Name: Yu Feng
Student No.: 22200055

Data Quality Report – Initial Findings

1. Overview

This report presents initial findings on the cleaned covid19 cdc dataset (cleaned-covid19-cdc-1_2-22200055.csv). The report will provide a summary of the dataset and detail any data quality issues that were identified during the analysis. Additionally, the report will explain how these issues will be addressed. The appendix of the report contains background information about the dataset, which includes definitions, assumptions, and explanations of the terms used. It also includes visual representations of the data, such as feature summaries, histograms, and charts. The data cleaning process is explained in detail, which includes the removal of rows or features with no data or redundant features, as well as replacing missing values with a reasonable estimate.

At first glance, this dataset appears to be missing a large amount of data and contains many unknowns and null values. The main problem observed is that the data for the continuous features are more anomalous, with many outliers and unreasonably negative values. In addition, there is a constant column containing only one constant value and redundant columns that need to be removed.

2. Summary

In order to clean the data, different ways of dealing with these null values are taken depending on the situation.

First, duplicate rows have been removed. Assume one row represents a single patient's information. When there are duplicate rows for the same patient, it could lead to inaccurate insights about covid death causes. Duplicate rows can also create inconsistencies in the data. When more than two rows are supposed to represent the same thing but have slightly different values, it can lead to confusion and errors in the following analysis.

Besides, unknown and missing values are transformed into null values to facilitate the calculation of the percentage of null values (%missing). After completing the conversion, the feature exposure_yn has only one constant value 'Yes', so we remove this constant column.

It is observed that res_state and state_fips_code represent the same feature, so the state_fips_code feature was removed from them for subsequent analysis.

For the features with different %missing, we take the corresponding way to process them to eliminate these null values. When we look into the details for the res_state feature, we find that the 0.005% missing it shows corresponds to only one row, which also contains null values for many other features, so we choose to delete this row. For features with a %missing of 90% or more, we usually choose to delete the feature. Whether we can delete the feature directly or not also depends on the impact of the feature on the target feature death_yn, for example, underlying_conditions_yn in this case we think the feature has a great impact on the target feature, so we decide to keep the feature and add a new category called " unknown" to fill these

null values. And we choose to remove the features icu_yn and process, which are also up to 90% because they seem to have only a minor impact on the target feature. For the features res_county and county_fips_code, we found that one res_county corresponds to several different county_fips_code, probably because the code of county is more specific to the location than the county name. We also found that some county names do not have corresponding county codes, which means patients living in these counties with NaN values for county code, so the rows where those county names do not contain any county codes were removed. Before imputing their null values, I chose to group all counties and then take the most frequent value of all county codes to fill in the null values, and fill in the null values of county names accordingly. For feature ethnicity, we also take this impute approach to fill the null values and replace the null values with the most frequent Non-Hispanic/Latino values.

For features age_group, race, sex, symptom_status, hosp_yn, and underlying_conditions_yn we derived a new feature from these features respectively with missing values. Because of these features, it is not appropriate to take other impute ways to deal with these null values, which would lead to large errors in the data.

For the two continuous features, although these two features have too many missing values (47.07% and 52.88%), I impute these null values instead of just dropping them, because these continuous features will largely influence the following analysis. From the histogram we can see that the most frequent value 0 occupies a large percentage of the total, so would replace the null values with 0 for both of them. Also, there are some special values for these two features, which is a negative value, because the interval should not be negative as neither the onset date nor the positive tested date should be earlier than the data first collected. For these negative values, I presume the dates were recorded reversely, so I just keep these negative values as absolute ones instead of dropping them.

There are a large number of outliers throughout the feature set. however, from initial indications, these values seem plausible but should be investigated further.

3. Review Continuous Features

3.1. Descriptive Statistics

There are 2 continuous features, case_positive_specimen_interval and case_onset_interval. Feature case_positive_specimen_interval has 47.07% missing values and case_onset_interval has 55.22% but I will not drop these two features because they will largely influence the following analysis. From the histograms we can see that the most frequent value 0 occupies a large percentage of the total, so would replace the null values with 0. This method was applied to both of these two continuous features. Many of these features have outliers. All seem plausible but should be investigated further.

3.2. Histograms

All histograms can be found in the appendix as a summary sheet. Individual plots can be found in the accompanying notebook. As we can see from the original histograms, most of the data in the histogram of these two features are concentrated around 0 weeks, with the number reaching nearly 10,000, while the other values distributed around 0 can only be seen in the figure with only a very slight height. In order to make their differences visible in the histograms, I added the parameter 'log=True' to each of the graphs. in the next graphs, looking at the overall picture, we find that the feature case_positive_specimen_ interval contains more values other than those around 0, and the number of positive values is

significantly more than the number of negative values. In other words, the feature case_onset_interval contains fewer values other than near 0, but the difference between the number of positive and negative values is not significant.

3.3 Box plots

All boxplots can be found in the appendix as a summary sheet. Individual plots can be found in the accompanying notebook. Again, outliers will be investigated further but no immediate action is expected. As we can see from the box plots, both of them appear as a line with many dots, which may indicate that the dataset has little variation or that the variation is concentrated around a single value. It may also indicate that the data points are too widely dispersed to be displayed in a box plot format when there are many outliers or if the distribution of the data is highly skewed.

For the irregular negative values, would convert these values to positive ones as this feature is calculated as (pos_date - earliest_date)/7, negative values mean pos_date is earlier than earliest_date which is unreasonable, presume that these dates were reversed in this case.

For outliers, will just keep it as is in this feature because: Weeks for onset since the earliest date: Max is 64 weeks = 1 year and 3 months. The data was lately collected from 2020.01 and the last updated date was 2022.10, with an interval of 2 years and 9 months. Weeks for positive cases since the earliest date: Max is 63 months = 1 year and 3 months. So these outliers are plausible.

4. Review Categorial Features

4.1. Descriptive Statistics

There are 17 Categorial features in the dataset, 1 of which is the target and will not be evaluated here. The 16 remainings are ‘case_month’, ‘res_state’, ‘state_fips_code’, ‘res_county’, ‘county_fips_code’, ‘age_group’, ‘sex’, ‘race’, ‘ethnicity’, ‘process’, ‘exposure_yn’, ‘current_status’, ‘symptom_status’, ‘hosp_yn’, ‘icu_yn’, ‘underlying_conditions_yn’.

The data cleaning process involves identifying and handling null values in the dataset. First, missing values are transformed into null values to calculate the percentage of null values in each feature. The constant feature 'exposure_yn' is removed, and the redundant feature 'state_fips_code' is removed as it represents the same feature as 'res_state'. For features with different percentages of missing values, appropriate methods are chosen to handle them. The 'res_state' feature with a low percentage of missing values is cleaned by deleting the row containing null values. For features with a high percentage of missing values, the decision to delete or keep the feature depends on its impact on the target feature 'death_yn'. The 'icu_yn' and 'process' features are removed as they have a minor impact on the target feature. The 'res_county' feature is grouped and imputed with the most frequent value of all county codes, and the 'ethnicity' feature is imputed with the most frequent Non-Hispanic/Latino values. For features 'age_group', 'race', 'sex', 'symptom_status', 'hosp_yn', and 'underlying_conditions_yn', new features are derived to handle the missing values. Overall, the data cleaning process involves identifying and handling null values using appropriate methods for each feature.

When checking for invalid or inconsistent values, I found that the cardinality of 'county_fips_code' is large, at 1208.0, but this is reasonable because the county code represents a more detailed location than the county name, and a res_county corresponds to

several different county_fips_codes.

After plotting a box to compare feature ‘state_fips_code’ and ‘res_state’, I foung they describe the same thing and feature ‘state_fips_code’ was dropped as feature ‘res_state’ would be more readable.

4.2. Histograms

The histograms can be found in the accompanying pdf.

5. Action to take

7 main actions will be taken to process missing values or irregular values, summarised below;

- Logical integrity:
  - Conduct logic tests of features based on given background information.
  - Drop rows which contain a logical error.
- Missing and unknown values:
  - Check which features contain these values.
  - Convert these values containing "Missing" and "Unknown" to NaN for proper subsequent analysis.
- Mode imputation:
  - See where imputation is possible.
  - Replacing missing values with the mode of the available values for that variable.
- Add new features:
  - Derive a missing indicator feature from features with missing values.
- Redundancy removal:
  - Remove features that repeatedly describe the same content.
- Continuous Features
  - Convert the irregular negative values to positive ones.
- Outliers
  - Review outliers, checking for validity

6. References

[1] The data comes from the Centers for Disease Control and Prevention (CDC: https://covid.cdc.gov/covid-data-tracker/).

7. Appendix

7.1. Terminology & Assumptions

- Each row represents one single patient's personal information.
- Negative values for continuous features mean the earlier date and collection date was reversed which is unacceptable.
- The dataset was collected from 2020.1 (given by DATA.CDC.gov) to 2022.10 (given by csv file "last update").
- "res_region" is a Categorial column that groups the states into five regions of the United States.
- "elderly" in the DataFrame indicates whether a patient is 65 years old or older.
- "time_diff" means the difference between two time intervals (case_positive_specimen_interval and case_onset_interval) in the dataframe.

7.2. Continuous Features

Descriptive Statistics

Feature	count	mean	std	min	25%	50%	75%	max
case_positive_specimen_interval	10000.0	0.1612	2.083758	-63.0	0.0	0.0	0.0	56.0
case_onset_interval	8460.0	-0.039716	1.604339	-58.0	0.0	0.0	0.0	64.0

7.3. Categorial Features

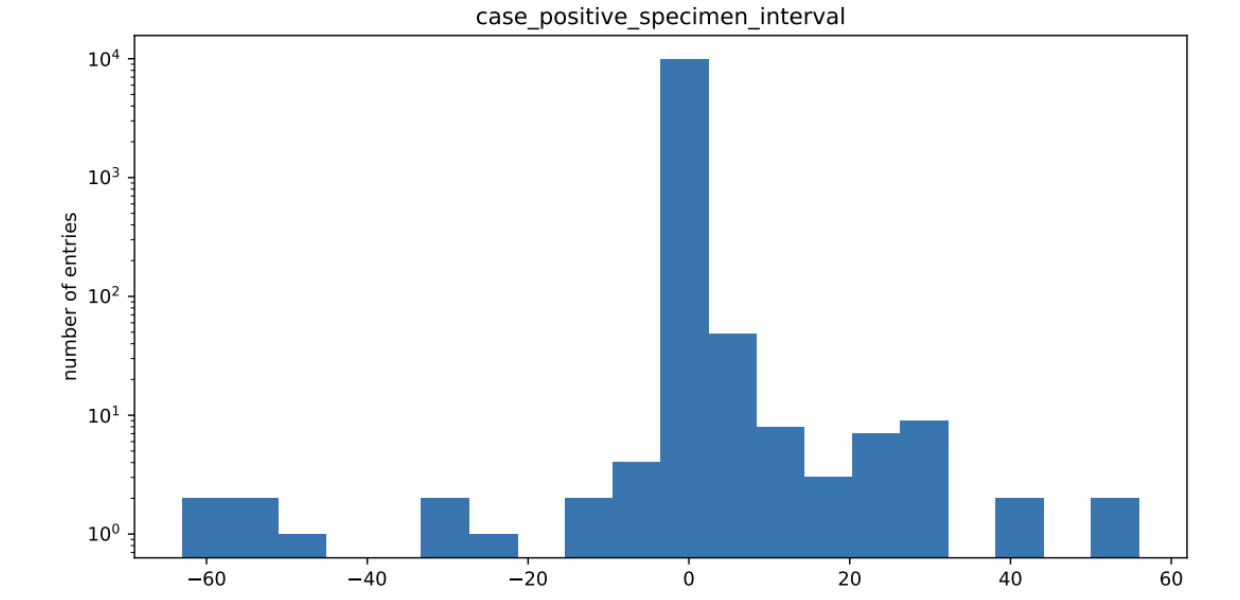
Descriptive Statistics

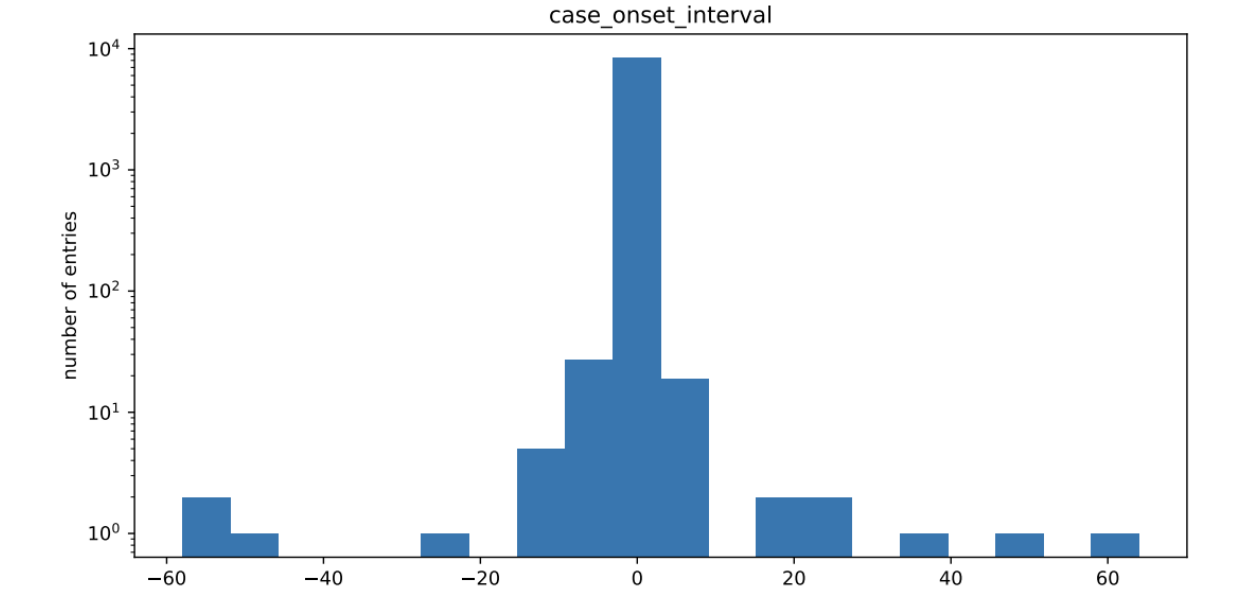
Feature	count	unique	top	freq
case_month	18892	35	2022-01	2349
res_state	18891	49	NY	1897
state_fips_code	18891.0	49.0	36.0	1897.0
res_county	17679	864	MIAMI-DADE	377
county_fips_code	17679.0	1208.0	12086.0	377.0
age_group	18705	4	18 to 49 years	7294
sex	18355	2	Female	9626
race	14341	6	White	11679
ethnicity	12948	2	Non-Hispanic/Latino	11324
process	1669	7	Clinical evaluation	807
exposure_yn	1863	1	Yes	1863
current_status	18892	2	Laboratory-confirmed case	15867
symptom_status	9207	2	Symptomatic	8912
hosp_yn	12633	2	No	9536
icu_yn	1641	2	No	1181
death_yn	18892	2	No	14354

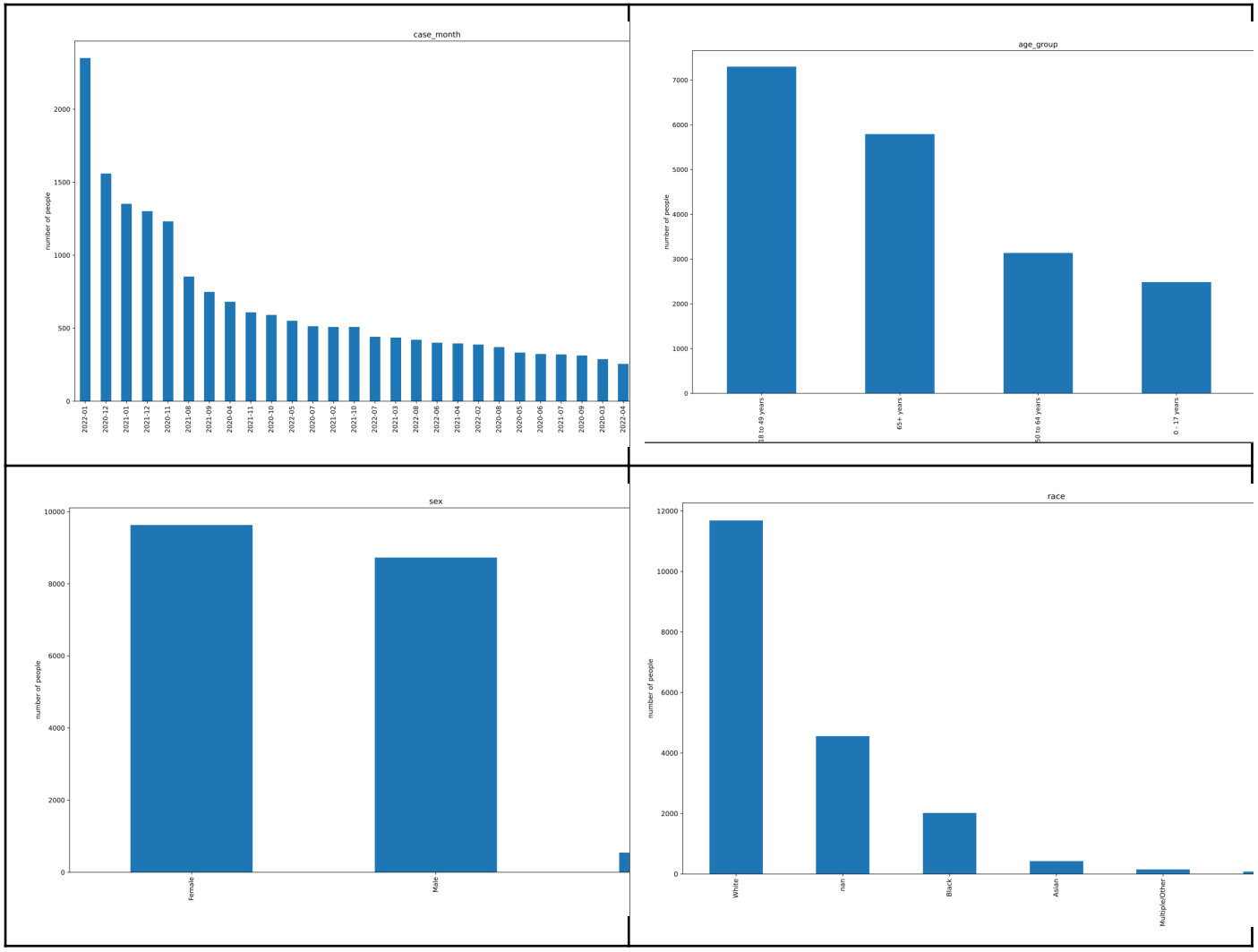
underlying_conditions_yn	1723	2	Yes	1701

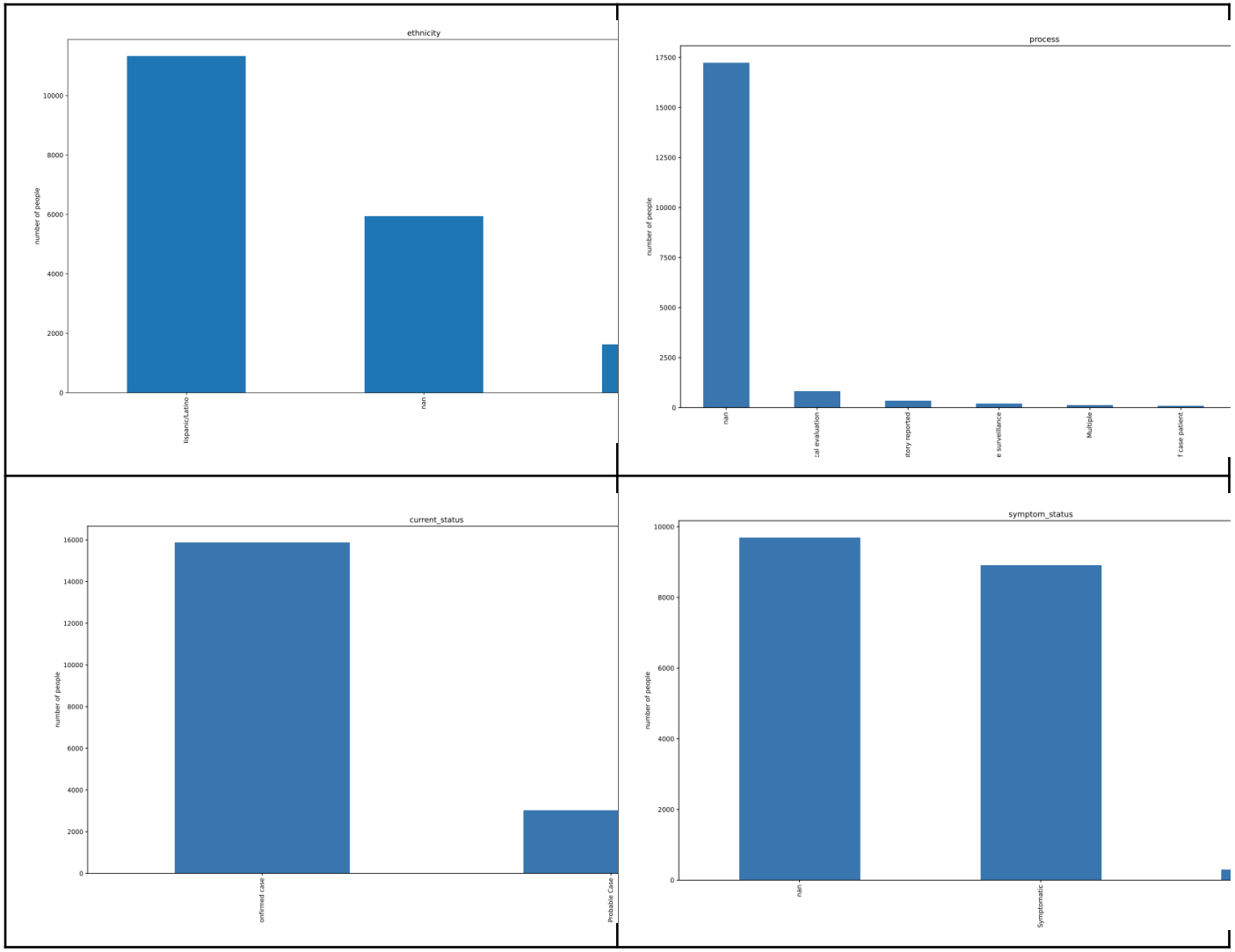
7.4. Box Plots & Histograms & Bar Plots

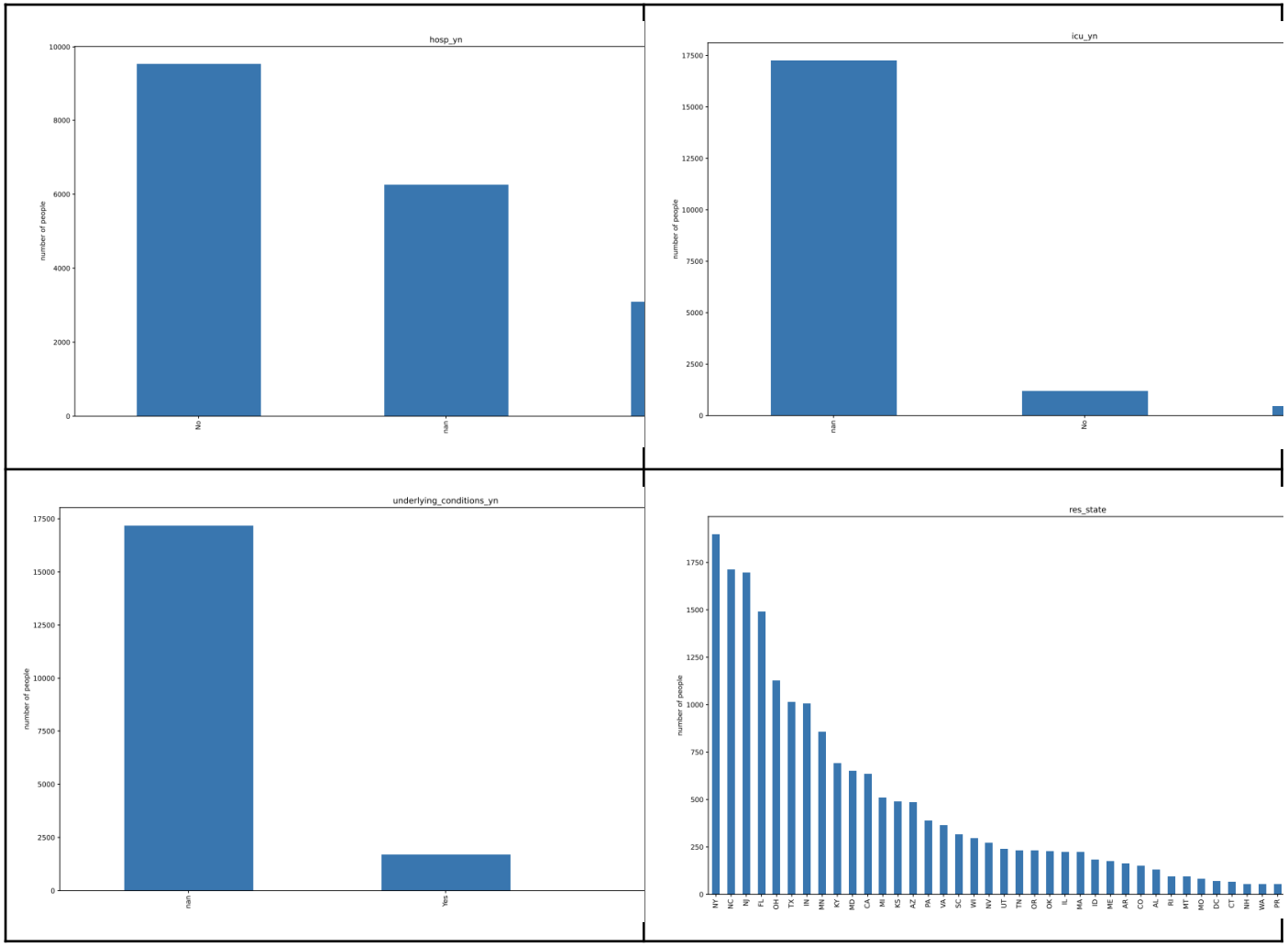
See below the summary of box plots, histograms and bar plots. Accompanying pdfs will show larger plots.

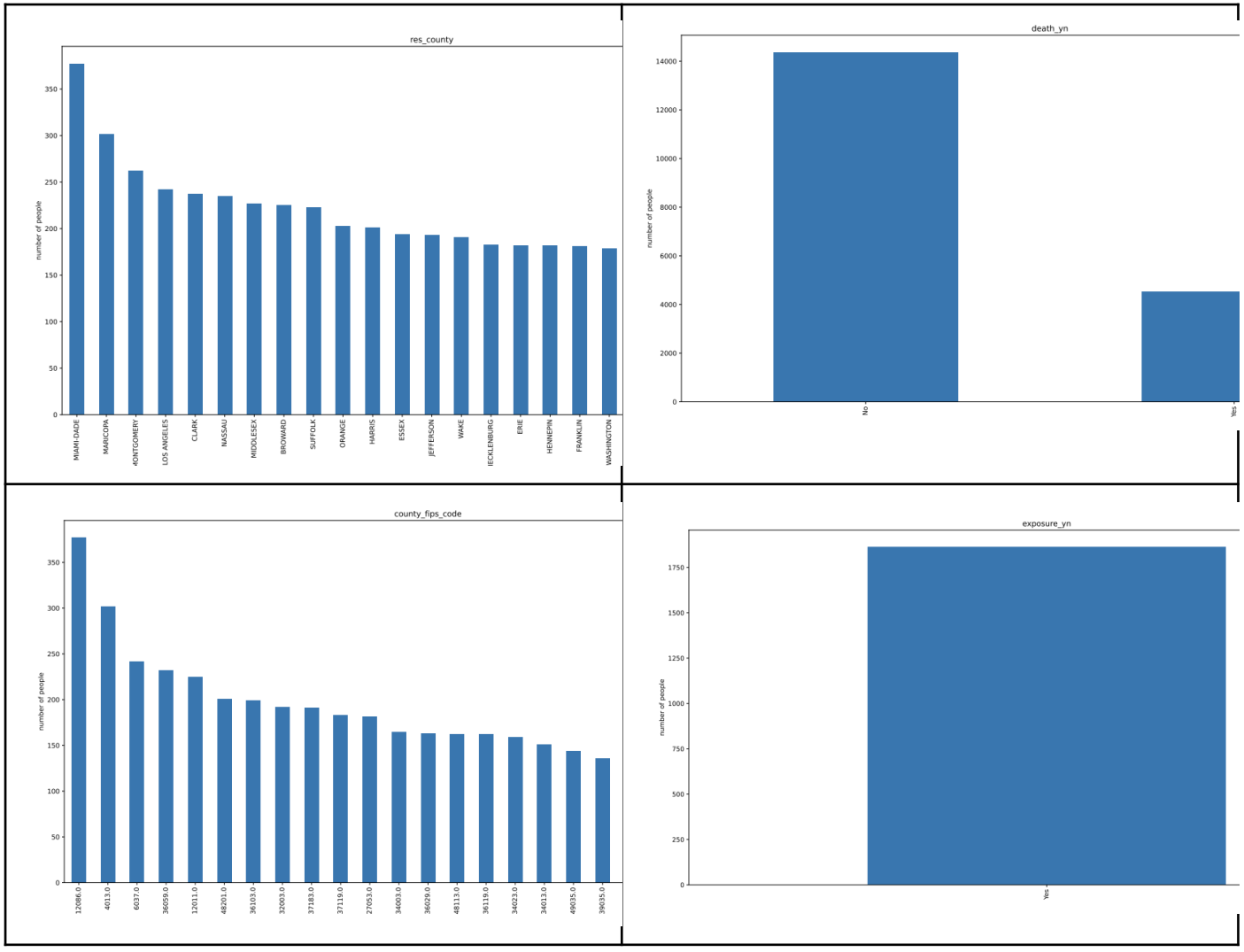

case_positive_specimen_interval
number of entries
10^4
10^3
10^2
10^1
10^0
−60
−40
−20
0
20
40
60

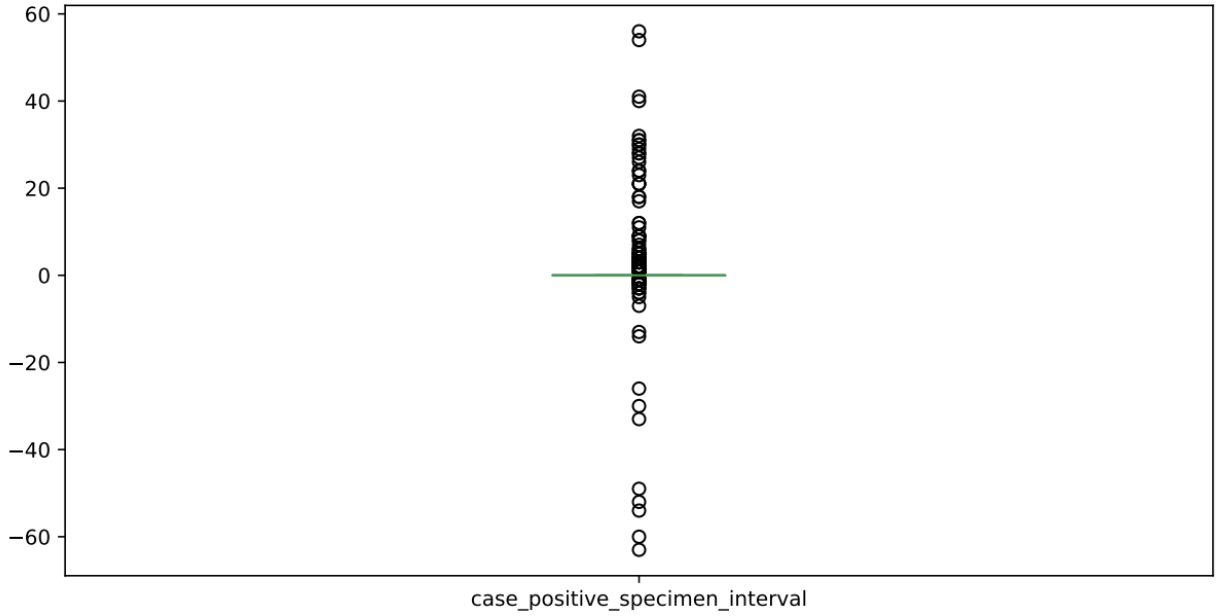

case_onset_interval
number of entries
10^4
10^3
10^2
10^1
10^0
−60
−40
−20
0
20
40
60

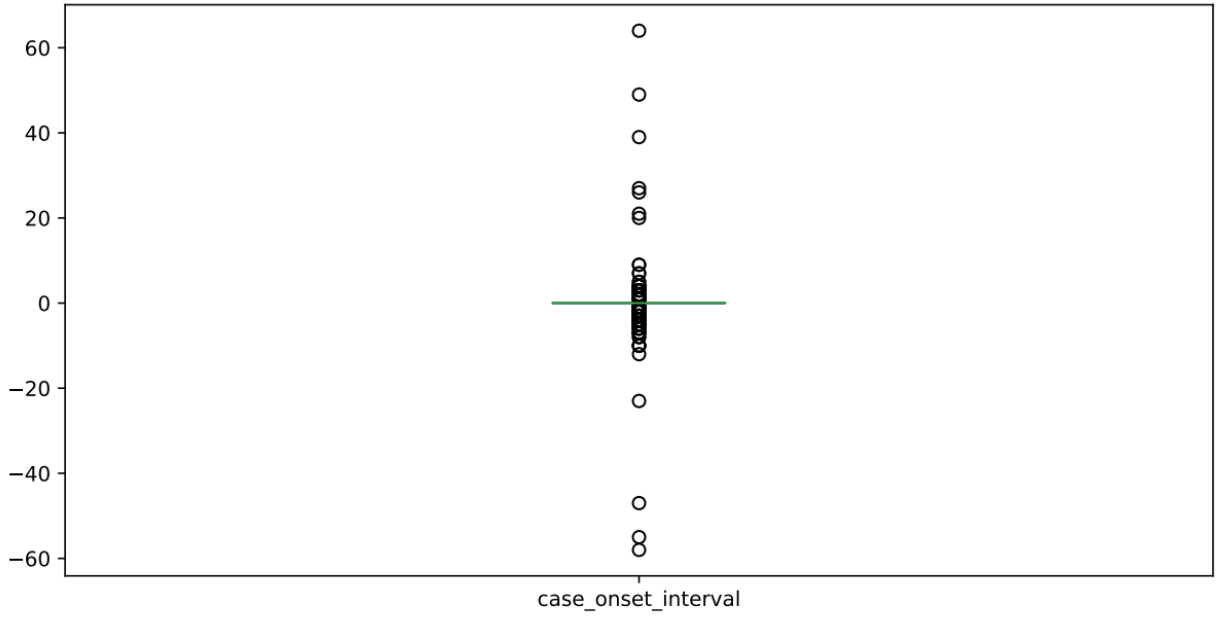

case_month
number of people
0
500
1000
1500
2000
2022-01
2020-12
2021-01
2021-12
2020-11
2021-08
2021-09
2020-04
2021-11
2020-10
2022-05
2020-07
2021-02
2021-10
2022-07
2021-03
2022-08
2022-06
2021-04
2022-02
2020-08
2020-05
2020-06
2021-07
2020-09
2020-03
2022-04
age_group
number of people
0
1000
2000
3000
4000
5000
6000
7000
18 to 49 years
65+ years
50 to 64 years
0 -17 years
sex
number of people
0
2000
4000
6000
8000
10000
Female
Male
race
number of people
0
2000
4000
6000
8000
10000
12000
White
nan
Black
Asian
Multiple/Other


ethnicity
number of people
0
2000
4000
6000
8000
10000
lispanic/Latino
nan
process
number of people
0
2500
5000
7500
10000
12500
15000
17500
nan
cal evaluation
story reported
e surveillance
Multiple
r case patient
current_status
number of people
0
2000
4000
6000
8000
10000
12000
14000
16000
onfirmed case
Probable Case
symptom_status
number of people
0
2000
4000
6000
8000
10000
nan
Symptomatic


hosp_yn
number of people
No
nan
icu_yn
number of people
nan
No
underlying_conditions_yn
number of people
nan
Yes
res_state
number of people
NY
NC
NJ
FL
OH
TX
IN
MN
KY
MD
CA
MI
KS
AZ
PA
VA
SC
WI
NV
UT
TN
OR
OK
IL
MA
ID
ME
AR
CO
AL
RI
MT
MO
DC
CT
NH
WA
PR


res_county
number of people
0
50
100
150
200
250
300
350
MIAMI-DADE
MARICOPA
MONTGOMERY
LOS ANGELES
CLARK
NASSAU
MIDDLESEX
BROWARD
SUFFOLK
ORANGE
HARRIS
ESSEX
JEFFERSON
WAKE
MECKLENBURG
ERIE
HENNEPIN
FRANKLIN
WASHINGTON
death_yn
number of people
0
2000
4000
6000
8000
10000
12000
14000
No
Yes
county_fips_code
number of people
0
50
100
150
200
250
300
350
12086.0
4013.0
6037.0
36059.0
12011.0
48201.0
36103.0
32003.0
37183.0
37119.0
27053.0
34003.0
36029.0
48113.0
36119.0
34023.0
34013.0
49035.0
39035.0
exposure_yn
number of people
0
250
500
750
1000
1250
1500
1750
Yes


60
40
20
0
−20
−40
−60
case_positive_specimen_interval


60
40
20
0
−20
−40
−60
case_onset_interval