Data Quality Report – Initial Findings

1. Overview

This report will outline the initial findings based on the cleaned dataset (CovidCleaned_2.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 to be missing quite a large amount of data. There were no duplicate columns or columns with irregular cardinalities, however there was 12,142 null values in total. There were a large number of rows with duplicate data that have been removed (count 821). After digging into this data, it appears impossible to tell from the dataset if these were legitimate duplicates or cases with similar credentials. It was therefore concluded that the safest way to deal with these duplicates was to remove them.

2. Summary

Several tests were carried out to check the logical integrity of the data. This test brought a greater understanding of the data. In total 7,914 instances of irrational data was observed. An example of one of the cases is as follows: It was discovered that in a number of cases, people did not have hospital admissions but had icu admissions This is clearly impossible, and may have been the result of an overworked medical care expert not completing a hospital admission form. This is to be checked with the domain expert. See logical integrity section below for further details.

For the continuous features, there were cases where the earliest clinical dates preceded the date of first positive specimen collection (76 failures – See logical integrity test 4). This is impossible and possible due to medical professional error once again. Similarly, there were instances where the date of positive specimen collection is prior to symptom onset date(77 failures – See logical integrity test 5). This is also impossible and these errors will need to be addressed.

For the categorical values, occurrences of “Unknown” and “Missing” are constant throughout. These will need to be combined into one and then investigated to see if it is worth imputing or dropping the data altogether. This will be done on a case-by-case basis, determining the percentage of each feature that is missing.

3. Review Logical Integrity

Tests were carried out. The failures are listed below:

- Test 1 - Check if there were icu admissions but not hospital admissions. (Impossible)
    - 1 case found
- Test 2 - Check if date of first positive specimen collection is not blank then it should show a Laboratory-confirmed case in the current status. (Impossible
    - 220 cases found
- Test 3 - Check if current status is Laboratory-confirmed case but first positive specimen collection is blank (Impossible)
    - 5910 cases found
- Test 4 - Check if date of positive specimen collection is not before the earliest available date for the record. (Impossible)
    - 76 cases found
- Test 5 - Check if date of positive specimen collection is prior to symptom onset date. (Impossible)
    - 77 cases found
- Test 6 - Check if earliest date is populated but initial date is blank. Based on the CDC report the "CDC recommends researchers use cdc_case_earliest_dt in time series". The cdc_report_dt should not be blank provided cdc_case_earliest_dt is populated.
    - 1598 cases found
- Test 7 - Check if earliest date is actually the earliest available date.
    - 32 cases found

4. Review Continuous Features

4.1 Descriptive Statistics

There are 4 continuous features. All continuous features are datetime64.

- cdc_case_earliest_dt: Based on the descriptive statistics, the ranges of dates seem plausible. The most frequent date of 5th January 2021seems likely also, as cases rose around that date period.
- cdc_report_dt: The ranges of dates once again seem plausible with a the most frequent being on 10th June 2020. This date being an outlier, this will need to be investigated further.
- pos_spec_dt: These ranges do make sense and there is no real outliers here. This will need to be investigated further.
- onset_dt: There is no real outliers here. These dates do seem plausible.

4.2 Histograms

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

4.3 Box plots

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

5. Review Categorical Features

5.1. Descriptive Statistics

There are 8 categorical features in the dataset and are as follows:

- current_status
- sex
- age_group
- race_ethnicity_combined
- hosp_yn
- icu_yn
- death_yn
- medcond_yn

There is a large number of results taken up by both unknown and missing throughout. This is shown as both “icu_yn” and “medcond_yn” have “Missing” as their most common column. This is an overwhelming majority and these columns will need to be looked at in depth.

There appears to be many unique values for what should be a relatively straight forward feature. For example, “sex” is showing 4 unique values. This will need to be investigated and perhaps combine the “Missing” and “Unknown” values into one.

A large number of null values may also result in a dropped column if it is greater than 60%. As a general rule we cannot impute the data once it is beyond 30%. This includes the columns “icu_yn” and “medcond_yn”. We will have to look into these more thoroughly.

5.2 Histograms

The histograms can be found in the accompanying pdf.

6. Action to take

Several main actions will be taken, summarised below:

- Joining ‘Unknown’ and ‘Missing
  - Combining the column names that are unknown and missing into one grouped category.
- Large Number of Duplicate Data
  - It was decided that the safest way to go about this from a machine learning perspective was to remove these duplicates.
- hosp_yn as "no" and icu_yn as "yes"
  - We will drop the row as it only accounts for 1 row (as per logical integrity test 1) and displays little information
- Large Number of Null values
  - 12,142 total null values were found in this dataset. We will be examining each column in a case by case basis to decide how to approach these null values.
- Removal of icu_yn column
  - There are far too many missing or unknowns in these columns. Completely removing these columns might be the best alternative.
- Removal of mecond_yn column
  - Similarly, there are an overwhelming number of missing and unknown columns in this category to make it worthwhile to keep.
- current_status cannot be "Probable Case" if it has a pos_spec_dt
  - This accounts for 220 results as per the logical integrity test 2, we will impute these results.
- pos_spec_dt should not be blank provided it is a laboratory confirmed case
  - It looks as though this column will need to be dropped as these null values account for 5,910 as per logical integrity test 3.
- pos_spec_dt before the earliest available date
  - There are 76 results failing this test as per logical integrity test 4. We will impute these results.
- pos_spec_dt should not be before the onset_dt
  - We will impute this onset_dt with the pos_spec_dt as there is only 77 (as per logical integrity test 5) results failing ths test as per logical integrity test 5.
- If cdc_case_earliest_dt is populated, the cdc_report_dt should not be blank
  - We will need to investigate this further to see if it is possible to impute this field.
- Earliest available date should be the cdc_case_earliest_dt column.
  - here are 32 results that are failing this as per logical integrity test 7. We will impute these results.
- cdc_report_dt is a depreciated column as per the CDC report
  - As this is a depreciated column, we will need to drop this column.

7. References

1. Fundamentals of Machine Learning for Predictive Data Analytics 2015 – John D.Kelleher.
2. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

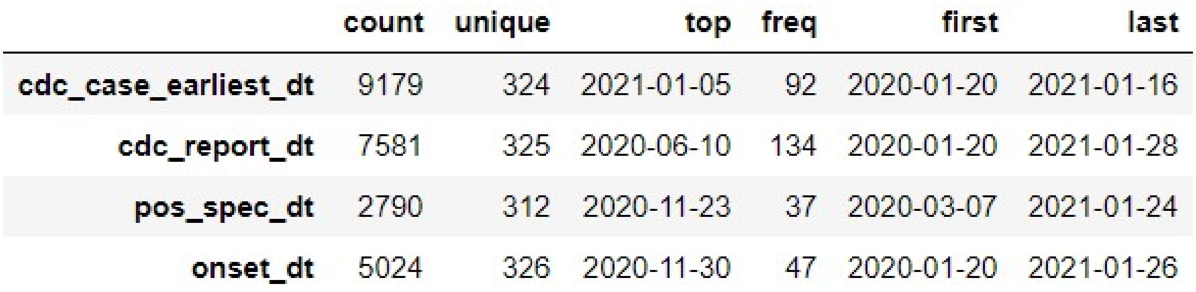
8. Appendix

8.1 Terminology & Assumptions:

From the CDC Report:

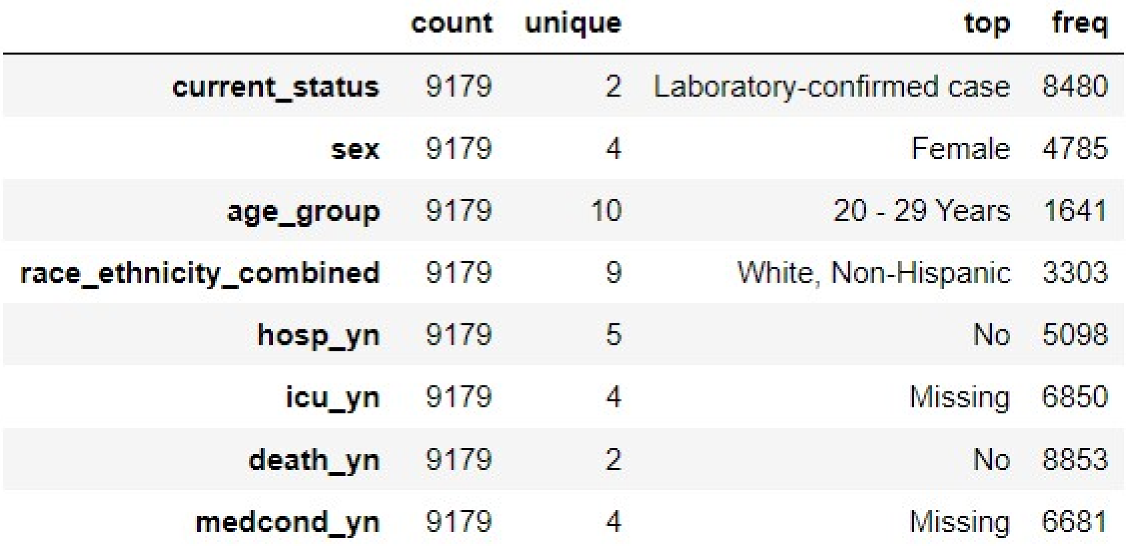
- cdc_case_earliest_dt - Calculated date--the earliest available date for the record, taken from either the available set of clinical dates (date related to the illness or specimen collection) or the calculated date representing initial date case was received by CDC. This variable is optimized for completeness and may change for a given record from time to time as new information is submitted about a case
- cdc_report_dt - Calculated date representing initial date case was reported to CDC. Depreciated; CDC recommends researchers use cdc_case_earliest_dt in time series and other time-based analyses.
- pos_spec_dt - Date of first positive specimen collection
- onset_dt - Symptom onset date, if symptomatic
- current_status - Case Status: Laboratory-confirmed case; Probable case
- sex - Sex: Male; Female; Unknown; Other
- age_group - Age Group: 0 - 9 Years; 10 - 19 Years; 20 - 39 Years; 40 - 49 Years; 50 - 59 Years; 60 - 69 Years; 70 - 79 Years; 80 + Years
- race_ethnicity_combined - Race and ethnicity (combined): Hispanic/Latino; American Indian / Alaska Native, Non-Hispanic; Asian, Non-Hispanic; Black, Non-Hispanic; Native Hawaiian / Other Pacific Islander, Non-Hispanic; White, Non-Hispanic; Multiple/Other, Non-Hispanic
- hosp_yn - Hospitalization status
- icu_yn - ICU admission status
- death_yn - Death status
- medcond_yn - Presence of underlying comorbidity or disease

8.3 Continuous Features
Descriptive Statistics



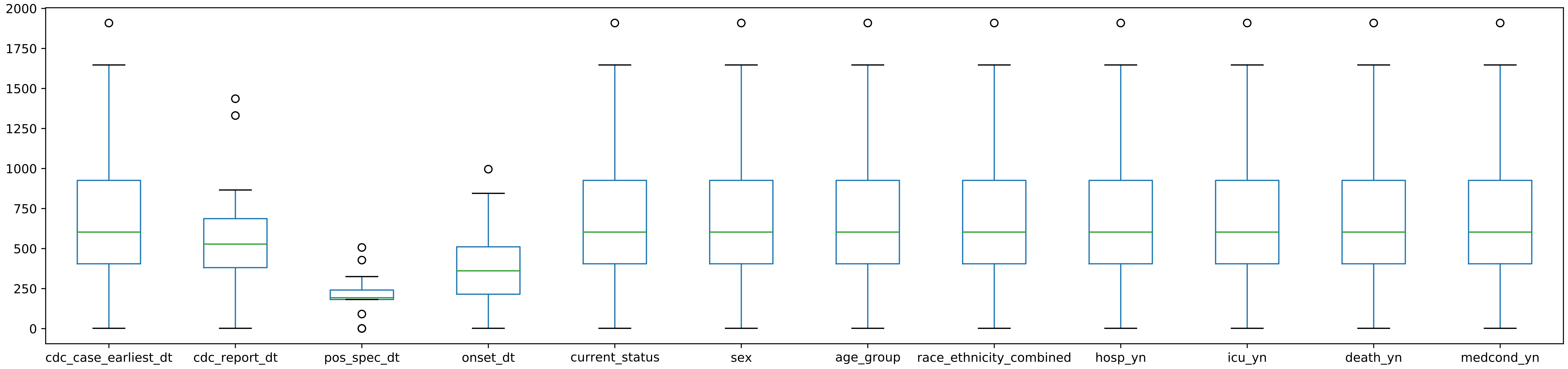
	count	unique	top	freq	first	last
cdc_case_earliest_dt	9179	324	2021-01-05	92	2020-01-20	2021-01-16
cdc_report_dt	7581	325	2020-06-10	134	2020-01-20	2021-01-28
pos_spec_dt	2790	312	2020-11-23	37	2020-03-07	2021-01-24
onset_dt	5024	326	2020-11-30	47	2020-01-20	2021-01-26

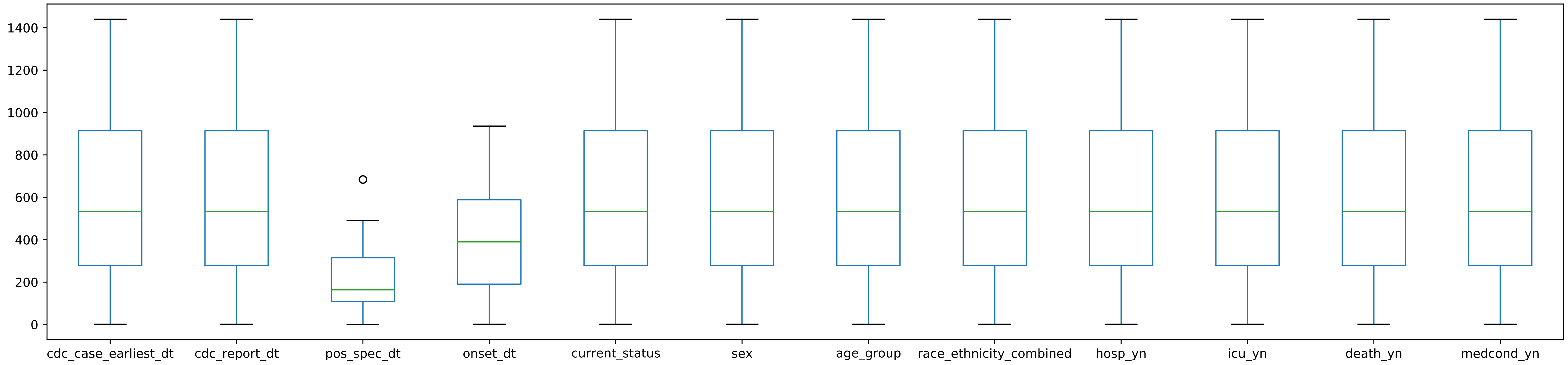
8.4 Categorical Features
Descriptive Statistics

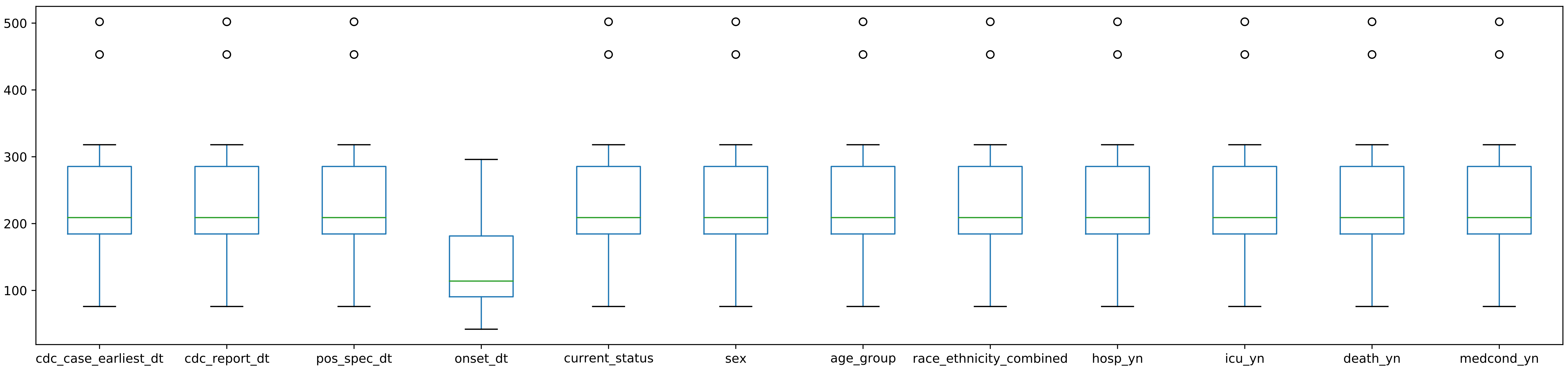


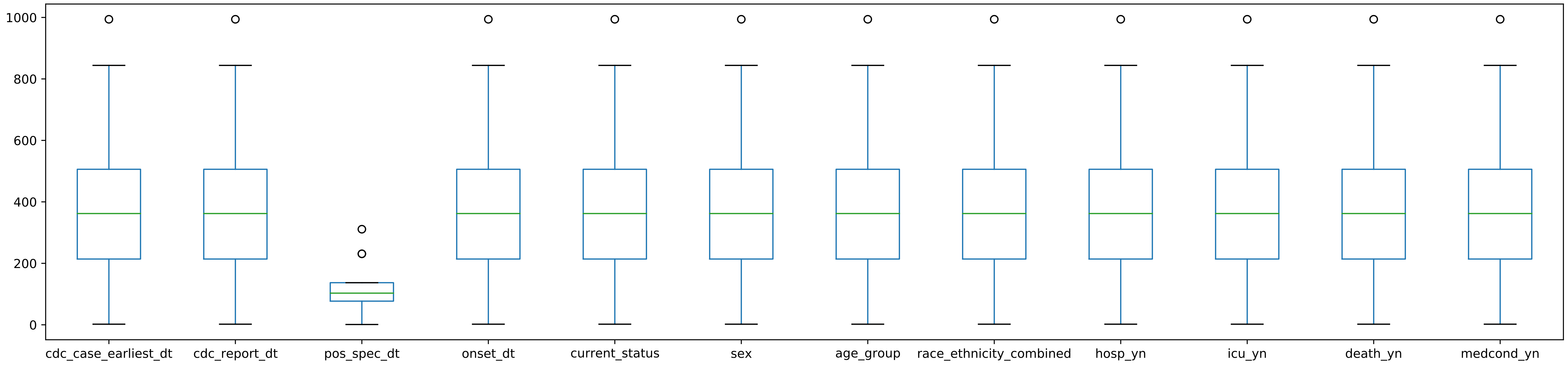
	count	unique	top	freq
current_status	9179	2	Laboratory-confirmed case	8480
sex	9179	4	Female	4785
age_group	9179	10	20 - 29 Years	1641
race_ethnicity_combined	9179	9	White, Non-Hispanic	3303
hosp_yn	9179	5	No	5098
icu_yn	9179	4	Missing	6850
death_yn	9179	2	No	8853
medcond_yn	9179	4	Missing	6681

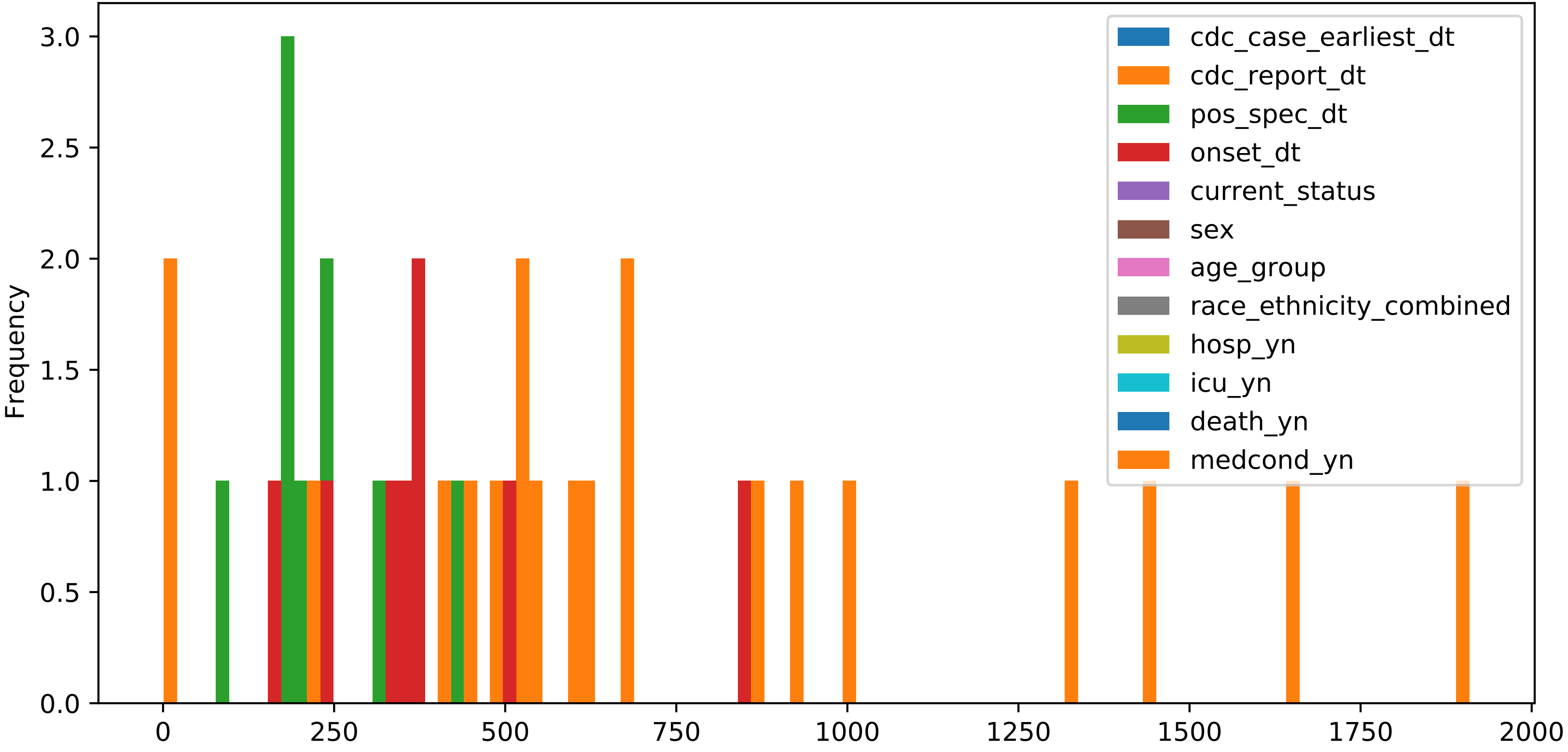
8.5 Box Plots & Histograms
Please see Box Plots & Histograms on the following pages.

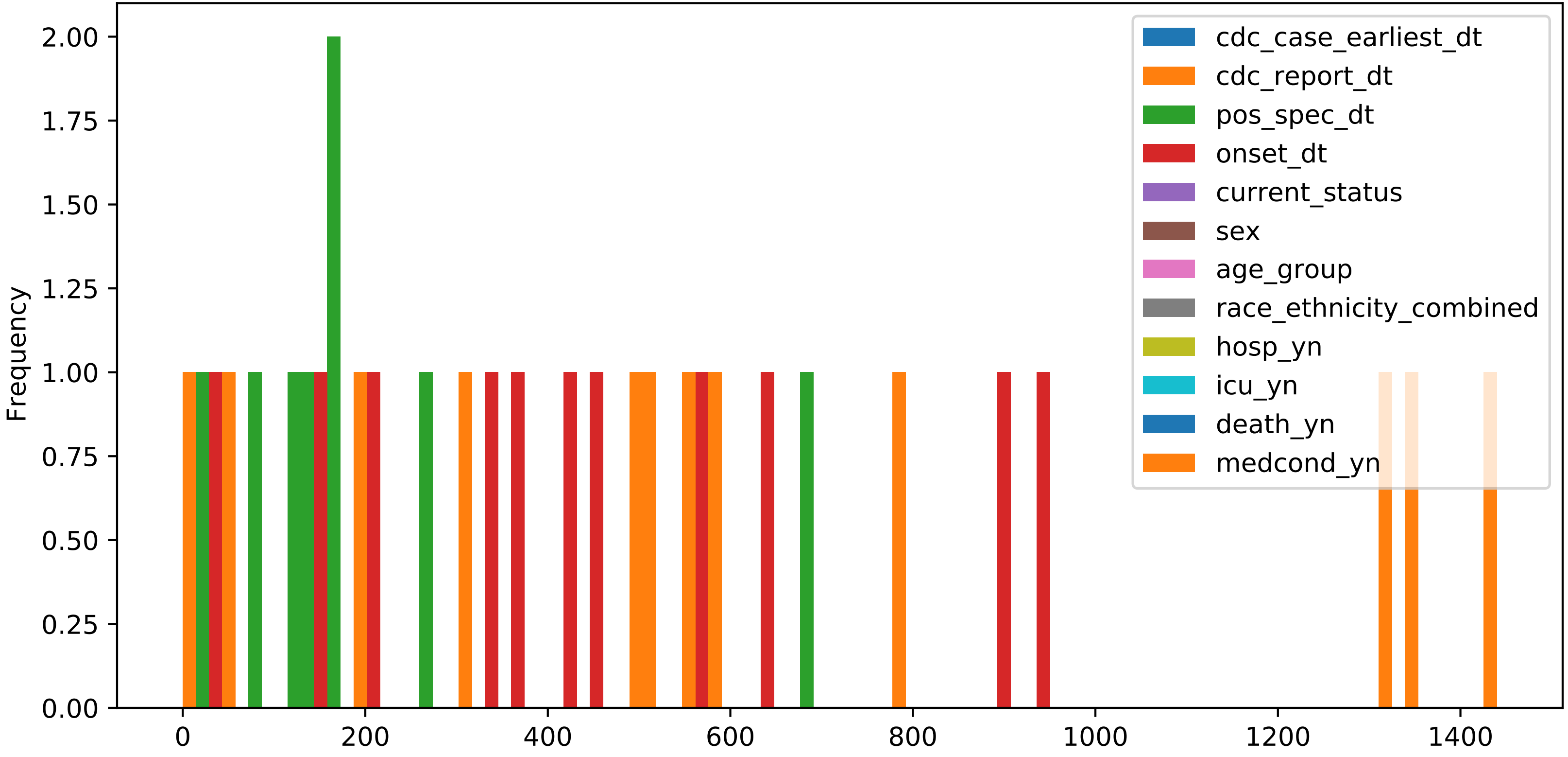

2000
1750
1500
1250
1000
750
500
250
0
cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn

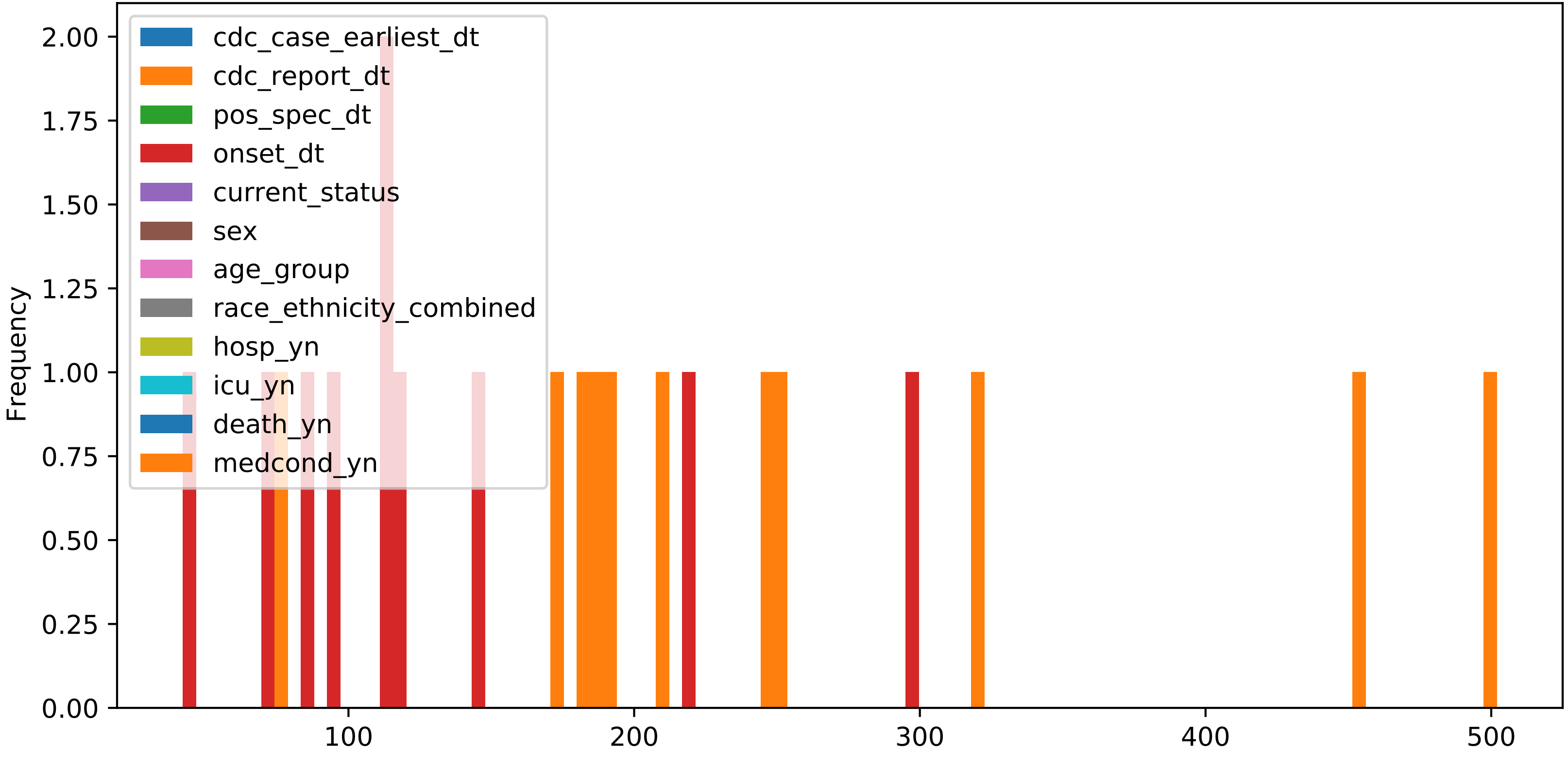

1400
1200
1000
800
600
400
200
0
cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn

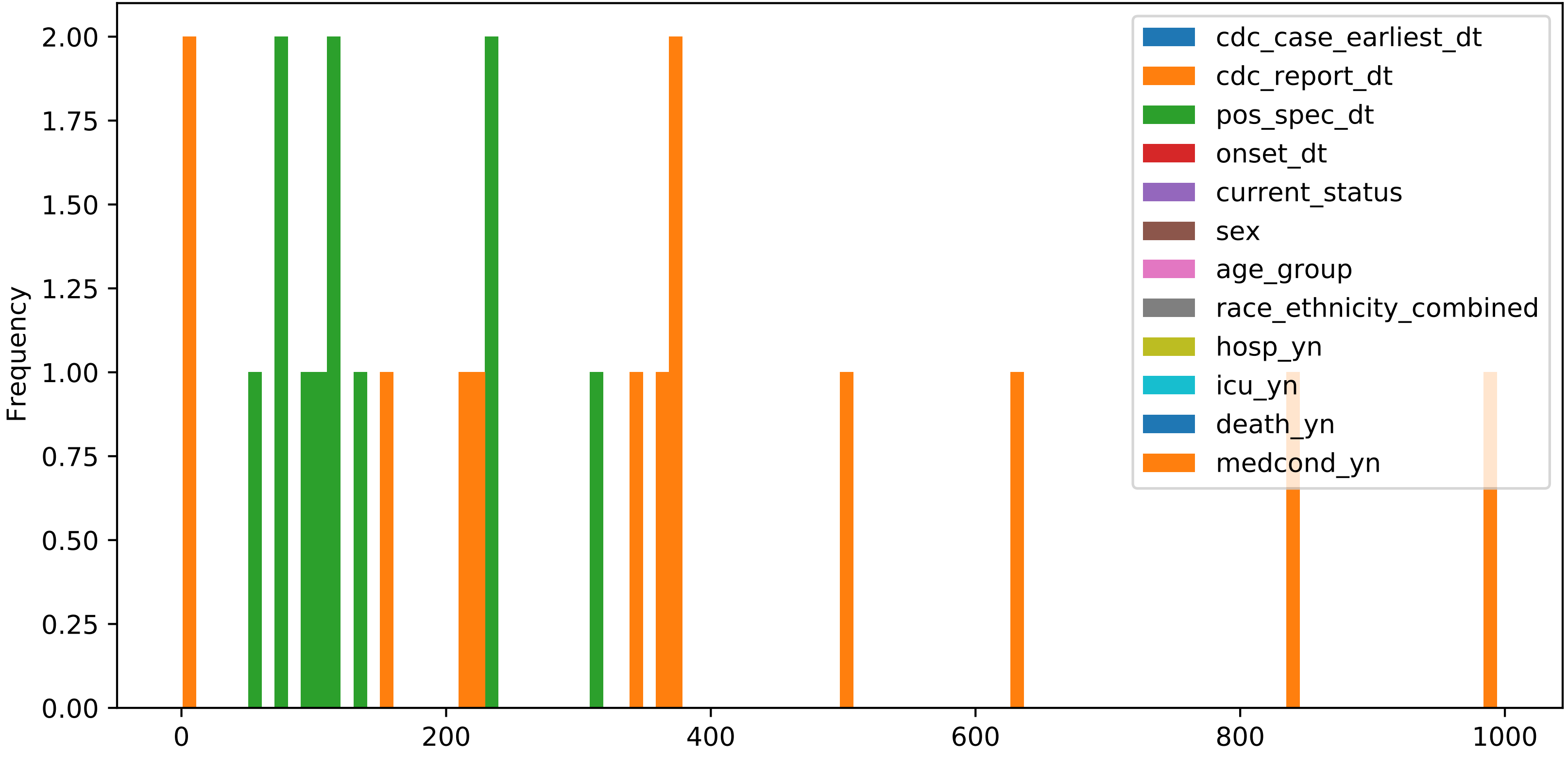

500
400
300
200
100
cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn

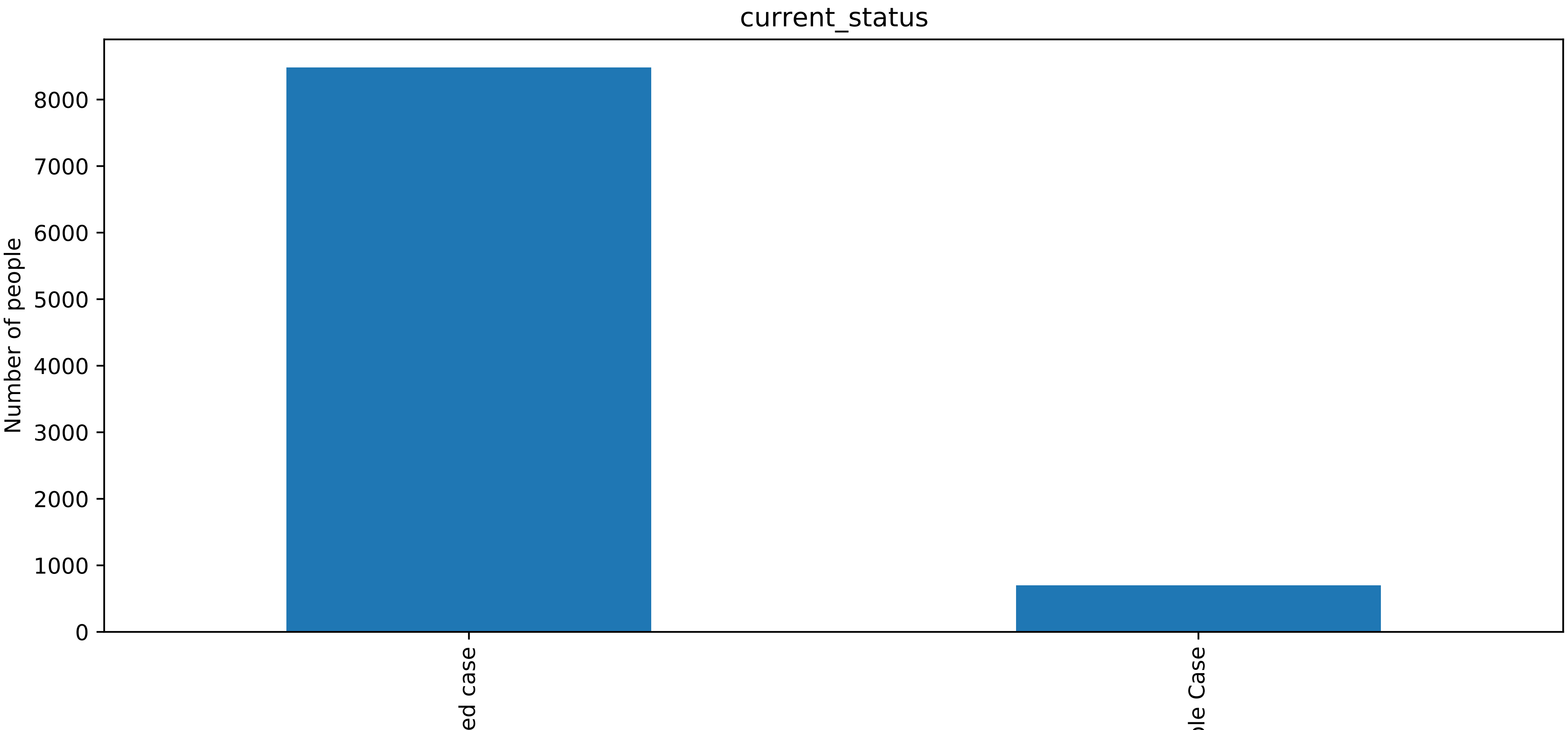

1000
800
600
400
200
0
cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn

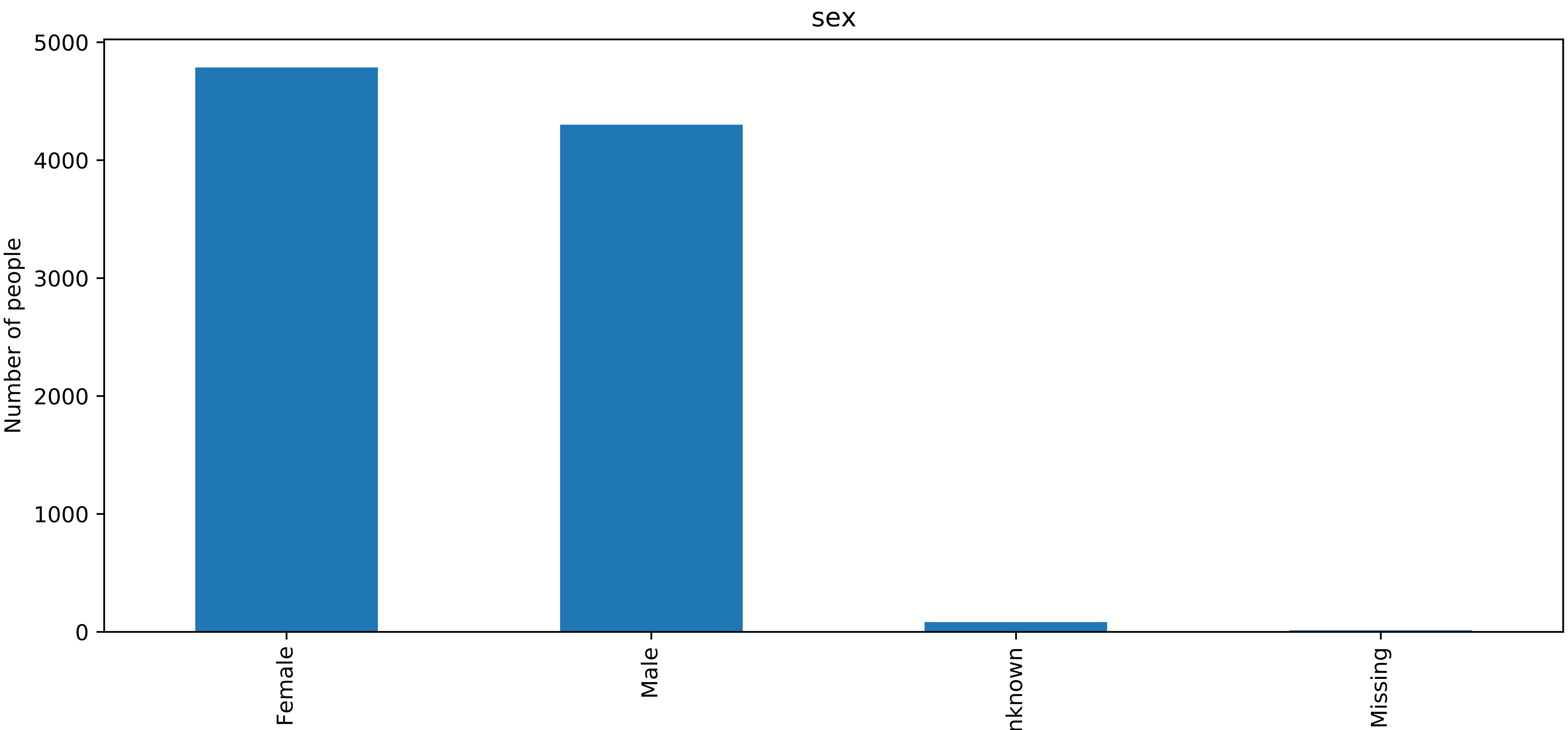

Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
250
500
750
1000
1250
1500
1750
2000
cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn

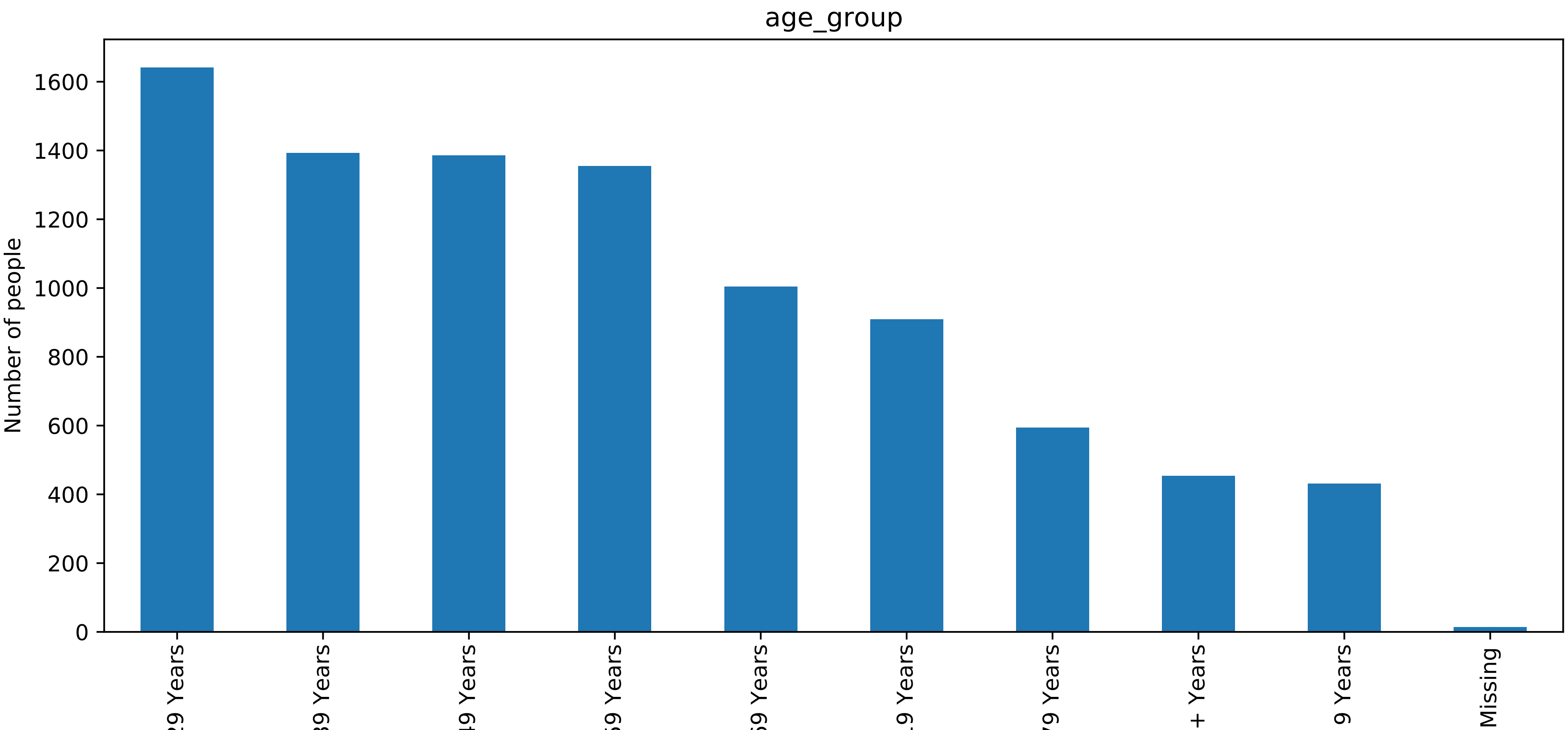

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Frequency
0
200
400
600
800
1000
1200
1400
cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn

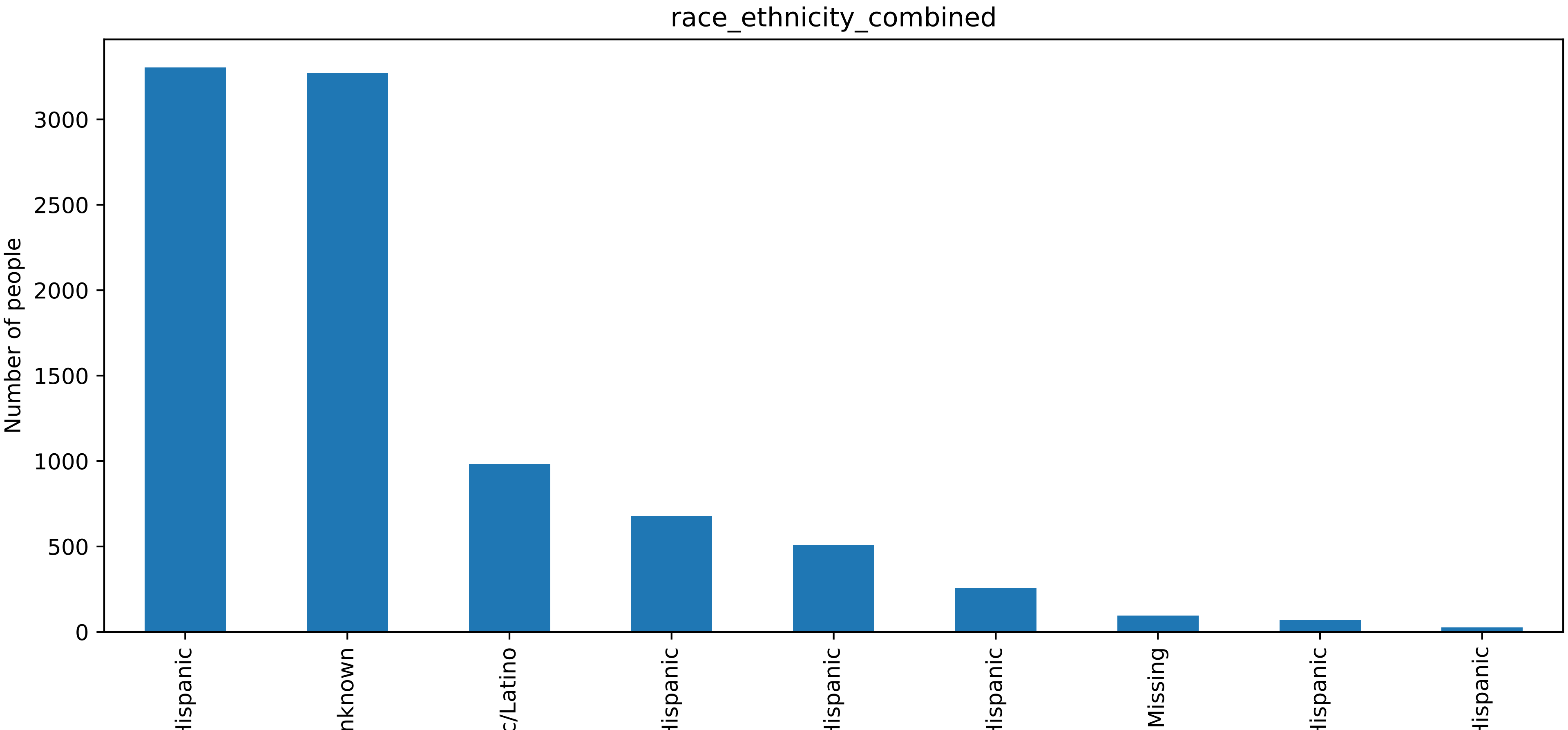

cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn
Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
100
200
300
400
500

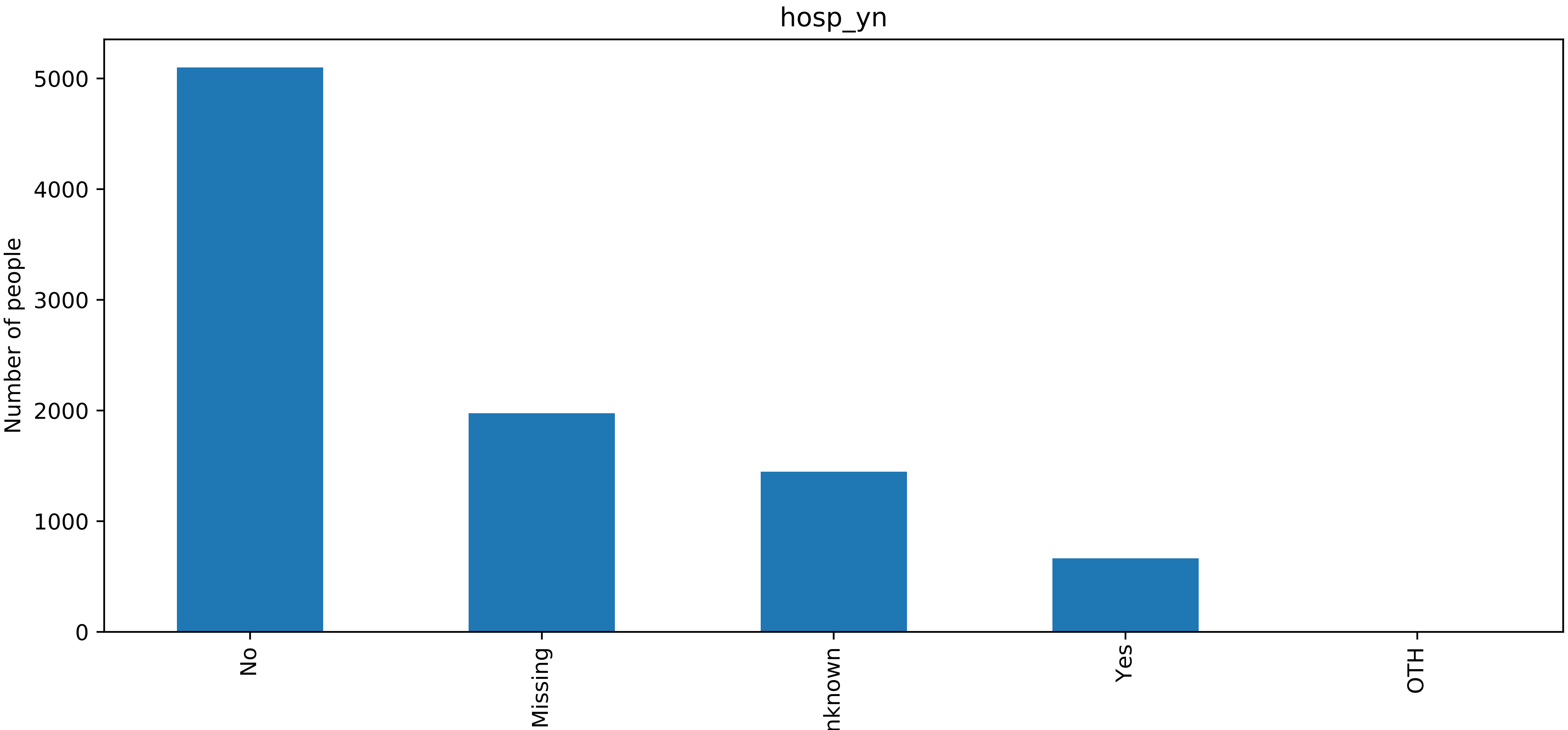

Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
200
400
600
800
1000
cdc_case_earliest_dt
cdc_report_dt
pos_spec_dt
onset_dt
current_status
sex
age_group
race_ethnicity_combined
hosp_yn
icu_yn
death_yn
medcond_yn

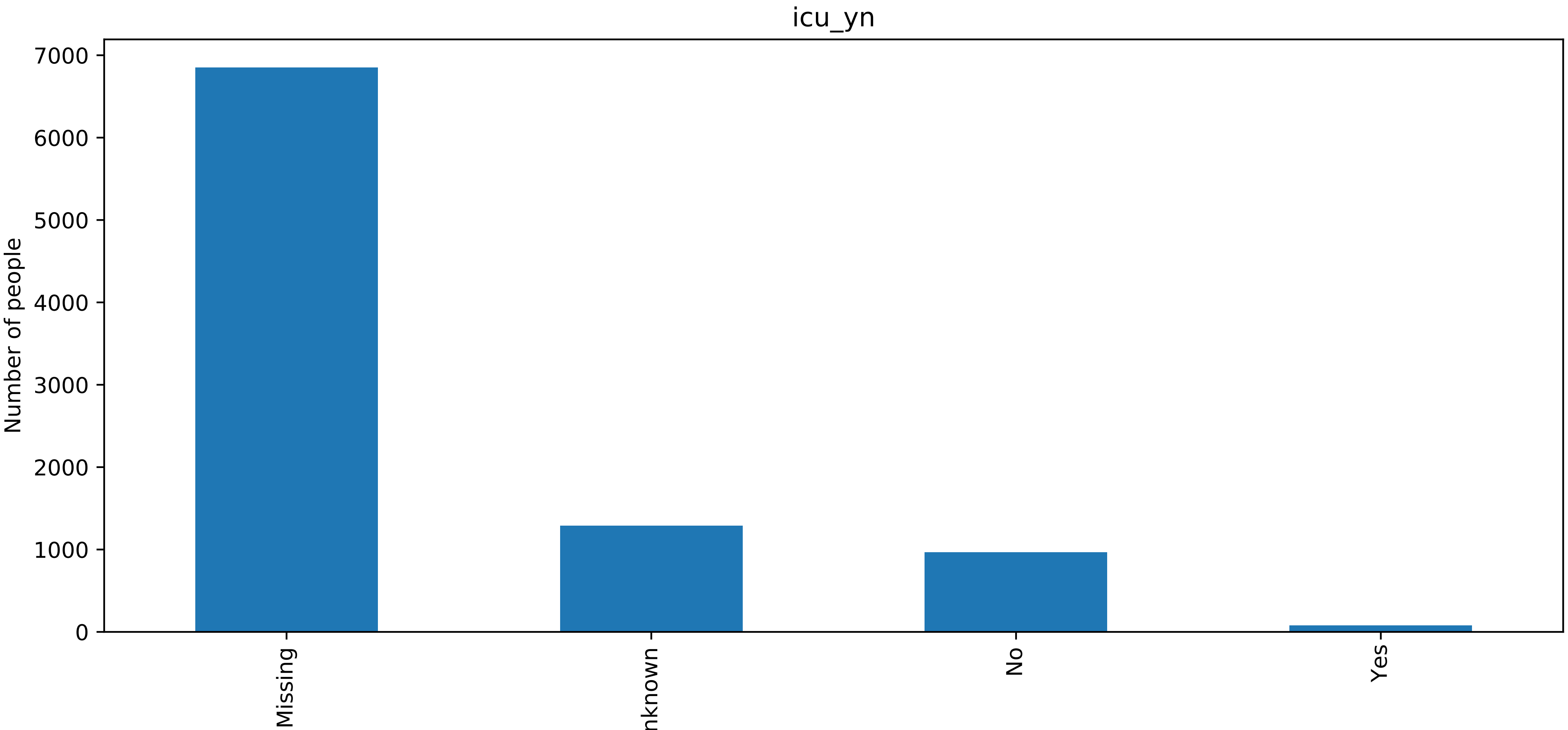

current_status
Number of people
0
1000
2000
3000
4000
5000
6000
7000
8000
ed case
ple Case

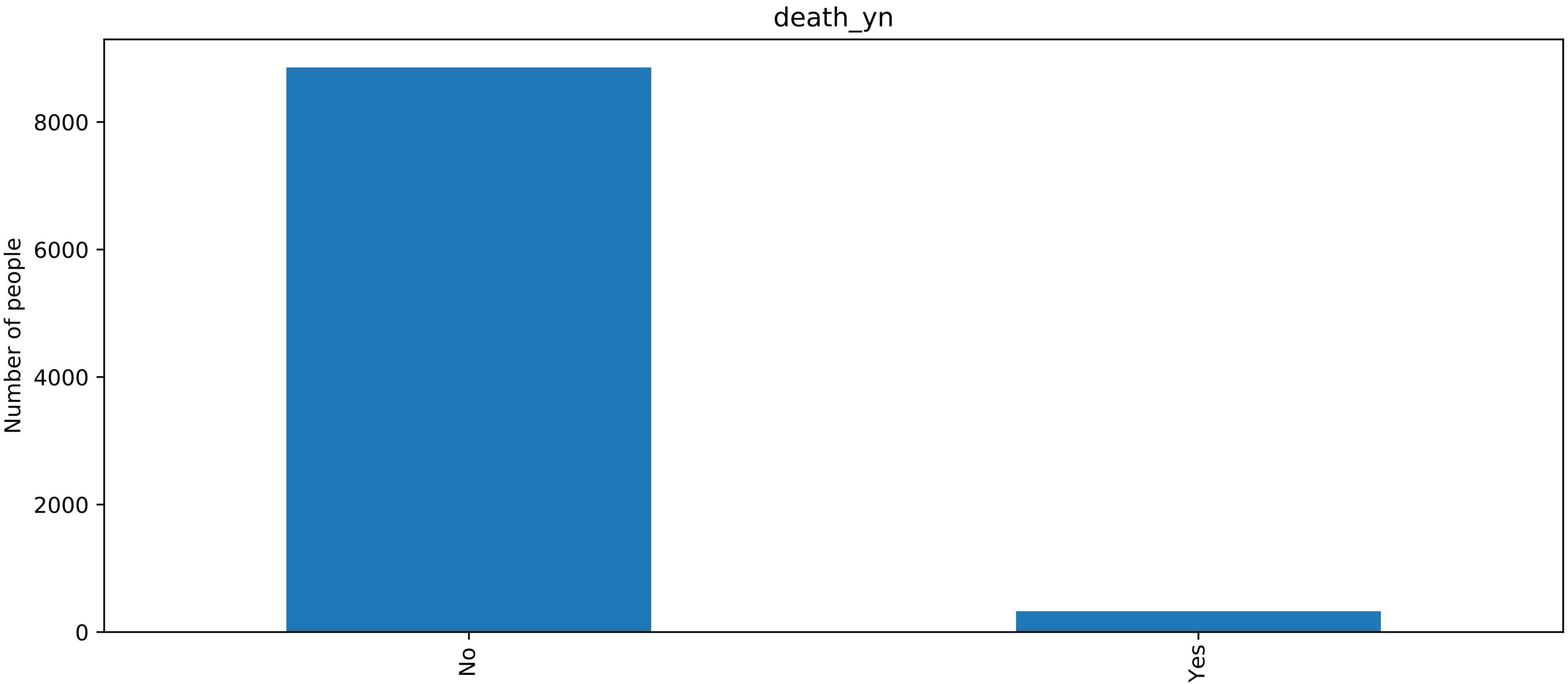

sex
Number of people
5000
4000
3000
2000
1000
0
Female
Male
nknown
Missing

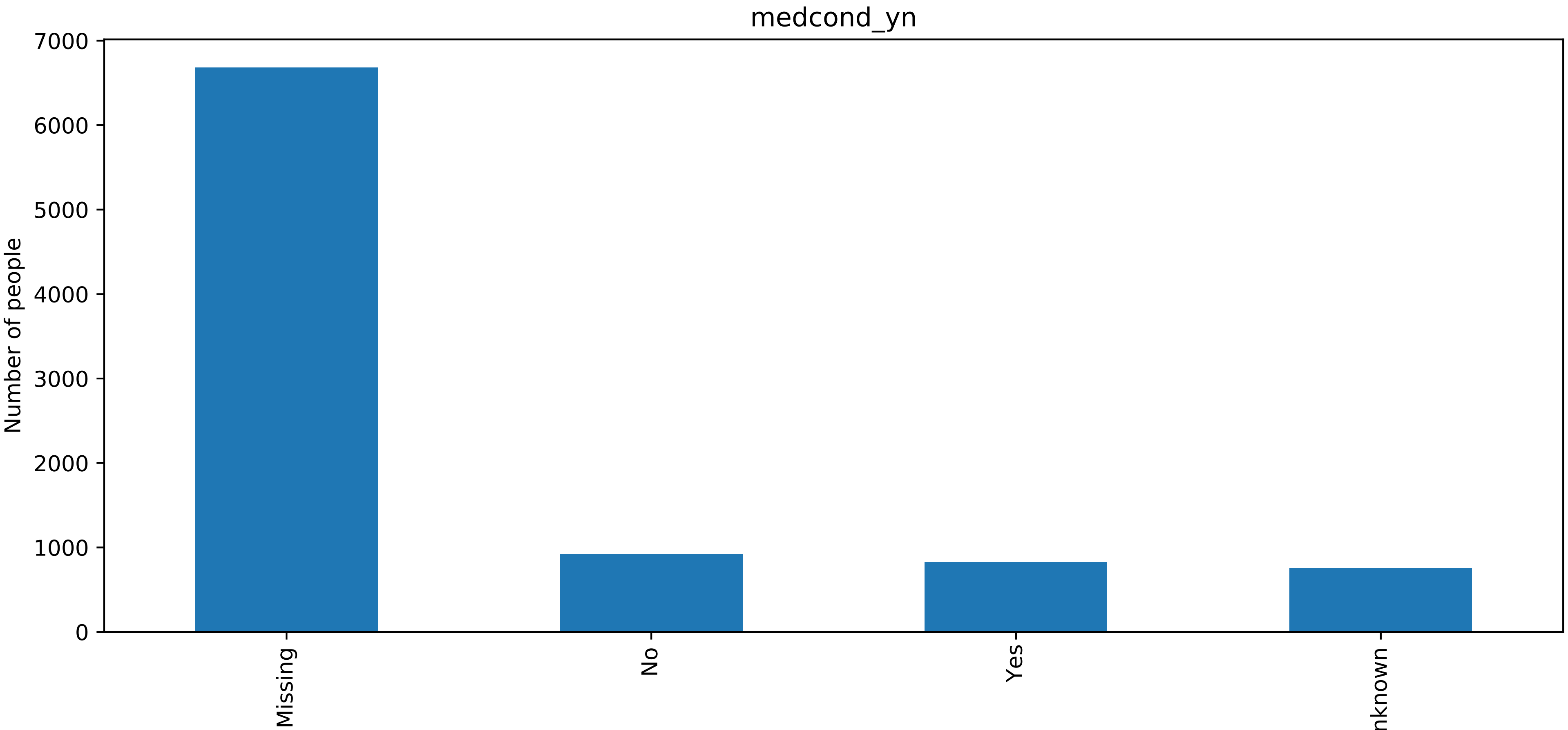

age_group
Number of people
0
200
400
600
800
1000
1200
1400
1600
29 Years
39 Years
49 Years
59 Years
69 Years
9 Years
79 Years
+ Years
9 Years
Missing


race_ethnicity_combined
Number of people
0
500
1000
1500
2000
2500
3000
Hispanic
nknown
c/Latino
Hispanic
Hispanic
Hispanic
Missing
Hispanic
Hispanic


hosp_yn
Number of people
0
1000
2000
3000
4000
5000
No
Missing
nknown
Yes
OTH


icu_yn
Number of people
7000
6000
5000
4000
3000
2000
1000
0
Missing
nknown
No
Yes


death_yn
Number of people
0
2000
4000
6000
8000
No
Yes


medcond_yn
Number of people
0
1000
2000
3000
4000
5000
6000
7000
Missing
No
Yes
nknown