Data Quality Report

Initial Findings

Cian Belton
19321726

COMP47350: Data Analytics

UCD School of Computer Science

13/03/2023

Data Quality Report- Initial Findings:

1) Overview:

This report will include my findings for the cleaned dataset (part_1_covid19-cdc-19321726.csv). It will summarise the data, describe the various issues observed with the data and how these issues will be addressed. Please see the appendix for some background on this dataset. The appendix includes terminology, assumptions, explanations and summary of changes made to the original dataset. It also includes feature summaries, histograms, bar charts, pie charts and box plots used to visualise the data.

On first indication the dataset seems to lack a significant amount of values. 3 features have a large amount of null values. There are only 8 features with no null values. Upon further inspection it is evident that other categorical features contain null values appearing as Missing/Unknown. In the end there were only 3 features that had 20000 rows of complete data: case_month, current_status and death_yn. Other issues observed were negative and extremely large values for case_positive_specimen_interval and case_onset_interval features as these should be positive and within a reasonable range (0-10 weeks).

There was 1064 duplicate rows in this dataset. I decided to not drop these duplicate rows because there is no unique identifier for each row and it is plausible for the following scenario to occur:
2 people who live in the same county catch Covid in the same month and they are of the same race, ethnicity, gender and age range. They also have the same underlying conditions, so it appears as a duplicate row. However, they are not the same person.

2) Summary:

I carried out 3 tests on the dataset to check it's logical integrity. Test 1 and 2 both had a large number of failures and this is one of the reasons why I decide to drop the features used in these tests. Test 3 had only 2 failures so I will drop the two rows that this failure occurs in.

In the categorical features there were the alternative versions of null such as:
- Missing: Values left unanswered when the form was filled out
-Unknown: A choice on the form i.e., for age group the options were: [0 - 17 years; 18 - 49 years; 50 - 64 years; 65 + years; Unknown; Missing; NA]

3) Review Logical Integrity:

3 tests were carried out to test the logical integrity of the data. The results are below:

1) Check the values for case_positive_specimen_interval:
   a) 9,546 rows failed this test.
   b) 85 of these values were the non-null values that I converted to -1 in order to store this feature as type 'int'.
   c) I am considering values inside of the range 0-10 weeks to be valid rows.

2) Check the values for case_onset_interval:
   a) 11,653 rows failed this test.
   b) I am considering values inside of the range 0-10 weeks to be valid rows.

3) Check that all rows that have yes for icu_yn also have yes for hosp_yn:
   a) 2 rows failed this test.
   b) If you are in an intensive care unit you must have been in a hospital.

4) Review Continuous Features:

4.1) Descriptive Statistics:

There are 2 continuous features in this dataset. They are:

- case_positive_specimen_interval:
  a. This feature has 9461 null values
  b. 85 values were outside of the range 0-10 weeks.
  c. I have decided to drop this feature for the following reasons:
     i. Large number of null values (47.3%)
     ii. Large number of 0 values (46.9%)
     iii. Failure of logical integrity test.
     iv. The specimen interval has no effect on the target outcome (death)

- case_onset_interval:
  a. This feature has 11365 null values
  b. 288 values were outside of the range 0-10 weeks.
  c. I have decided to drop this feature for the following reasons:
     i. Large number of null values (56.8%)
     ii. Large number of 0 values (41.6%)
     iii. Failure of logical integrity test.
     iv. The onset interval has no effect on the target outcome (death)

4.2) Charts:

See histograms and box plots for the continuous features in the Appendix.

5) Review Categorical Features:

5.1) Descriptive Statistics:

There are 17 categorical features in this dataset. They are:

- case_month:
  a. This feature has no issues (keep)

- res_state:
  a. This feature has only 1 null value
  b. Drop this row as it is only 1 row
  c. No major issues identified here (keep)

- state_fips_code:
  a. This feature has only 1 null value
  b. Drop feature as it acts like a duplicate of res_state

- res_county:
  a. This feature has 1195 null values
  b. No other major issues identified here (keep)

- county_fips_code:
  a. This feature has 1195 null values
  b. Drop feature as it acts like a duplicate of res_county

- age_group:
  a. This feature has 150 null values
  b. 34 Missing Values
  c. No major issues identified here (keep)

- sex:
  a. This feature has 377 null values
  b. 22 Missing values
  c. 85 Unknown values
  d. No major issues identified here (keep)

- race:
  a. This feature has 2297 null values
  b. 732 Missing values
  c. 1715 Unknown values
  d. I have decided to keep this feature as it may be useful for the prediction of death risk.

- ethnicity:
  a. This feature has 2490 null values
  b. 1081 Missing values
  c. 2736 Unknown values
  d. I have decided to keep this feature as it may be useful for the prediction of death risk.

- process:
  a. This feature has 18311 Missing values
  b. 57 Unknown values
  c. I have decided to drop this feature for the following reasons:
     i. Large number of missing and unknown values (91.8%)
     ii. The process under which the case was identified has no effect on the target outcome (death).

- exposure_yn:
  a. This feature has 17250 Missing values
  b. 799 Unknown values
  c. I have decided to drop this feature for the following reasons:
     i. Large number of missing and unknown values (90.2%)
     ii. The way in which the person might have been exposed to the virus has no effect on the target outcome (death).

- current_status:
  a. This feature has no issues (keep)

- symptom_status:
  a. This feature has 8358 Missing values
  b. 2233 Unknown values
  c. I have decided to drop this feature because of its:
     i. Large number of missing and unknown values (53%). If there is more than 50% of the data missing I will drop the feature.

- hosp_yn:
  a. This feature has 4336 Missing values
  b. 2338 Unknown values
  c. I have decided to keep this feature as it may be useful for the prediction of death risk.

- icu_yn:
  a. This feature has 15446 Missing values
  b. 2776 Unknown values

c. 2 rows failed logical test 3

d. I have decided to not drop this feature even though 91% of data is missing because I expect this to be a good predictor of death risk. So I will take it into account when analysing.

- death_yn:

  a. This feature has no issues (keep)

- underlying_conditions_yn:

  a. This feature has 18299 null values

  a. I have decided to not drop this feature even though 91% of data is missing because I expect this to be a good predictor of death risk. So I will take it into account when analysing.

5.2) Charts:

See bar plots and pie charts for the categorical features in the Appendix.

6) Actions to Take:

The following actions will be taken:

1. Dropping the following features for the reasons outlined above:
   i. case_positive_specimen_interval
   ii. case_onset_interval
   iii. state_fips_code
   iv. county_fips_code
   v. process
   vi. exposure_yn
   vii. symptom_status

2. Drop null rows in:
   i. res_state (1 row)

3. Rows failing logical tests:
   i. Rows that failed test 1 and 2 will be kept due to the large amount of data that would be discarded if I dropped these rows.
   ii. Rows that failed test 3 will be dropped.

4. Missing/Unknown Values:
   i. For features that contain large amounts of missing/unknown values I will not drop these rows as they still contain valid information for other features.
   ii. Replace all Unknown and null values with Missing to represent one value
   iii. Will remove missing values when studying individual features to give the maximum amount of information. This is much better than removing all the rows containing missing/unknown values.

5. Imputation:
   i. I will not carry out imputation due to the difficulty in using it for categorical features.
   ii. Instead I will analyse the features with and without the missing values to show how they relate to death risk.

7) References:

1) CDC Description of data
https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

8) Appendix:

a. Data Dictionary:

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

-res_state: State of residence

-state_fips_code: State FIPS code

-res_county: County of residence

-county_fips_code: County FIPS code

-age_group: Age group [0 - 17 years; 18 - 49 years; 50 - 64 years; 65 + years; Unknown; Missing; NA, if value suppressed for privacy protection.]

-sex: Sex [Female; Male; Other; Unknown; Missing; NA, if value suppressed for privacy protection.]

-race: Race [American Indian/Alaska Native; Asian; Black; Multiple/Other; Native Hawaiian/Other Pacific Islander; White; Unknown; Missing; NA, if value suppressed for privacy protection.]

-ethnicity: Ethnicity [Hispanic; Non-Hispanic; Unknown; Missing; NA, if value suppressed for privacy protection.]

-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? [Clinical evaluation; Routine surveillance; Contact tracing of case patient; Multiple; Other; Unknown; Missing]

-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? [Yes, Unknown, Missing]

-current_status: What is the current status of this person? [Laboratory-confirmed case, Probable case]

-symptom_status: What is the symptom status of this person? [Asymptomatic, Symptomatic, Unknown, Missing]

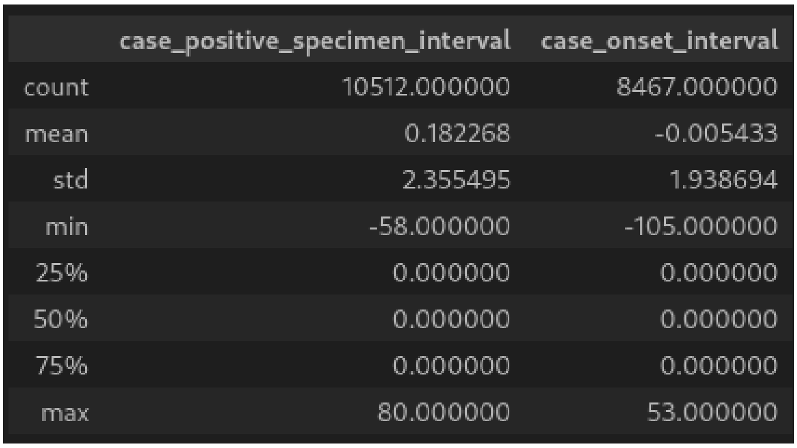
-hosp_yn: Was the patient hospitalized? [Yes, No, Unknown, Missing]

-icu_yn: Was the patient admitted to an intensive care unit (ICU)? [Yes, No, Unknown, Missing]

-death_yn: Did the patient die as a result of this illness? [Yes; No; Unknown; Missing; NA, if value suppressed for privacy protection.]

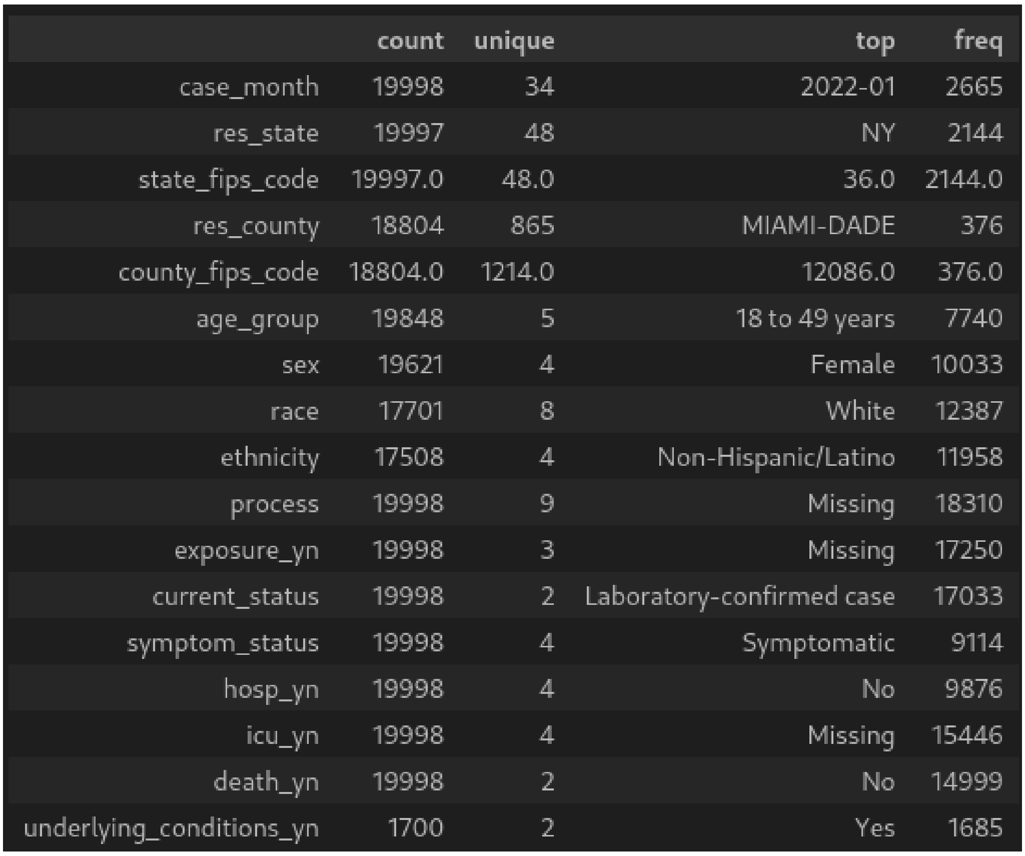
-underlying_conditions_yn: Did the patient have one or more of the underlying medical conditions and risk behaviors: diabetes mellitus, hypertension, severe obesity (BMI>40), cardiovascular disease, chronic renal disease, chronic liver disease, chronic lung disease, other chronic diseases, immunosuppressive condition, autoimmune condition, current smoker, former smoker, substance abuse or misuse, disability, psychological/psychiatric, pregnancy, other. [Yes, No, blank]

b. Continuous Features:



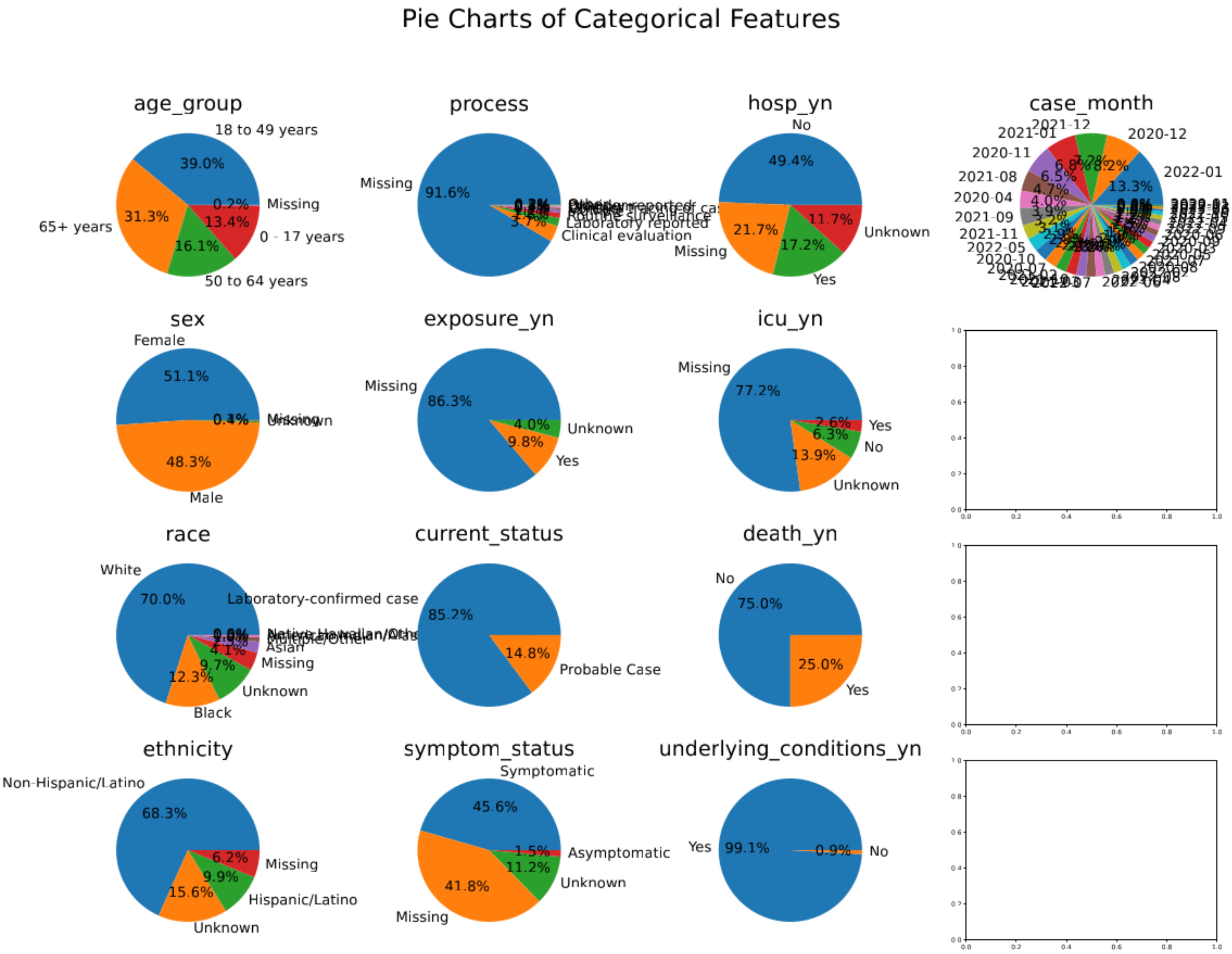
	case_positive_specimen_interval	case_onset_interval
count	10512.000000	8467.000000
mean	0.182268	-0.005433
std	2.355495	1.938694
min	-58.000000	-105.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	80.000000	53.000000

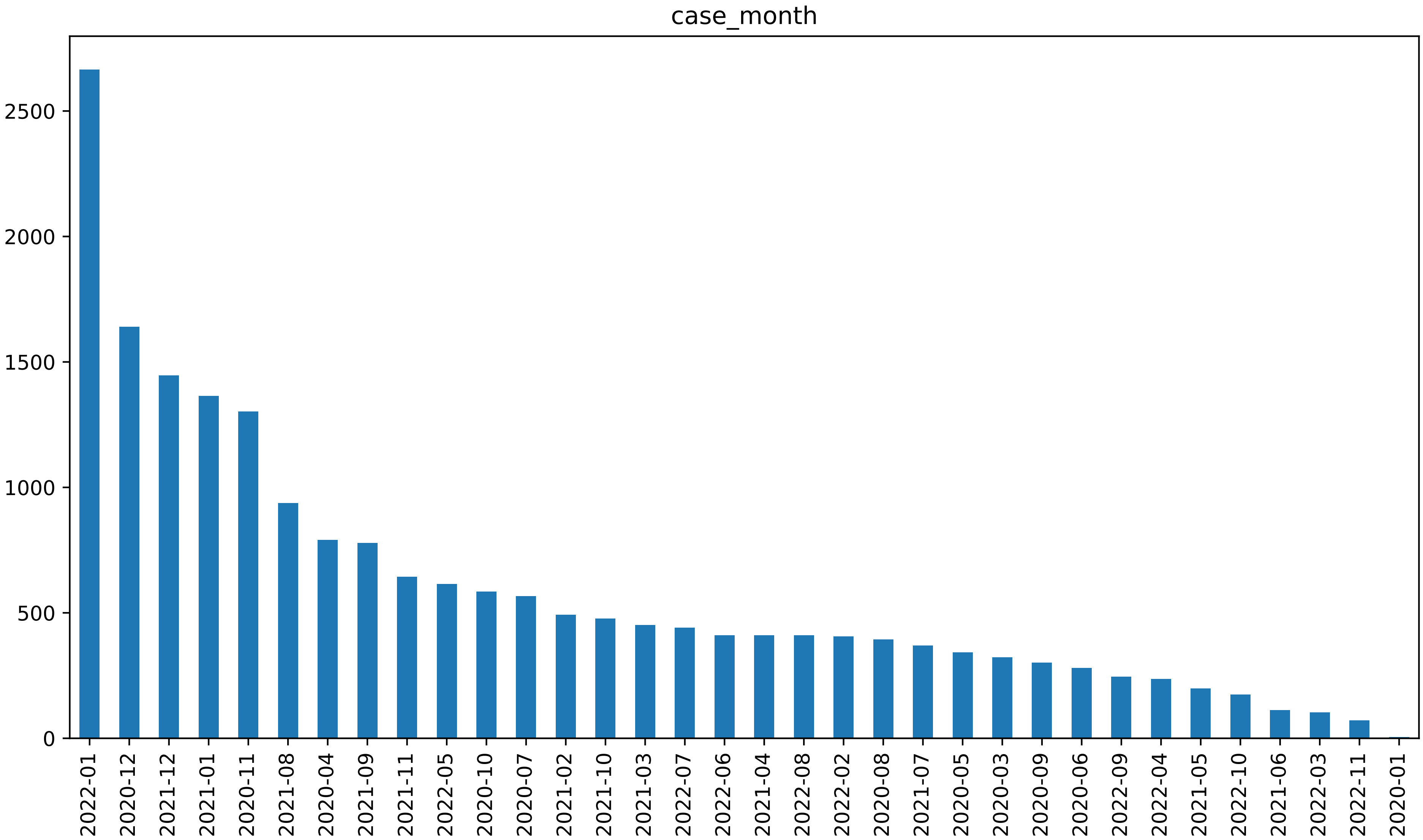
c. Categorical Features:
(after I have already dropped the 2 rows that failed test 3)

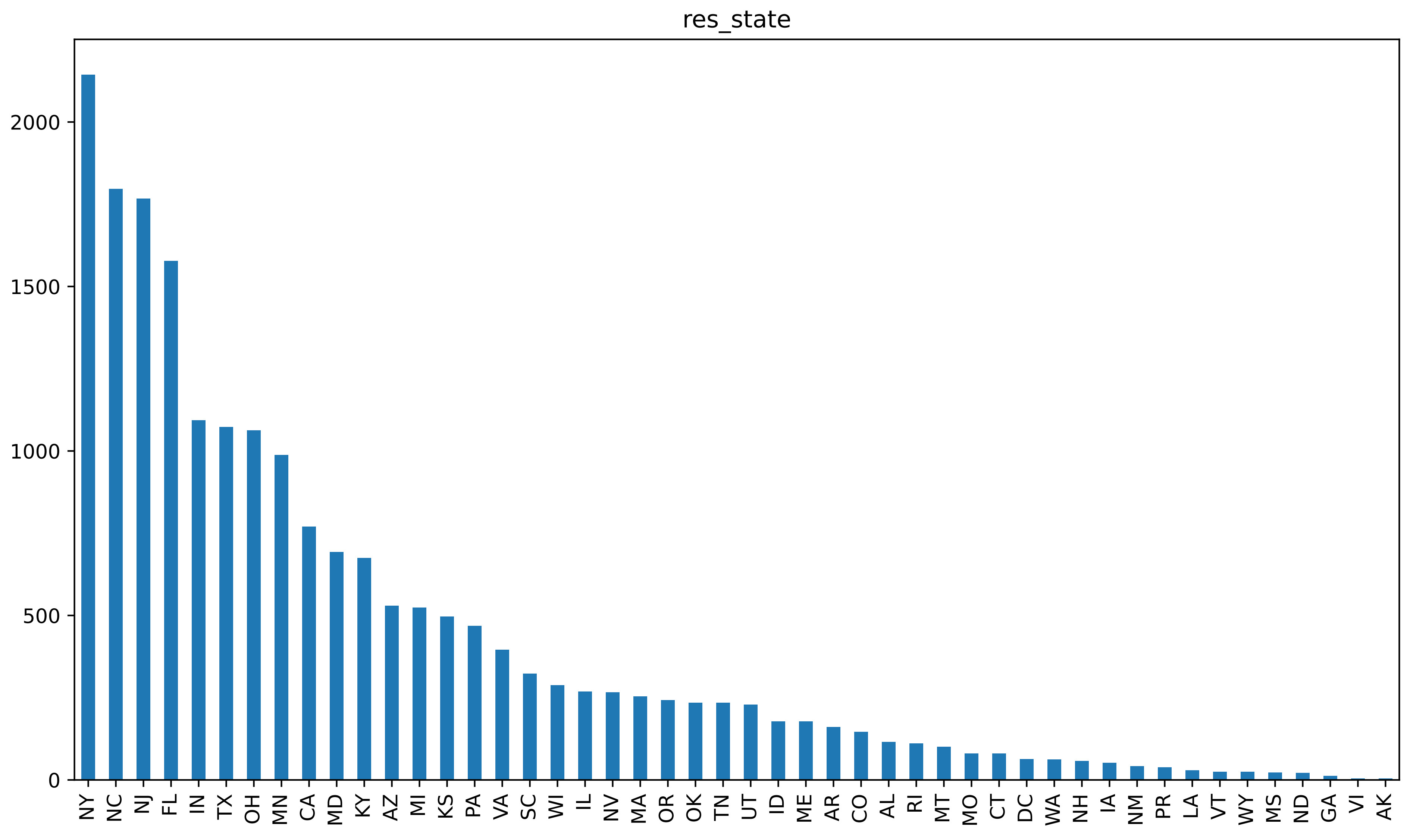


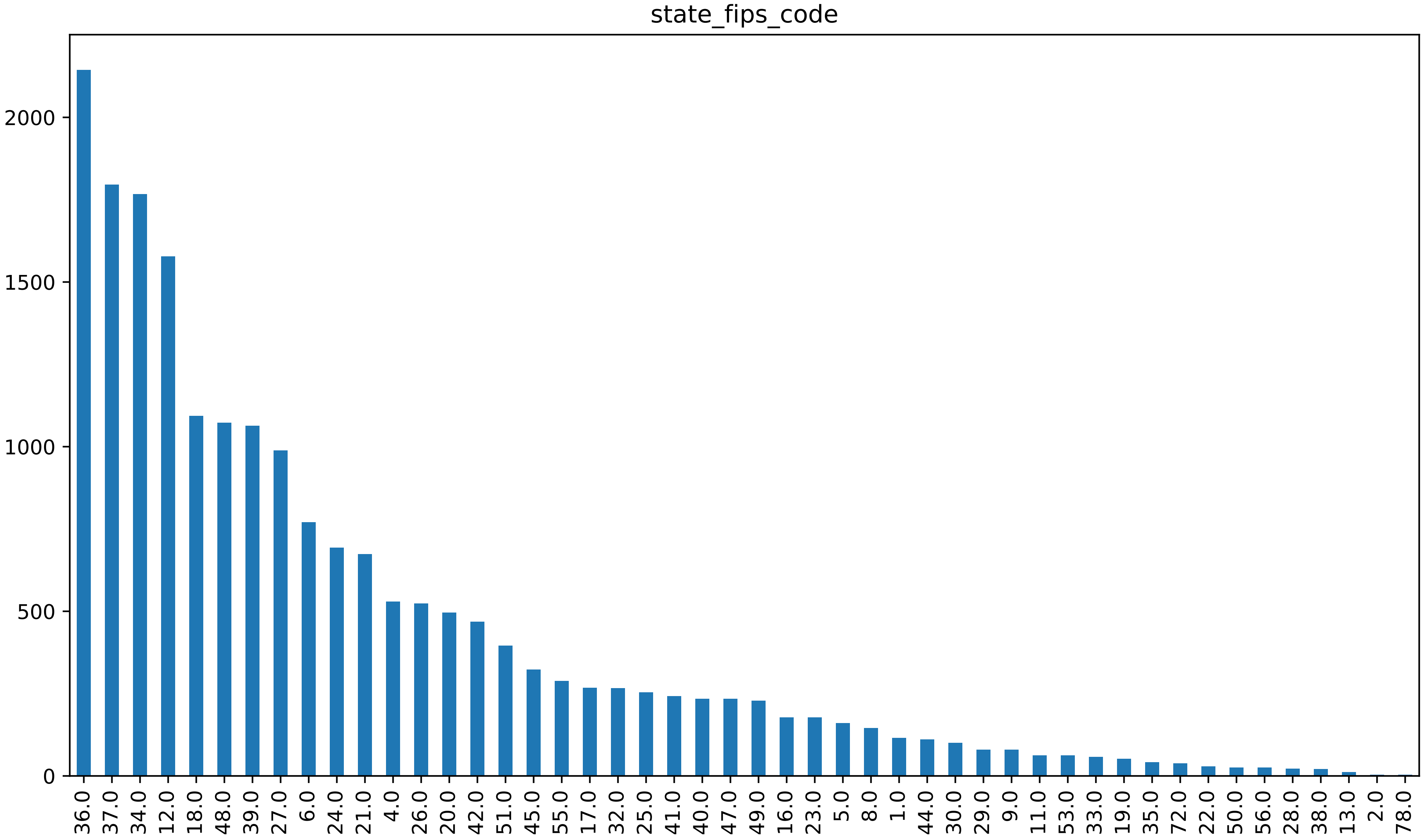
	count	unique	top	freq
case_month	19998	34	2022-01	2665
res_state	19997	48	NY	2144
state_fips_code	19997.0	48.0	36.0	2144.0
res_county	18804	865	MIAMI-DADE	376
county_fips_code	18804.0	1214.0	12086.0	376.0
age_group	19848	5	18 to 49 years	7740
sex	19621	4	Female	10033
race	17701	8	White	12387
ethnicity	17508	4	Non-Hispanic/Latino	11958
process	19998	9	Missing	18310
exposure_yn	19998	3	Missing	17250
current_status	19998	2	Laboratory-confirmed case	17033
symptom_status	19998	4	Symptomatic	9114
hosp_yn	19998	4	No	9876
icu_yn	19998	4	Missing	15446
death_yn	19998	2	No	14999
underlying_conditions_yn	1700	2	Yes	1685

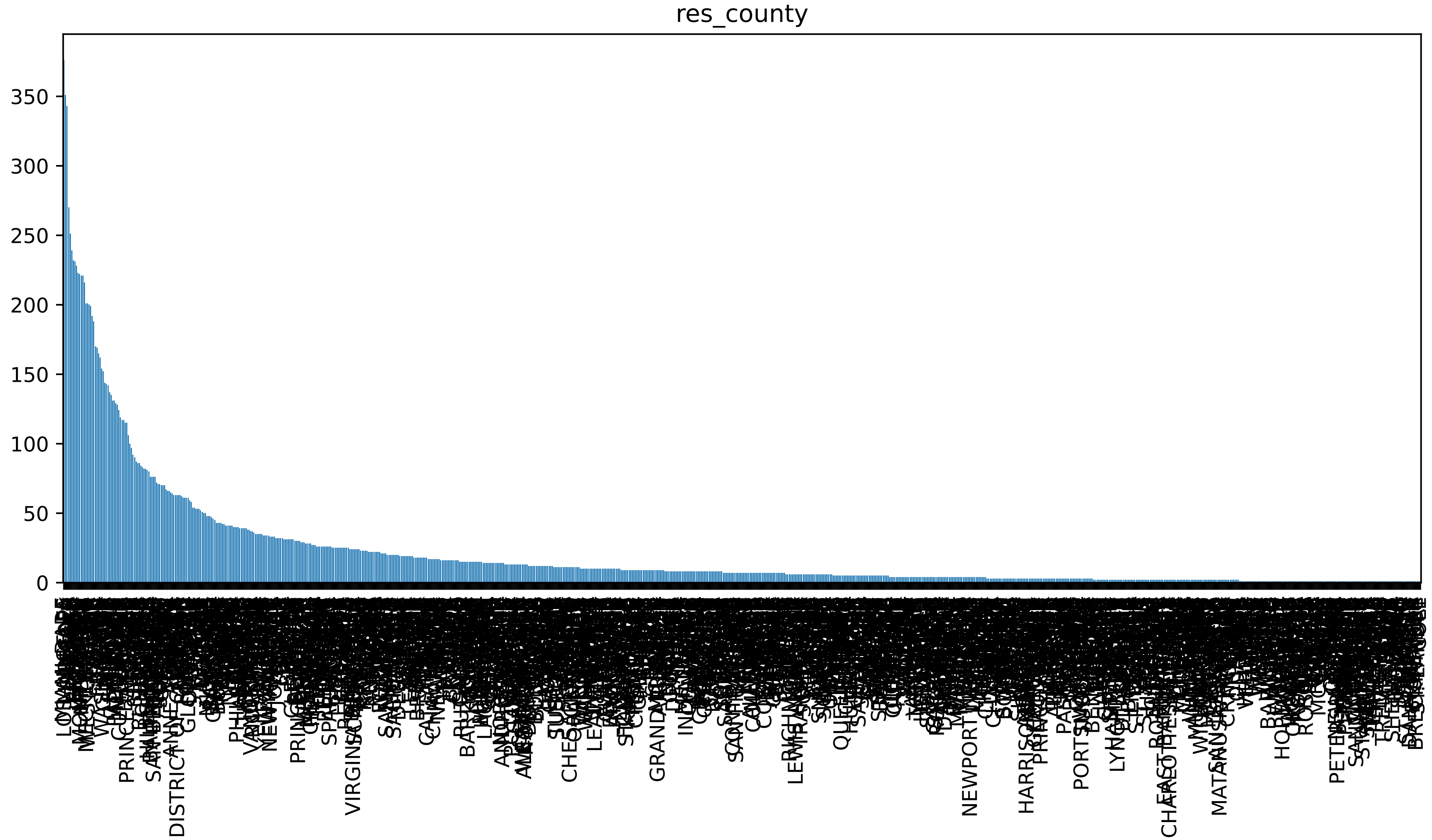
d. Box Plots, Bar Plots, Pie Charts & Histograms:
See below for summary plots and histograms. PDF files included will show plots in more detail.

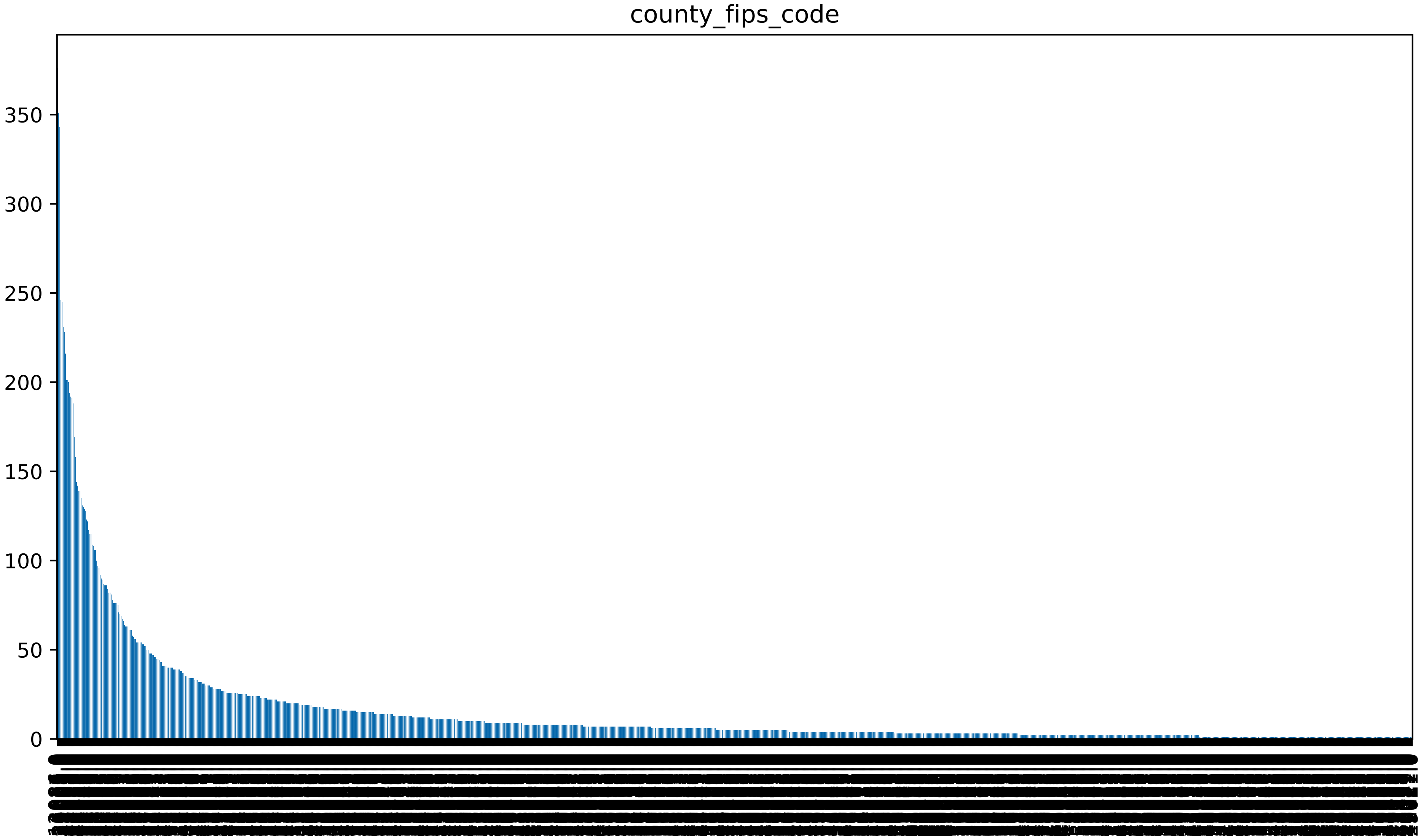


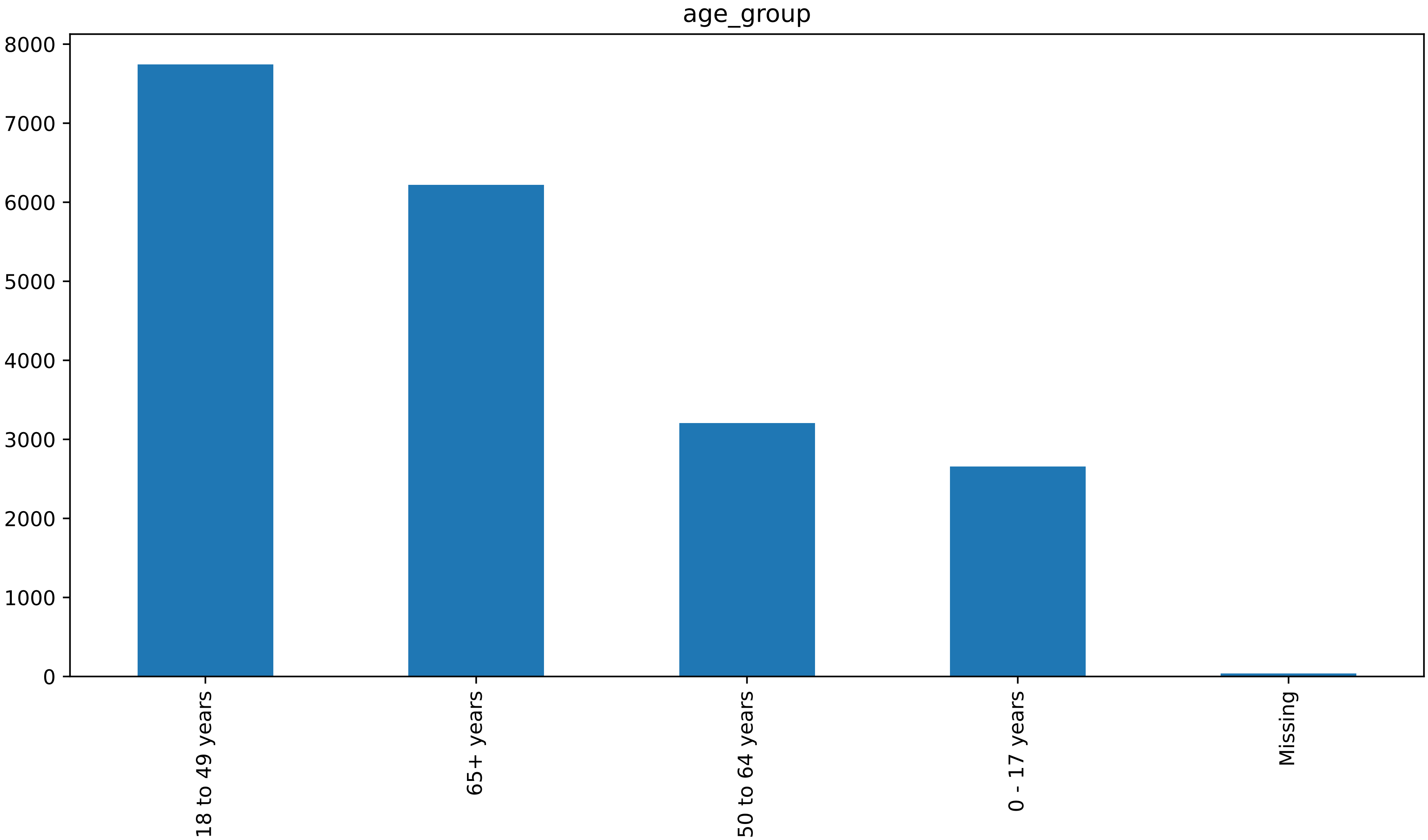

case_month
2500
2000
1500
1000
500
0
2022-01
2020-12
2021-12
2021-01
2020-11
2021-08
2020-04
2021-09
2021-11
2022-05
2020-10
2020-07
2021-02
2021-10
2021-03
2022-07
2022-06
2021-04
2022-08
2022-02
2020-08
2021-07
2020-05
2020-03
2020-09
2020-06
2022-09
2022-04
2021-05
2022-10
2021-06
2022-03
2022-11
2020-01

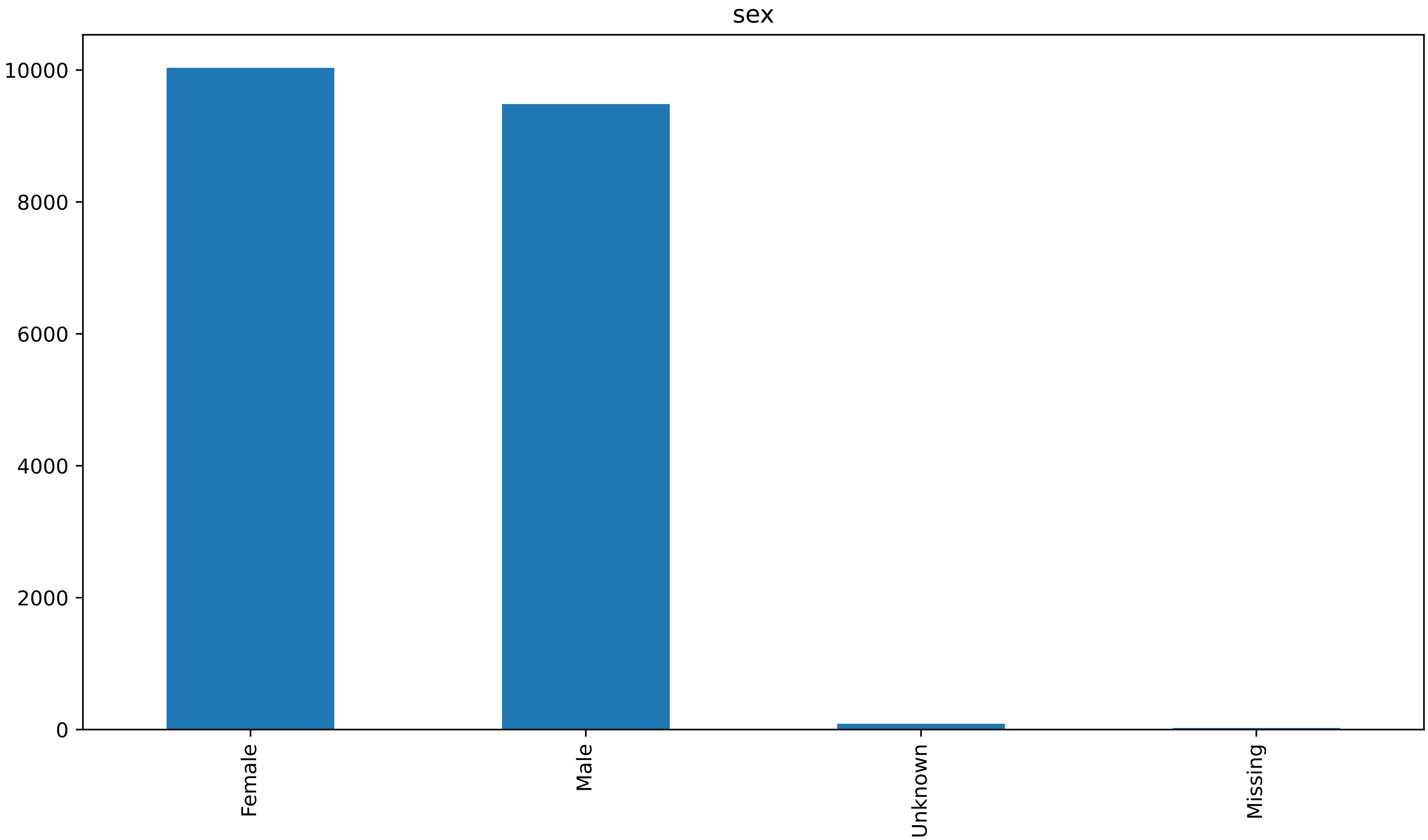

res_state
2000
1500
1000
500
0
NY
NC
NJ
FL
IN
TX
OH
MN
CA
MD
KY
AZ
MI
KS
PA
VA
SC
WI
IL
NV
MA
OR
OK
TN
UT
ID
ME
AR
CO
AL
RI
MT
MO
CT
DC
WA
NH
IA
NM
PR
LA
VT
WY
MS
ND
GA
VI
AK

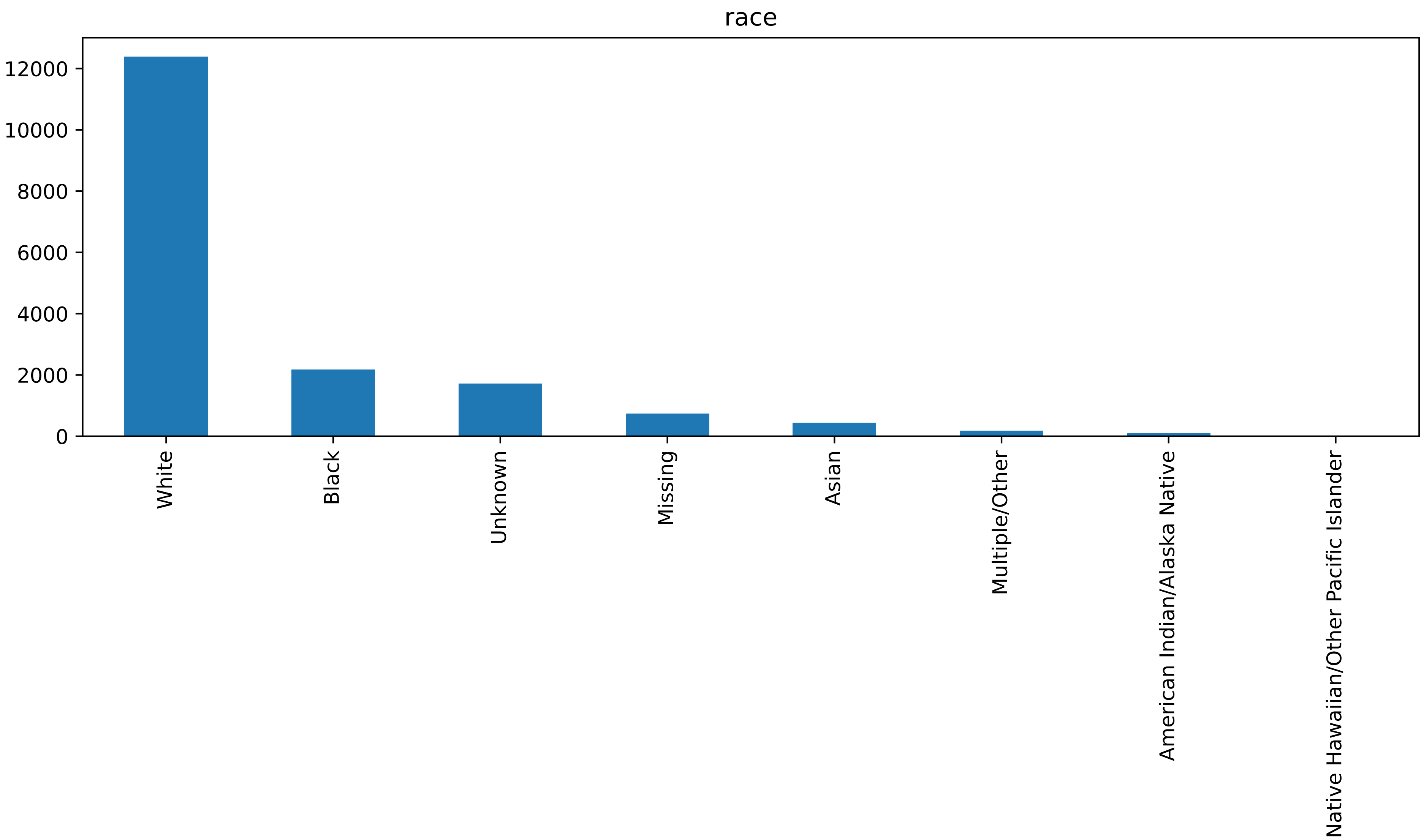

state_fips_code
2000
1500
1000
500
0
36.0
37.0
34.0
12.0
18.0
48.0
39.0
27.0
6.0
24.0
21.0
4.0
26.0
20.0
42.0
51.0
45.0
55.0
17.0
32.0
25.0
41.0
40.0
47.0
49.0
16.0
23.0
5.0
8.0
1.0
44.0
30.0
29.0
9.0
11.0
53.0
33.0
19.0
35.0
72.0
22.0
50.0
56.0
28.0
38.0
13.0
2.0
78.0

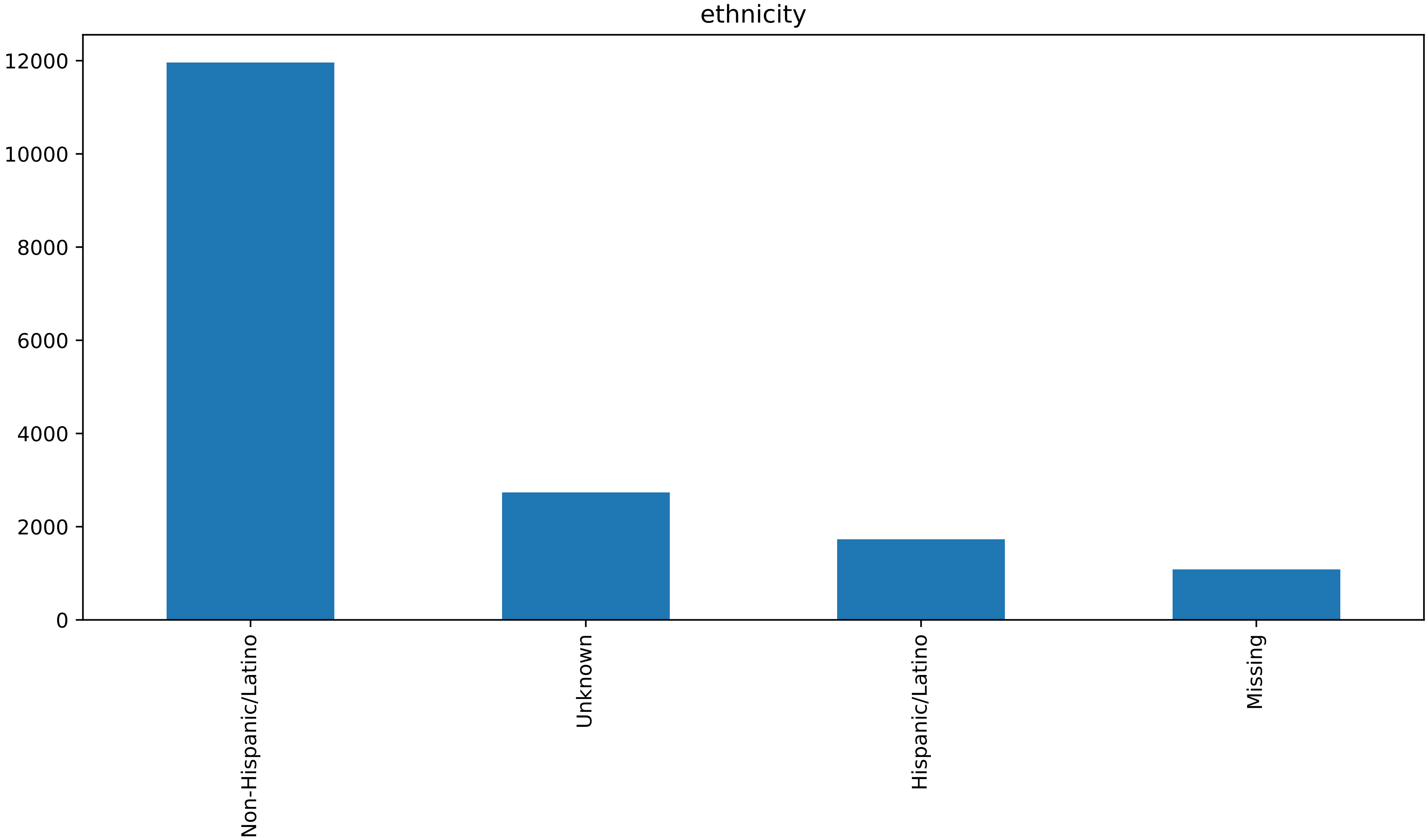

res_county
0
50
100
150
200
250
300
350
DISTRICT O
VIRGINIA
NEWPORT
HARRISON
PORTSMO
CHARLOTTE
MATAN
PETERSB

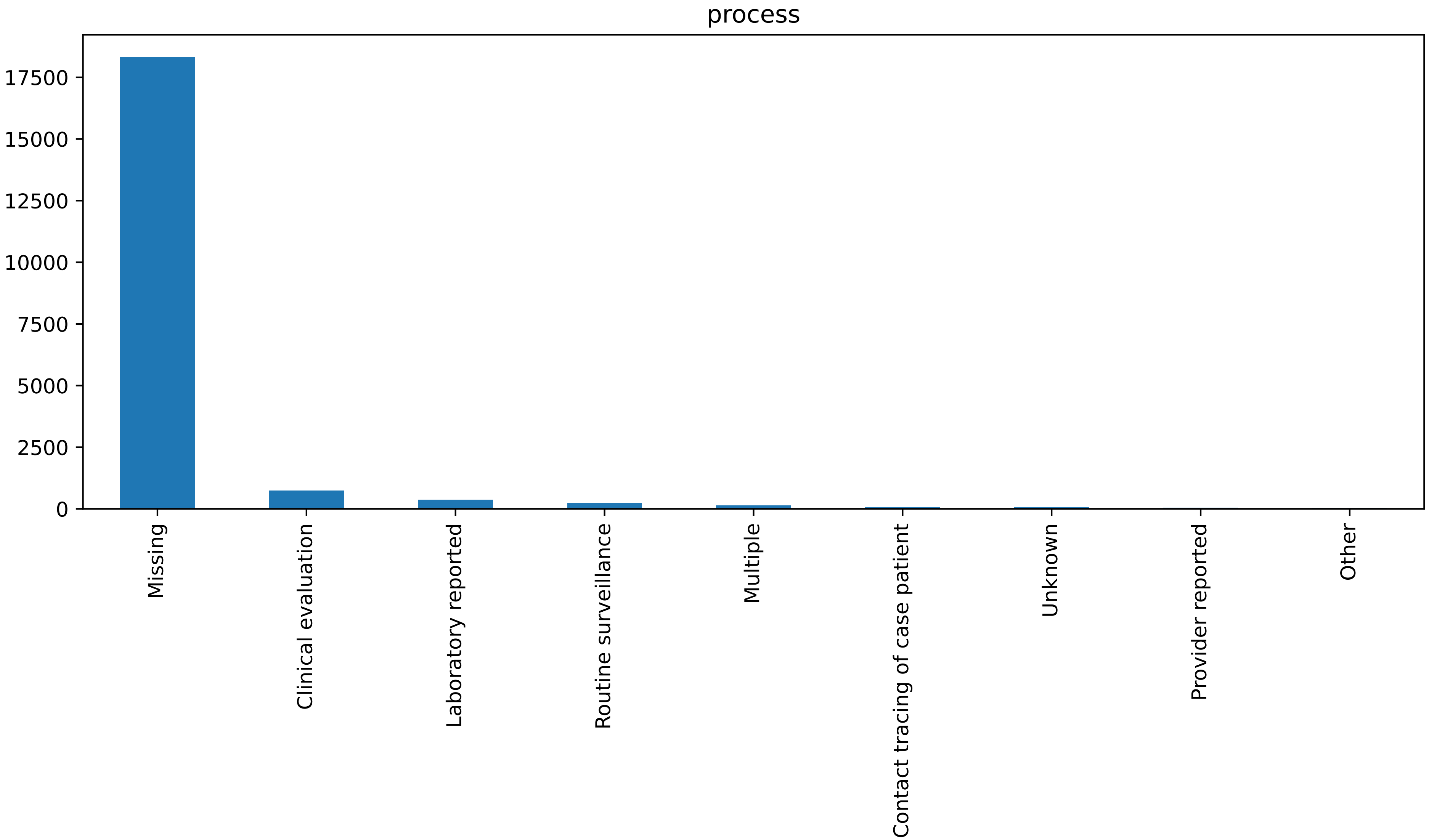

county_fips_code
350
300
250
200
150
100
50
0

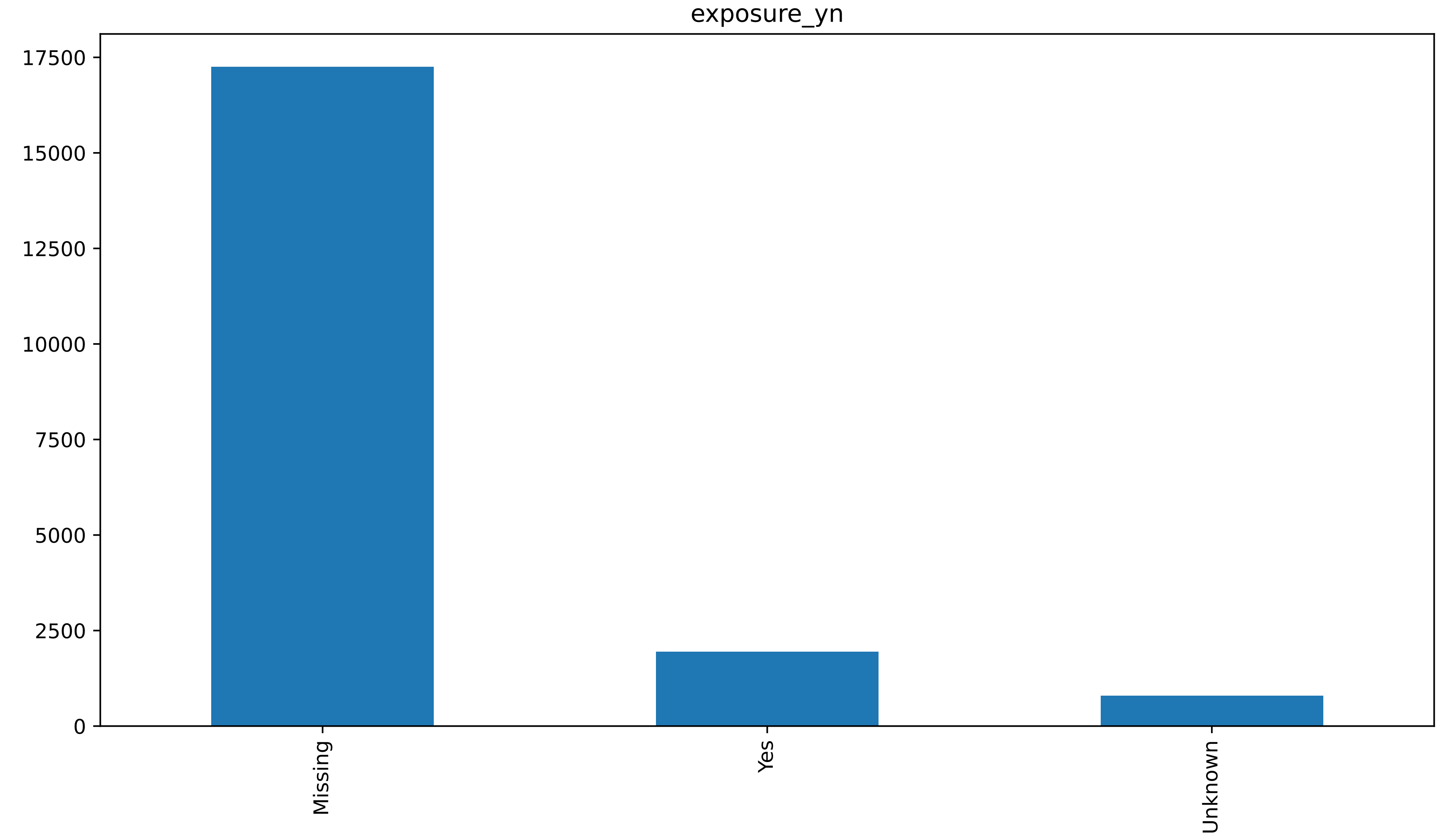

age_group
8000
7000
6000
5000
4000
3000
2000
1000
0
18 to 49 years
65+ years
50 to 64 years
0 - 17 years
Missing

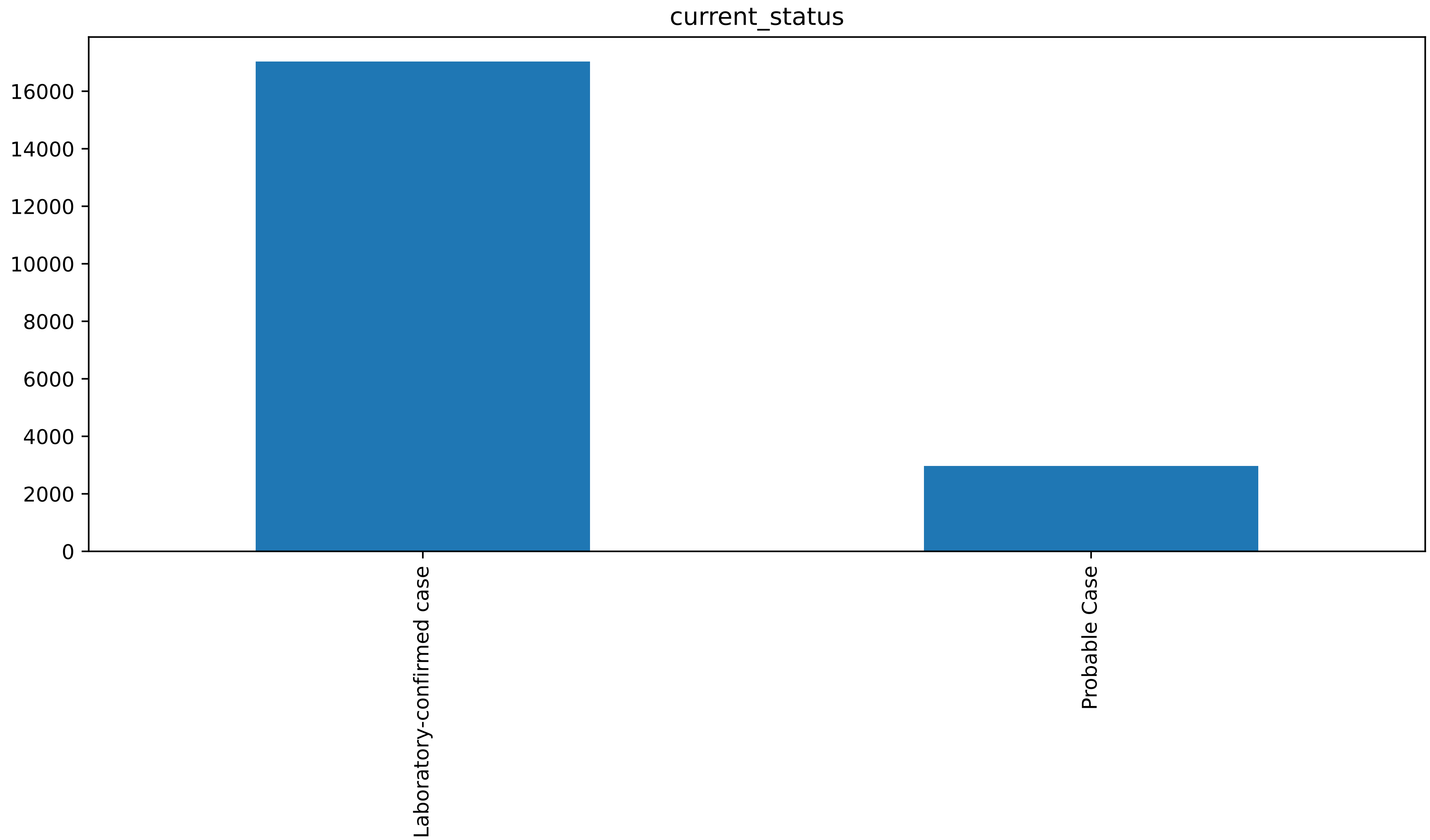

sex
10000
8000
6000
4000
2000
0
Female
Male
Unknown
Missing

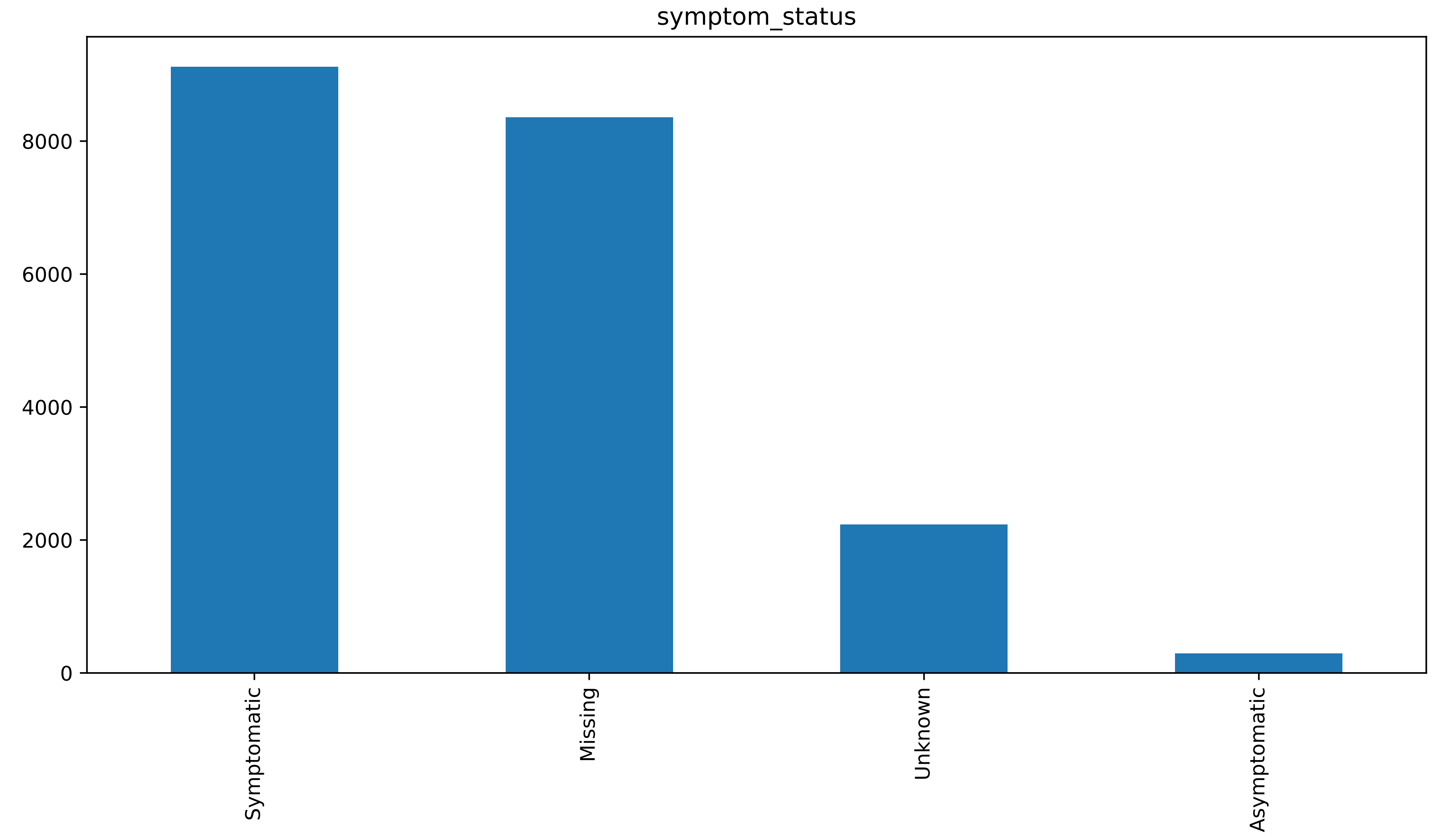

race
0
2000
4000
6000
8000
10000
12000
White
Black
Unknown
Missing
Asian
Multiple/Other
American Indian/Alaska Native
Native Hawaiian/Other Pacific Islander

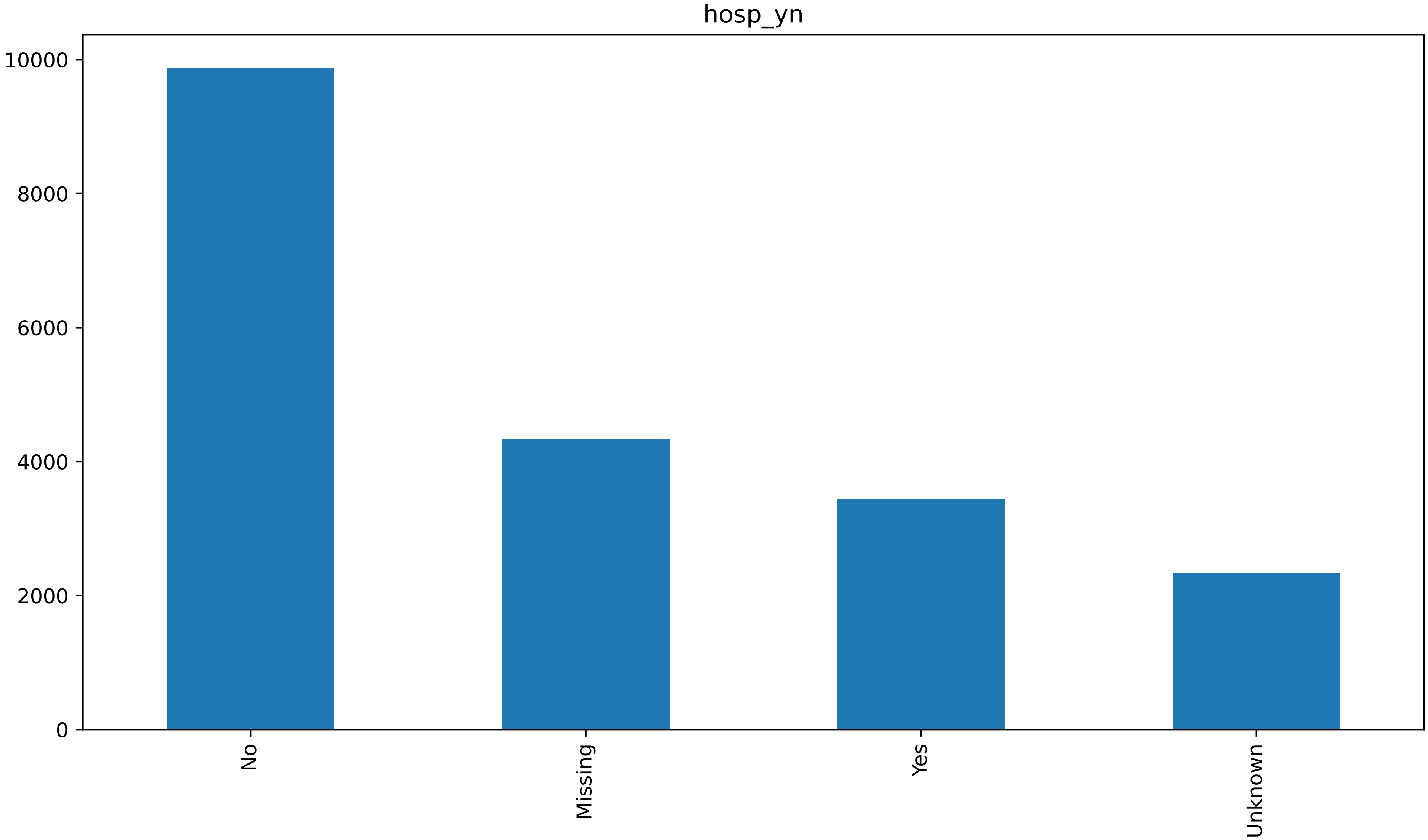

ethnicity
12000
10000
8000
6000
4000
2000
0
Non-Hispanic/Latino
Unknown
Hispanic/Latino
Missing

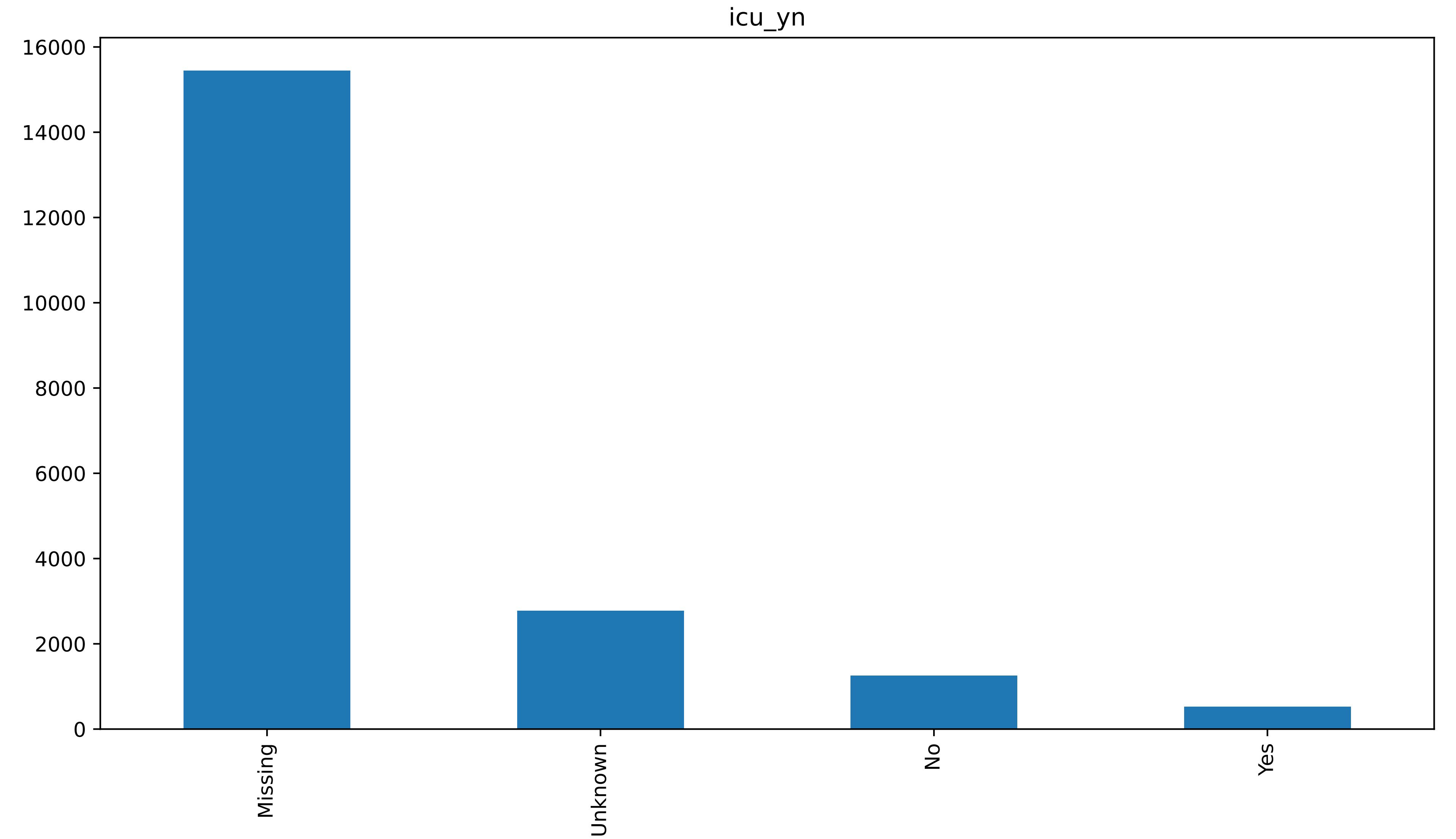

process
17500
15000
12500
10000
7500
5000
2500
0
Missing
Clinical evaluation
Laboratory reported
Routine surveillance
Multiple
Contact tracing of case patient
Unknown
Provider reported
Other

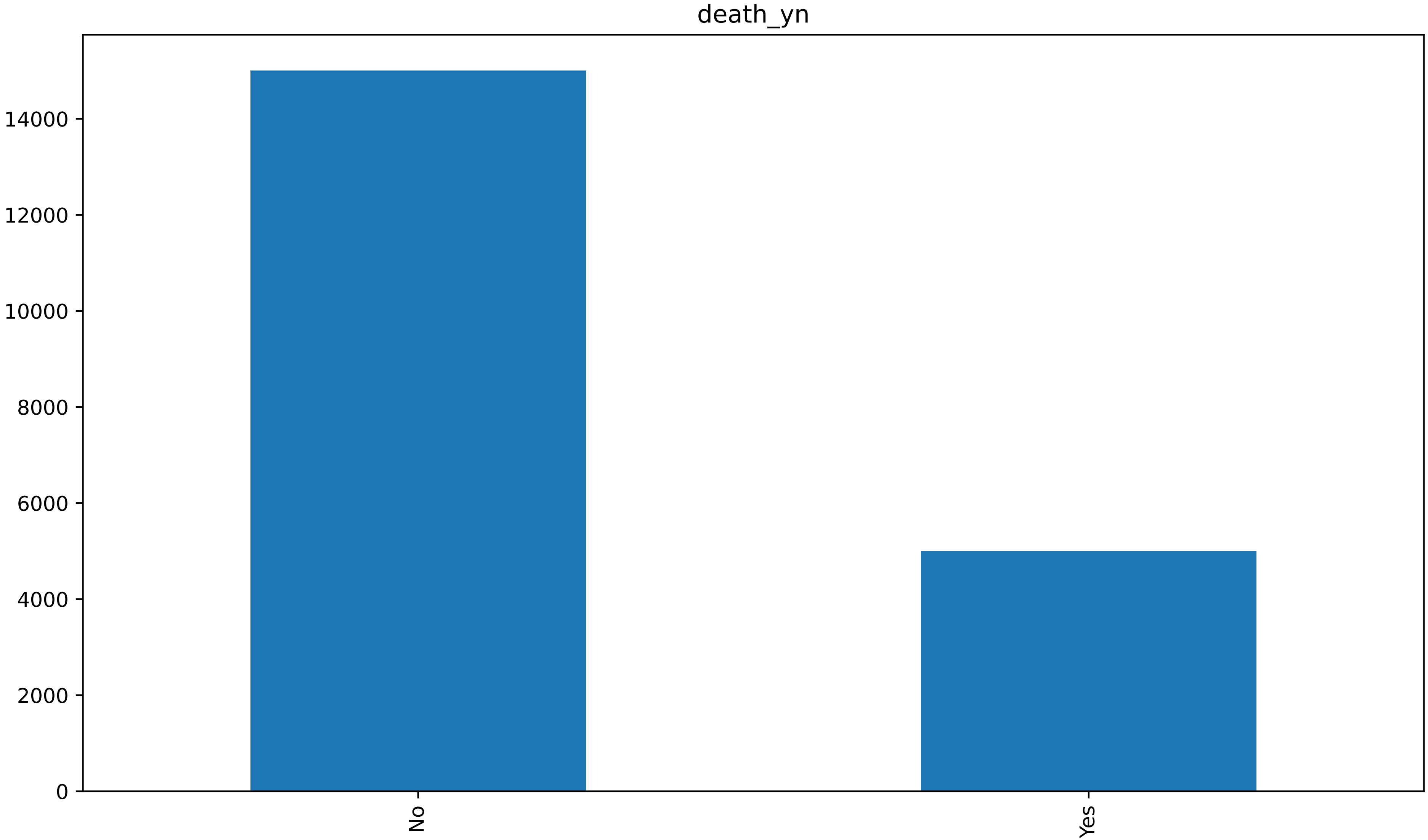

exposure_yn
17500
15000
12500
10000
7500
5000
2500
0
Missing
Yes
Unknown

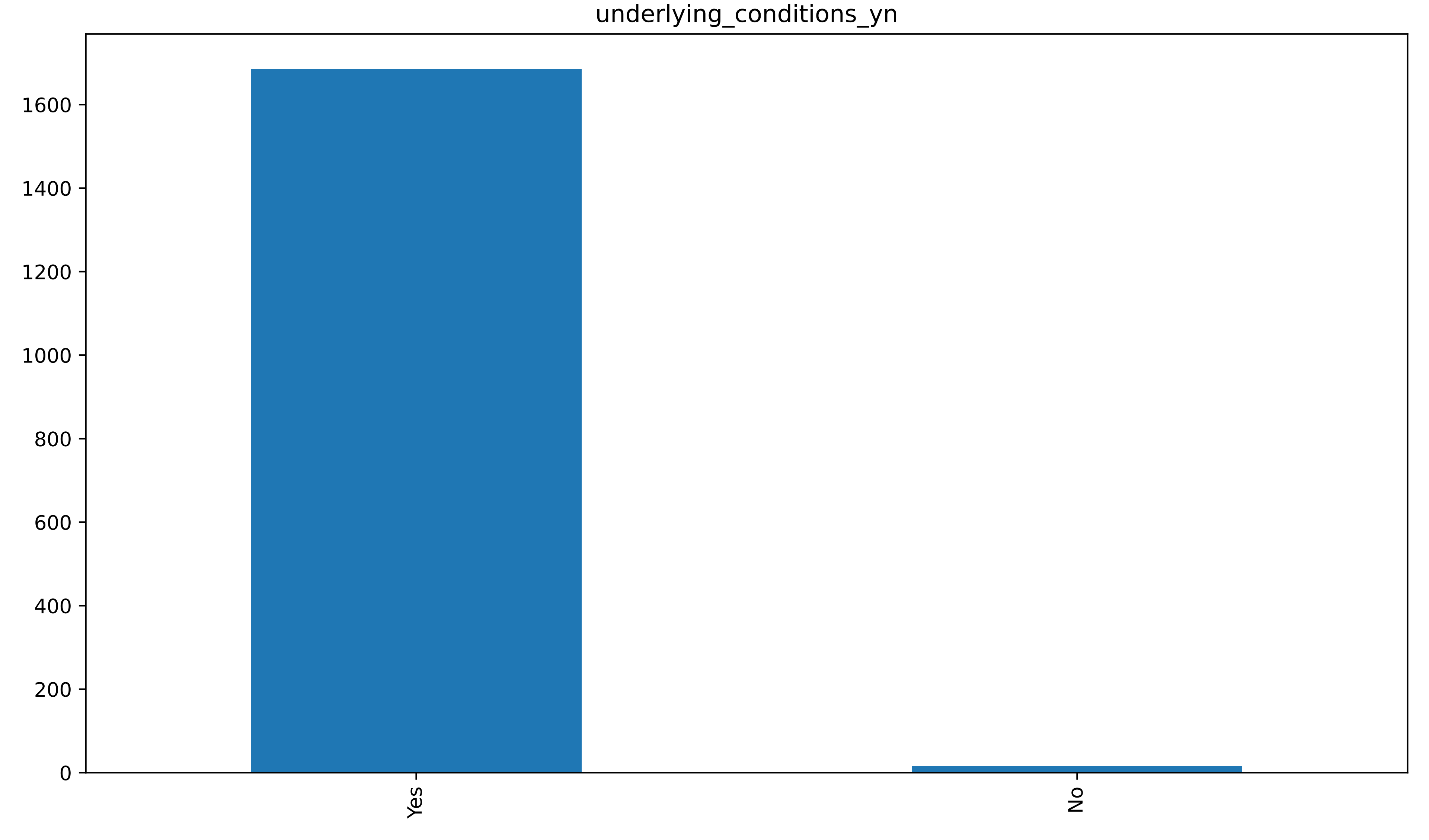

current_status
16000
14000
12000
10000
8000
6000
4000
2000
0
Laboratory-confirmed case
Probable Case

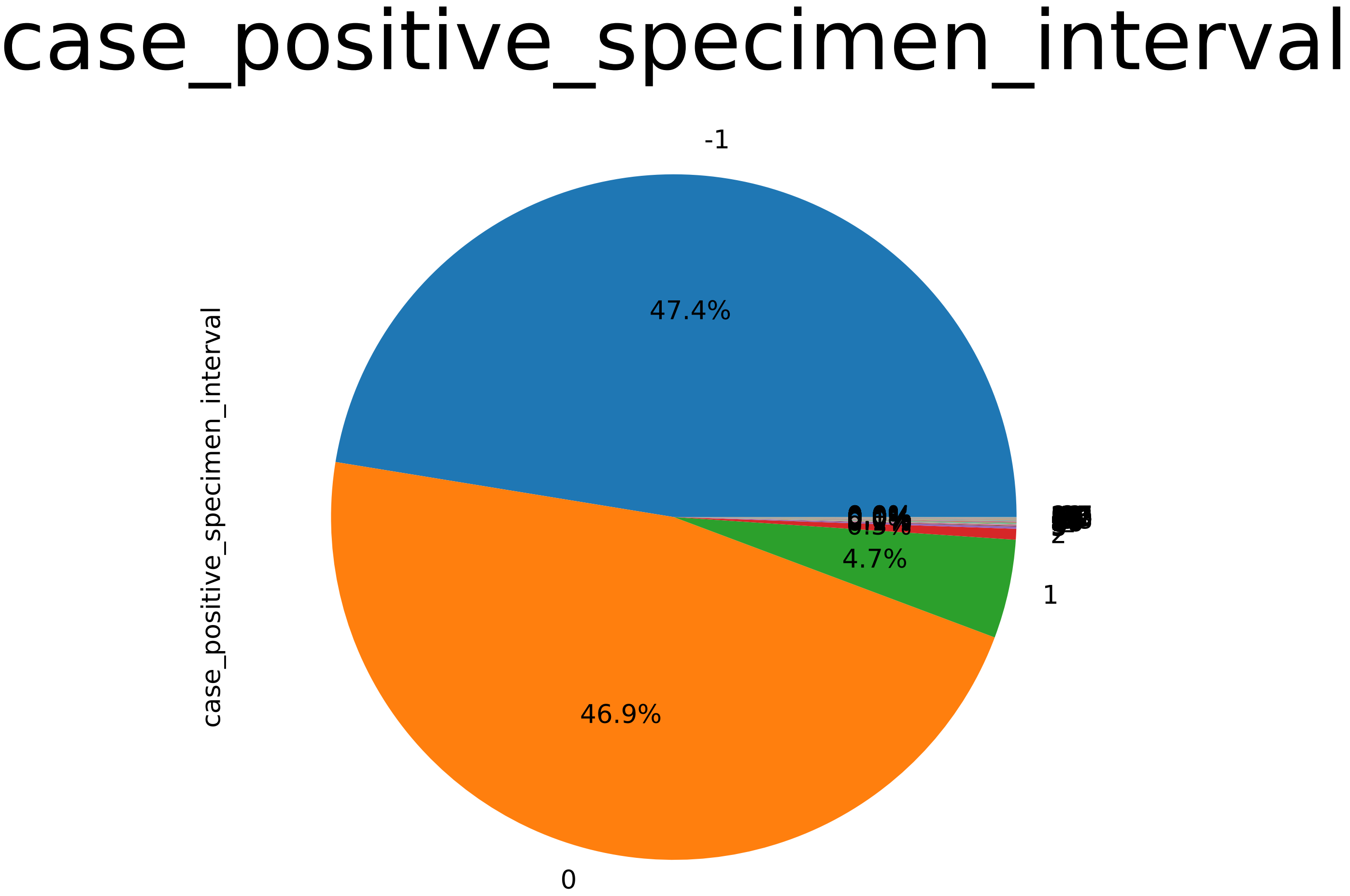

symptom_status
8000
6000
4000
2000
0
Symptomatic
Missing
Unknown
Asymptomatic

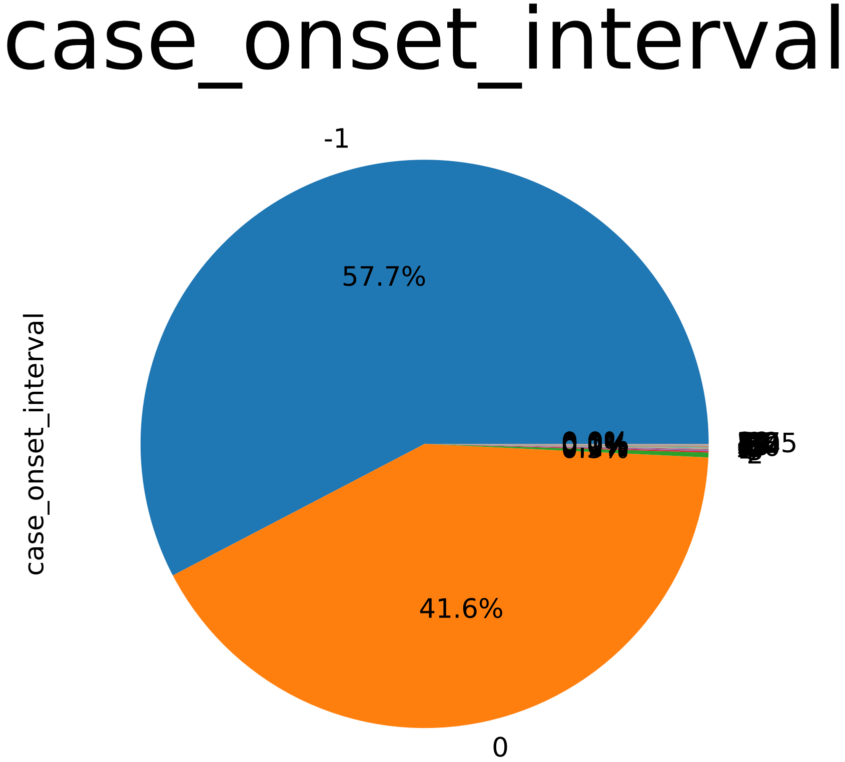

hosp_yn
10000
8000
6000
4000
2000
0
No
Missing
Yes
Unknown

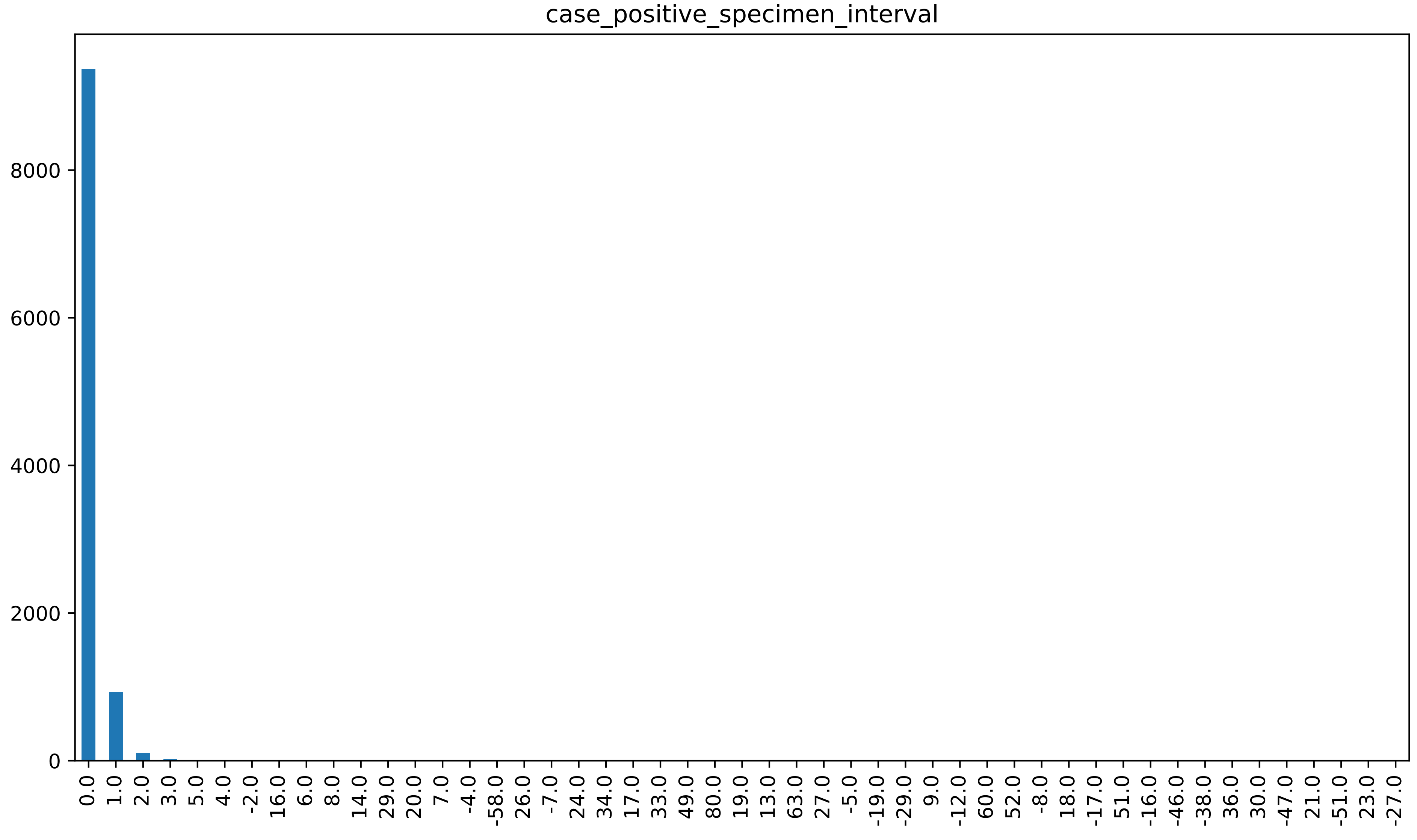

icu_yn
16000
14000
12000
10000
8000
6000
4000
2000
0
Missing
Unknown
No
Yes

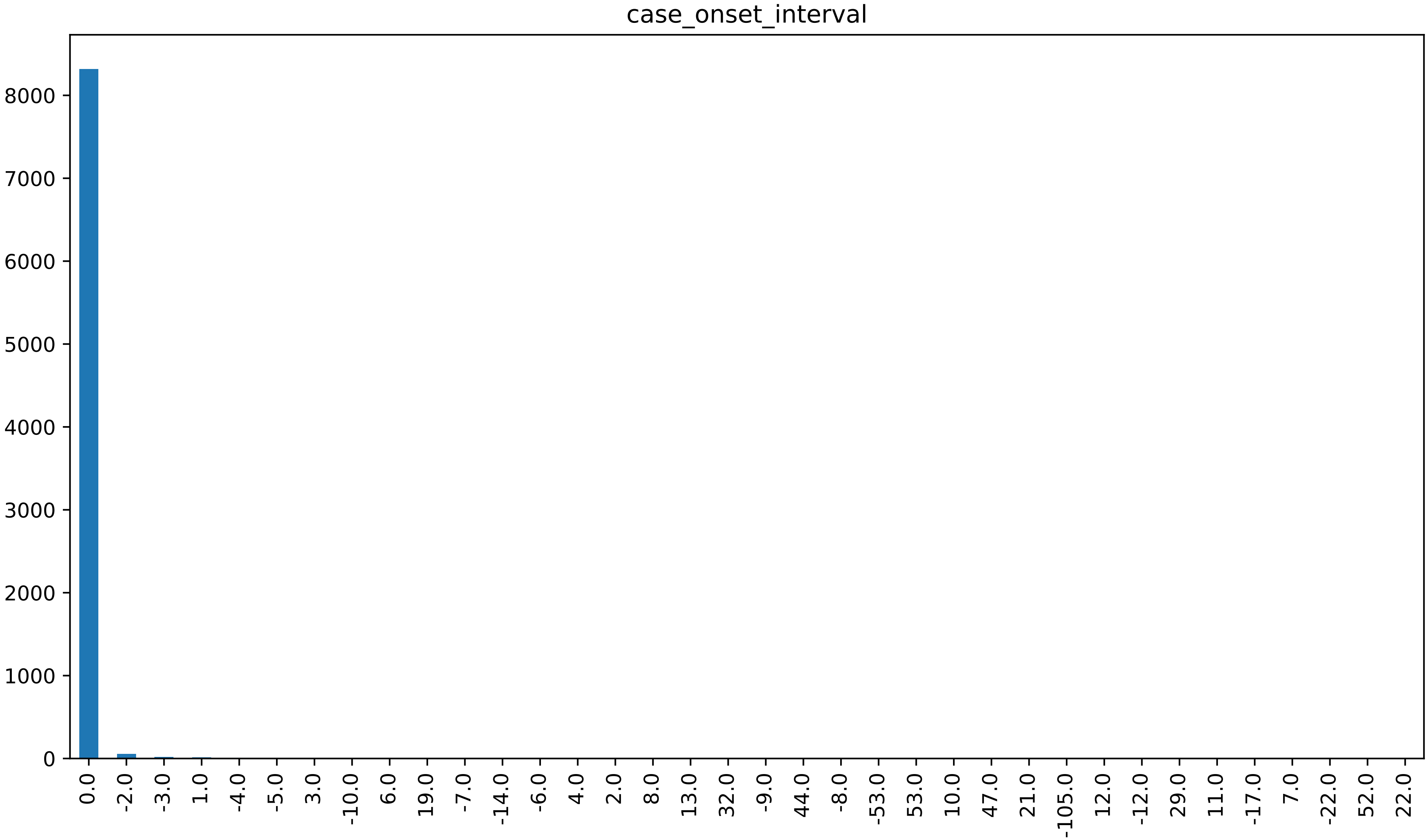

death_yn
0
2000
4000
6000
8000
10000
12000
14000
No
Yes

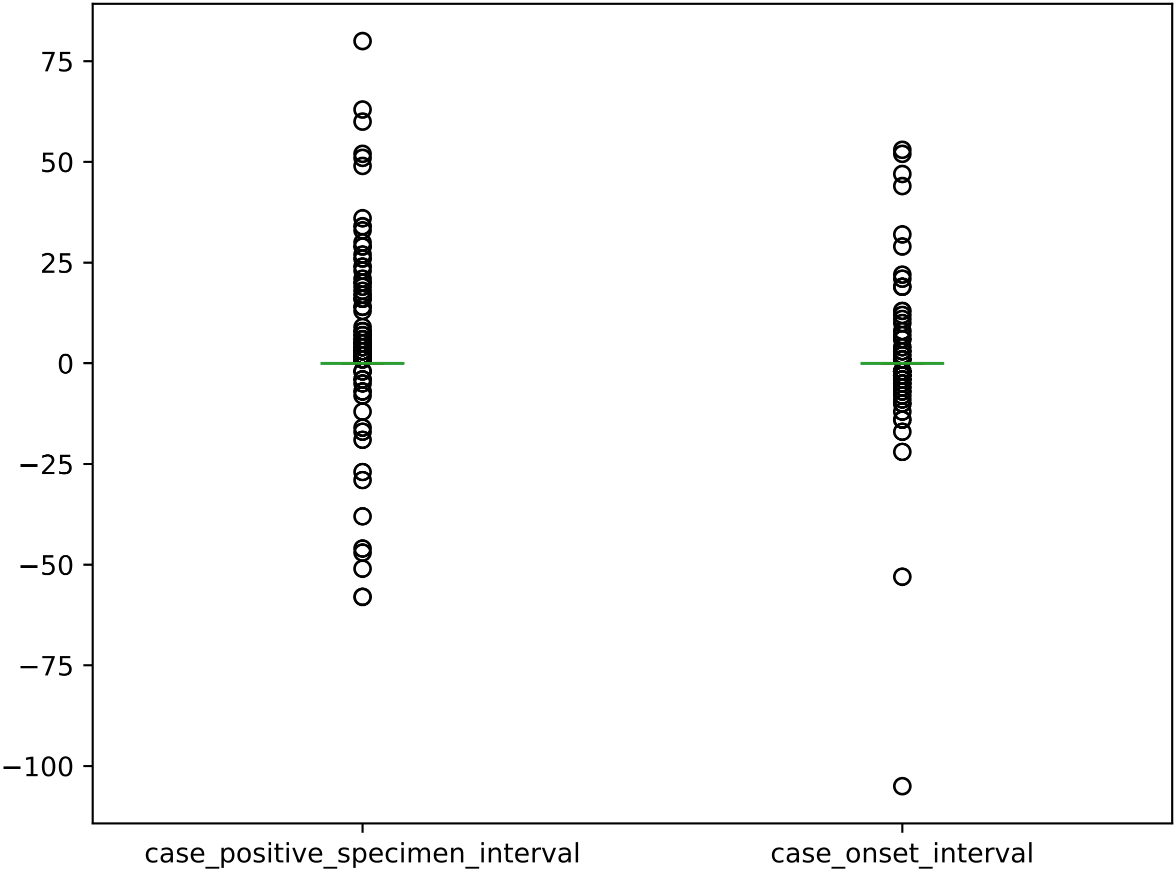

underlying_conditions_yn
1600
1400
1200
1000
800
600
400
200
0
Yes
No


case_positive_specimen_interval
case_positive_specimen_interval
-1
47.4%
46.9%
0
4.7%
1
2
0.9%


case_onset_interval
case_onset_interval
-1
57.7%
41.6%
0


case_positive_specimen_interval
8000
6000
4000
2000
0
0.0
1.0
2.0
3.0
5.0
4.0
-2.0
16.0
6.0
8.0
14.0
29.0
20.0
7.0
-4.0
-58.0
26.0
-7.0
24.0
34.0
17.0
33.0
49.0
80.0
19.0
13.0
63.0
27.0
-5.0
-19.0
-29.0
9.0
-12.0
60.0
52.0
-8.0
18.0
-17.0
51.0
-16.0
-46.0
-38.0
36.0
30.0
-47.0
21.0
-51.0
23.0
-27.0


case_onset_interval
0
1000
2000
3000
4000
5000
6000
7000
8000
0.0
-2.0
-3.0
1.0
-4.0
-5.0
3.0
-10.0
6.0
19.0
-7.0
-14.0
-6.0
4.0
2.0
8.0
13.0
32.0
-9.0
44.0
-8.0
-53.0
53.0
10.0
47.0
21.0
-105.0
12.0
-12.0
29.0
11.0
-17.0
7.0
-22.0
52.0
22.0


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