Data Quality Report – Initial Findings

Data Quality Report_22200374

1. Overview

This report will outline the initial findings based on the cleaned dataset (covid19-cdc-22200374-updated&cleaned-data.csv). It will summarise the data, describe the various data quality issues observed and how they will be addressed. Please see 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, histograms and boxplots used to visualise the data.

On first indication the dataset appears relatively clean.There are no duplicate columns, or columns with irregular cardinalities, or constant columns. However, there are 1052 rows of duplicate data here, and I decided to drop them after comparison.The main issues observed were regarding special values for continuous data and numerical scales used for categorical data. In addition, a significant number of outliers were present. Also, several logical tests were carried out on the data a significant number of inconsistencies were found. For the column underlying_conditions_yn, we have a missing rate of 91.054465%, which is very high, but I don't want to drop them, because their value is still very important, but to reassign them while waiting for the Data Quality Plan.

2. Summary

Several tests were carried out to check the logical integrity of the data. This brought about a significant number of failures of the data. In total 8763 instances of irrational data was observed. For example, in 665 instances patients' symptom_status is symptomatic, but their case_onset_interval is empty, therefore we can assign some meaning value to their empty case_onset_interval. In in 3 instances patients weren't hospitalized, but they have been to the ICU. This irrational data will need to be dealt with and should be checked with the domain expert. See logical integrity section for further details.

For the continuous features there was the inclusion of several special values. The negative values of 'case_positive_specimen_interval' and 'case_onset_interval' are generally impossible, and the reason may be that the data positions of 'onset_dt' and 'cdc_case_earliest' or 'pos_spec_dt' and 'cdc_case_earliest_dt' are reversed when recording. Then the result of their subtraction is negative. This needs to be dealt with later, and the absolute value is taken to fix it. Some missing data in 'case_positive_specimen_interval' and 'case_onset_interval' can be calculated and compensated according to the values of the rest of the categorical columns.

For the categorical values several changes are recommended. The missing rate of 'underlying_conditions_yn' is very high, but his remaining data is very precious. I don't want to drop its only remaining data, and then may change the missing/empty data to unknown values for further research. Other categorical values are quite satisfactory. and you can see Review Categorical Features section for further details.

There was a significant number of outliers present across the feature set. However, on first indication these values appear to be plausible but should be investigated further. Other categorical values

3. Review Logical Integrity

11 tests were carried out. The results are below;

- Test 1 - Check if any entries have 'case_positive_specimen_interval's value < 0 (impossible)
  - 56 cases found

- Test 2 - Check if any entries have 'case_onset_interval's value < 0 (impossible)
  - 285 cases found

- Test 3 - Check if any entries have 'symptom_status' = 'Symptomatic' ,and their 'case_onset_interval's value is empty
  (Data can then be estimated and supplemented)
  - 665 cases found

- Test 4 - Check if any entries have 'symptom_status' = 'Asymptomatic', and their 'case_onset_interval's value isn't empty

  (impossible)

  - 0 cases found

- Test 5_1 - Check if any entries have 'hosp_yn' = 'No' ,and their 'icu_yn' = 'Yes' ,which means the patient wasn't hospitalized but went to the ICU, which feels like a contradictory case

  (impossible)

  - 3 cases found

- Test 5_2 - Check if any entries have 'hosp_yn' = 'No' ,and their 'death_yn' = 'No' ,and 'icu_yn'='Missing'

  (Patients who have not been hospitalized and have not died, it is theoretically impossible for them to receive ICU)

  - 7425 cases found

- Test 6 - Check if any entries have 'symptom_status' = 'Asymptomatic' , their 'death_yn' = 'Yes' , and 'current_status' = 'Laboratory-confirmed case'. Patient's death must have been symptomatic.

  (impossible)

  - 54 cases found

- Test 7 - Check if any entries have 'case_positive_specimen_interval's value is empty, and 'current_status' = 'Laboratory-confirmed case'.

  (impossible, Data can then be estimated and supplemented)

  - 7697 cases found

- Test 8_1 - Check if any entries have 'current_status' = 'Probable case' , their 'process' = 'Laboratory reported' or 'Autopsy' or 'Clinical evaluation' or 'Routine surveillance' or 'Routine physical examination'

  (impossible)

  - 0 cases found

- Test 8_2 - Check if any entries have 'current_status' = 'Laboratory-confirmed case' , their 'process' = 'Laboratory reported' or 'Autopsy' or 'Clinical evaluation' or 'Routine surveillance' or 'Routine physical examination'

  (just 1185 cases found , so we know the data of 'process' is missing and damaged seriously.)

  - 1185 cases found

- Test 9 - Check if any entries have 'symptom_status' = 'Asymptomatic' ,and their 'icu_yn' = 'Yes'

  (impossible)

  - 3 cases found

- Test 10 - Check if any entries have 'symptom_status' = 'Asymptomatic' ,and their 'hosp_yn' = 'Yes' , and their 'underlying_conditions_yn' = 'No'

  (impossible)

  - 0 cases found

- Test 11 - Check if any entries have 'symptom_status' = 'Asymptomatic' ,and their 'death_yn' = 'Yes' , and their 'underlying_conditions_yn' = 'No'

  (impossible)

  - 0 cases found

- Test 12 - Check if any entries have 'symptom_status' ='Missing' or 'Unknown' or is empty, and their 'death_yn' = 'Yes'

  (impossible, People who died must have symptoms, so their missing, unknown, or empty 'symptom_status' can be changed to 'Symptomatic'.)

  - 9711 cases found

- Test 13_1 - Check the death cases who have been hospitalized and received ICU, and are there many whose "underlying_conditions_yn" is empty?

  (In this case their "underlying_conditions_yn" tends to be 'Yes', but the number is very small.)

  - 261 cases found

- Test 13_2 - Check how many "underlying_conditions_yn" of patients who have not died, have not been hospitalized, and are asymptomatic are empty?

  (In this case their "underlying_conditions_yn" tends to be 'No', but the number is very small.)

  - 163 cases found

4. Review Continuous Features

4.1. Descriptive Statistics

There are 2 continuous features, which will be summarised below:

case_positive_specimen_interval:

The mean of 'case_positive_specimen_interval' is 0.175236. Among the values it effectively records, there are 53 special values. Among the 53 special values, 0 appears most frequently, 8804 times, accounting for 46.464% of the total (including both missing and not missing).

There are only 9958 entries identified from the data with nearly 47.445641% of data missing (from a total of 18948 entries with duplicates removed). Most of the entries have a value of "0" , with tiny amount of outliers having a max value of 100 and mix value of -76.

'case_positive_specimen_interval' generally does not have negative values. The reason for the negative value may be that the operator has switched the values of 'pos_spec_dt' and 'cdc_case_earliest_dt' during calculations accidentally, thus resulting in error entries. A total of 56 pieces of data have negative values. **Later, The negative numbers will be addressed by switching back to positive number.**

According to Test 7, if the 'case_positive_specimen_interval' of entries are empty, and their 'current_status'='Laboratory-confirmed case' at the same time, we can give some estimated value later. **The missing values will be replaced by "0" , indicating that the positive specimen is collected in the same week the case is identified to CDC.** Because the laboratory-confirmed case must have a valid 'case_positive_specimen_interval'.

case_onset_interval:

The mean of 'case_onset_interval' is -0.041848, and there are 42 special values. Among the 42 special values, 0 appears most frequently, 8146 times, accounting for 42.991% of the total (including both missing and not missing).

There are only 8483 entries identified from the data with nearly 55.230103% of data missing (from a total of 18948 entries with duplicates removed). Most of the entries have a value of "0", with tiny amount of outliers having a max value of 54 and mix value of -58.

'case_onset_interal' generally does not have negative values. The reason for the negative value may be that the operator has switched the values of 'onset_dt' and 'cdc_case_earliest' during calculations accidentally, thus resulting in error entries. A total of 285 pieces of data have negative values. **Later, The negative numbers will be addressed by switching back to positive number.**

According to Test 3, if the 'case_onset_interal' value of entries is empty, and their 'symptom_status' = 'Symptomatic' at the same time, we can give some estimated value later. **The missing value will be replaced by "0", indicating that the collected date when symptom happened is in the same week the case is identified to CDC**. Because 'symptom_status' = 'Symptomatic' case must have a valid 'case_onset_interal', and 0 is the most frequently occurring value.

It's 55% missing (over 50%), but considering Test3, we will convert 665 'NaN' values to 0, which can reduce its missing rate to 51.72% ((10465-665)/18948=51.72%). And our original continuous features are very It is rare, so it can no longer be discarded decisively. **So, I will still retain it.**

4.2. Histograms

All histograms can be found on the appendix as summary sheet. Individual plots can be found in the accompanying notebook. From the graph we can observe that the extreme high frequency of values "0" has somehow detorted the graph, therefore a log graph is used to display other values in the features. Overall, these are plausible distribution. The outliers will be investigated further but no immediate action expected.

4.3. Box plots

All boxplots can be found on the appendix as summary sheet. Individual plots can be found in the accompanying notebook. Again, outliers will be investigated further but no immediate action expected. There are a lot of outliers, which we will discuss and analyze in detail later.

5. Review Categorical Features

5.1. Descriptive Statistics

There is a total of 17 categorical features in the dataset:

- **case_month:**

  All good and 0 missing. The top 5 values are 2022-01(2289), 2020-12(1573), 2021-01(1413), 2021-12(1304), 2020-11(1274).

- **res_state:**

  All good and 0 missing. The top value is NY.

- **state_fips_code:**

  All good and 0 missing. The highest value (36) properly aligned to the highest value in Res_state (NY). The second highest value (34) properly aligned to the second highest value in Res_state (NJ).

https://www.mercercountypa.gov/dps/state_fips_code_listing.htm

- **res_county:**

  There are NaN values, but the necessary relevant information is missing, so it is difficult to predict.

  I will just change 'NaN' values to 'Unknown'.

- **county_fips_code:**

  There are NaN values, but the necessary relevant information is missing, so it is difficult to predict.

  I will just change 'NaN' values to 'Unknown'.

- **age_group:**

  There are NaN values, but the necessary relevant information is missing, so it is difficult to predict.

  I will just change 'NaN' and 'Missing' values to 'Unknown'.

  The highest value is “18 to 49 years” which is probably because the range covers the majority of the populations.

- **sex:**

  There are NaN values, but the necessary relevant information is missing, so it is difficult to predict.

  I will just change 'NaN' and 'Missing' values to 'Unknown'.

  “Male” and “Female” has the similar proportion in this dataset.

- **race:**

  There are NaN values, but the necessary relevant information is missing, so it is difficult to predict.

  I will just change 'NaN' and 'Missing' values to 'Unknown'.

  Majorities of the values are “white”.

- **ethnicity:**

  There are NaN values, but the necessary relevant information is missing, so it is difficult to predict.

  I will just change 'NaN' and 'Missing' values to 'Unknown'.

  “Non-Hispanic/Latino” accounts for more than half of the values, which makes sense from the result of previous feature “race”.

- **process:**

  There are nearly 90.77% of “missing” values. Although the rest of values represent different categories (for eg. “autopsy”, “clinical evaluation” etc.), the samples are too small compared to the overall dataset.

  Useless to my test.

  So, I decided to drop this features. Test8_2 perfectly confirmed my idea.

- **exposure_yn:**

  Since there are more than 85.88% of "missing" values present within this features, I decided to drop this features.

  I decided to drop this features.

- **current_status:**

  All good. Most of the values are "laboratory-confirmed case".

- **symptom_status:**

  "Missing" values will be addressed.

  I will just change 'Missing' values to 'Unknown'.

  According to Test12, 9711 'Missing' or 'Unknown' or empty cases will be changed to 'Symptomatic'

  47.148% of the patients are "symptomatic".

- **hosp_yn:**

  There are "Missing" values will be addressed.

  The highest value is "No" (50.29%).

  I will just change 'Missing' values to 'Unknown'.

  You can see my Test 5. There should be a correlation between this features and "icu_yn", ie entries with "yes" to "icu_yn" should have "yes" to this feature as well. In contrast, entries with "no" to this feature should have "no" to "icu_yn". Actions will be taken to rectify those error values.

- **icu_yn:**

  Since there are more than 78.11% of "missing" entries, Test5_2 shows that there are 7425 entries whose 'icu_yn'='Missing',while their 'death_yn'='No', and 'hosp_yn'='No'.

  In that situation, **their missing 'icu_yn' can be replaced by 'No'**. Then, the missing rate will reduce to (18948-1649-7425)/18948=52.11%, which is nearly 50%. **Considering the diversity of data, I will still retain it.**

  Also I used this feature to get some bad paradoxical data to drop. Test5_1 and Test9 perfectly confirmed my idea.

Test9: Asymptomatic patients are unlikely to be admitted to an intensive care unit (ICU) without life-threatening symptoms.

Test5_1: It is impossible for patients admitted to an intensive care unit (ICU) not to be hospitalized.

- **death_yn:**

  All good

  Approximately 75.8% of values are "no", meaning most of the patients recovered from COVID-19.

- **underlying_conditions_yn:**

  There are only 1695 input out of 18938 entries due to large amount of 'NaN' value. Although ,after test13_1 and test13_2, I want to convert some of its 'Missing' values into effective values, the final result of effective values (1695+261+163)/18938=11.2% is far less than 50%

  I decided to drop this features, after using it to drop some entries.

5.2. Histograms

All boxplots can be found on the appendix as summary sheet. Individual plots can be found in the accompanying notebook.

6. Action to take

- **Logical integrity**
  - Test1: The negative 'case_positive_specimen_interval's value will be addressed by switching back to positive number.
  - Test2: The negative 'case_onset_interval's value will be addressed by switching back to positive number.
  - Test3: If the 'case_onset_interal' value of entries is empty, and their 'symptom_status' = 'Symptomatic' at the same time, the missing 'case_onset_interal' will be replaced by "0".
  - Test5_1: Rows failing this logical test will need to be dropped.

  - Test5_2: If the 'icu_yn' value of entries is 'Missing', and their 'death_yn' = 'No' and 'hosp_yn'='No' at the same time, the missing 'icu_yn' value will be replaced by "No".
  - Test6: Rows failing this logical test will need to be dropped.
  - Test7: If the 'case_positive_specimen_interval' value of entries is empty, and their 'current_status' = 'Laboratory-confirmed case' at the same time, the missing 'case_onset_interal' will be replaced by "0".
  - Test9: Rows failing this logical test will need to be dropped.
  - Test12: When some entries''symptom_status' ='Missing' or 'Unknown' or is empty, and their 'death_yn' = 'Yes' . The 'symptom_status's value can be changed to 'Symptomatic'.
- **Continuous Features**
  - case_positive_specimen_interval: Same action as Test1 and Test7
  - case_onset_interval: Same action as Test2 and Test3
- **Categorical Features**
  - res_county: 'NaN' values needed to be changed to 'Unknown'.
  - county_fips_code: 'NaN' values needed to be changed to 'Unknown'.
  - age_group: 'NaN' values and 'Missing' values needed to be changed to 'Unknown'.
  - sex: 'NaN' values and 'Missing' values needed to be changed to 'Unknown'.
  - race: 'NaN' values and 'Missing' values needed to be changed to 'Unknown'.
  - ethnicity: 'NaN' values and 'Missing' values needed to be changed to 'Unknown'.
  - process: This column will need to be dropped.
  - exposure_yn: This column will need to be dropped.
  - symptom_status: 'Missing' values needed to be changed to 'Unknown'.
  - hosp_yn: 'Missing' values needed to be changed to 'Unknown'.
  - icu_yn: 'Missing' values needed to be changed to 'Unknown'.
  - underlying_conditions_yn: This column will need to be dropped.
- **Outliers**
  - Review outliers, checking for validity

7. References

[1] CDC COVID Data Tracker

https://covid.cdc.gov/covid-data-tracker/#datatracker-home

[2] COVID-19 Case Surveillance Public Use Data with Geography

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

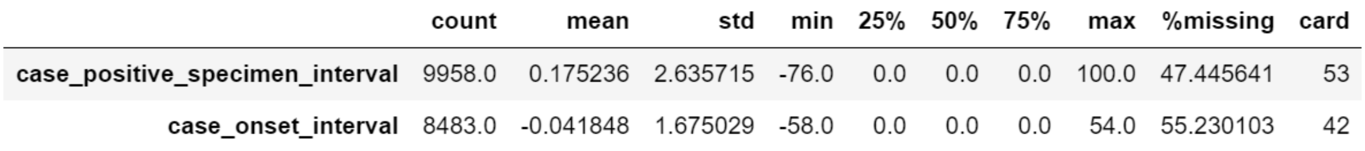
8. Appendix

8.1. Terminology & Assumptions

- case_month: The earlier month of the Clinical Date (date related to the illness or specimen collection) or the Date Received by CDC.
- case_positive_specimen_interval: Weeks between earliest date and date of first positive specimen collection.
- case_onset_interval: Weeks between earliest date and date of symptom onset.
- process: Under what process was the case first identified?
- exposure_yn: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/airplane, adult congregate living facility (nursing, assisted living, or long-term care facility), school/university/childcare center, correctional facility, community event/mass gathering, animal with confirmed or suspected COVID-19, other exposure, contact with a known COVID-19 case?
- underlying_conditions_yn: Did the patient have one or more of the underlying medical conditions and risk behaviors: 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?

8.2. Continuous Features

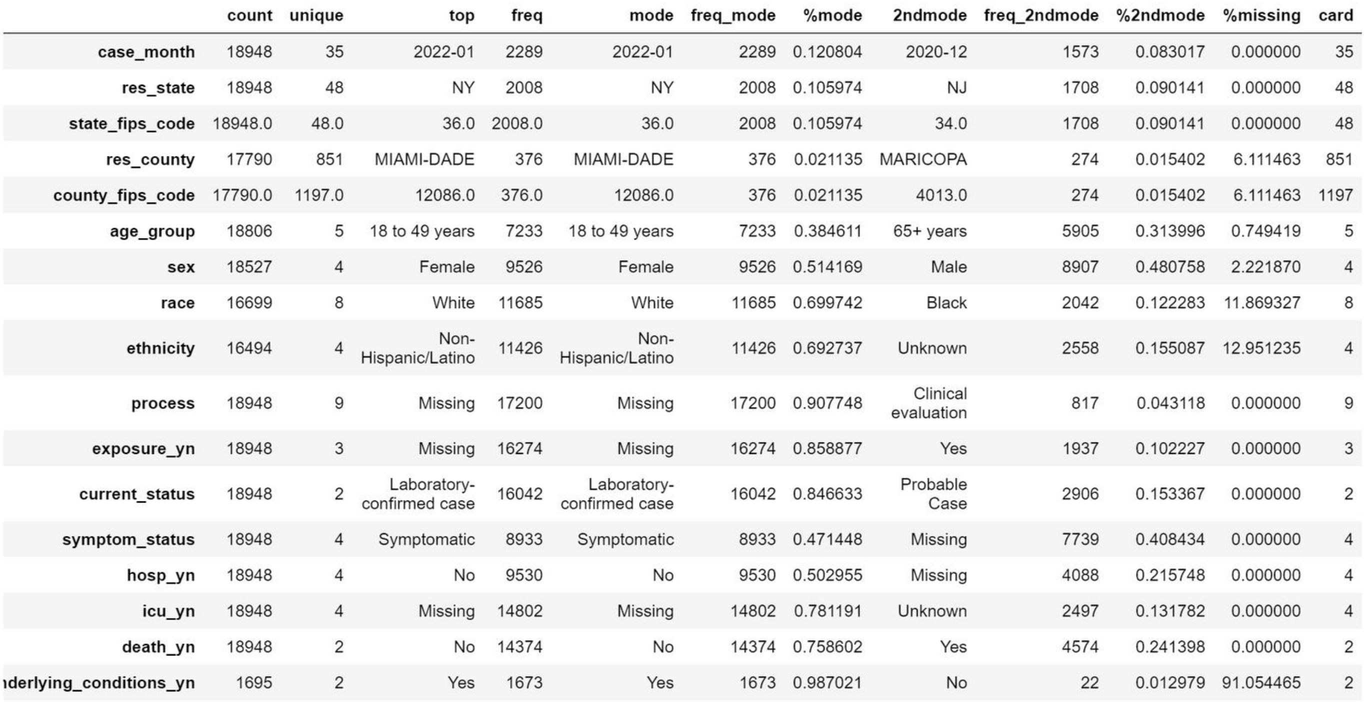
Descriptive Statistics



	count	mean	std	min	25%	50%	75%	max	%missing	card
case_positive_specimen_interval	9958.0	0.175236	2.635715	-76.0	0.0	0.0	0.0	100.0	47.445641	53
case_onset_interval	8483.0	-0.041848	1.675029	-58.0	0.0	0.0	0.0	54.0	55.230103	42

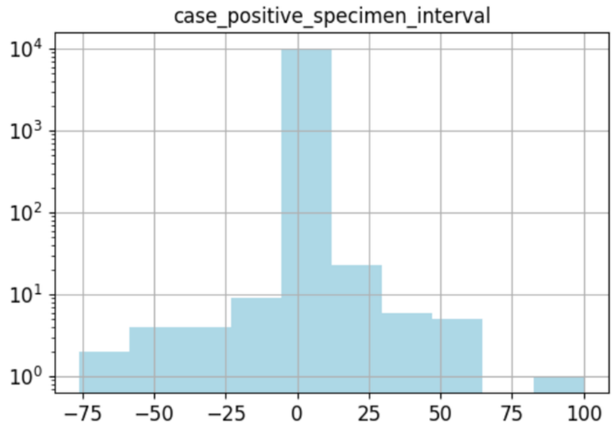
8.3. Categorical Features

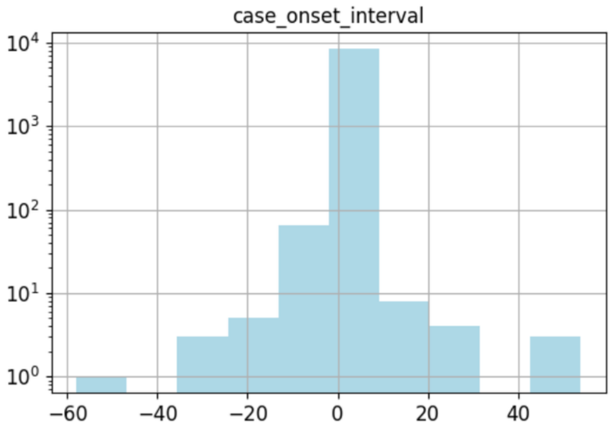
Descriptive Statistics

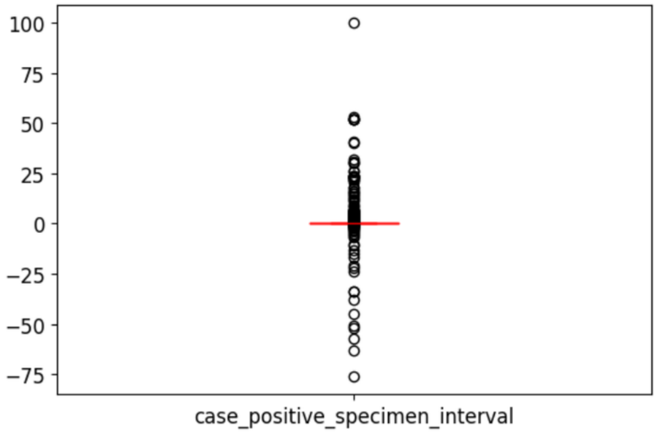


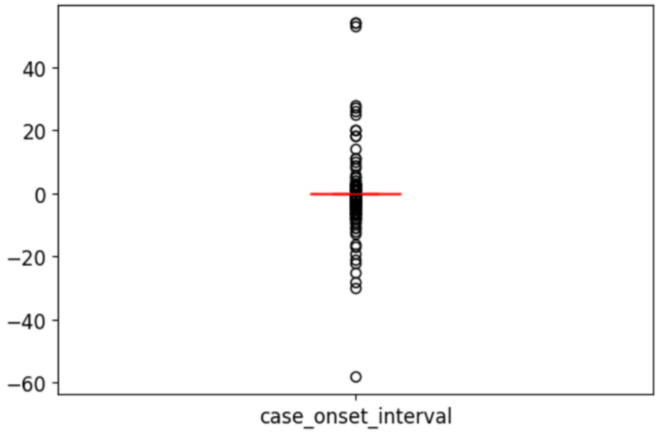
	count	unique	top	freq	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode	%missing	card
case_month	18948	35	2022-01	2289	2022-01	2289	0.120804	2020-12	1573	0.083017	0.000000	35
res_state	18948	48	NY	2008	NY	2008	0.105974	NJ	1708	0.090141	0.000000	48
state_fips_code	18948.0	48.0	36.0	2008.0	36.0	2008	0.105974	34.0	1708	0.090141	0.000000	48
res_county	17790	851	MIAMI-DADE	376	MIAMI-DADE	376	0.021135	MARICOPA	274	0.015402	6.111463	851
county_fips_code	17790.0	1197.0	12086.0	376.0	12086.0	376	0.021135	4013.0	274	0.015402	6.111463	1197
age_group	18806	5	18 to 49 years	7233	18 to 49 years	7233	0.384611	65+ years	5905	0.313996	0.749419	5
sex	18527	4	Female	9526	Female	9526	0.514169	Male	8907	0.480758	2.221870	4
race	16699	8	White	11685	White	11685	0.699742	Black	2042	0.122283	11.869327	8
ethnicity	16494	4	Non-Hispanic/Latino	11426	Non-Hispanic/Latino	11426	0.692737	Unknown	2558	0.155087	12.951235	4
process	18948	9	Missing	17200	Missing	17200	0.907748	Clinical evaluation	817	0.043118	0.000000	9
exposure_yn	18948	3	Missing	16274	Missing	16274	0.858877	Yes	1937	0.102227	0.000000	3
current_status	18948	2	Laboratory-confirmed case	16042	Laboratory-confirmed case	16042	0.846633	Probable Case	2906	0.153367	0.000000	2
symptom_status	18948	4	Symptomatic	8933	Symptomatic	8933	0.471448	Missing	7739	0.408434	0.000000	4
hosp_yn	18948	4	No	9530	No	9530	0.502955	Missing	4088	0.215748	0.000000	4
icu_yn	18948	4	Missing	14802	Missing	14802	0.781191	Unknown	2497	0.131782	0.000000	4
death_yn	18948	2	No	14374	No	14374	0.758602	Yes	4574	0.241398	0.000000	2
nderlying_conditions_yn	1695	2	Yes	1673	Yes	1673	0.987021	No	22	0.012979	91.054465	2

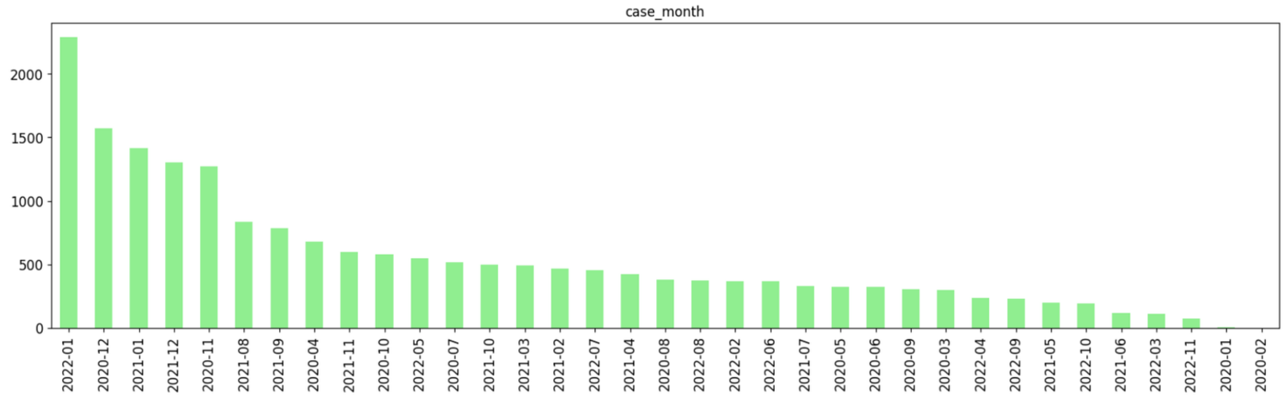
8.4. Box Plots & Histograms

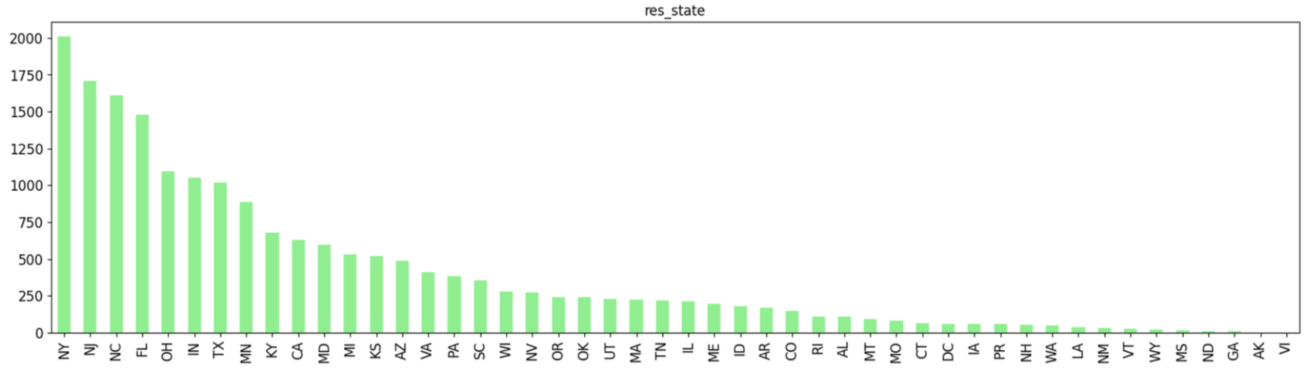


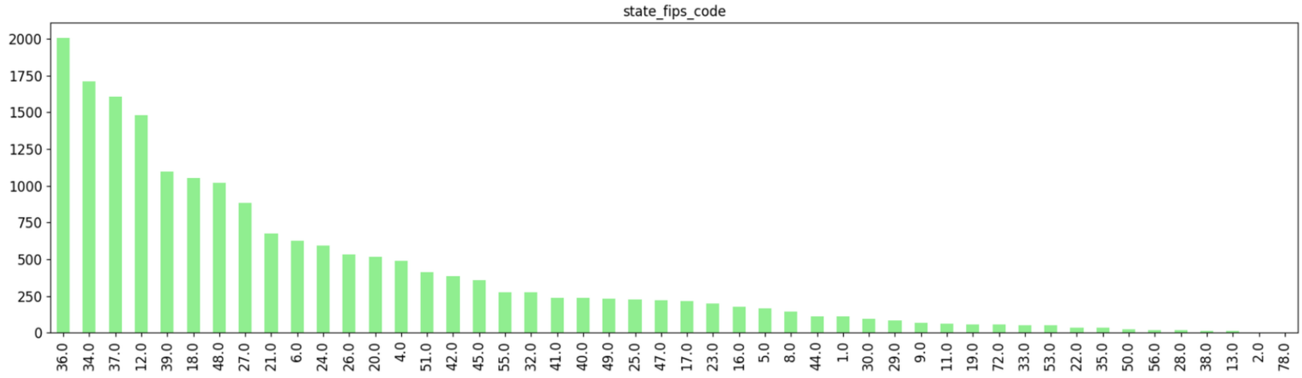


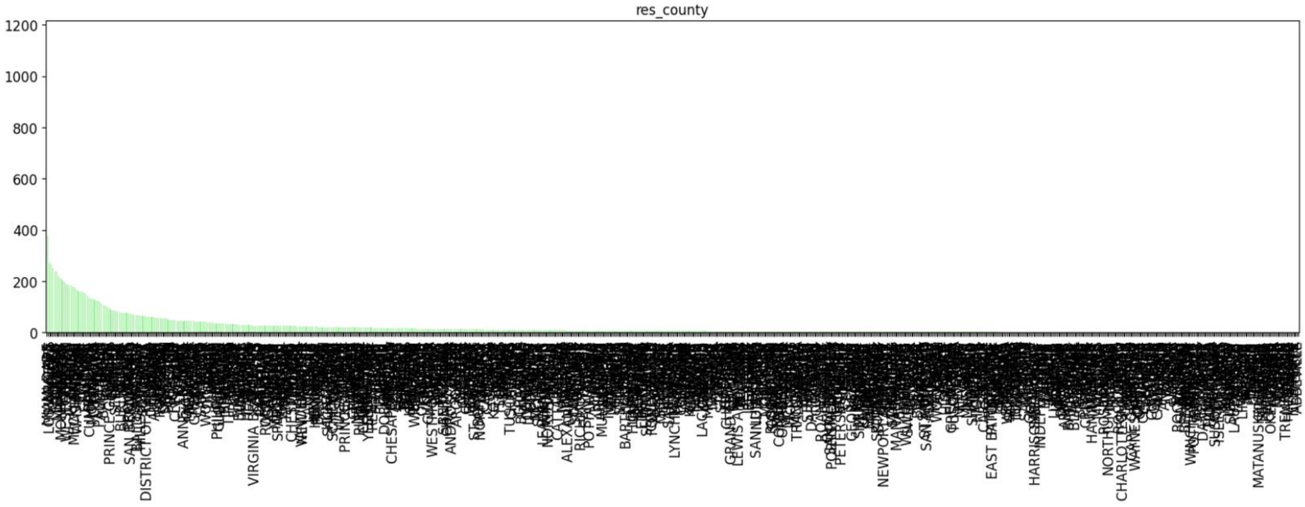

100
75
50
25
0
−25
−50
−75
case_positive_specimen_interval

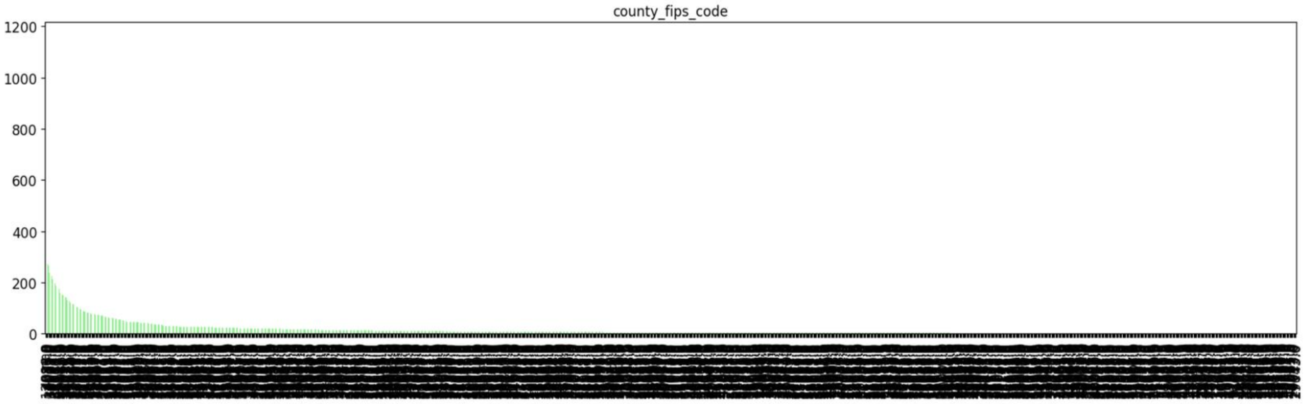

40
20
0
−20
−40
−60
case_onset_interval

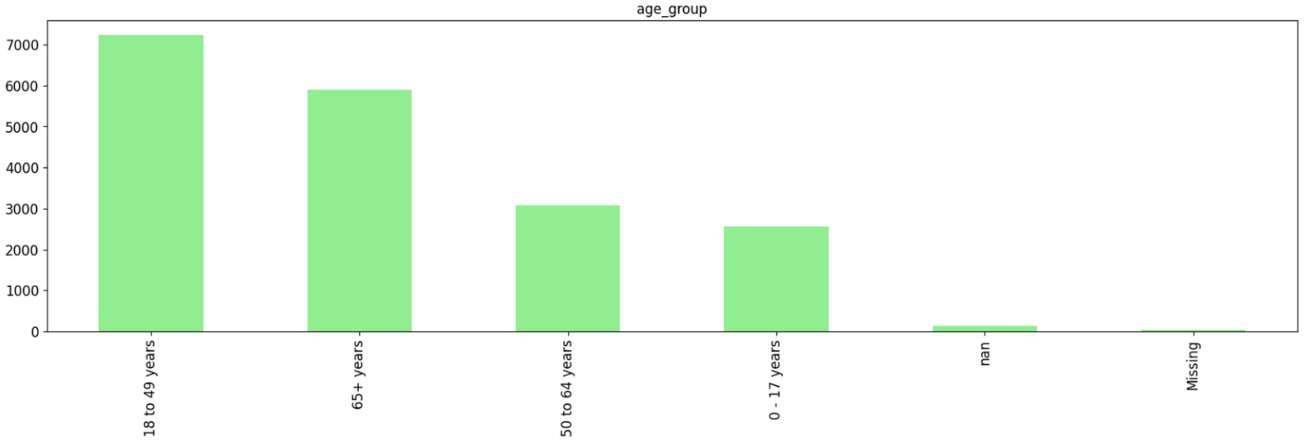

case_month
2000
1500
1000
500
0
2022-01
2020-12
2021-01
2021-12
2020-11
2021-08
2021-09
2020-04
2021-11
2020-10
2022-05
2020-07
2021-10
2021-03
2021-02
2022-07
2021-04
2020-08
2022-08
2022-02
2022-06
2021-07
2020-05
2020-06
2020-09
2020-03
2022-04
2022-09
2021-05
2022-10
2021-06
2022-03
2022-11
2020-01
2020-02

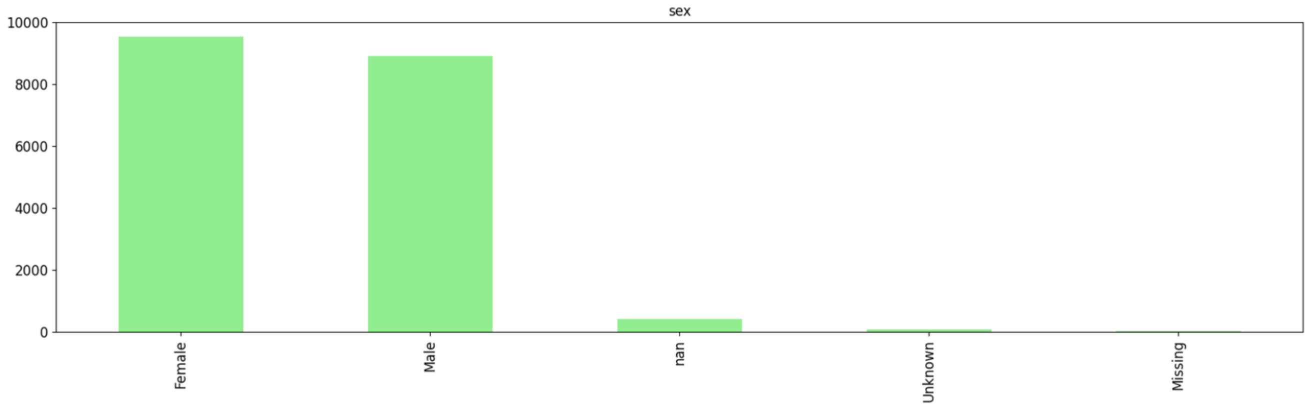

res_state
2000
1750
1500
1250
1000
750
500
250
0
NY
NJ
NC
FL
OH
IN
TX
MN
KY
CA
MD
MI
KS
AZ
VA
PA
SC
WI
NV
OR
OK
UT
MA
TN
IL
ME
ID
AR
CO
RI
AL
MT
MO
CT
DC
IA
PR
NH
WA
LA
NM
VT
WY
MS
ND
GA
AK
VI

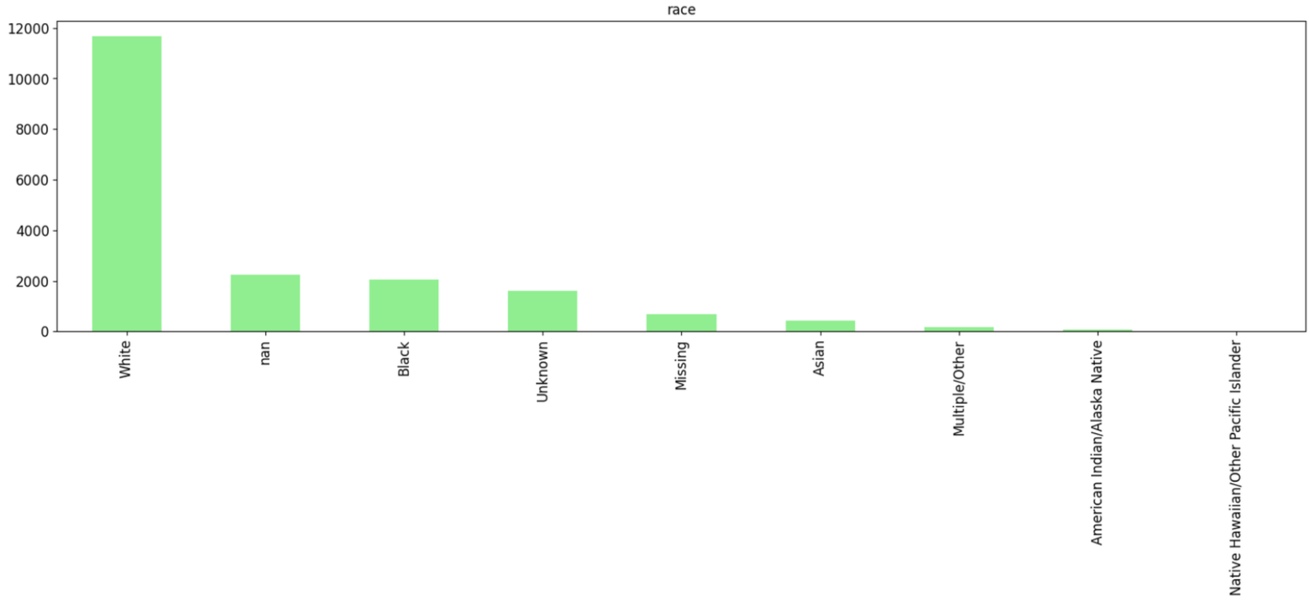

state_fips_code
2000
1750
1500
1250
1000
750
500
250
0
36.0
34.0
37.0
12.0
39.0
18.0
48.0
27.0
21.0
6.0
24.0
26.0
20.0
4.0
51.0
42.0
45.0
55.0
32.0
41.0
40.0
49.0
25.0
47.0
17.0
23.0
16.0
5.0
8.0
44.0
1.0
30.0
29.0
9.0
11.0
19.0
72.0
33.0
53.0
22.0
35.0
50.0
56.0
28.0
38.0
13.0
2.0
78.0

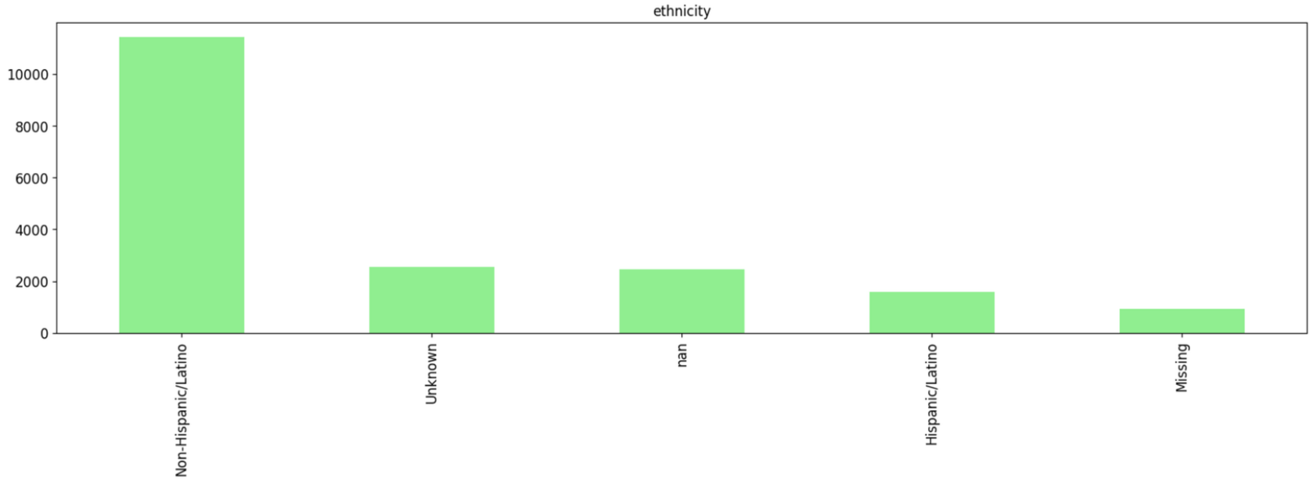

res_county
0
200
400
600
800
1000
1200
PRINC
DISTRICT OF
ANN
VIRGINIA
CHESAP
WESTM
ANDER
ALEXAN
BARB
LYNCH
GRANG
SANND
PETERS
NEWPOR
EAST BAT
HARRISON
NORTHUM
CHARLOTT
WAYNE
MATANUSK
TREM

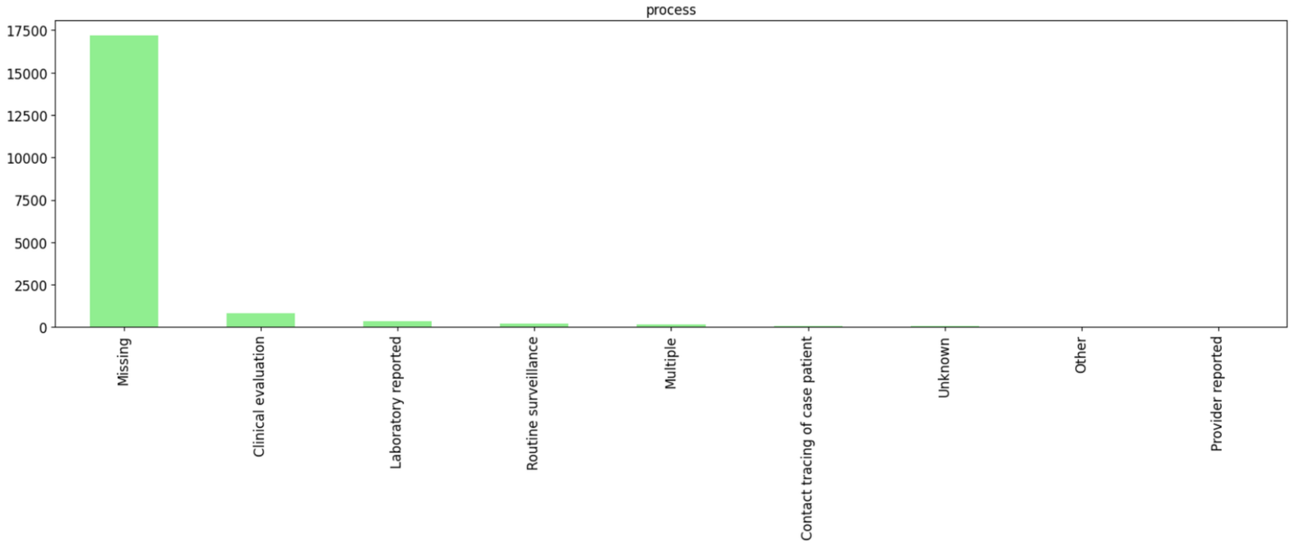

county_fips_code
0
200
400
600
800
1000
1200

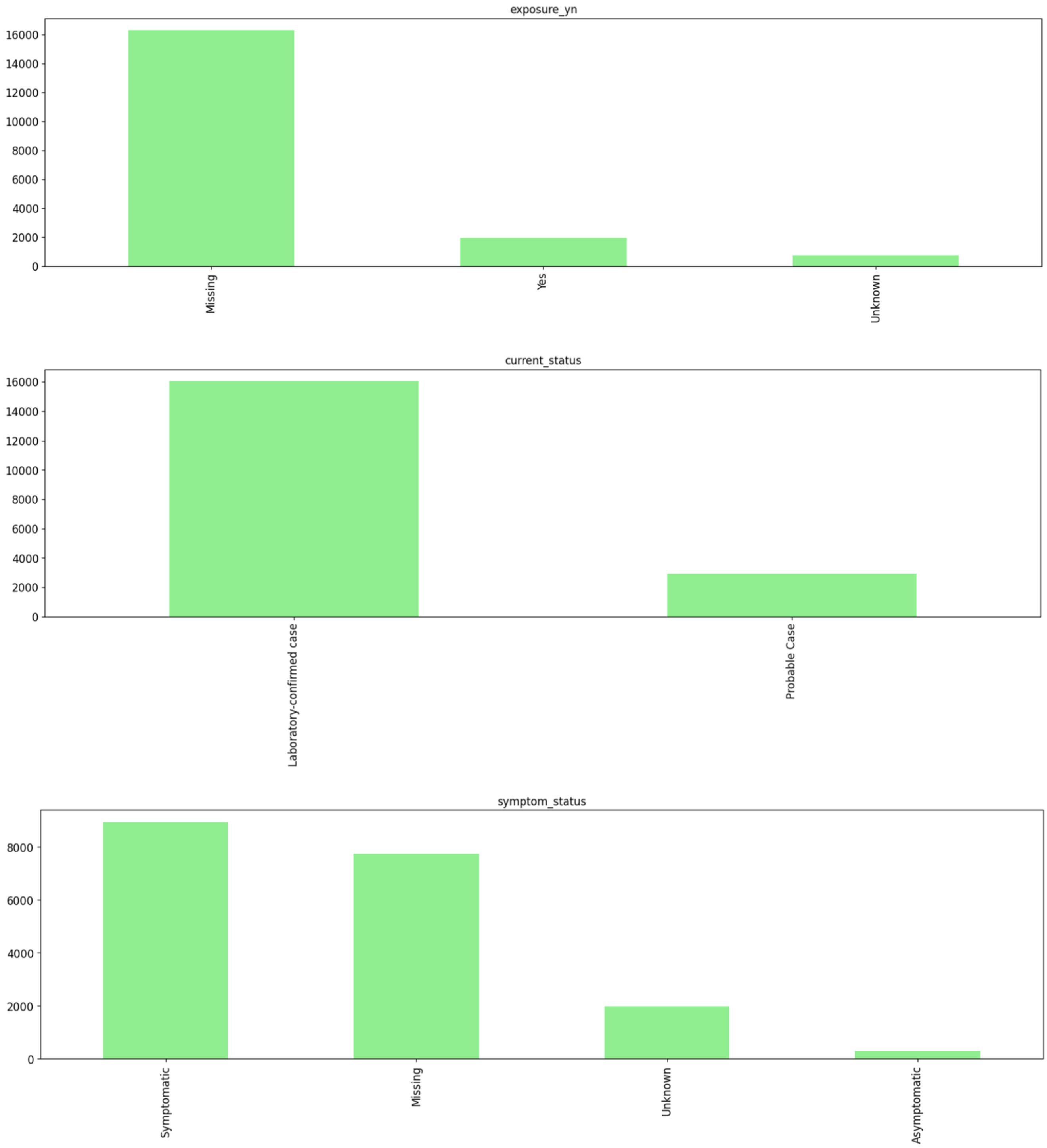

age_group
0
1000
2000
3000
4000
5000
6000
7000
18 to 49 years
65+ years
50 to 64 years
0 - 17 years
nan
Missing

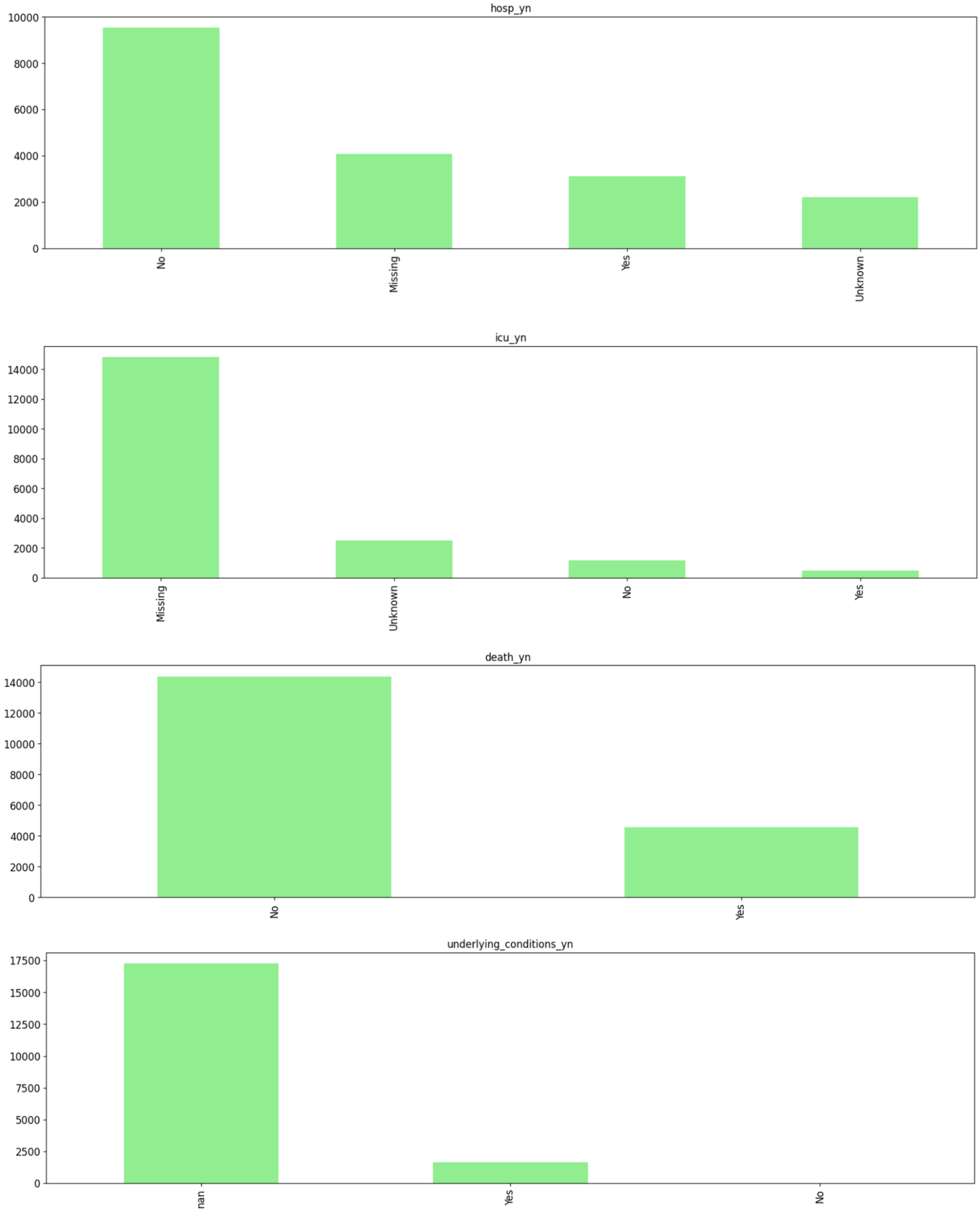

sex
0
2000
4000
6000
8000
10000
Female
Male
nan
Unknown
Missing


race
0
2000
4000
6000
8000
10000
12000
White
nan
Black
Unknown
Missing
Asian
Multiple/Other
American Indian/Alaska Native
Native Hawaiian/Other Pacific Islander


ethnicity
0
2000
4000
6000
8000
10000
Non-Hispanic/Latino
Unknown
nan
Hispanic/Latino
Missing


process
0
2500
5000
7500
10000
12500
15000
17500
Missing
Clinical evaluation
Laboratory reported
Routine surveillance
Multiple
Contact tracing of case patient
Unknown
Other
Provider reported


exposure_yn
0
2000
4000
6000
8000
10000
12000
14000
16000
Missing
Yes
Unknown
current_status
0
2000
4000
6000
8000
10000
12000
14000
16000
Laboratory-confirmed case
Probable Case
symptom_status
0
2000
4000
6000
8000
Symptomatic
Missing
Unknown
Asymptomatic


hosp_yn
10000
8000
6000
4000
2000
0
No
Missing
Yes
Unknown
icu_yn
14000
12000
10000
8000
6000
4000
2000
0
Missing
Unknown
No
Yes
death_yn
14000
12000
10000
8000
6000
4000
2000
0
No
Yes
underlying_conditions_yn
17500
15000
12500
10000
7500
5000
2500
0
nan
Yes
No