were revealed to have high rates of missing data. The specific values can be perused below.
- res_county: 2830, county_fips_code: 2830, age_group: 387, sex: 1095, race: 6205, ethnicity: 6798, case_positive_specimen_interval: 23802, case_onset_interval: 28678, underlying_conditions_yn: 45952.

Test 3: Checking for duplicated columns using the duplicate() method.

- No duplicate columns found.

Test 4: Checking for duplicate rows, using the duplicate() method.

- 4675 rows containing duplicate data were found.
- The decision was made to keep the first instance of each duplicated piece of data and drop the others.

5. Review Continuous Features.

There are four continuous features - case_positive_specimen_interval, state_fips_code, county_fips_code and case_onset_interval. Further explanation of these can be found in appendix 9.3., if required.

5.1. Descriptive Statistics.

The descriptive statistics chart can be viewed in appendix 9.7. of this report. The chart clearly indicates that all four features contain missing data. Observations on the generated chart have been separated by their respective columns.

Case_positive_specimen_interval:

- The minimum value -115 implies negative values in the column, which could be erroneous or show issues within the dataset.

- The max value of 111 appears unusually large.
- The standard deviation implies a large variance.
- The mean value is low and could indicate that the time between gathering the data and confirming the case was quite short.
- 46.52% of this column is missing data, which indicates that analysis of this variable is likely to be difficult to do in a reliable and valid manner.
- The cardinality of this column is 80, indicating 80 distinct intervals of returning a positive specimen.

Case_onset_interval:

- The mean value is slightly negative, which appears to mean that the cases were confirmed before symptoms began. Considering how widespread COVID testing was at the time of the data collection, this is not unexpected.
- The standard deviation is large, implying a wide range in the data.
- The minimum value (-88) is unexpected and probably implies some erroneous data entry.
- The maximum value (99) seems quite large.
- A cardinality of 64 is visible in the chart, indicating a relatively low range of intervals for case onset.
- At around 55.65%, the missing data rate in this column is high. As over half the data of this set is missing, that could severely limit the usability of the data.

State_fips_code:

- There is no missing data in this column, which is ideal as it renders all data available for analysis.
- With a cardinality of 52, it can be assumed that this dataset contains information from each state of the USA.

County_fips_code:

- This has a relatively low percentage of missing data (approximately 5.85%). This implies that the majority of the data is available.
- The cardinality of this column (1376) suggests that it covers a wide range of counties within the states.

5.2. Histograms.

The histograms generated from the dataset can be viewed in appendix 9.5. of this report. While some successful histograms have been generated, two of them are not returning useful data. As can be seen in the appendix, two of the graphs are generating one solid column that projects directly up from 0 on the y-axis. This shows that how the data is being processed is problematic.

Both of these columns have a low count of unique values (case_positive_specimen_interval has 80 and case_onset_interval has 64 - see appendix 9.2. for individual column value count). It is possible that the bin width for the graph is too large or too small for this histogram to generate correctly with the set.
The successfully generated histograms are discussed below. It must be noted that both would have benefited from having labelled axes. The bin width for both is appropriate,

rendering the data relatively easy to read. Both histograms serve as a way to visualise the completeness of the dataset and the suitability of it for further processing.

state_fips_code:

- All values appear present on the histogram, showing that this is a complete data set.
- The distribution has a moderate left skew, with fewer values appearing towards the right of the graph.
- There is a mild outlier on the rightmost side of the graph.

county_fips_code:

- This appears to be a complete graph, with all values present on the histogram.
- The distribution is relatively symmetrical, albeit with a slight skew to the left.

5.3. Box plots.

The box plots for these continuous features can be viewed in appendix 9.5. of this report. The box plots returned varying results. While two box plots are visualising correctly, two have collapsed and are returning a multitude of dots in a straight line. Both of these dataset columns contain a limited range of data values (as per the previous section on the histograms), which can cause this collapsed plot. It may be better to visualise this data in another way.

Comments on the successful box plots are detailed below.

county_fips_code:

- The median of this box plot is relatively high, appearing to be approximately 3,500. The first quartile begins at approximately 2000 and the third quartile is at approximately 3800.
- Therefore, this dataset is positively skewed with a concentration of values towards the higher end of the dataset. This may indicate that there are more counties with higher FIPS codes, or perhaps that people resident in these counties are presenting with confirmed COVID cases for unknown reasons (e.g. population density).

state_fips_code:

- The first quartile of this dataset begins at 20 and the third quartile terminates at just under 40. The median is approximately 35. The data is relatively symmetrically distributed.
- There are outliers present in the dataset, as can be seen from the presence of two separate dots above the box plot. While the majority of the data follows a relatively normal distribution, these outliers may skew results and must be kept in mind. The outliers may represent states with special characteristics or anomalies in their FIPS codes - however, they do not necessarily indicate errors in the dataset.

6. Review Categorical Features.

There are fifteen variables of categorical data - case_month, res_state, res_county, age_group, sex, race, ethnicity, process, exposure_yn, current_status, symptom_status, hosp_yn, death_yn, icu_yn and underlying_conditions_yn.

6.1. Descriptive Statistics.

The descriptive statistics for this data can be viewed in appendix 9.6. of this report. As there are fifteen separate variables, this report will not detail every instance visible in the chart. Some key points to remark on are below:

- The unique count of ‘res_state’ (52) and ‘res_county’ (962) indicate that the dataset covers a wide geographical across the USA.
- ‘Case_month’ spans 40 unique values, indicating that the dataset comprises values of 40 months. This indicates that this dataset represents a broad timespan across the pandemic.
- The 18-49 age group is most frequently represented in this dataset.
- With 23,087 cases, the majority of patients in this dataset are female (although it must be noted that this does not account for those patients whose sex data was missing or redacted).
- The majority of cases in this database were laboratory confirmed and at time of recording, the patients were symptomatic.
- The majority of patients were not hospitalised and did not die with COVID, as indicated by the high incidence of ‘No’ in those columns. The most frequent category in ‘icu_yn’ is Missing. While strictly speaking, we do not know if these patients were hospitalised, the fact that they were predominately not hospitalised would indicate that they were also not in the ICU.
- At first glance, the majority of patients had some form of underlying condition, as indicated by the ‘Yes’ value being most frequent in that column. However, there is an extremely high rate of missing values in this column (in excess of 91%).
- Demographic information such as age and sex is largely available, as indicated by the low count of missing data in those columns. However, race and ethnicity have higher percentages of missing data. This may impact the ability to analyse the impact of COVID on different racial and ethnic groups.
- With five distinct categories, age is separated into relatively few age bands. This may impact the possibility of how the patient’s age interplays with patient outcomes.

6.2. Bar Plots of Categorical Features.

The bar plots for this report can be viewed in appendix 9.4. of the report. As there are fifteen bar plots, the analysis will be kept appropriately brief. It must be noted that the bar plots almost universally show a high rate of ‘nan’, ‘missing’ and ‘unknown’ values, which cause the charts to skew and render the known data more difficult to read.

The bar plot for ‘case_month’ illustrates that the highest incidence of the data comes from January 2022. However, while the resulting frequency descendance does not occur in chronological order, it is notable that the highest five values all occur in winter months. This will be useful in further analysis of the dataset.

All states are not equally represented in the ‘res_state’ graph. This may indicate that people in certain regions were more at risk of contracting COVID and may also impact the potential for assessing the medical outcomes of people from areas with less documentation.

The bar plot for ‘res_county’ was entirely unable to generate correctly and may need to be represented in an alternate format. As this column has a high cardinality, this is not surprising.

The vast majority of the process and exposure data is missing, which may hamper further analysis.

7. Action To Take.

These points will be reviewed in greater detail in the data quality plan. However, for the completeness of this report, I will briefly touch on them now.

High rate of missing/nan/unknown data:

- Use predictive analysis to model them (in cases where the lost data is relatively low).
- Explore multiple imputation techniques to generate data for cases where the lost data rate is high.

High rate of duplicate row data found:

- Much of this has already been corrected (which was done via the pandas drop() function to remove duplicate data and maintain the initial instance). Some further rows may need to be merged to consolidate the data.

Collapsed box plot correction:

- Display the data from the collapsed box plots in a different manner. Use of a strip plot or a density plot may be better suited to this narrow range of values.

Uni-line histogram correction:

- Adjust the bin width of the histograms to visualise the data in separate columns.

Visualisation of ‘res_county’ data:

- Aggregate the counties by corresponding state in order to achieve a readable chart.

8. References.

The dataset used in this report is an extract from the work of the Centers for Disease Control and Prevention (https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4).

Prof. Georgiana Ifrim provided a sample data quality report (titled 'COMP47350-sample_solution_DQR_DQP_DataExploration_DataPrep_CreditRiskPrediction', available in the learning materials for COMP47350). This was of assistance in structuring and understanding the requirements of this report.

The following link was of use in understanding box plots:
https://mathsathome.com/understand-and-compare-box-plots/

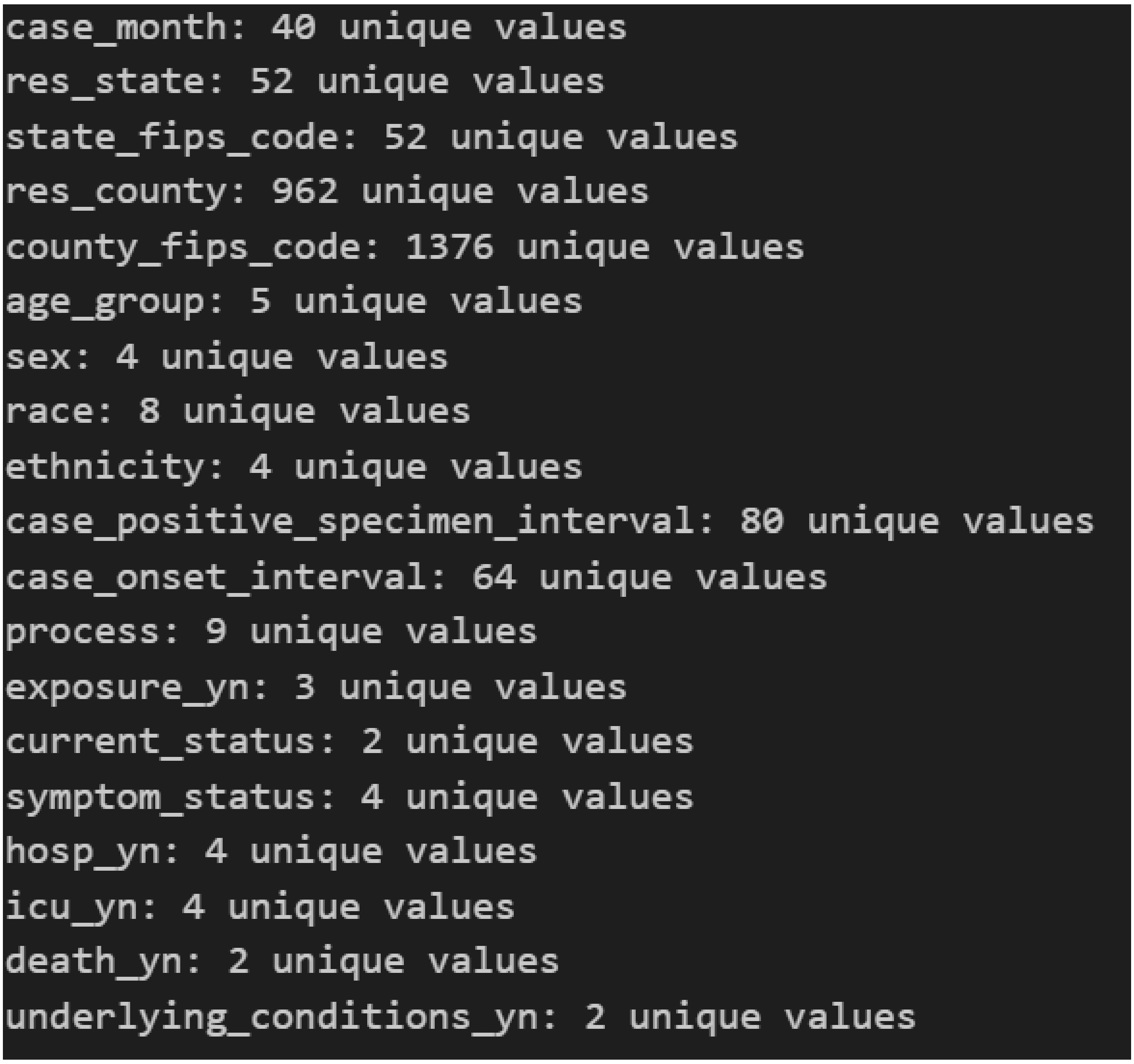
The following link was of use in understanding histograms:
https://www.labxchange.org/library/items/lb:LabXchange:10d3270e:html:1

9. Appendix.

9.1. Terminology and Assumptions.

- Categorical: variables that can take on a limited number of distinct values, used to classify data into specific groups.
- Continuous: variables that can take on an infinite number of values in a certain range.
- Nan: stands for Not A Number. A floating-point value used to represent undefined or unrepresentable numerical data.

9.2. Special Values.



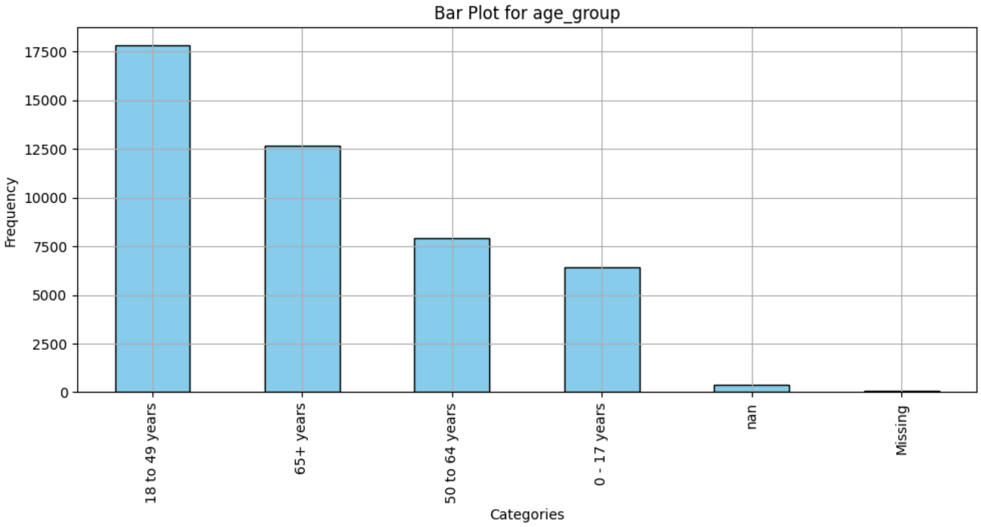
```
case_month: 40 unique values
res_state: 52 unique values
state_fips_code: 52 unique values
res_county: 962 unique values
county_fips_code: 1376 unique values
age_group: 5 unique values
sex: 4 unique values
race: 8 unique values
ethnicity: 4 unique values
case_positive_specimen_interval: 80 unique values
case_onset_interval: 64 unique values
process: 9 unique values
exposure_yn: 3 unique values
current_status: 2 unique values
symptom_status: 4 unique values
hosp_yn: 4 unique values
icu_yn: 4 unique values
death_yn: 2 unique values
underlying_conditions_yn: 2 unique values
```

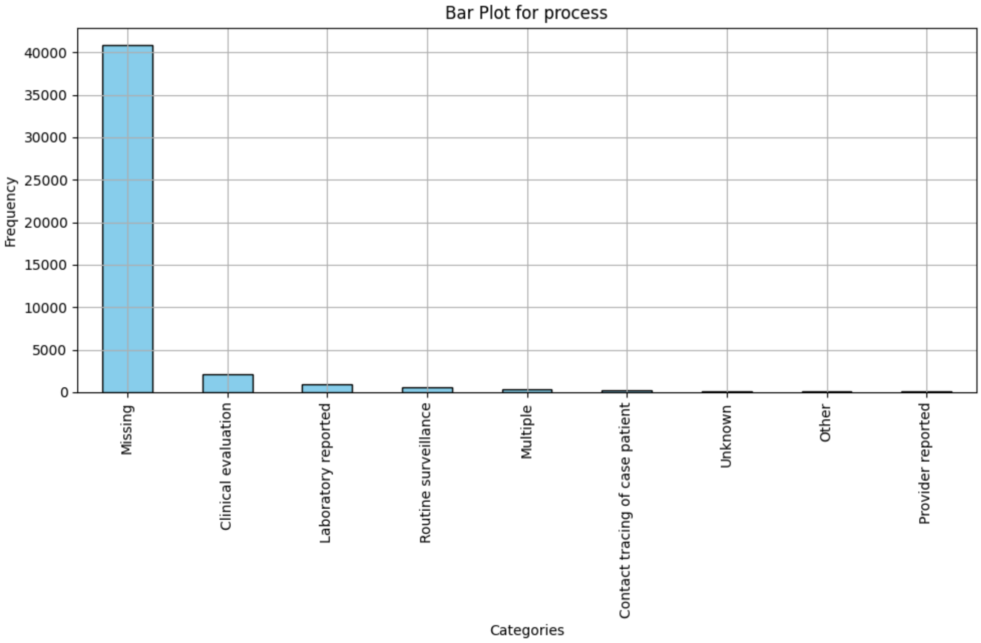
9.3. Data Column Values.

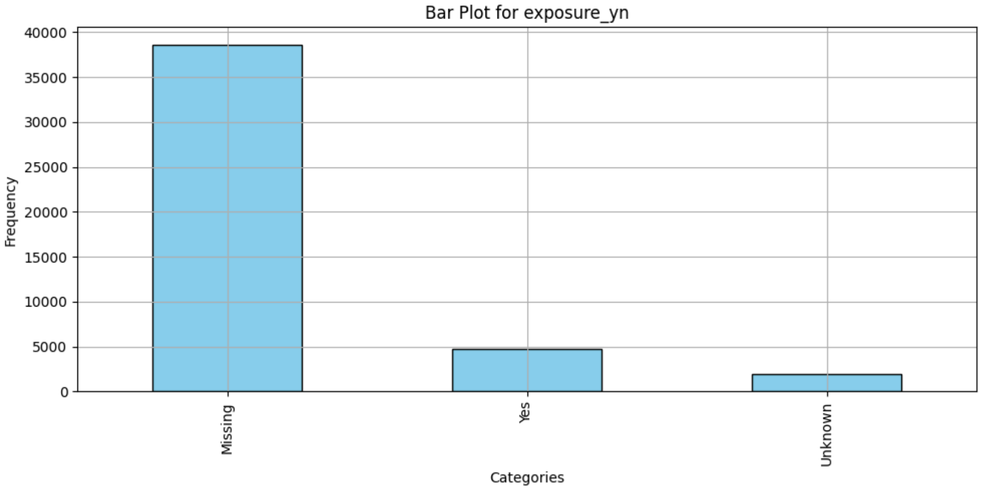
It should be noted that the values below detail the converted values as per 'cleaned_data.csv' and not the initial values of 'covid19-cdc-23222213.csv'.

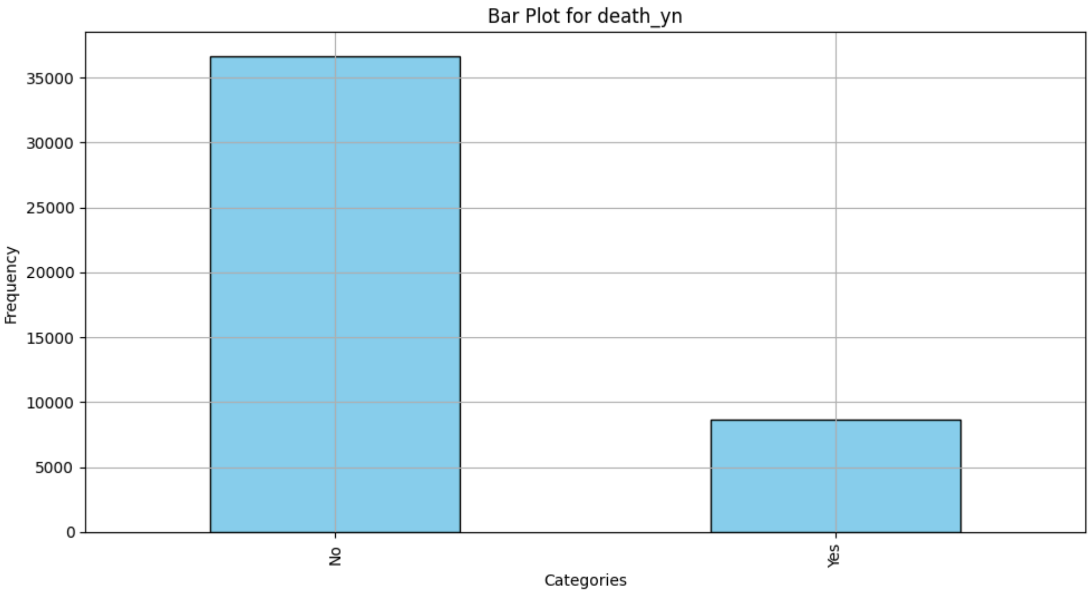
Column Name	Type	Description
case_month	Categorical	This column details the date (in year-month format) that the patient information form was completed.
res_state	Categorical	The state (in two-letter abbreviation) of the case.
state_fips_code	Continuous	The unique numerical identifier for the state.
res_county	Categorical	The specific state county(subdivision) of the case.
county_fips_code	Continuous	The unique numerical identifier for the county.
age_group	Categorical	The age band (of four options) that the patient falls into.
sex	Categorical	This notes whether the patient is female or male.
race	Categorical	This notes the race of the patient.
ethnicity	Categorical	This indicates the ethnicity of the patient.
case_positive_specimen_interval	Continuous	Weeks between earliest date and date of first positive specimen collection.
case_onset_interval	Continuous	Weeks between earliest date and date of symptoms commencing.
process	Categorical	How case was identified (contact tracing, clinical evaluation etc).
exposure_yn	Categorical	For 14 days prior to illness onset, did the patient have exposures (travel, contact with confirmed case etc).
current_status	Categorical	Current patient status (confirmed/probable case)
symptom_status	Categorical	Asymptomatic/symptomatic.
hosp_yn	Categorical	Whether the patient was hospitalised.
icu_yn	Categorical	Whether the patient was in ICU.
death_yn	Categorical	Whether the patient survived.
underlying_conditions_yn	Categorical	Did the patient have a known underlying medical condition or risk behaviour.

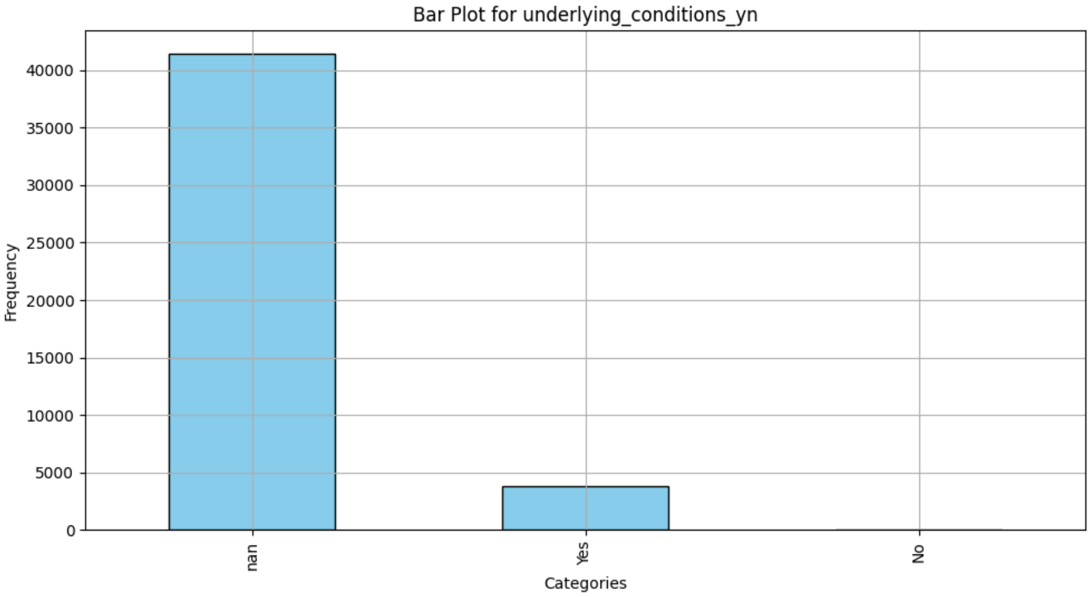
9.4. Continuous & Categorical Features (Bar Plots)

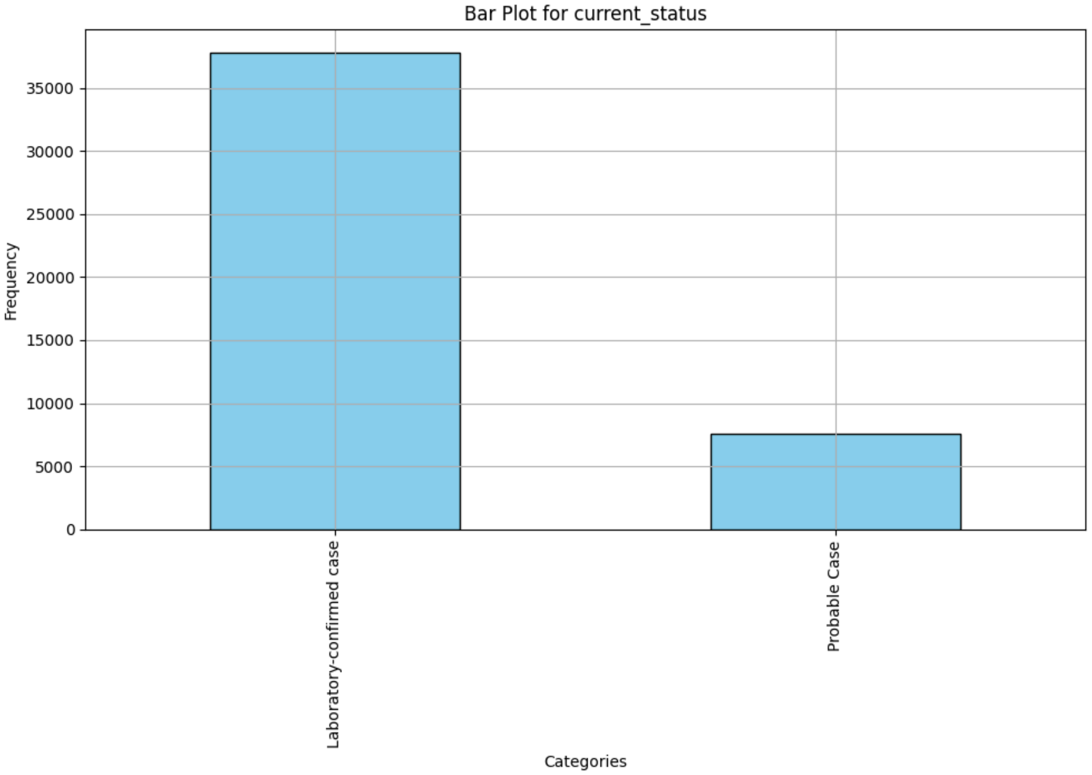


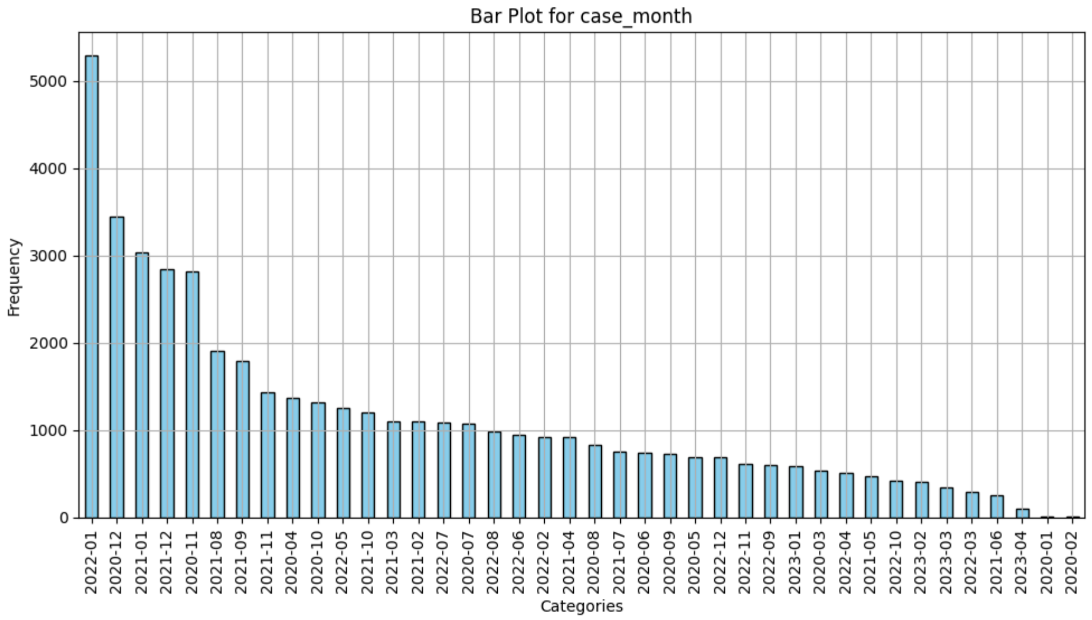


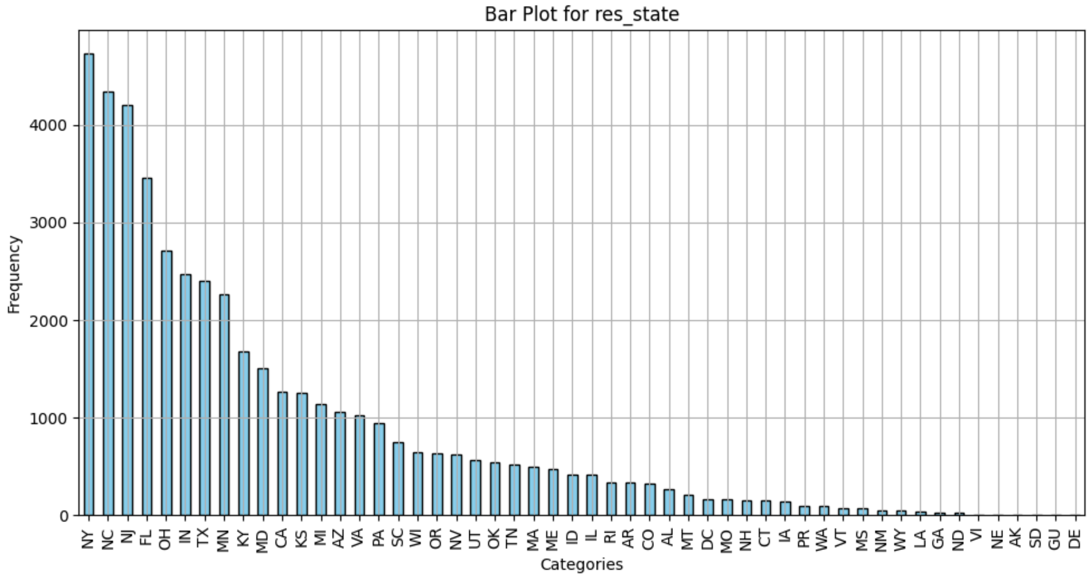


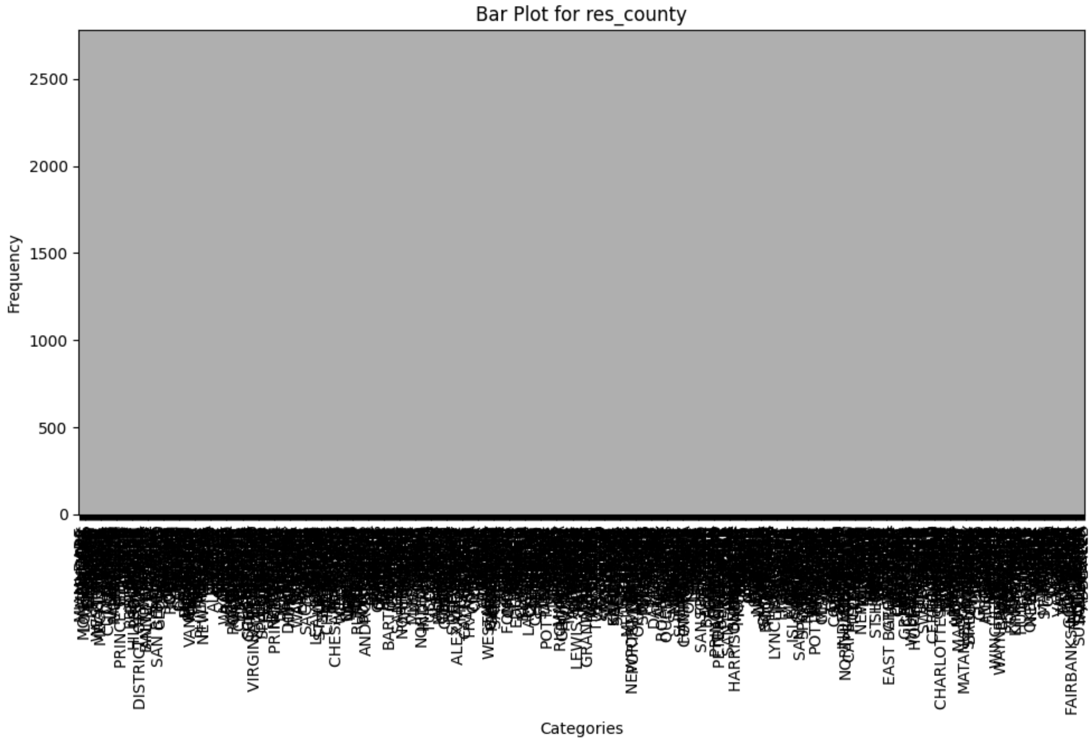

Bar Plot for death_yn
Frequency
0
5000
10000
15000
20000
25000
30000
35000
No
Yes
Categories

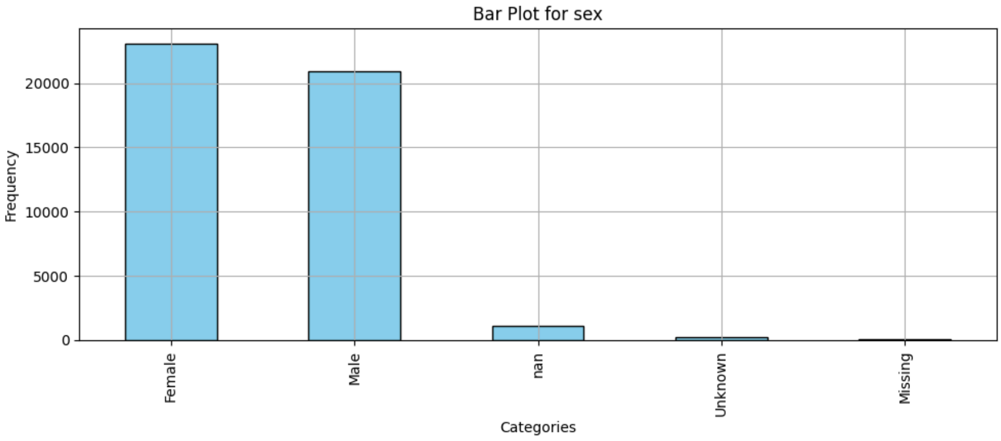

Bar Plot for underlying_conditions_yn
Frequency
0
5000
10000
15000
20000
25000
30000
35000
40000
nan
Yes
No
Categories

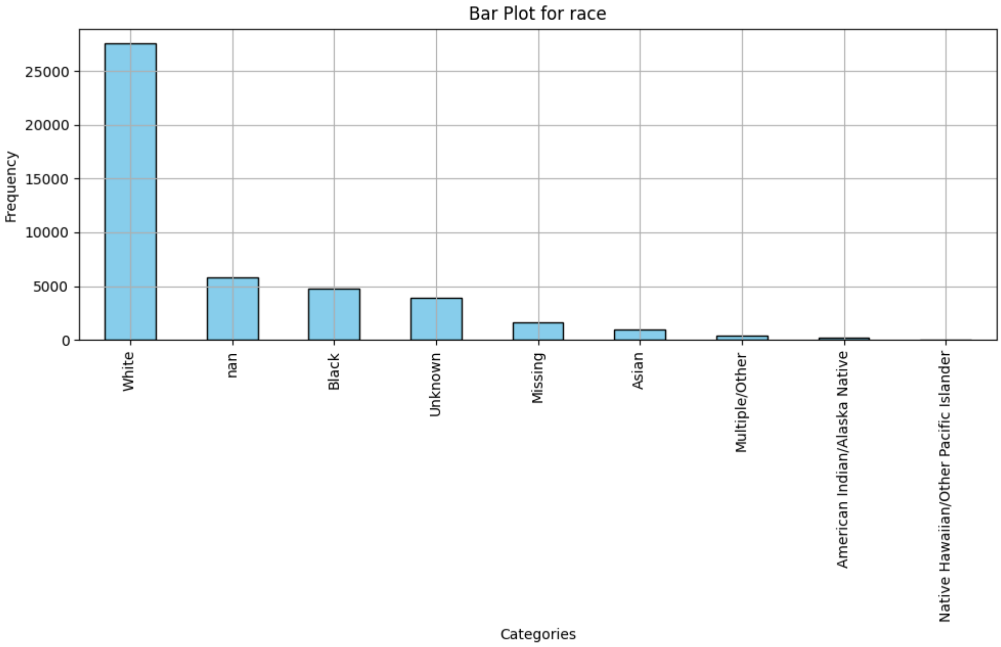

Bar Plot for current_status
Frequency
0
5000
10000
15000
20000
25000
30000
35000
Laboratory-confirmed case
Probable Case
Categories

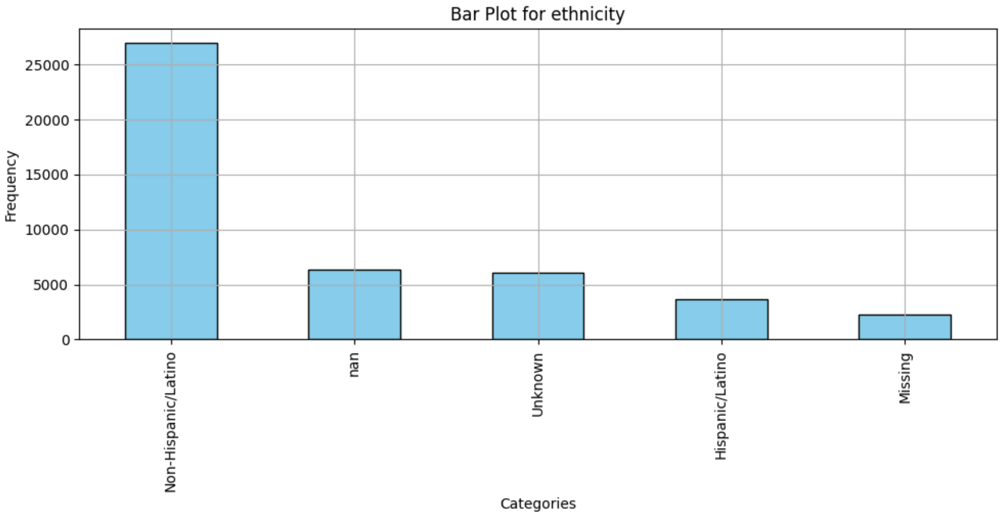

Bar Plot for case_month

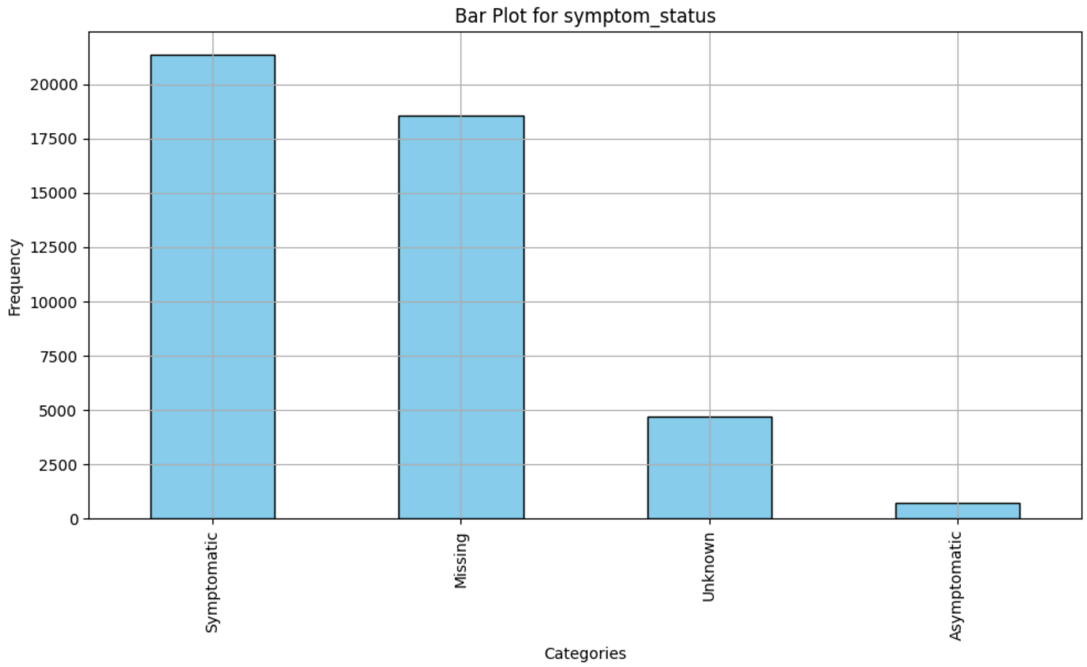

Bar Plot for res_state

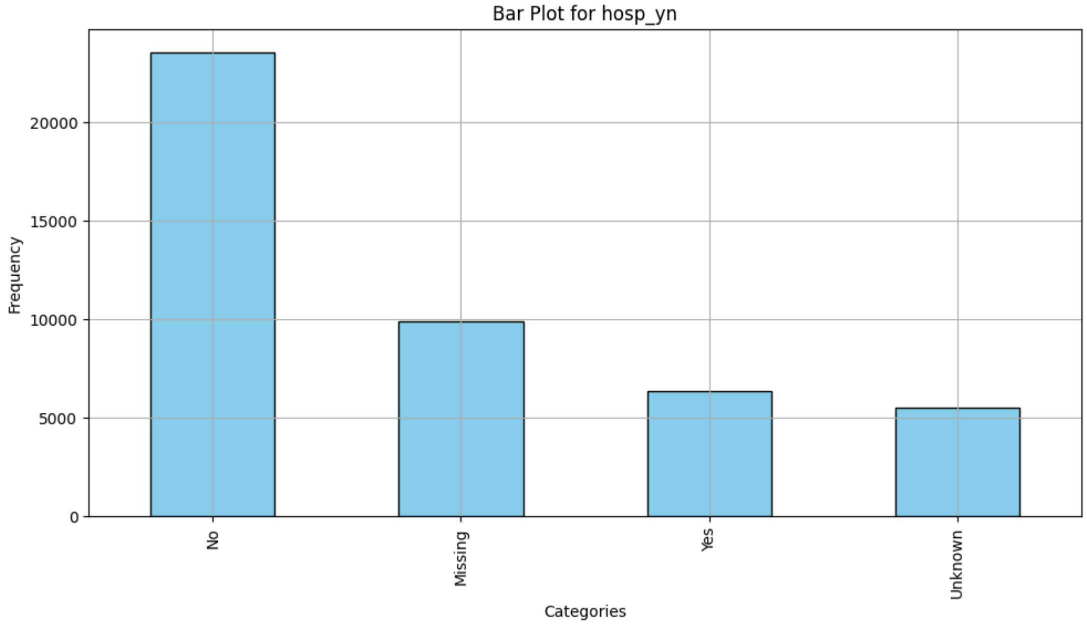

Bar Plot for res_county

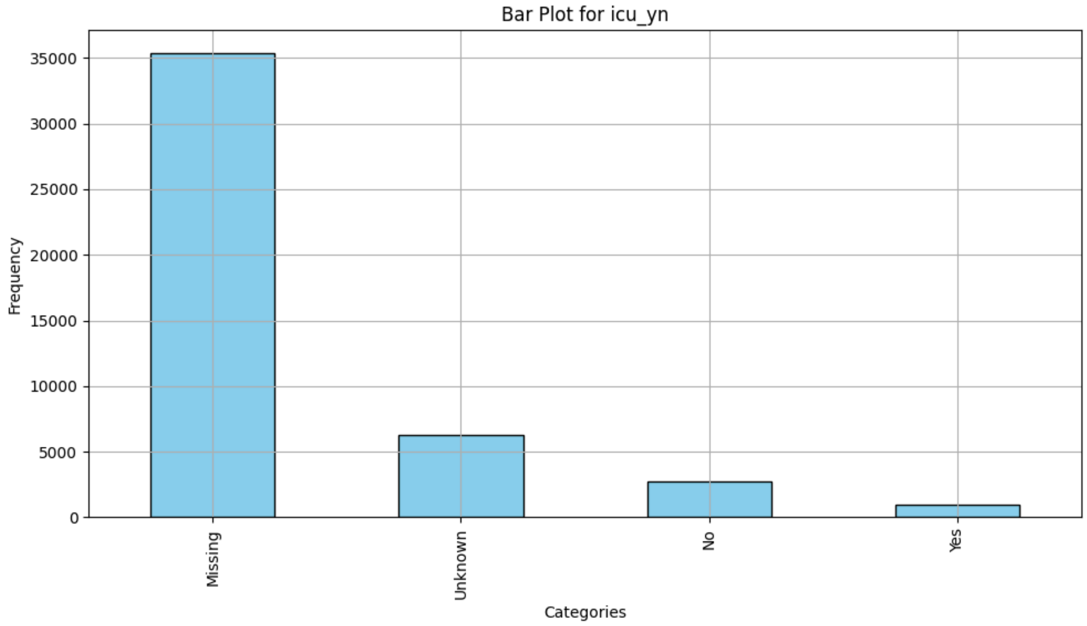

Bar Plot for sex
Frequency
0
5000
10000
15000
20000
Female
Male
nan
Unknown
Missing
Categories


Bar Plot for race
Frequency
0
5000
10000
15000
20000
25000
White
nan
Black
Unknown
Missing
Asian
Multiple/Other
American Indian/Alaska Native
Native Hawaiian/Other Pacific Islander
Categories

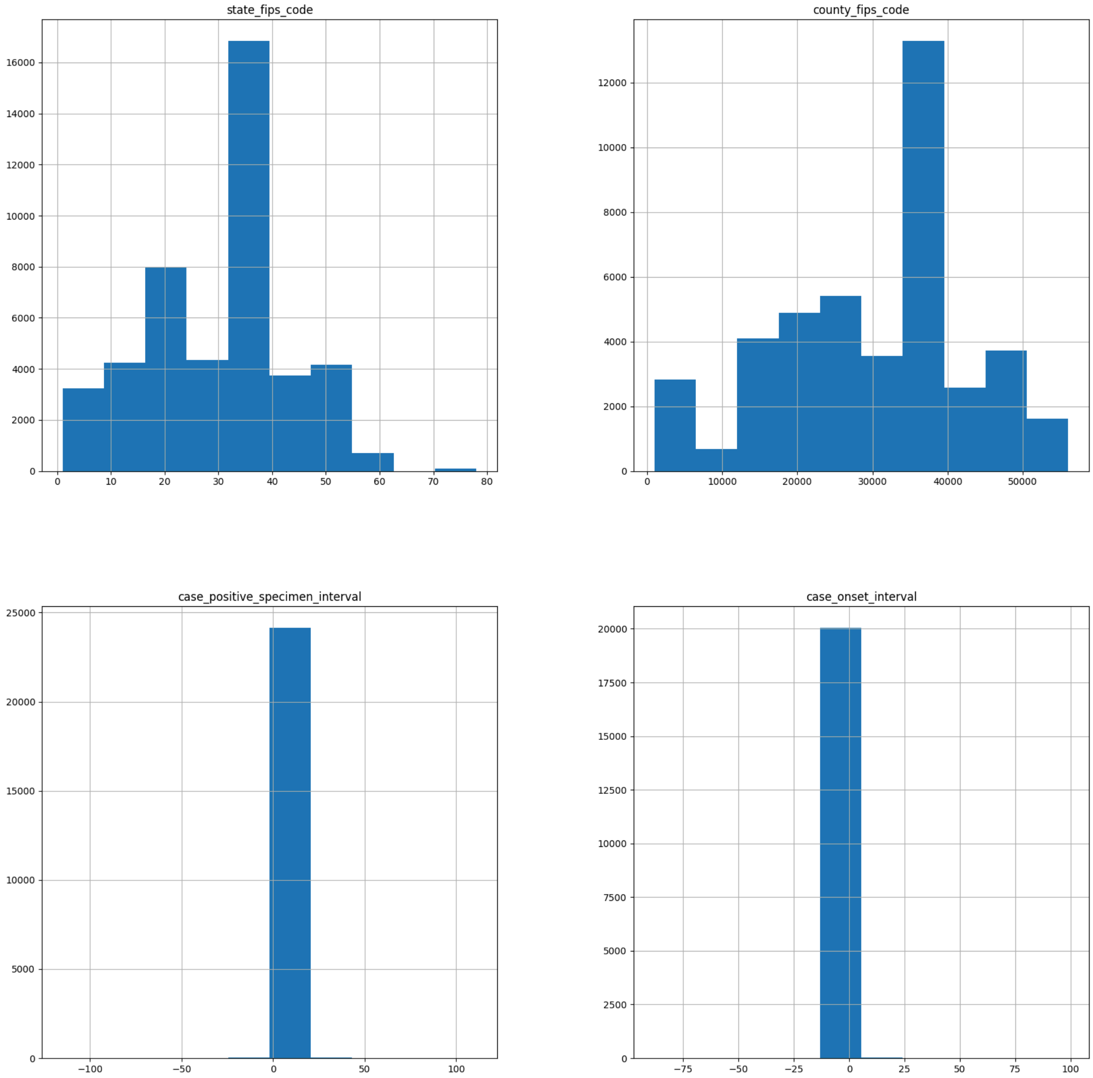

Bar Plot for ethnicity
Frequency
0
5000
10000
15000
20000
25000
Non-Hispanic/Latino
nan
Unknown
Hispanic/Latino
Missing
Categories

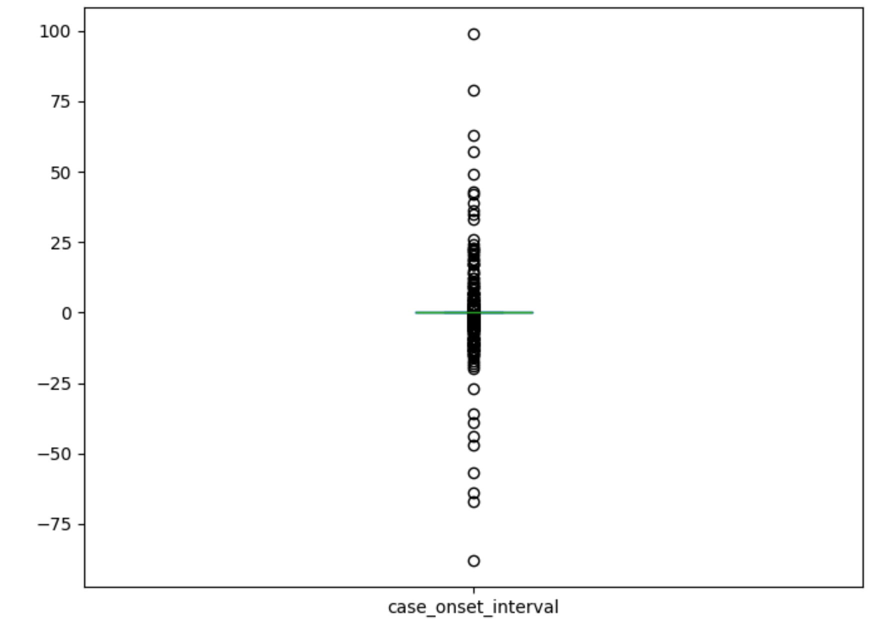

Bar Plot for symptom_status
Frequency
0
2500
5000
7500
10000
12500
15000
17500
20000
Symptomatic
Missing
Unknown
Asymptomatic
Categories

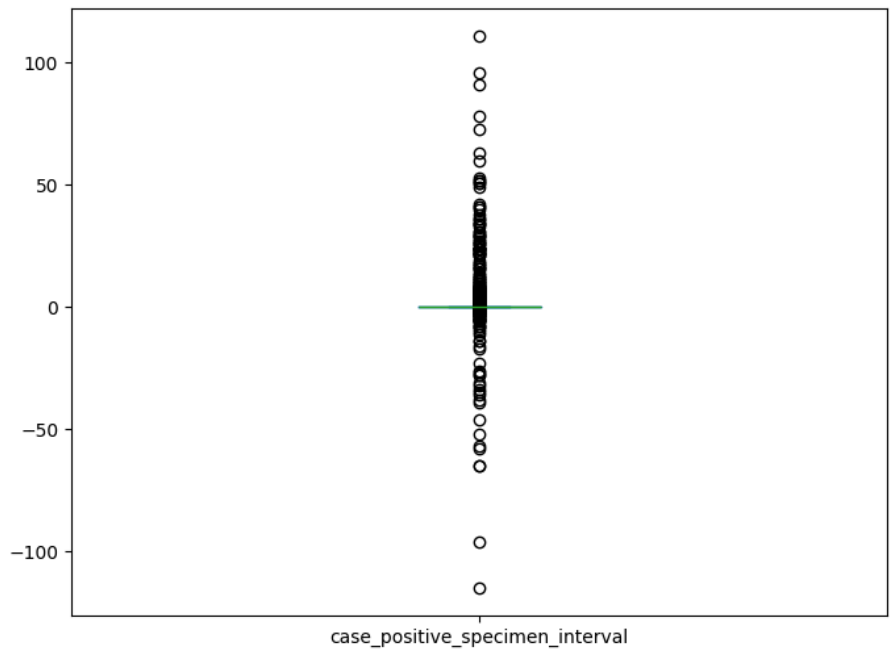

Bar Plot for hosp_yn
Frequency
0
5000
10000
15000
20000
No
Missing
Yes
Unknown
Categories

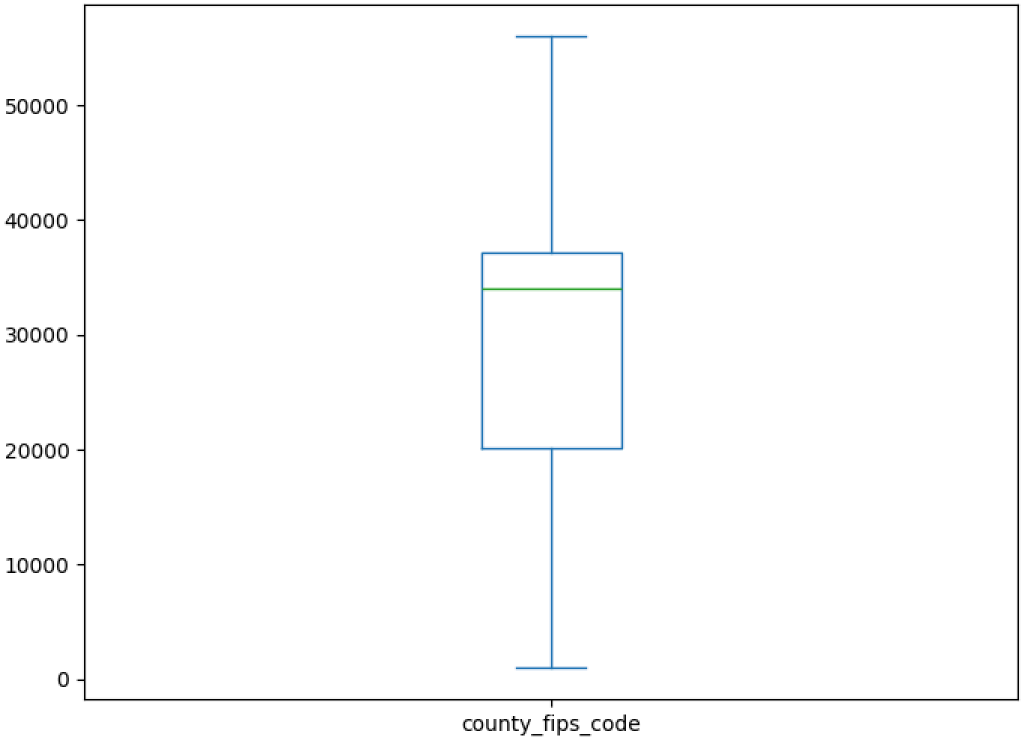

Bar Plot for icu_yn
Frequency
0
5000
10000
15000
20000
25000
30000
35000
Missing
Unknown
No
Yes
Categories

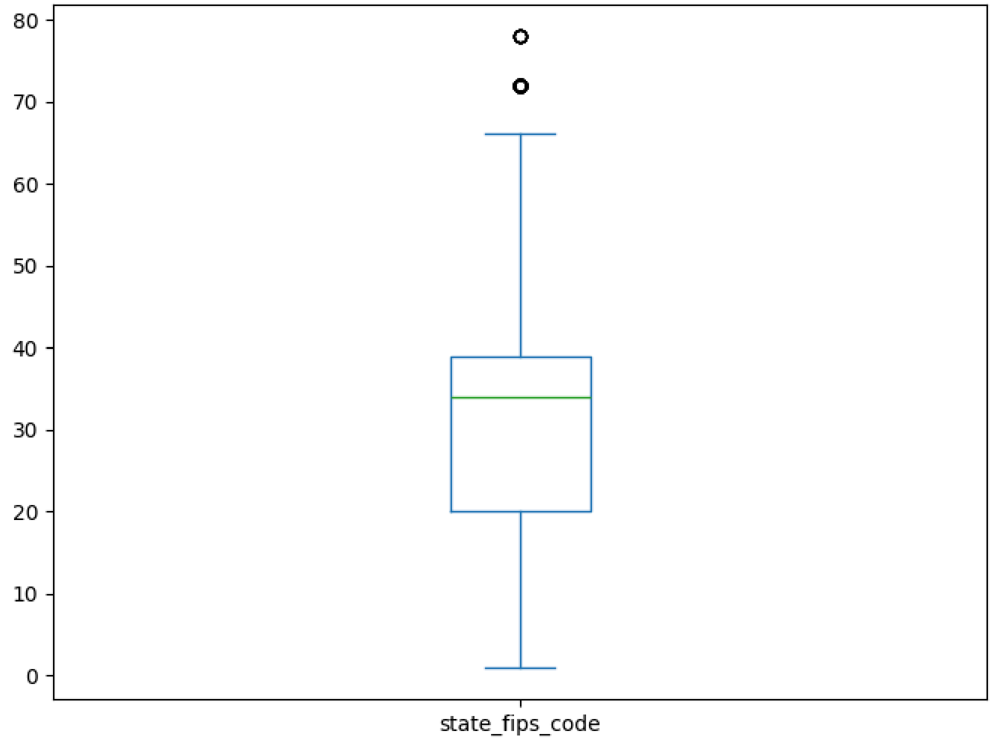
9.5. Box Plots & Histograms.



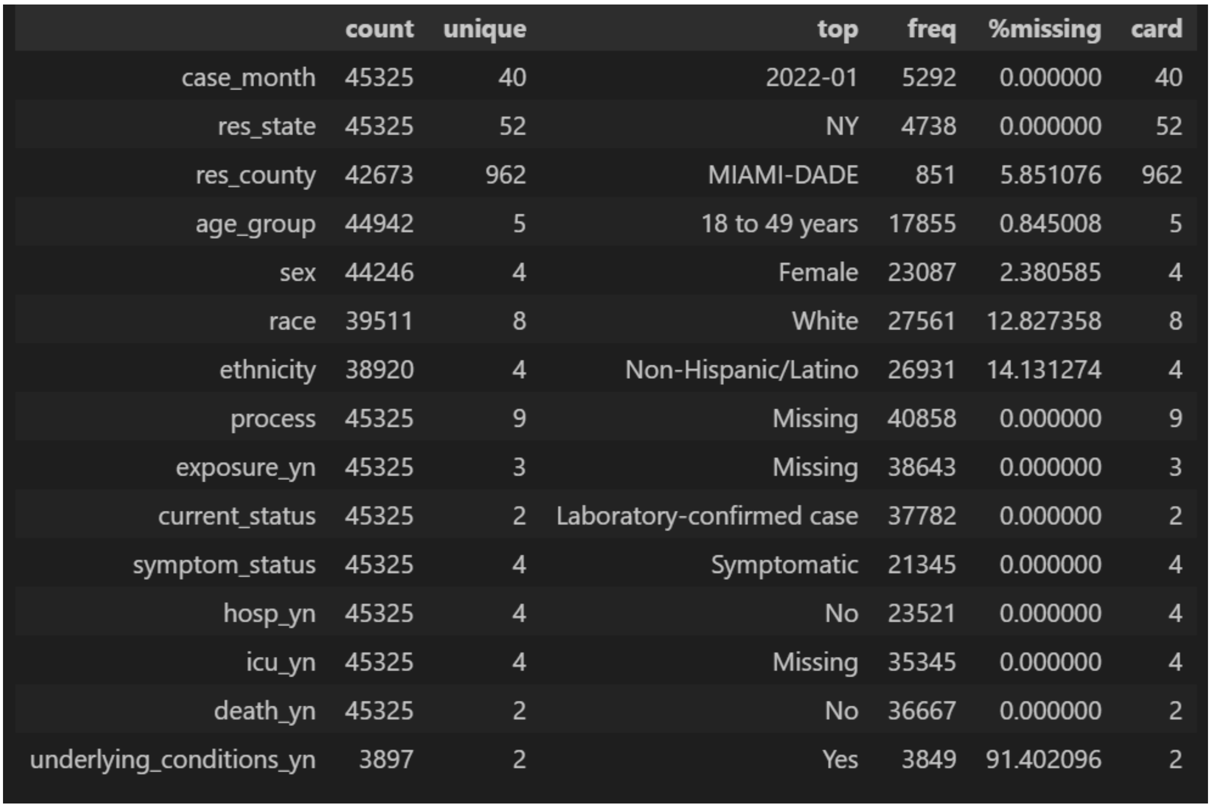

100
75
50
25
0
−25
−50
−75
case_onset_interval


100
50
0
−50
−100
case_positive_specimen_interval


50000
40000
30000
20000
10000
0
county_fips_code

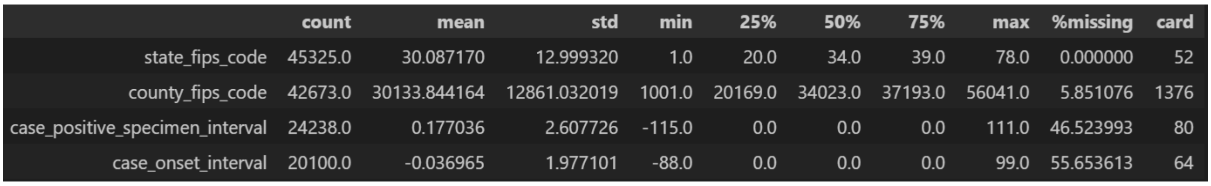

80
70
60
50
40
30
20
10
0
state_fips_code

9.6. Descriptive Statistics for Categorical Features.



	count	unique	top	freq	%missing	card
case_month	45325	40	2022-01	5292	0.000000	40
res_state	45325	52	NY	4738	0.000000	52
res_county	42673	962	MIAMI-DADE	851	5.851076	962
age_group	44942	5	18 to 49 years	17855	0.845008	5
sex	44246	4	Female	23087	2.380585	4
race	39511	8	White	27561	12.827358	8
ethnicity	38920	4	Non-Hispanic/Latino	26931	14.131274	4
process	45325	9	Missing	40858	0.000000	9
exposure_yn	45325	3	Missing	38643	0.000000	3
current_status	45325	2	Laboratory-confirmed case	37782	0.000000	2
symptom_status	45325	4	Symptomatic	21345	0.000000	4
hosp_yn	45325	4	No	23521	0.000000	4
icu_yn	45325	4	Missing	35345	0.000000	4
death_yn	45325	2	No	36667	0.000000	2
underlying_conditions_yn	3897	2	Yes	3849	91.402096	2

9.7. Descriptive Statistics for Continuous Features.



	count	mean	std	min	25%	50%	75%	max	%missing	card
state_fips_code	45325.0	30.087170	12.999320	1.0	20.0	34.0	39.0	78.0	0.000000	52
county_fips_code	42673.0	30133.844164	12861.032019	1001.0	20169.0	34023.0	37193.0	56041.0	5.851076	1376
case_positive_specimen_interval	24238.0	0.177036	2.607726	-115.0	0.0	0.0	0.0	111.0	46.523993	80
case_onset_interval	20100.0	-0.036965	1.977101	-88.0	0.0	0.0	0.0	99.0	55.653613	64