Data Quality Report - Initial Findings

1. Overview

This report will outline the initial findings based on the provided sample of the Covid-19 dataset (covid19-cdc-23221675.csv) It will summarize the data, present high-level observations on any obvious trends, and address any apparent data quality concerns. The appendix contains all of the descriptive plots of the data.

The first pass through the data revealed a large amount of missing, suppressed (for privacy reasons) or NaN values. There were also a significant number of outlier values in the continuous features that were otherwise a near constant feature (a value of **0**).

Some logical integrity checks were also performed to identify contradictory or potentially misleading observations.

2. Summary

This report provides an overview of the Covid-19 dataset, highlighting significant issues such as missing, suppressed, and outlier values. Logical integrity tests reveal inconsistencies in data with respect primarily to symptom status and hospitalization. Descriptive statistics expose extensive missing values across categorical and continuous features. Recommendations include removing rows failing logical tests, potentially imputing values for specific cases, and addressing outliers.

Pairwise interactions were explored with a number of features and the proportion of ‘death_yn’ as an outcome. Theses were the most pertinent interactions to explore as they directly address the problem statement:

> “**build a data analytics solution for death risk prediction”**

Additionally, a number of new features were imputed: year and month were split into independent features, and new values for ‘seasonality’, ‘political_affiliation’ and (is_)‘rural’. These features were chosen based on research into the problem domain and the availability of ‘state’ and ‘county_fips_code’ values allowing for straightforward imputing of values.

These new features showed clear trends that would aid with **death risk prediction**.

3. Review Logical Integrity

- Test 1 - Symptom status and hosp_yn/icu_yn
    - Number of rows where symptom status was 'asymptomatic' but 'hosp_yn' or 'icu_yn' were ’Yes’:

    - Found observations: **47**
- Test 2 - Did an asymptomatic observation die?
    - If so, is it reasonable to attribute this to Covid-19 exposure?
    - Number of rows where ‘symptom status’ was 'Asymptomatic' and death_yn was 'Yes':
    - Found observations: **96**
- Test 3 - icu_yn was “Yes” but hosp_yn was “No
    - Number of observations where 'icu_yn' was 'Yes' but 'hosp_yn' was 'No':
    - Found observations: **4**
- Test 4 - valid dates
    - All years and months were valid:
    - No years before 2020 (first recorded case in the United States) or in the future, and no months less than 1 or greater than 12.
- Test 5 - duplicate columns
    - According to the data dictionary provided, ‘state_fips_code’ is wholly derivative of ‘res_state’ and therefore redundant.

4. Review Continuous Features

4.1. Descriptive Statistics

The summaries reveal a huge number of missing values for both continuous features: ‘case_positive_specimen_interval’ and ‘case_onset_interval’. Of the values that are present, the vast majority of them, ~89% and ~96% respectively have the same value: **0**.

4.2. Histograms

Considering the overwhelming clustering around **0**, the histograms for these features tell us next to nothing about the distribution other than the fact they are nearly constant columns.

4.3. Box plots

The boxplots aren’t much more descriptive and further highlight the fact these features are almost constant: every value that isn’t **0** is considered an outlier.

5. Review Categorical Features

5.1. Descriptive Statistics

Several features have missing values: ‘res_county’, ‘county_fips_code’, ‘age_group’, ‘sex’, ‘race’, ‘ethnicity’ and ‘underlying_condition_yn’. There is also a significant number of observations for these features that are either ‘Missing’ or ‘Unknown’.

Beyond this, the descriptive statistics revealed other features contain a significant number of ‘Missing’ and ‘Unknown’ values: ‘process’, ‘exposure_yn’, ‘sympyom_status’, ‘hosp_yn’ and ‘icu_yn’.

5.2. Histograms

The histograms reveal the extent of these missing values (including ‘Missing’ and ‘Unknown’ values). A small number of values for ‘sex’ are coded additionally as ‘Unknown.

Additionally, ‘race’ and ‘ethnicity’ also contain a significant number of ‘Unknown’ and ‘Missing’ values.

The vast majority of values for ‘process’, ‘icu_yn’ and ‘exposure_yn’ are coded as ‘Missing’ or ‘Unknown’, as are a significant number for ‘symptom_status’ and ‘hosp_yn’.

6. Action to take

1. Remove rows that fail the logical integrity tests 1 and 2.
2. *Potentially* Impute values for rows that fail logical integrity test 3
    - (hosp_yn == “yes” but icu_yn == “no”)
    - However, feature is missing too many values and will most likely be dropped
3. Clamp outliers for continuous variables.

6.1 Data Quality Plan

Comprehensive identification of data quality issues and appropriate actions:

Data Quality Plan

Feature	Data Quality Issue	Potential Handling Strategy (Choice)
res_state	No issues	
state_fips_code	Redundant	Drop feature
res_county	Missing features	Despite uncertainty in imputing as 'Missing', some of value is likely preserved in its derivative 'res_state' feature and therefore would prefer not to drop the rows Drop rows
age_group	Missing features	Impossible to impute: unreasonable to suggest a mean from the intervals Drop rows
sex	Missing features	Small number missing Drop rows
race	Missing features	No sensible basis to impute Drop rows
ethnicity	Missing features	To sensible basis to impute Drop rows
case_positive_specimen_interval	Outliers Missing features / invalid data	Outliers need to be handled Clamping Even in absence of outliers, too many values are missing Drop feature
case_onset_interval	Outliers Missing features / invalid data	Outliers need to be handled Clamping

		Even in absence of outliers, too many values are missing Drop feature
process	Missing features	Impute? No, too many missing values Drop feature
exposure_yn	Missing features	Impute? No, too many missing values Drop feature
current_status	No issues	Disproportionate (due care sampling)
Logical Test 1	Missing features	Drop rows
Logical Test 2	Missing features	Drop rows
Logical Test 3	Missing features	Impute inconsistency
Logical Test 4	No issues	
Logical Test 5	Missing features	Drop duplicate column
symptom_status	Missing features	Impute? No, too many missing values Drop feature
hosp_yn	Missing features	Impute? Possible, check *icu_yn* to impute but that feature missing too many values Drop feature
icu_yn	Missing features	Impute? No, too many missing values Drop feature
death_yn	No issues	Disproportionate (due care sampling)
underlying_conditions_yn	Missing features	Despite large number of missing values, I think it's reasonable to impute missing values as *No* considering underlying conditions if present *would* be recorded Impute

year	No issues	
month	No issues	

7. References

[1] - Kelleher, John D., Brian Mac Namee, and Aoife D'arcy. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press, 2020.

[2] - (Health Status Males v Females)
https://www.health.harvard.edu/newsletter_article/mars-vs-venus-the-gender-gap-in-health

[3] - (Health Status of Hispanic Americans)
https://www.pewresearch.org/short-reads/2023/10/30/5-facts-about-hispanic-americans-and-health-care/

[4] - (Delta Variant of Covid-19) https://elifesciences.org/articles/73584

[5] - (Seasonality of the State)
https://worldpopulationreview.com/state-rankings/all-four-seasons-states

[6] - Political Polarization and Covid-19
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8631569/#:~:text=For%20example%2C%20a%20national%20survey,of%20a%20vaccination%20conspiracy%20(YouGov%2C

[7] - (Presidential Election votes 2020)
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

[8] - (Rural Counties)
https://www.consumerfinance.gov/compliance/compliance-resources/mortgage-resources/rural-and-underserved-counties-list/

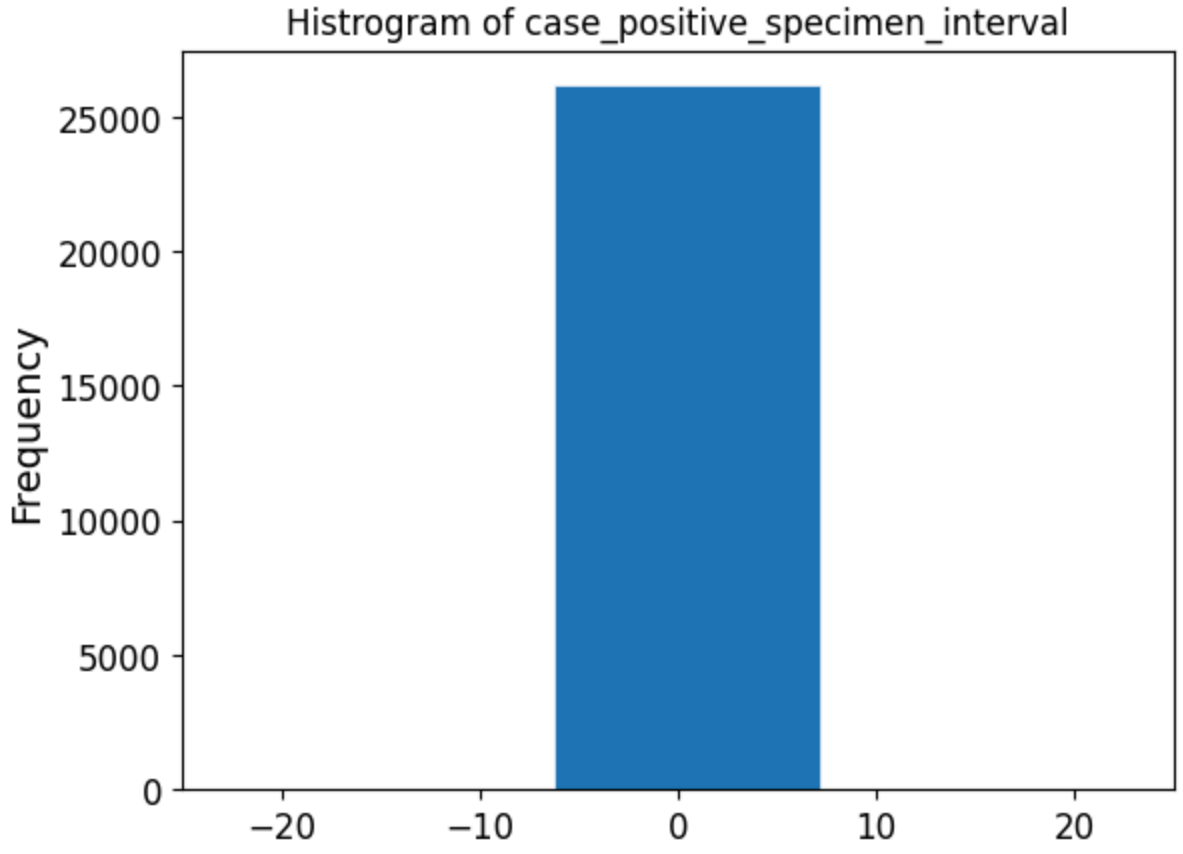
8. Appendix

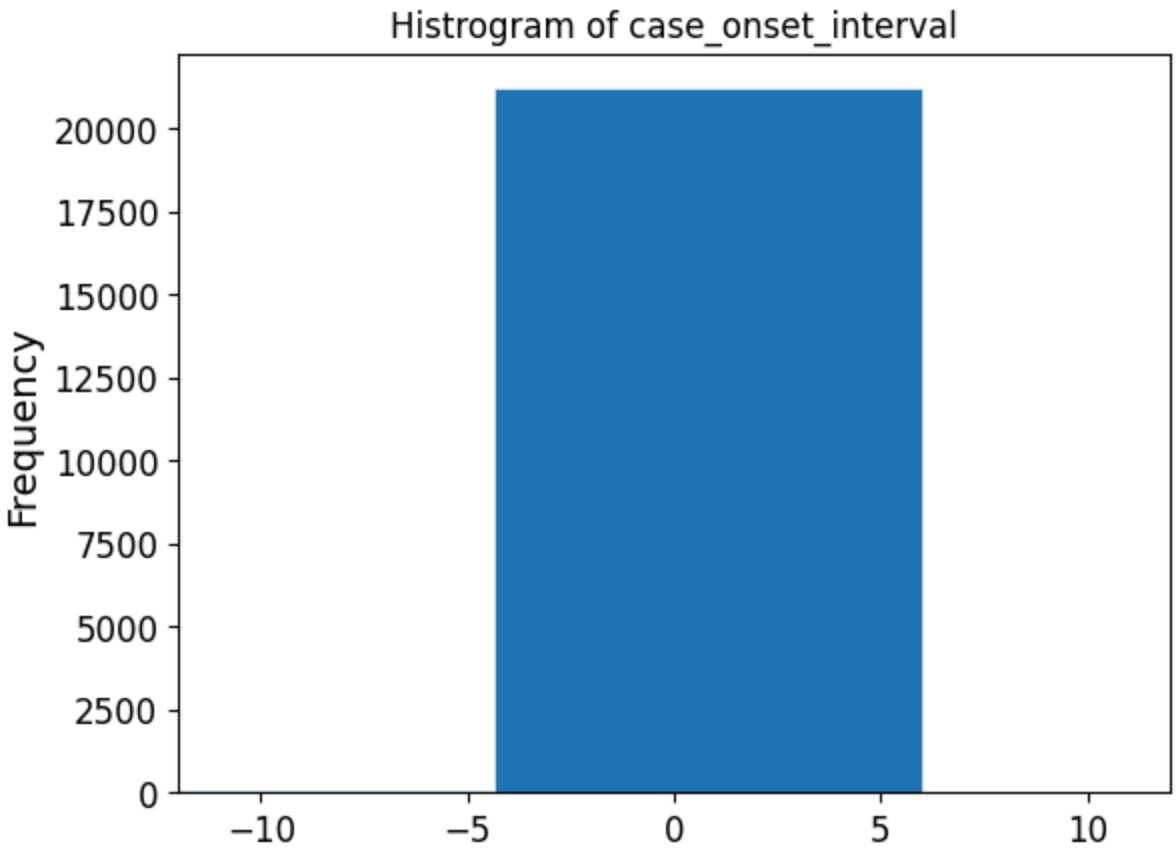
	count	mean	std	min	25%	50%	75%	max
case_positive_specimen_interval	26268.0	0.154104	2.385045	-100.0	0.0	0.0	0.0	101.0
case_onset_interval	21322.0	-0.062799	2.097310	-87.0	0.0	0.0	0.0	68.0

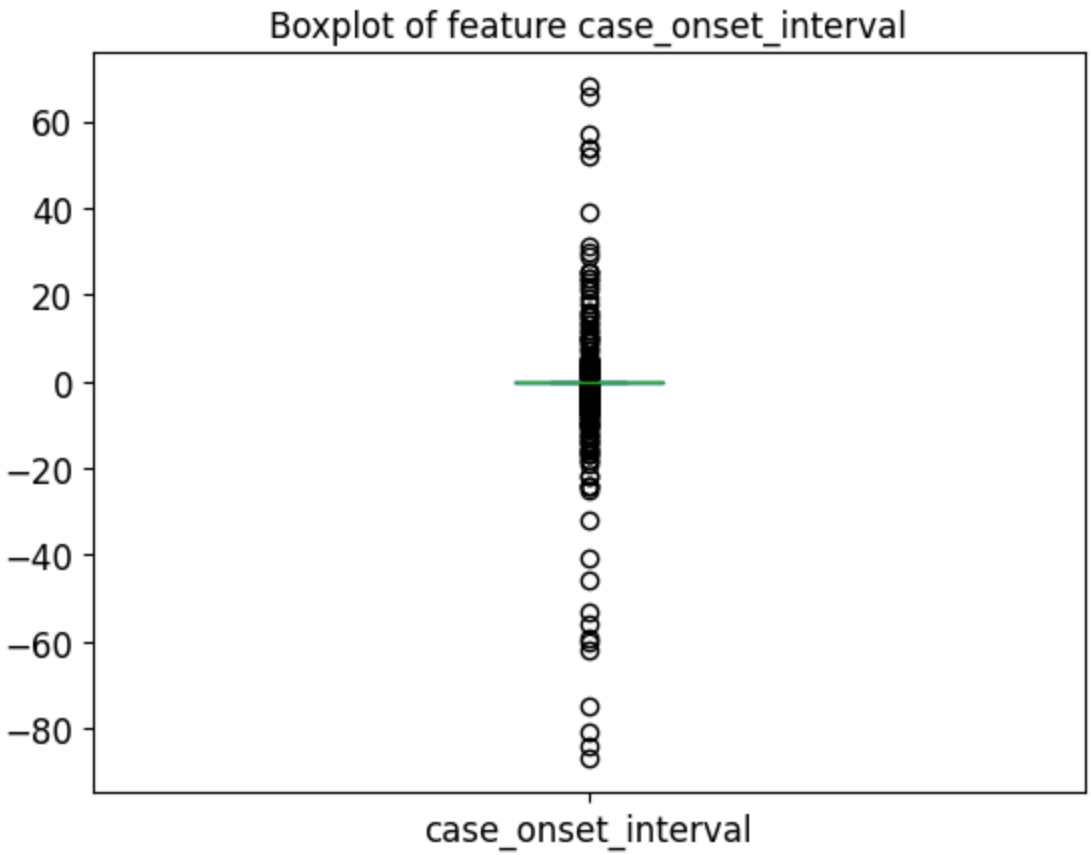
	%missing
case_positive_specimen_interval	47.464
case_onset_interval	57.356

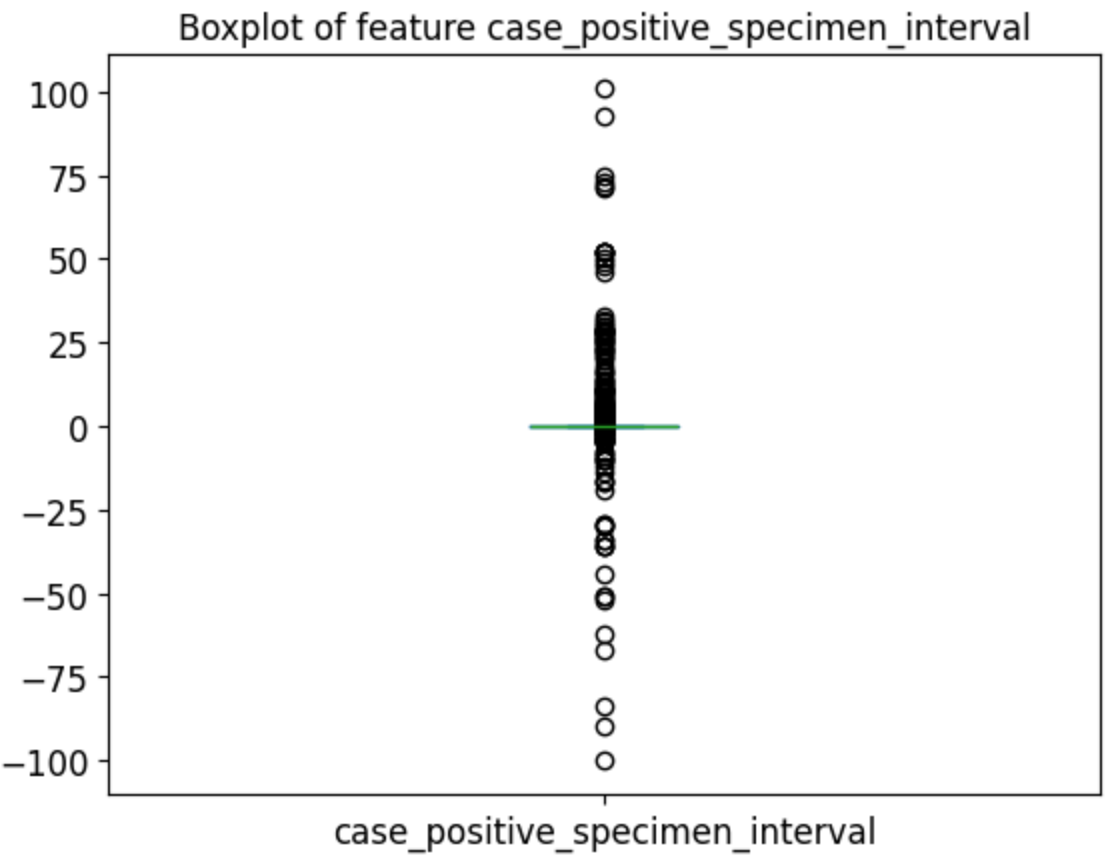
	count	unique	top	freq	%missing
res_state	50000	50	NY	5628	0.000
res_county	47114	936	MIAMI-DADE	994	5.772
county_fips_code	47114.0	1336.0	12086.0	994.0	5.772
age_group	49580	5	18 to 49 years	20028	0.840
sex	48850	4	Female	25179	2.300
race	43811	8	White	30496	12.378
ethnicity	43236	4	Non-Hispanic/Latino	29632	13.528
process	50000	10	Missing	45449	0.000
exposure_yn	50000	3	Missing	43114	0.000
current_status	50000	2	Laboratory-confirmed case	42245	0.000
symptom_status	50000	4	Symptomatic	22488	0.000
hosp_yn	50000	4	No	25335	0.000
icu_yn	50000	4	Missing	39200	0.000
death_yn	50000	2	No	40000	0.000
underlying_conditions_yn	4157	2	Yes	4093	91.686
year	50000	4	2021	18077	0.000
month	50000	12	01	10406	0.000

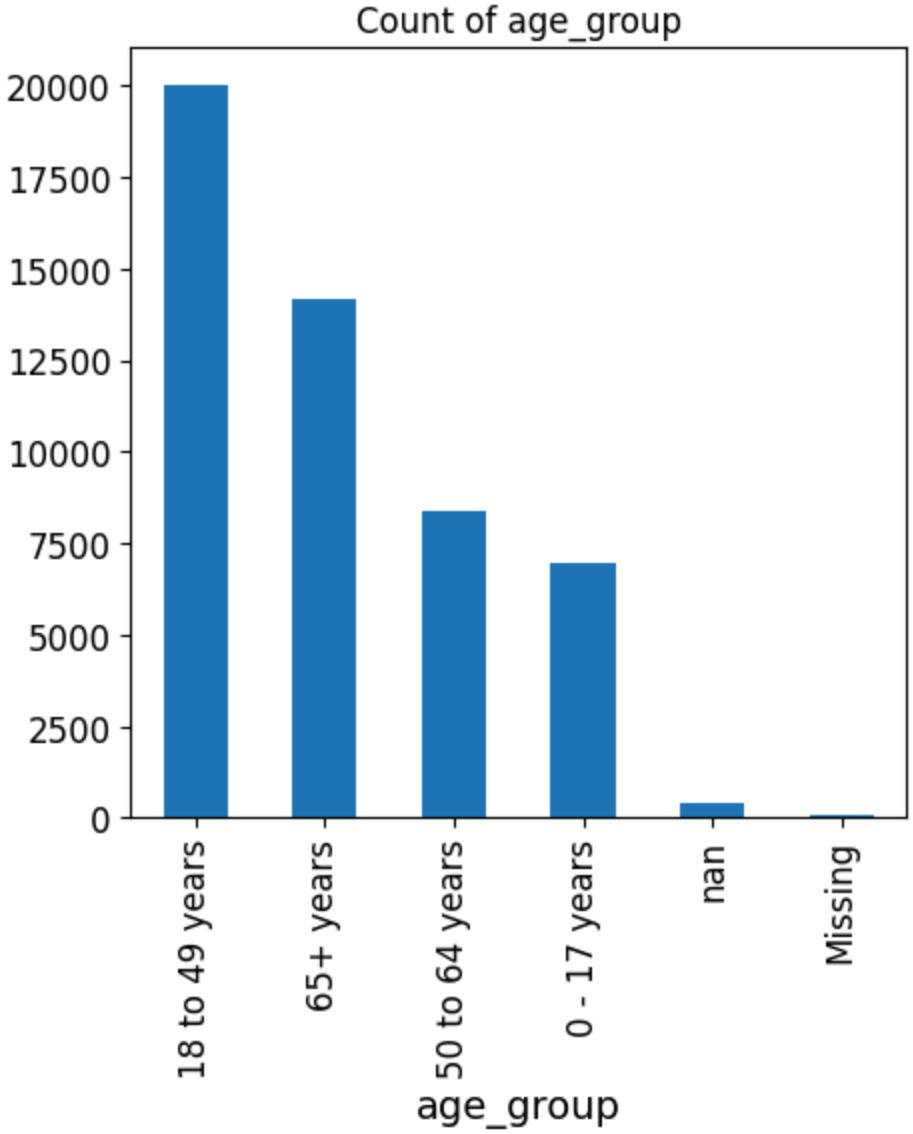
	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode
res_state	NY	5628	11.256	NC	4678	9.356
res_county	MIAMI-DADE	994	1.988	MARICOPA	771	1.542
county_fips_code	12086.0	994	1.988	4013.0	771	1.542
age_group	18 to 49 years	20028	40.056	65+ years	14145	28.29
sex	Female	25179	50.358	Male	23437	46.874
race	White	30496	60.992	Black	5294	10.588
ethnicity	Non-Hispanic/Latino	29632	59.264	Unknown	6796	13.592
process	Missing	45449	90.898	Clinical evaluation	2178	4.356
exposure_yn	Missing	43114	86.228	Yes	4868	9.736
current_status	Laboratory-confirmed case	42245	84.49	Probable Case	7755	15.51
symptom_status	Symptomatic	22488	44.976	Missing	21299	42.598
hosp_yn	No	25335	50.67	Missing	11398	22.796
icu_yn	Missing	39200	78.4	Unknown	6742	13.484
death_yn	No	40000	80.0	Yes	10000	20.0
underlying_conditions_yn	Yes	4093	8.186	No	64	0.128
year	2021	18077	36.154	2022	15468	30.936
month	01	10406	20.812	12	7956	15.912

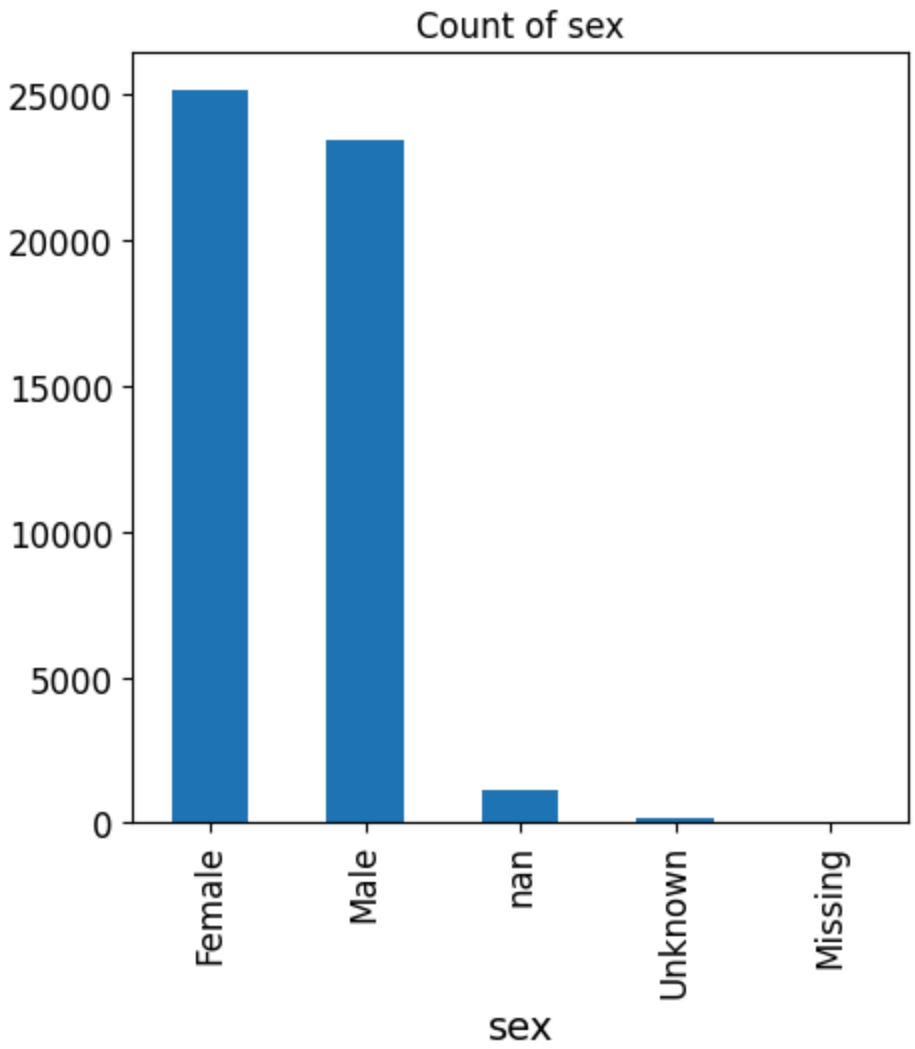


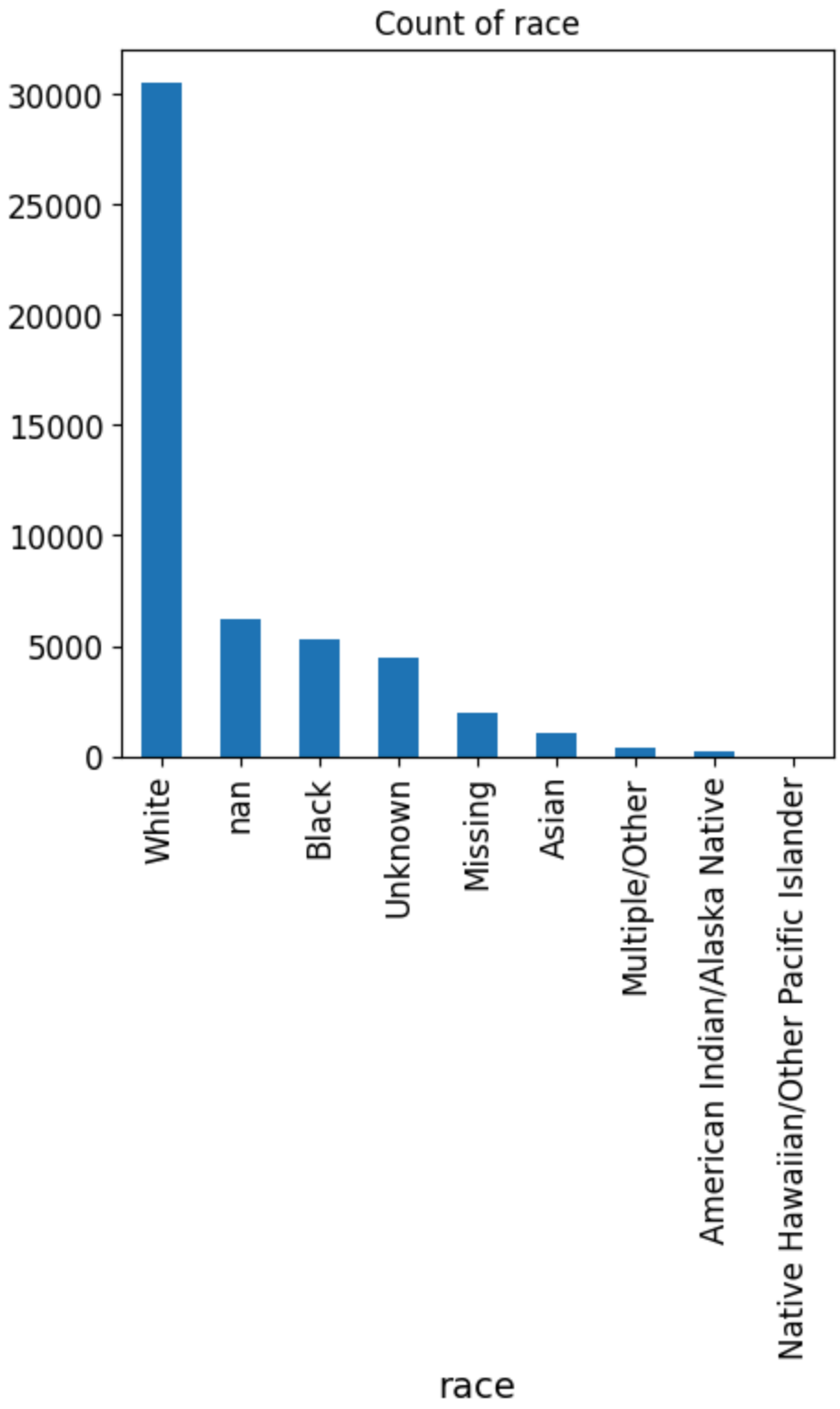

Histrogram of case_onset_interval
Frequency
20000
17500
15000
12500
10000
7500
5000
2500
0
−10
−5
0
5
10

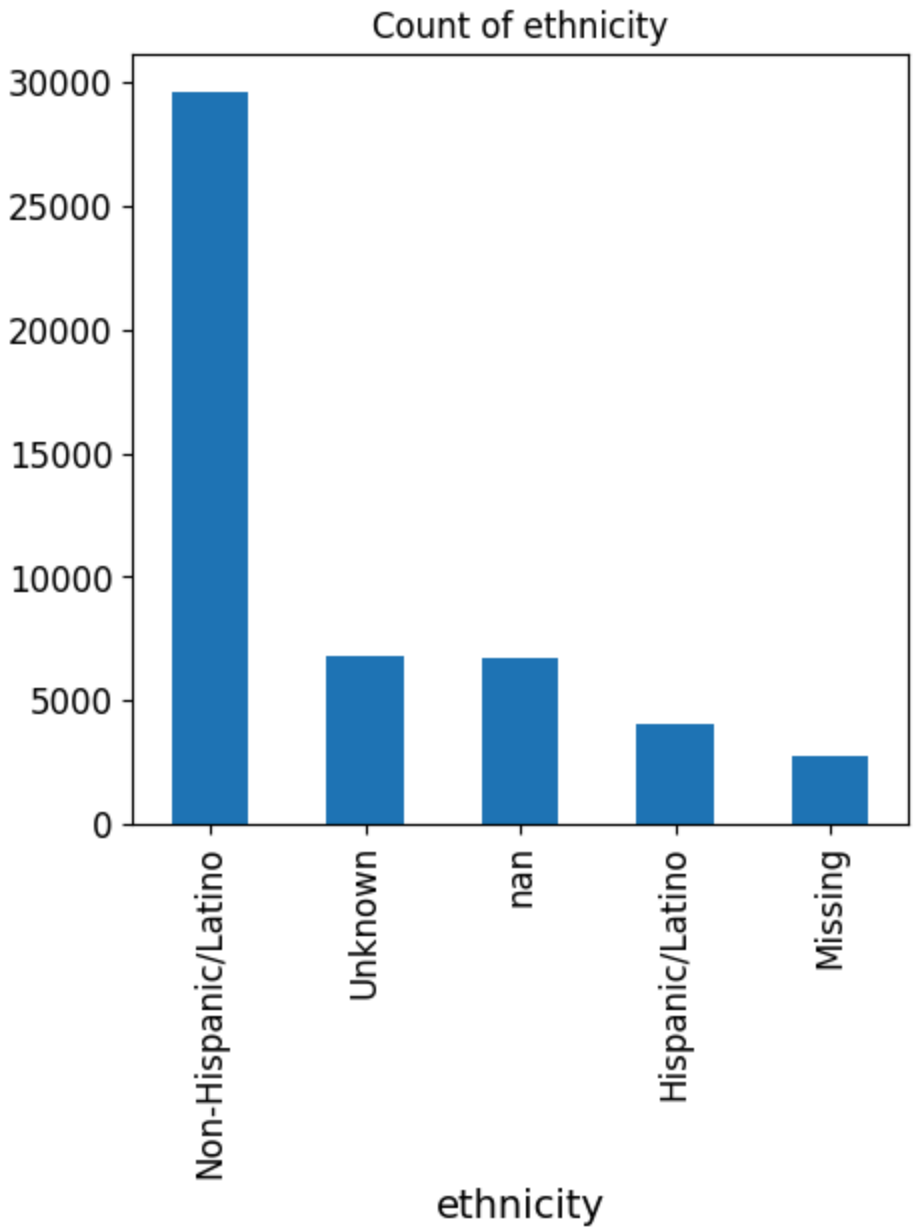

Boxplot of feature case_onset_interval
60
40
20
0
−20
−40
−60
−80
case_onset_interval

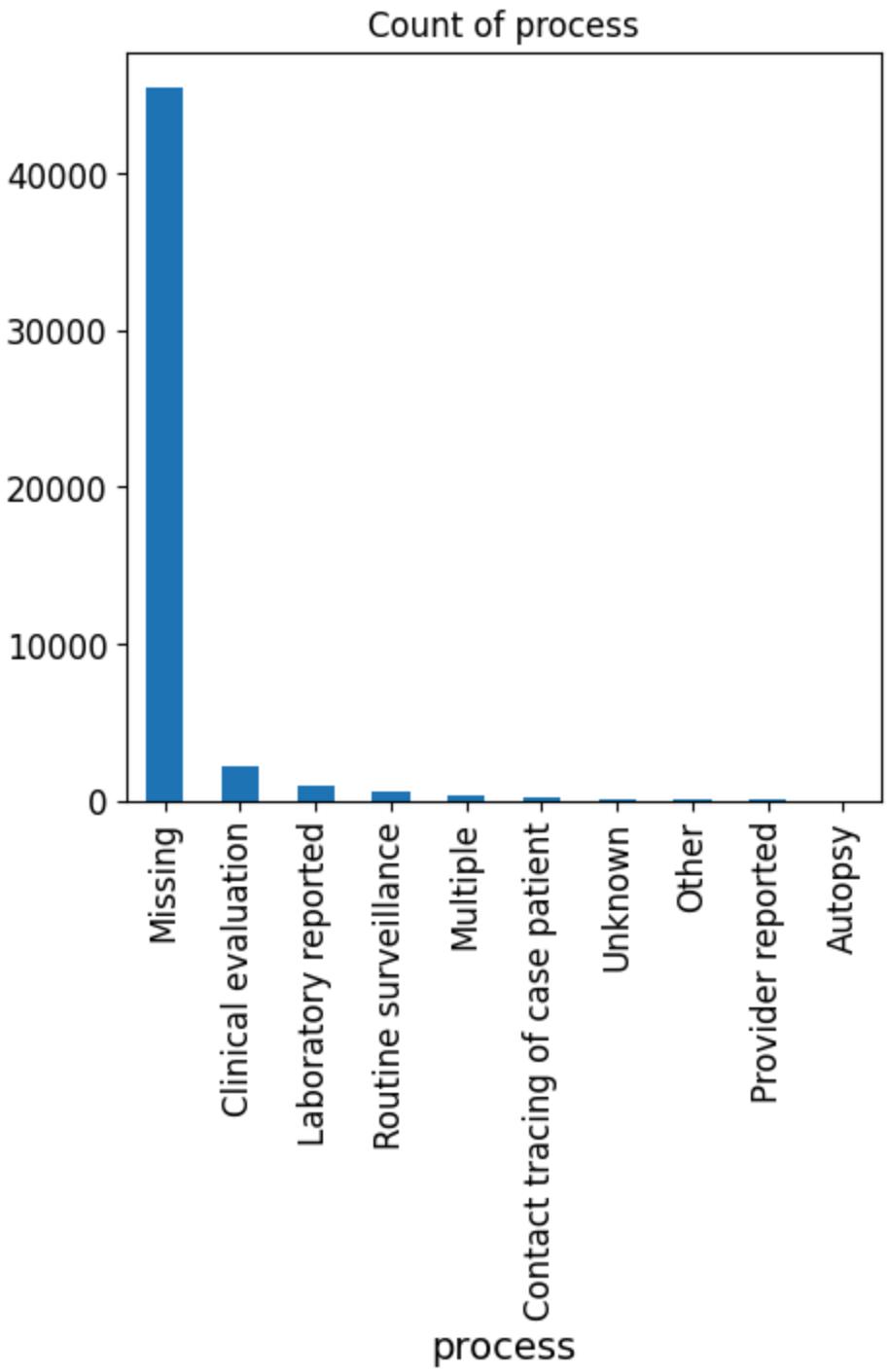

Boxplot of feature case_positive_specimen_interval
100
75
50
25
0
−25
−50
−75
−100
case_positive_specimen_interval

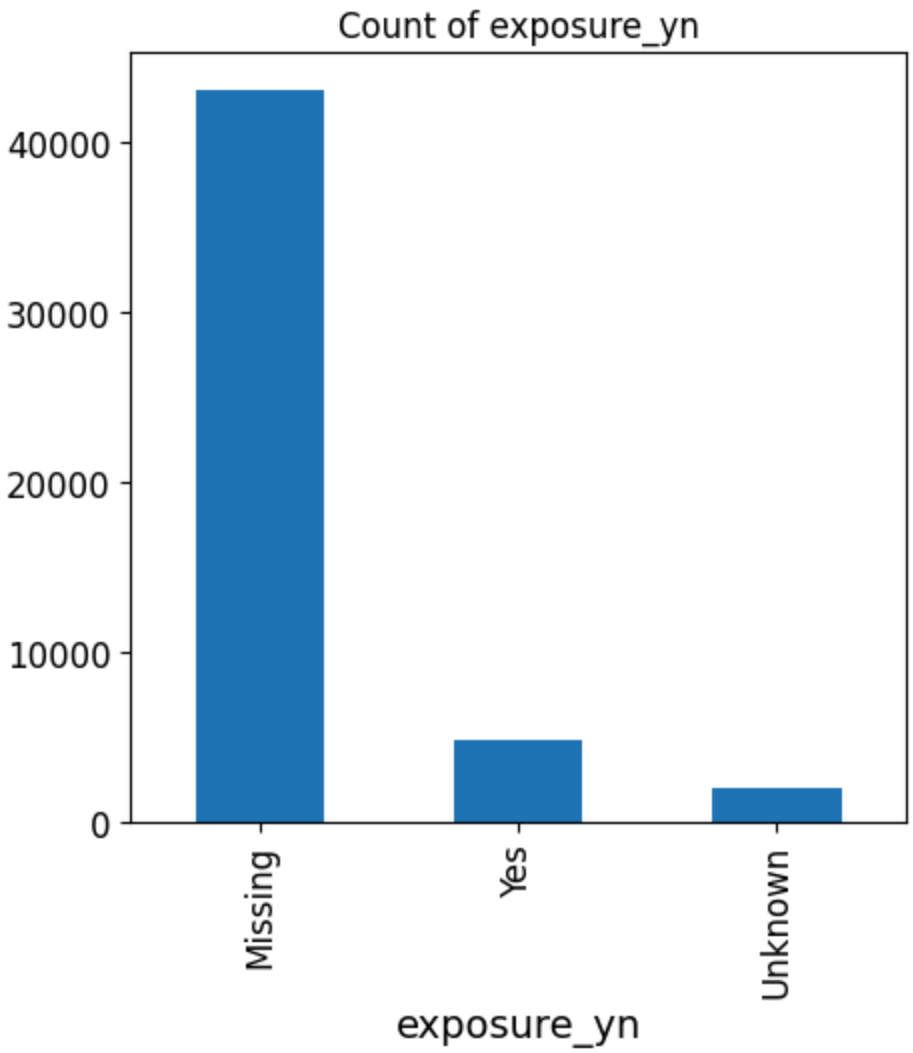

Count of age_group
20000
17500
15000
12500
10000
7500
5000
2500
0
18 to 49 years
65+ years
50 to 64 years
0 - 17 years
nan
Missing
age_group

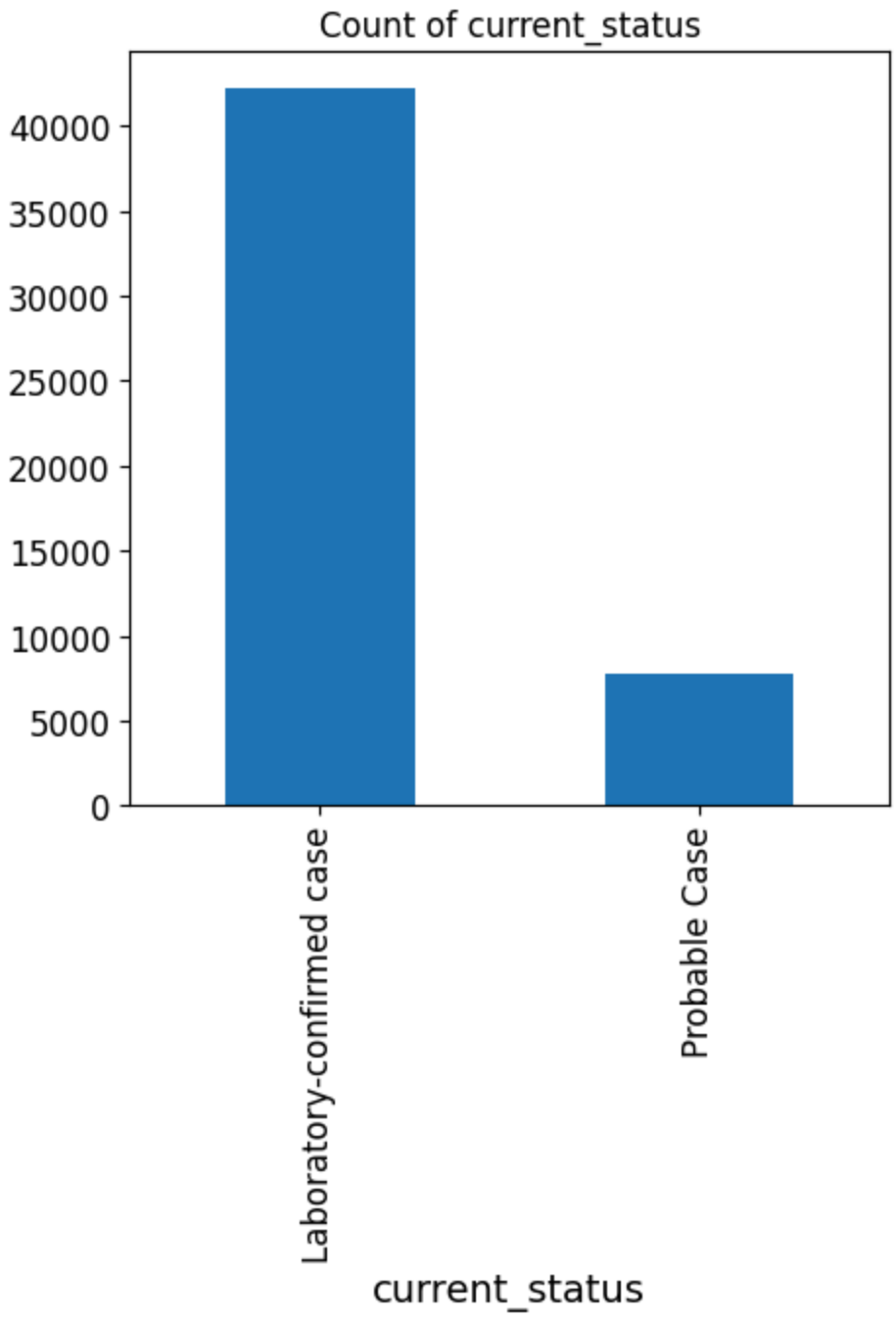

Count of sex
25000
20000
15000
10000
5000
0
Female
Male
nan
Unknown
Missing
sex

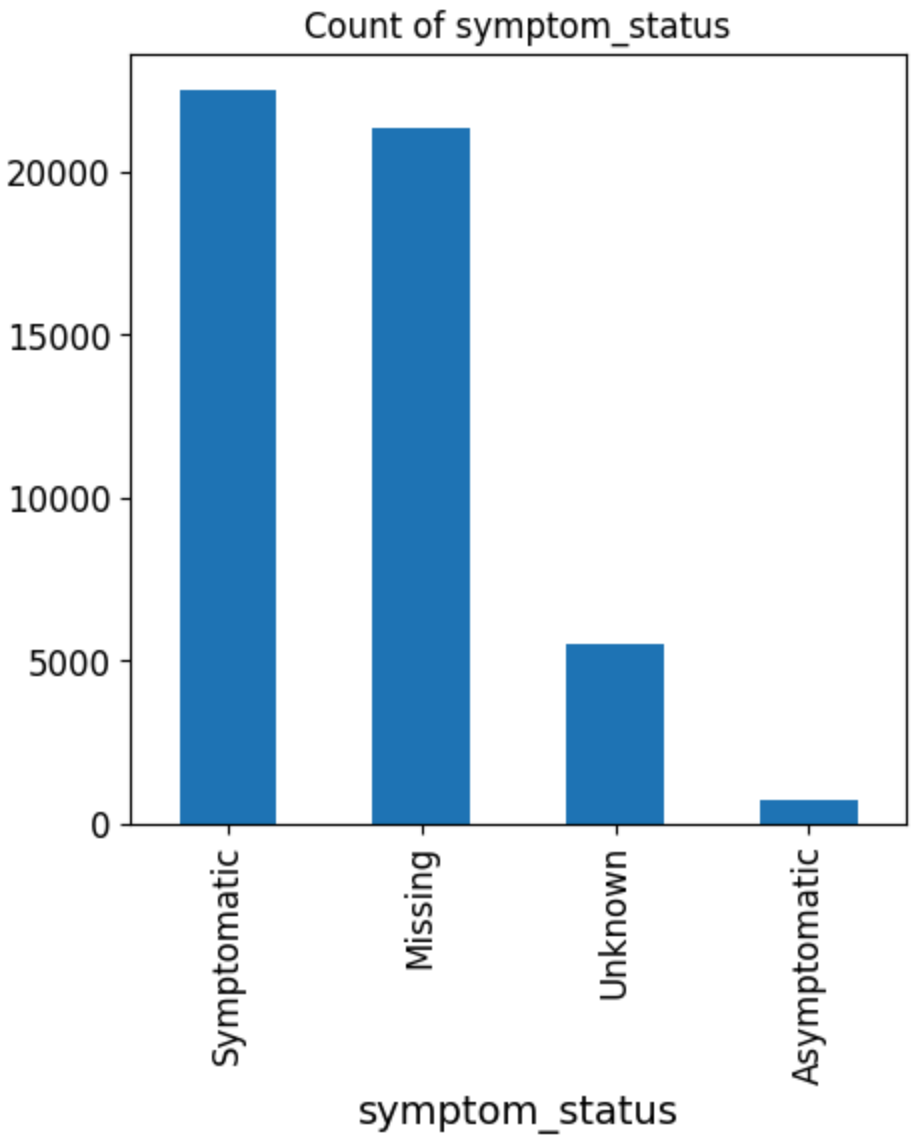

Count of race
30000
25000
20000
15000
10000
5000
0
White
nan
Black
Unknown
Missing
Asian
Multiple/Other
American Indian/Alaska Native
Native Hawaiian/Other Pacific Islander
race

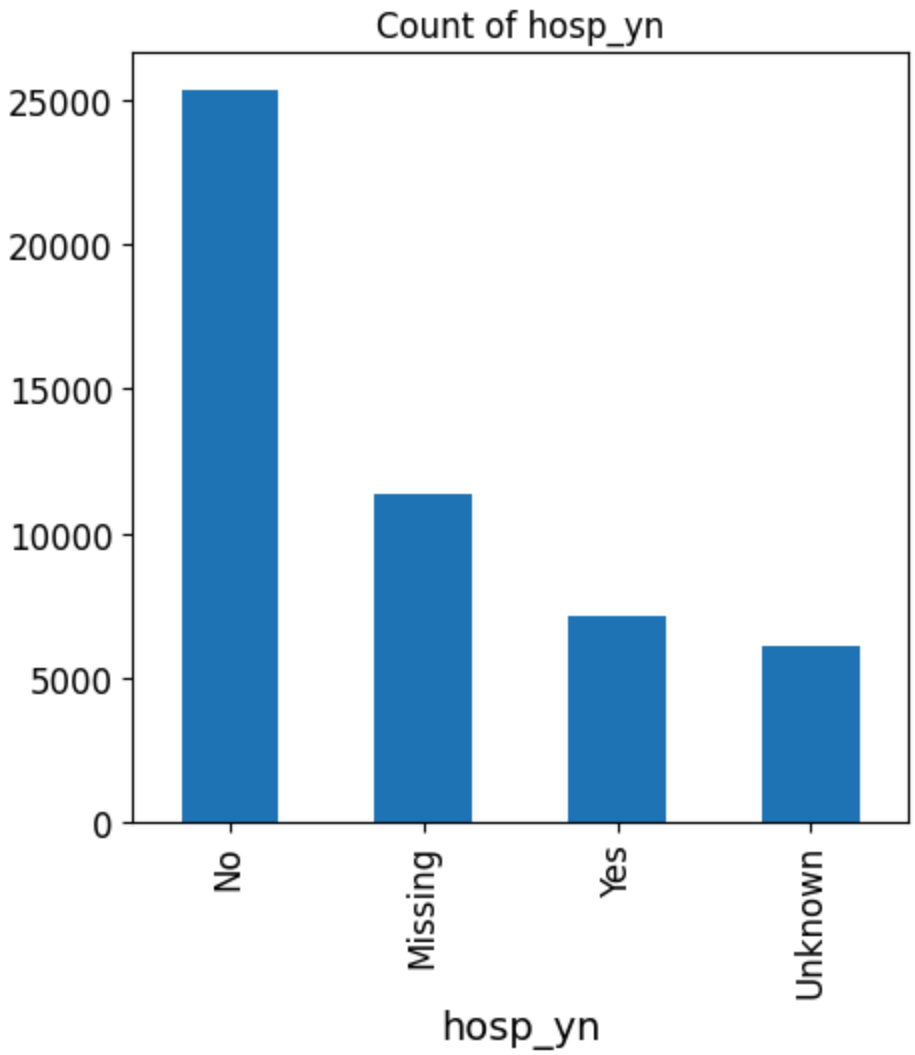

Count of ethnicity
30000
25000
20000
15000
10000
5000
0
Non-Hispanic/Latino
Unknown
nan
Hispanic/Latino
Missing
ethnicity

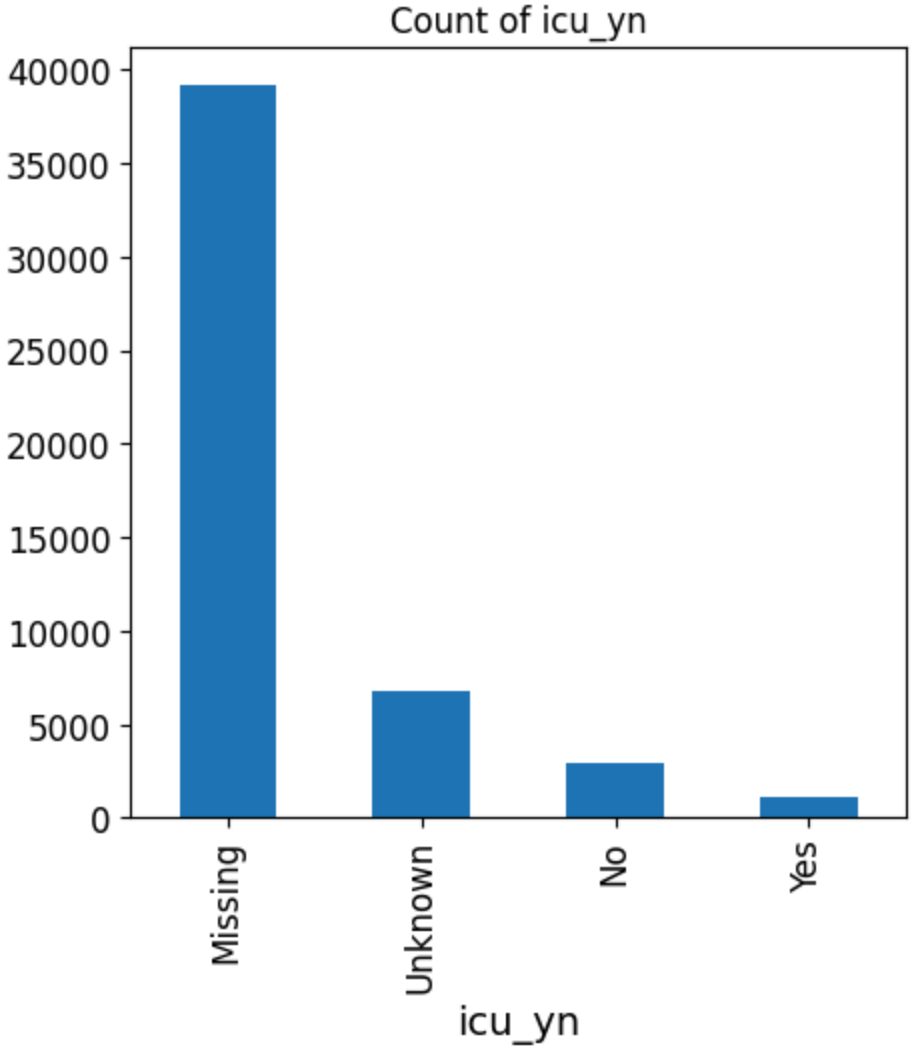

Count of process
40000
30000
20000
10000
0
Missing
Clinical evaluation
Laboratory reported
Routine surveillance
Multiple
Contact tracing of case patient
Unknown
Other
Provider reported
Autopsy
process

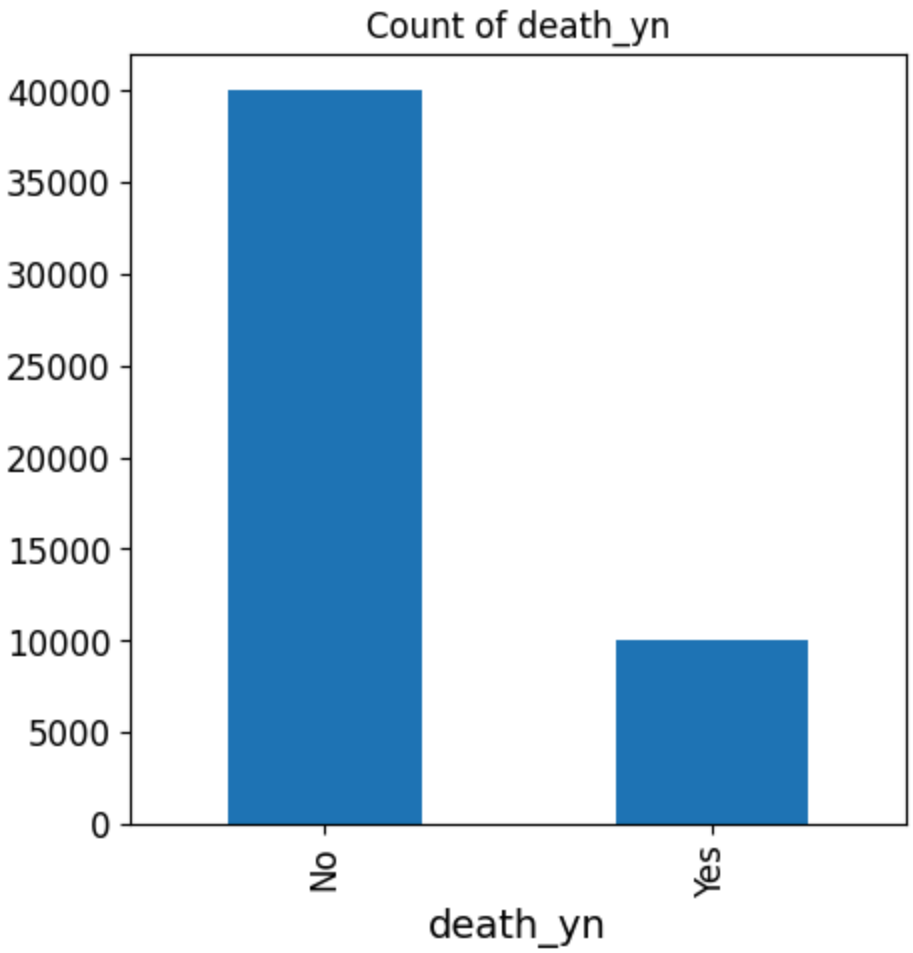

Count of exposure_yn
40000
30000
20000
10000
0
Missing
Yes
Unknown
exposure_yn

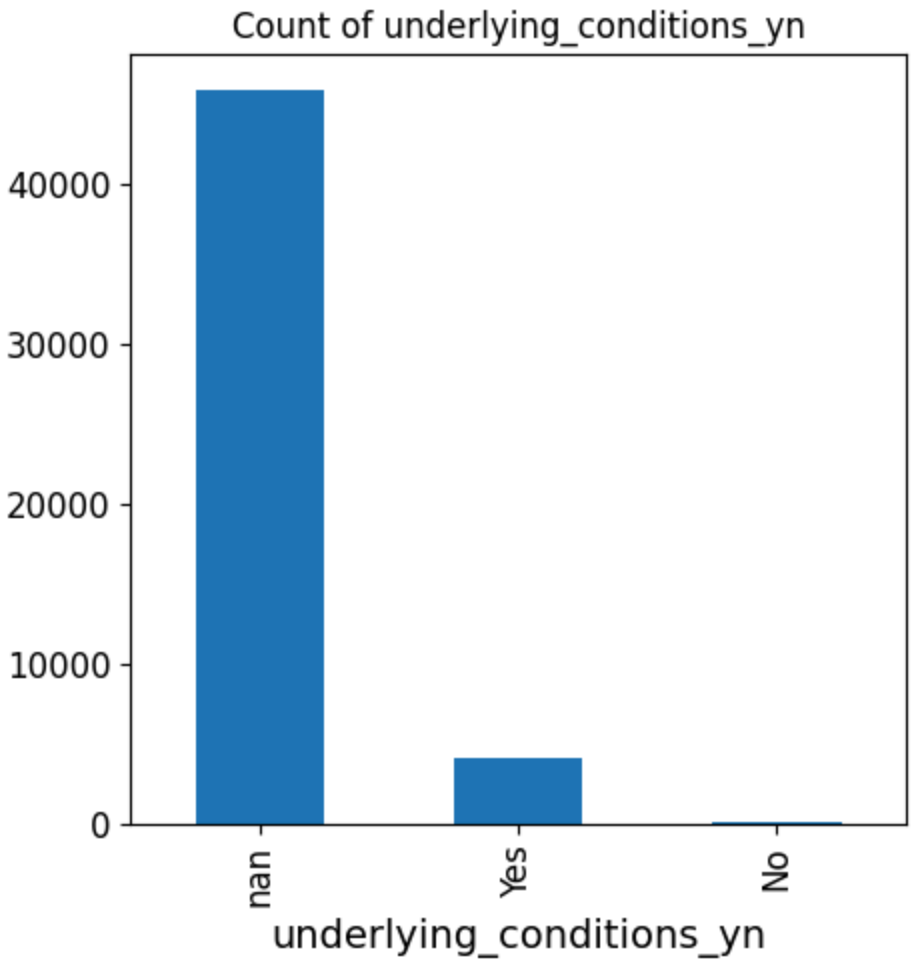

Count of current_status
40000
35000
30000
25000
20000
15000
10000
5000
0
Laboratory-confirmed case
Probable Case
current_status

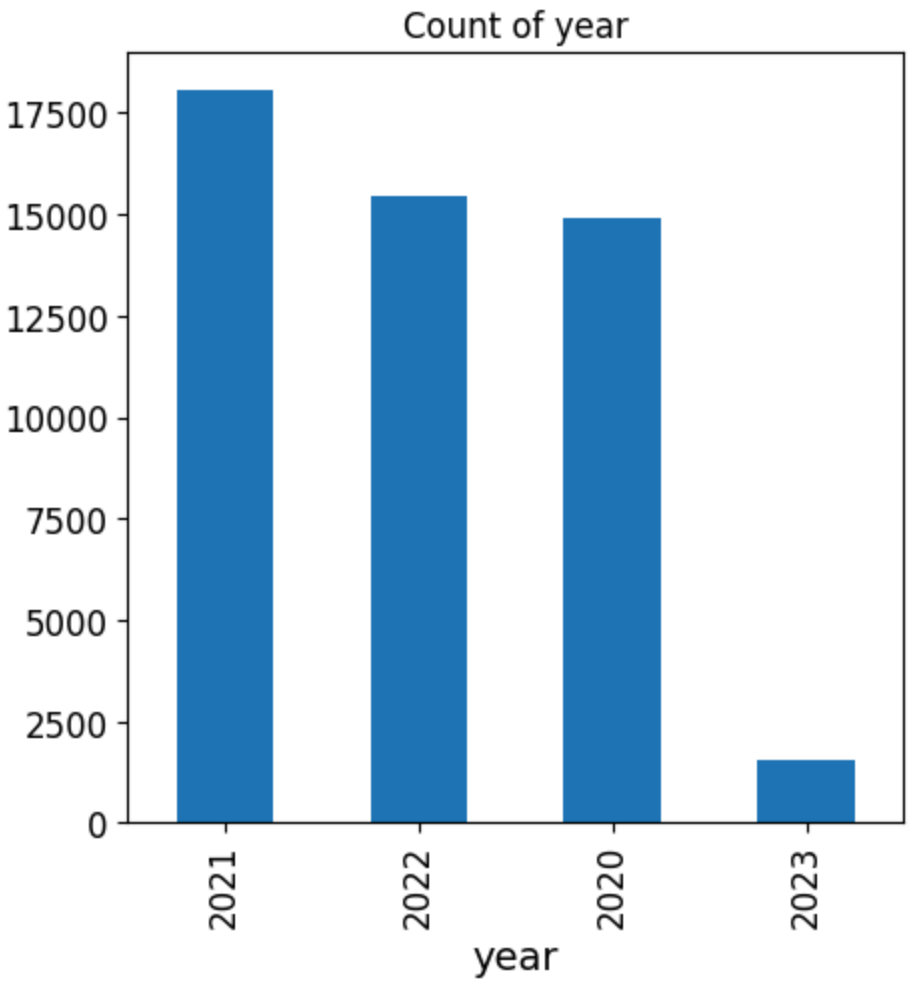

Count of symptom_status
20000
15000
10000
5000
0
Symptomatic
Missing
Unknown
Asymptomatic
symptom_status

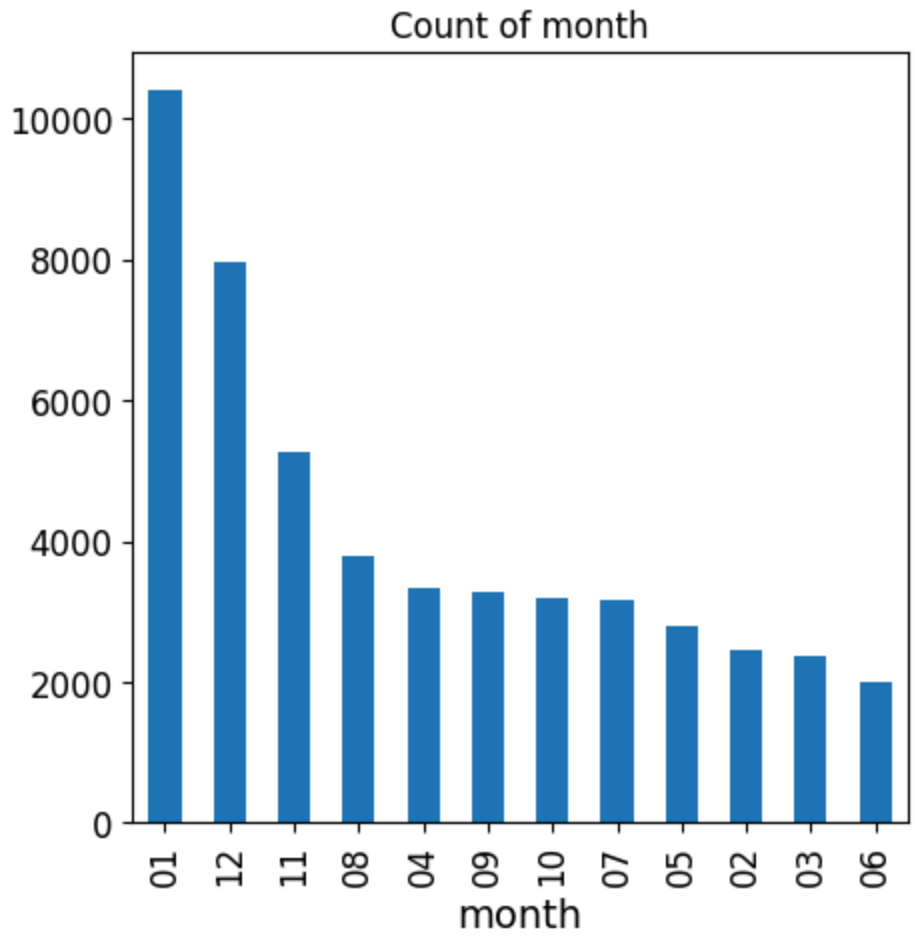

Count of hosp_yn
25000
20000
15000
10000
5000
0
No
Missing
Yes
Unknown
hosp_yn

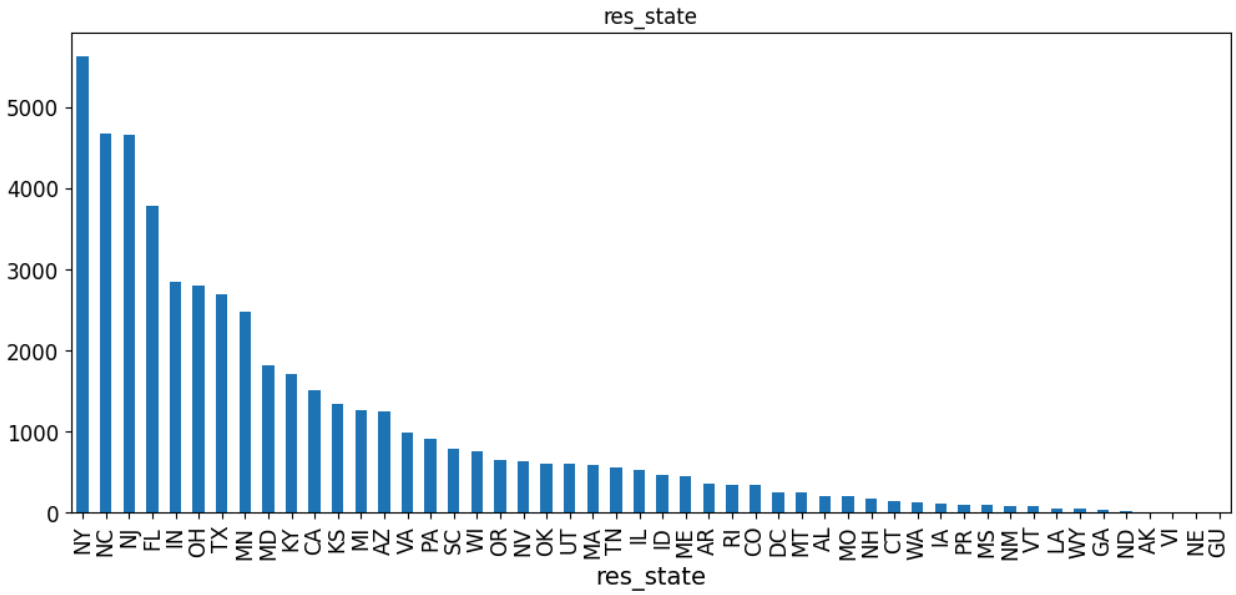

Count of icu_yn
40000
35000
30000
25000
20000
15000
10000
5000
0
Missing
Unknown
No
Yes
icu_yn

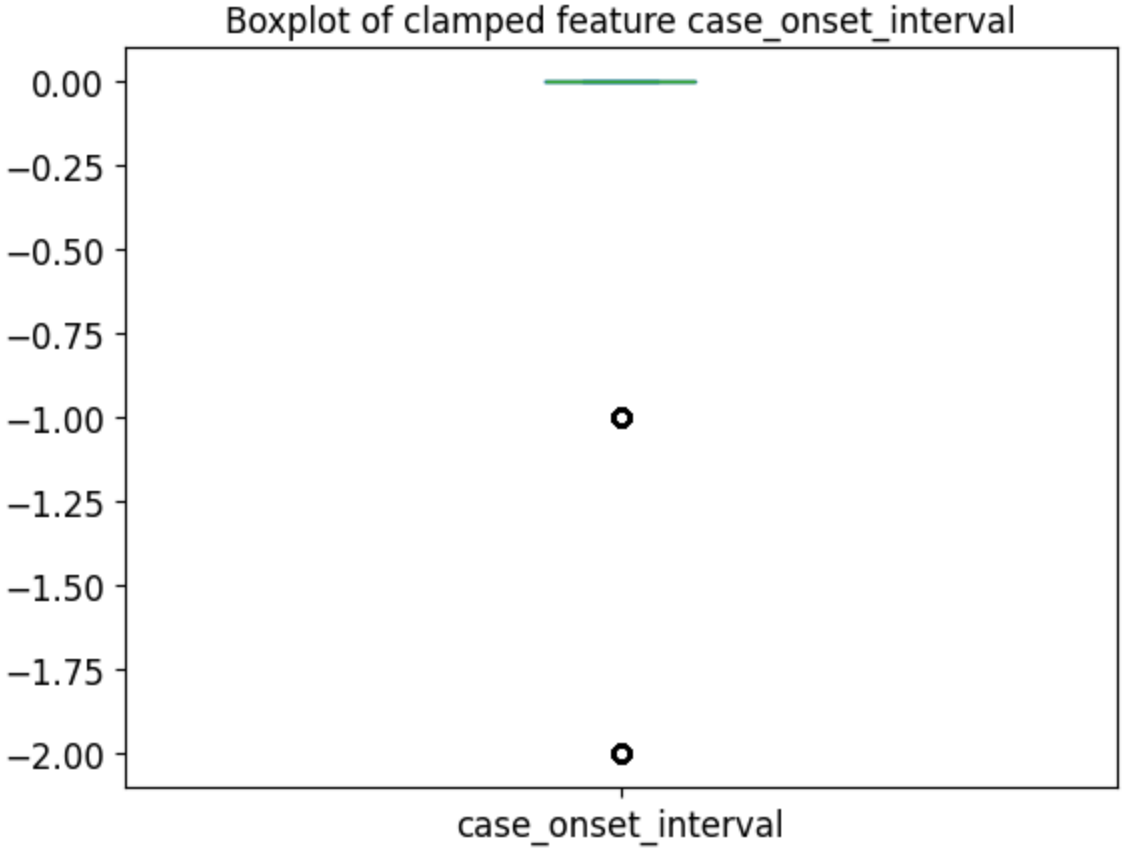

Count of death_yn
40000
35000
30000
25000
20000
15000
10000
5000
0
No
Yes
death_yn

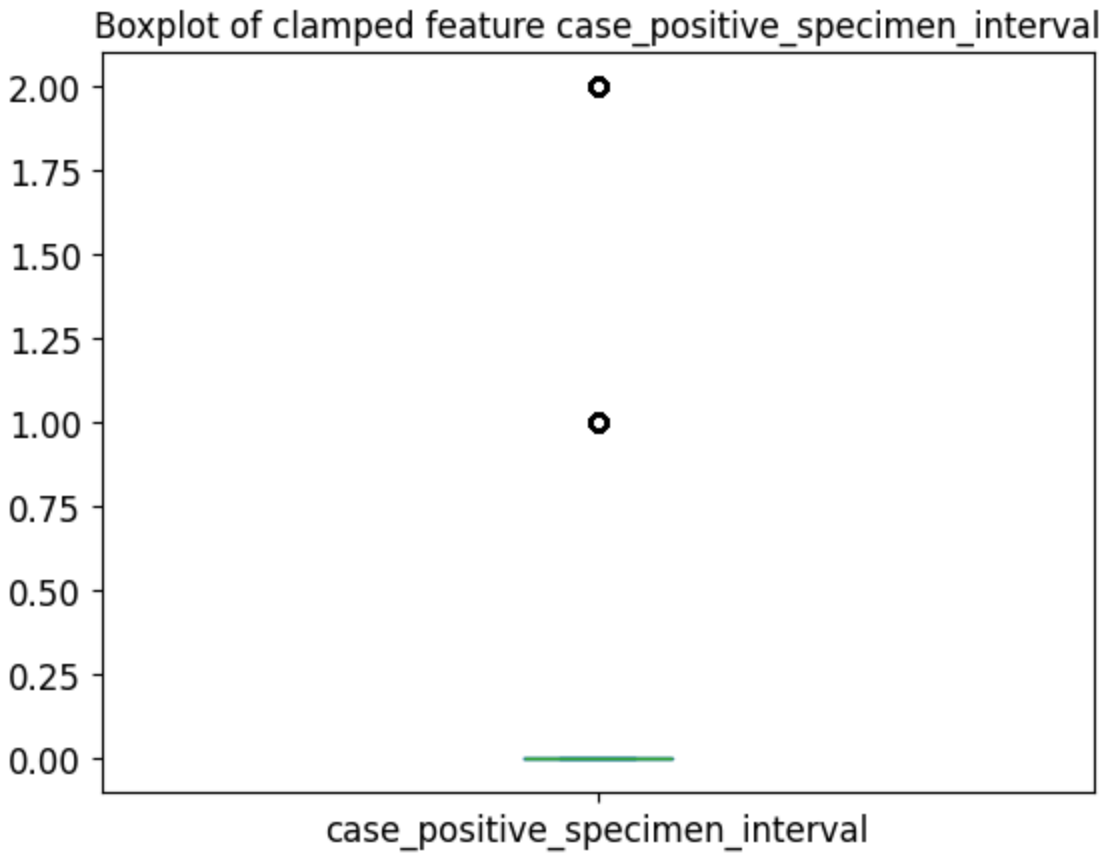

Count of underlying_conditions_yn
40000
30000
20000
10000
0
nan
Yes
No
underlying_conditions_yn

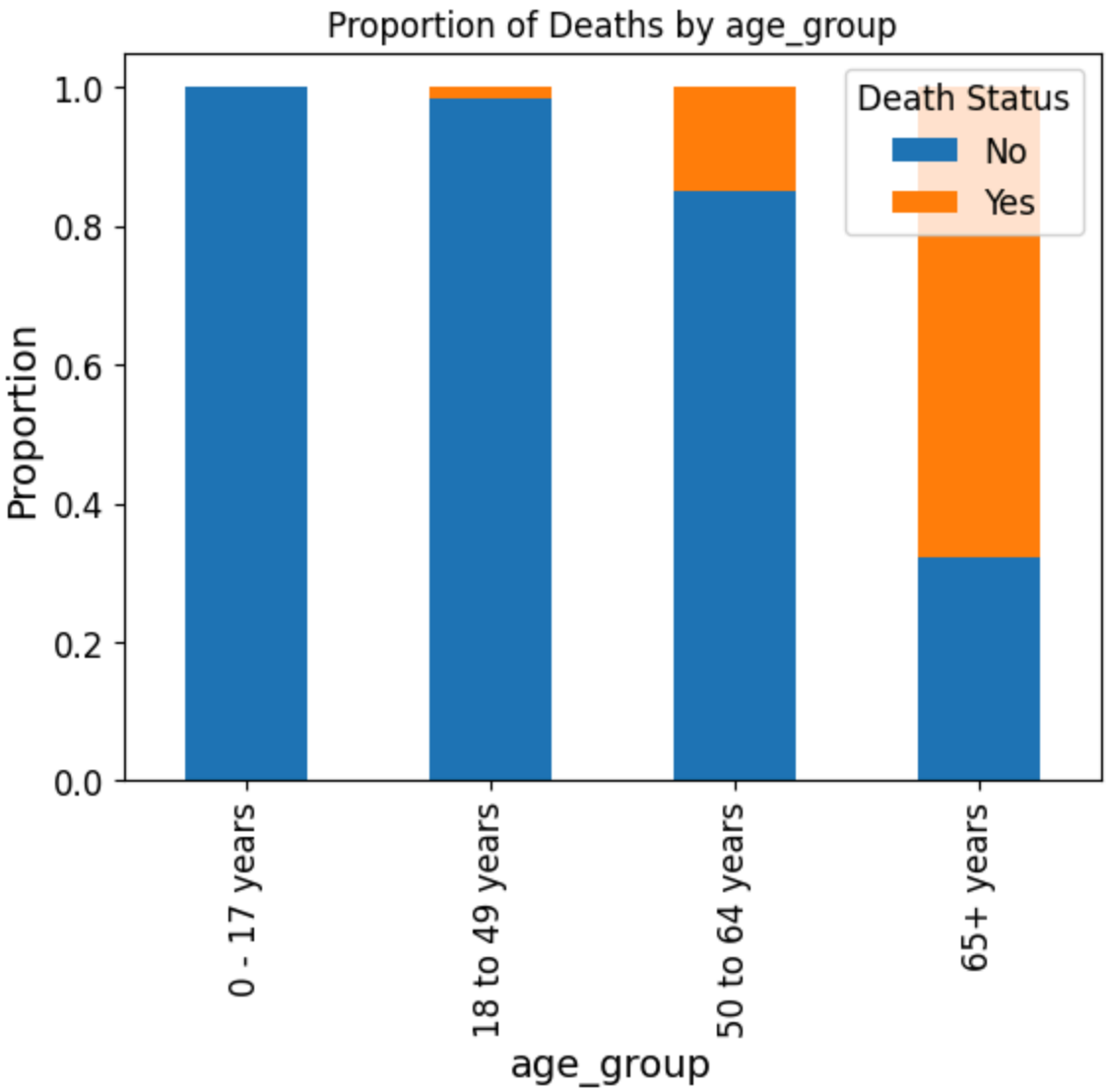

Count of year
17500
15000
12500
10000
7500
5000
2500
0
2021
2022
2020
2023
year

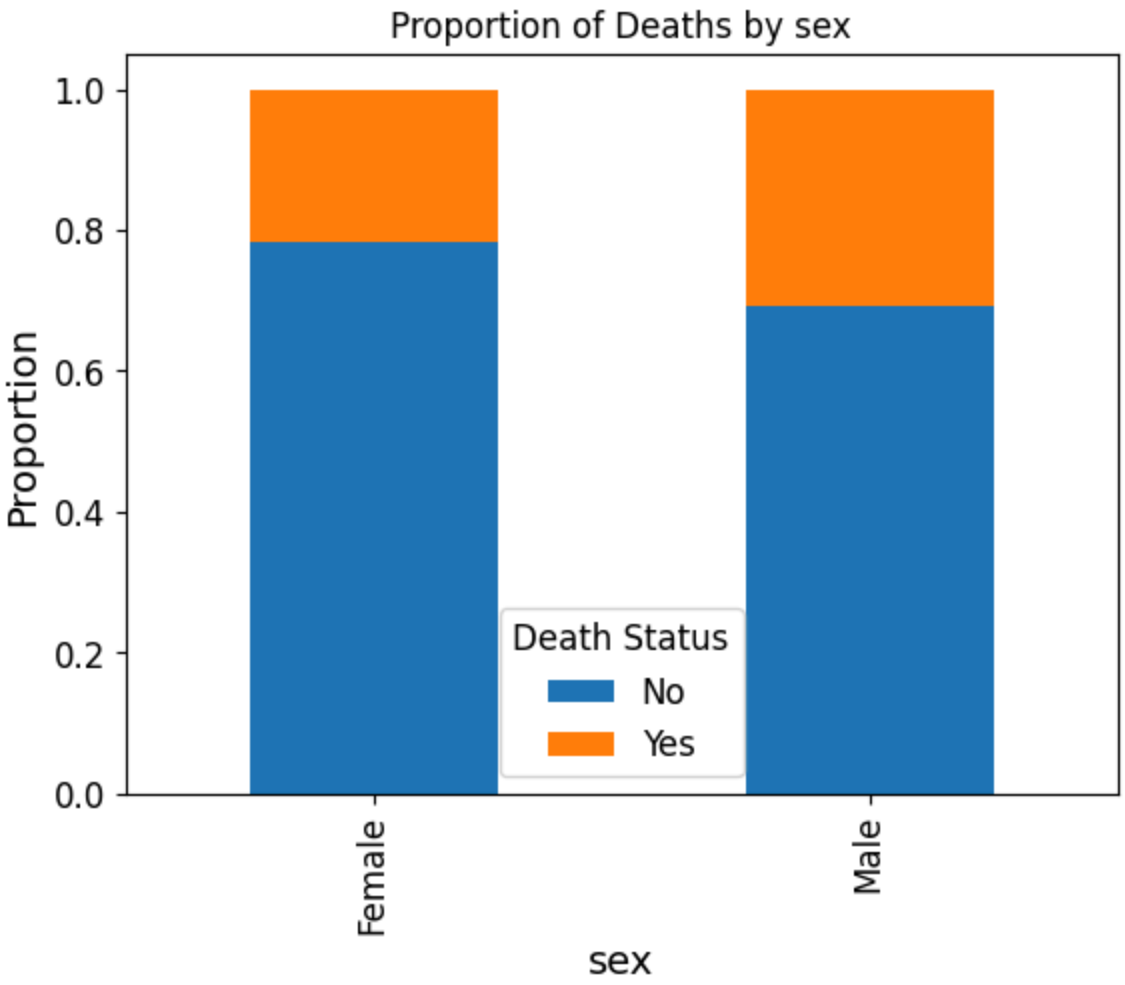

Count of month
10000
8000
6000
4000
2000
0
01
12
11
08
04
09
10
07
05
02
03
06
month

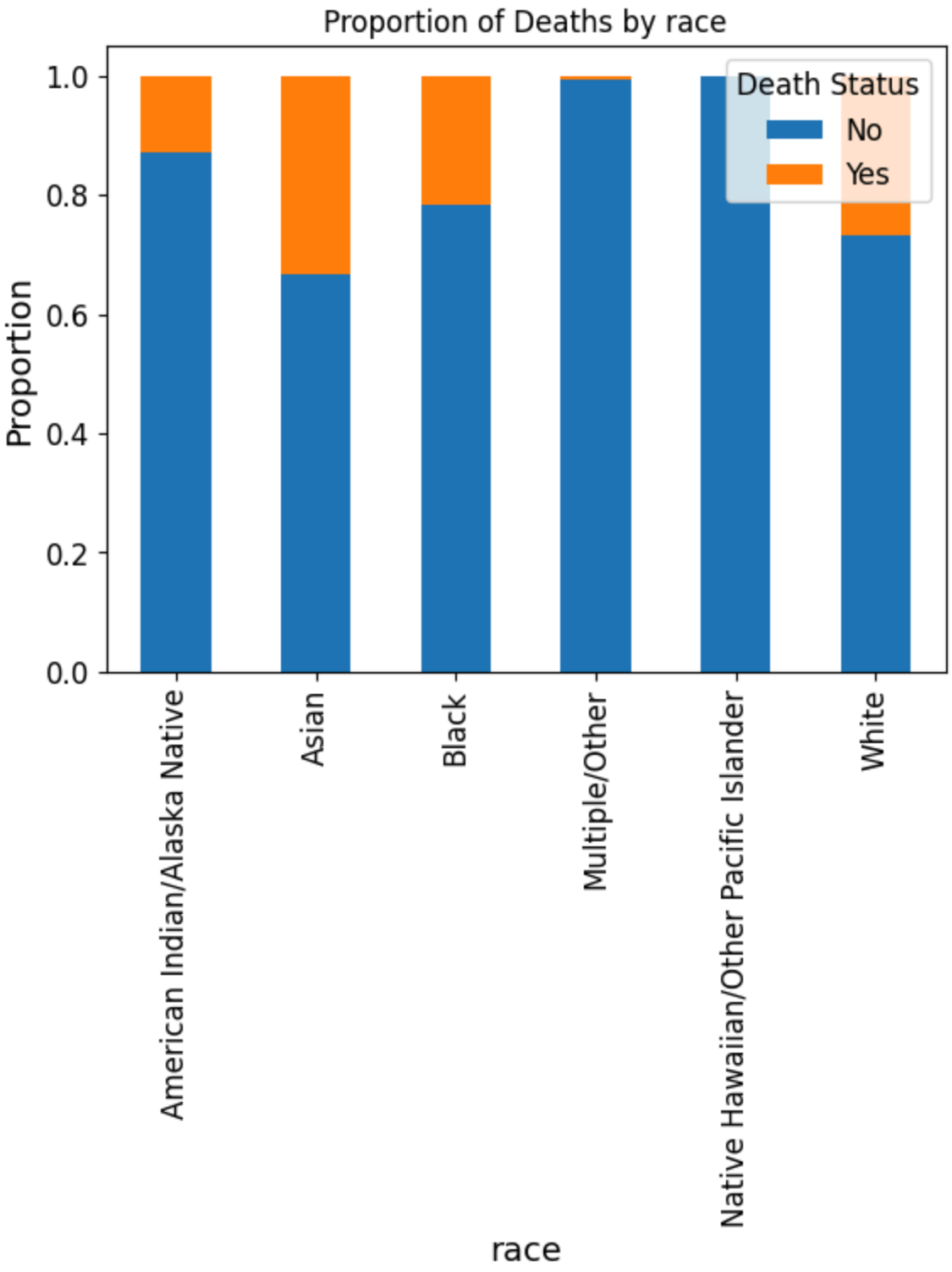

res_state
5000
4000
3000
2000
1000
0
NY
NC
NJ
FL
IN
OH
TX
MN
MD
KY
CA
KS
MI
AZ
VA
PA
SC
WI
OR
NV
OK
UT
MA
TN
IL
ID
ME
AR
RI
CO
DC
MT
AL
MO
NH
CT
WA
IA
PR
MS
NM
VT
LA
WY
GA
ND
AK
VI
NE
GU
res_state

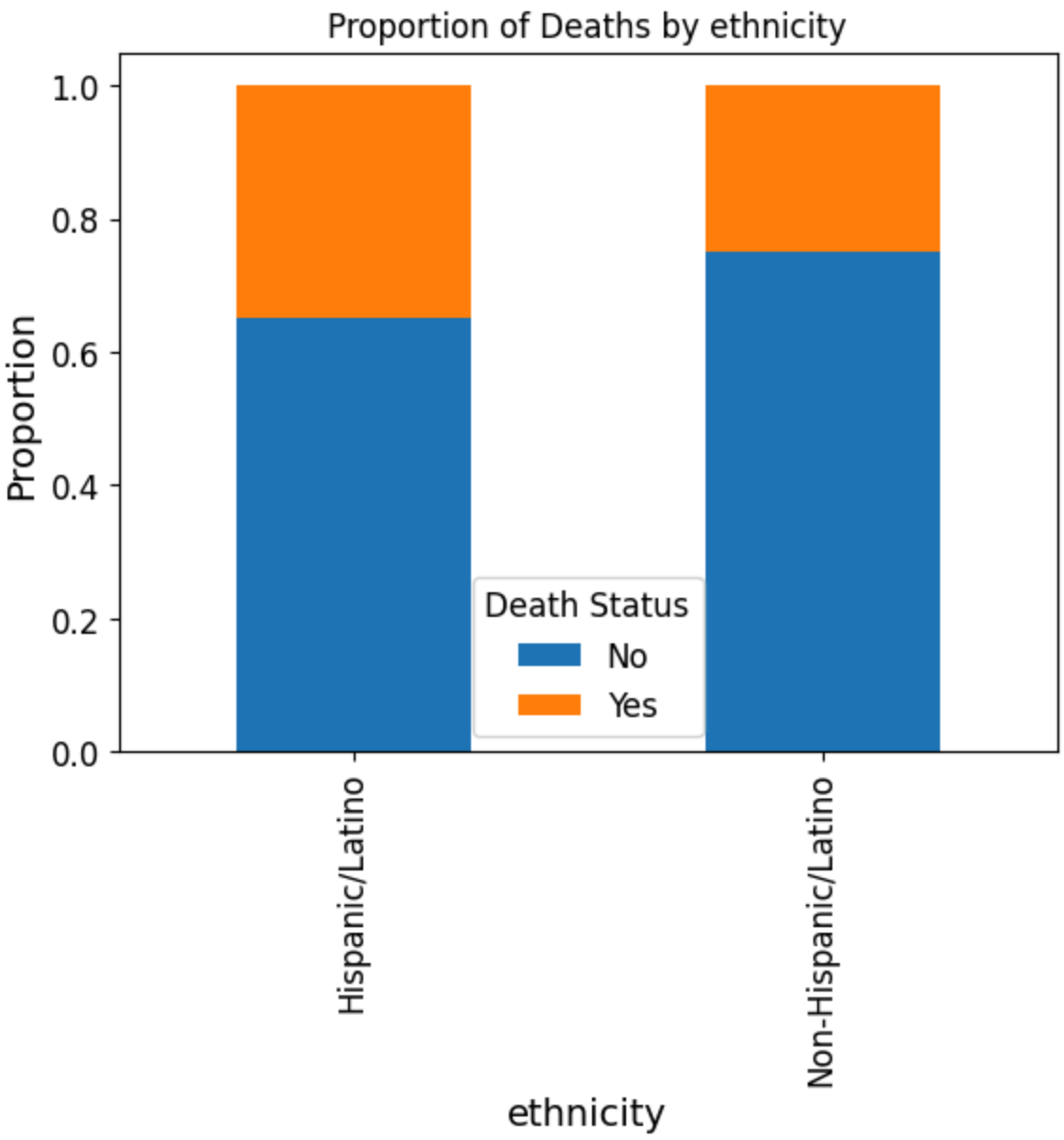

Boxplot of clamped feature case_onset_interval
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
−1.50
−1.75
−2.00
case_onset_interval

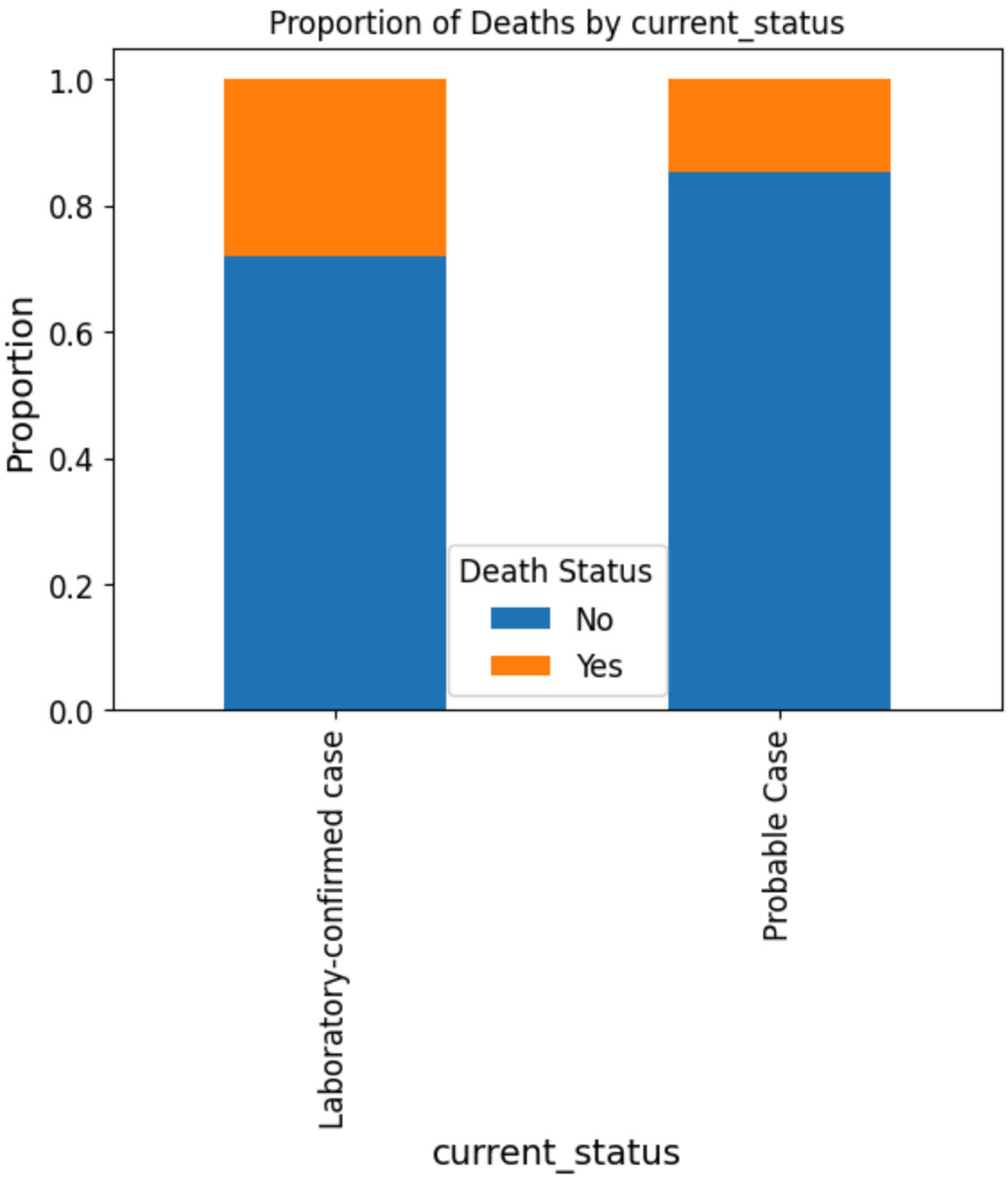

Boxplot of clamped feature case_positive_specimen_interval
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
case_positive_specimen_interval

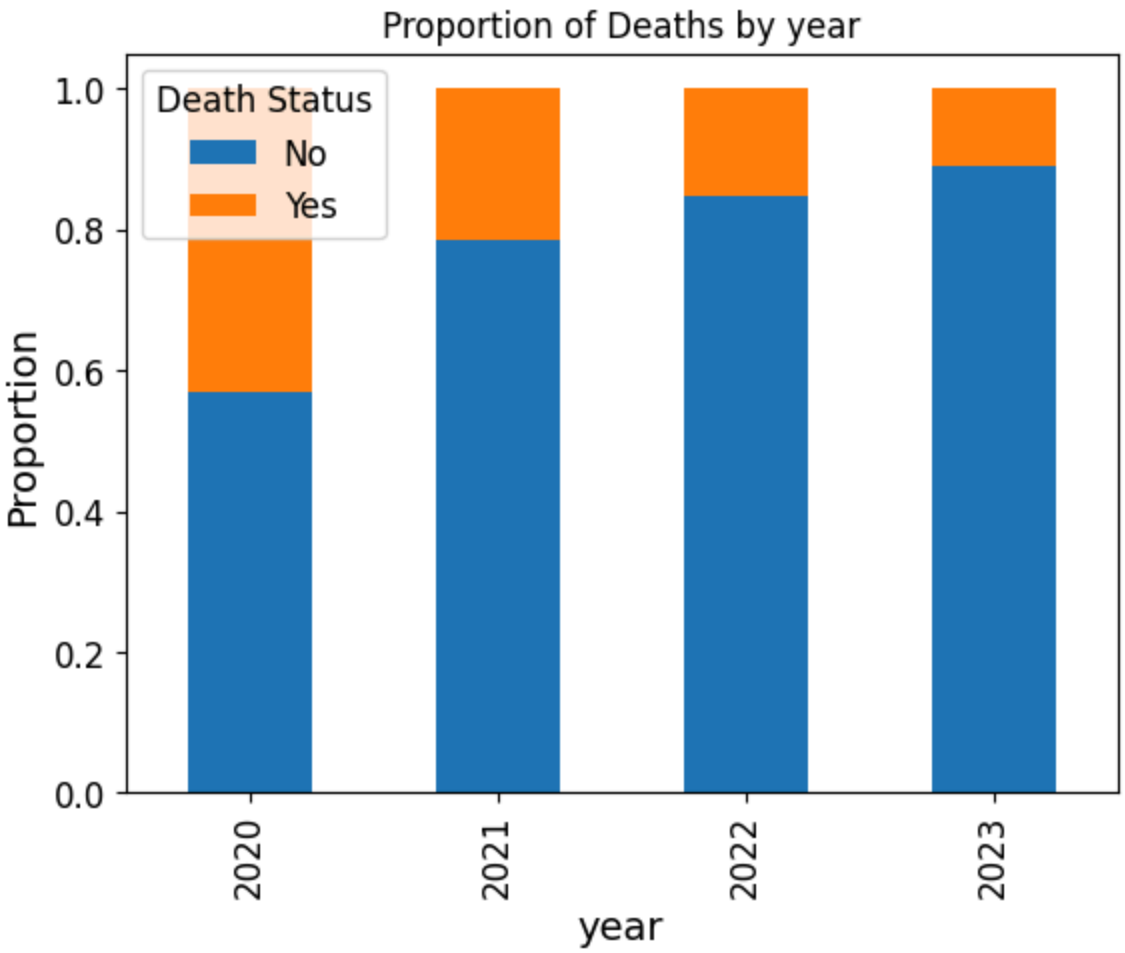

Proportion of Deaths by age_group
Death Status
No
Yes
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0 - 17 years
18 to 49 years
50 to 64 years
65+ years
age_group

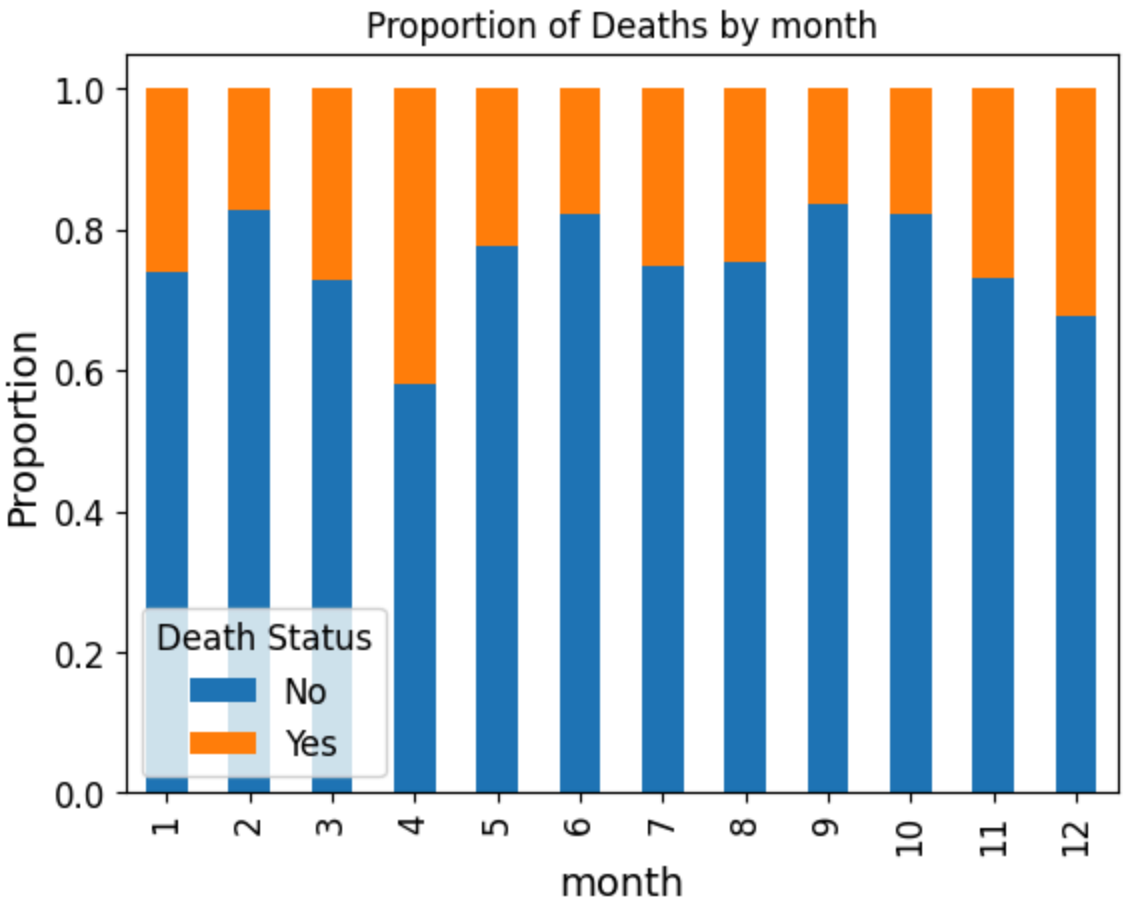

Proportion of Deaths by sex
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
Death Status
No
Yes
Female
Male
sex

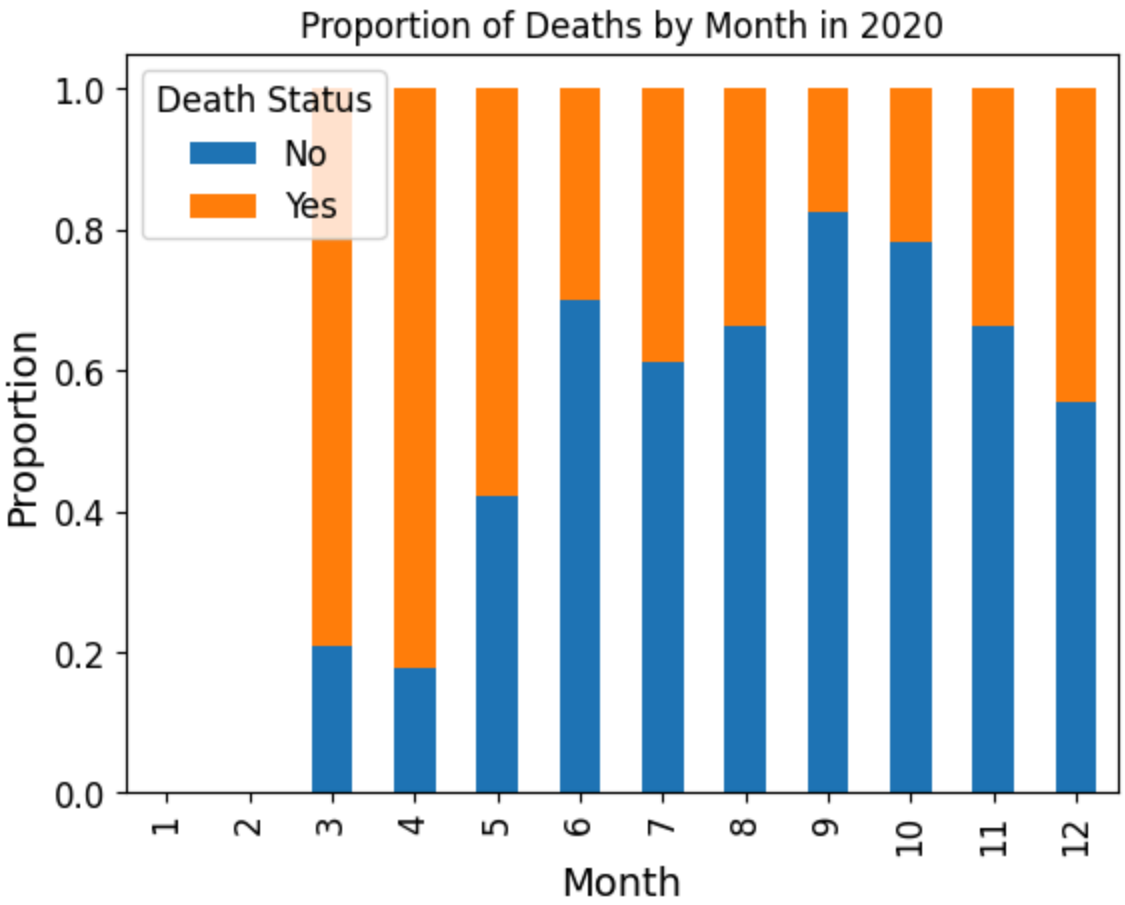

Proportion of Deaths by race
Death Status
No
Yes
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
American Indian/Alaska Native
Asian
Black
Multiple/Other
Native Hawaiian/Other Pacific Islander
White
race

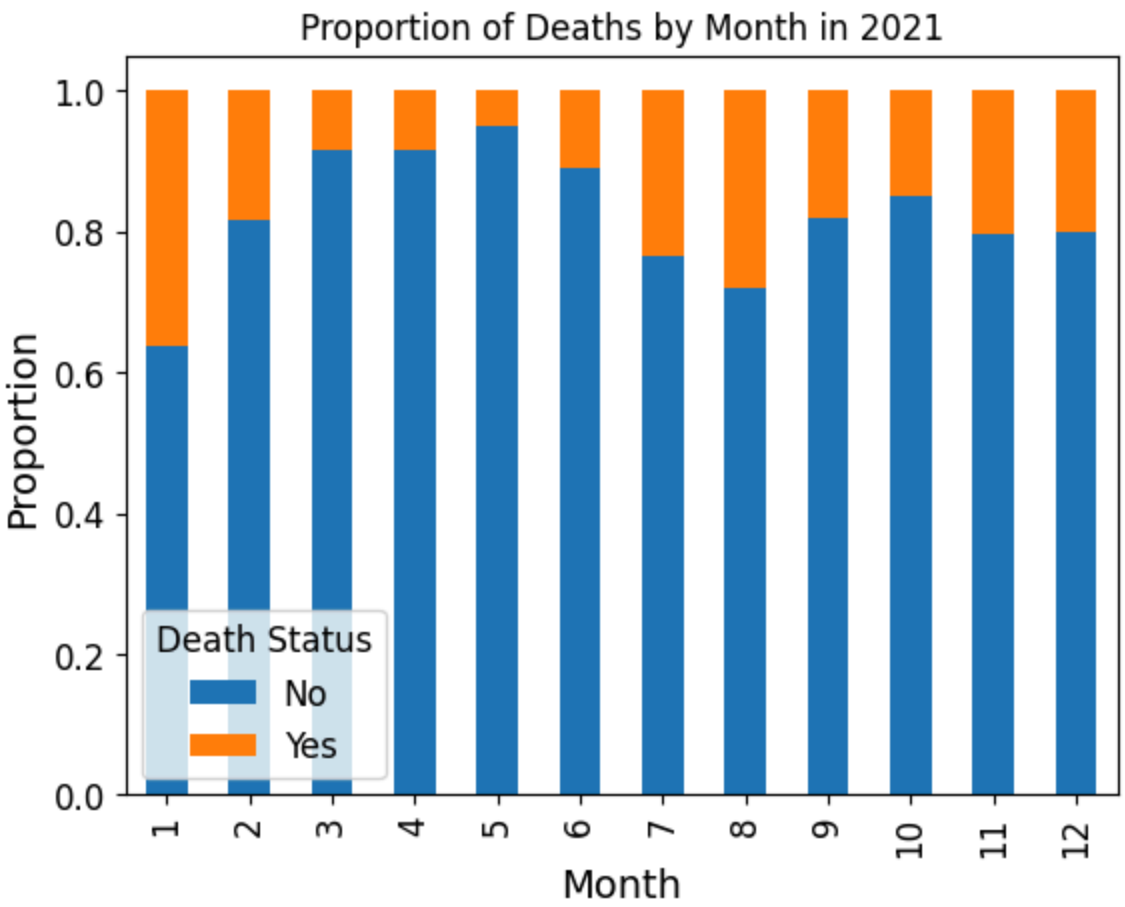

Proportion of Deaths by ethnicity
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Death Status
No
Yes
Hispanic/Latino
Non-Hispanic/Latino
ethnicity

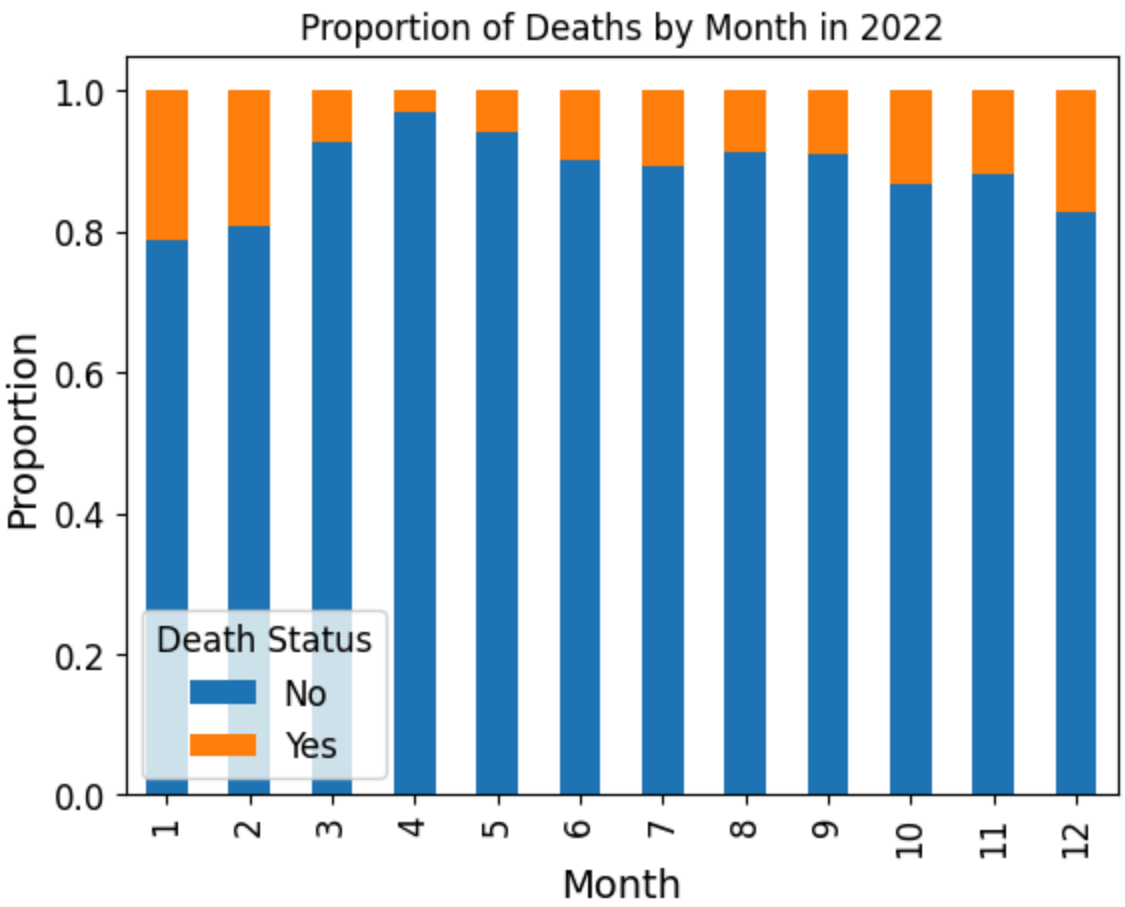

Proportion of Deaths by current_status
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Death Status
No
Yes
Laboratory-confirmed case
Probable Case
current_status

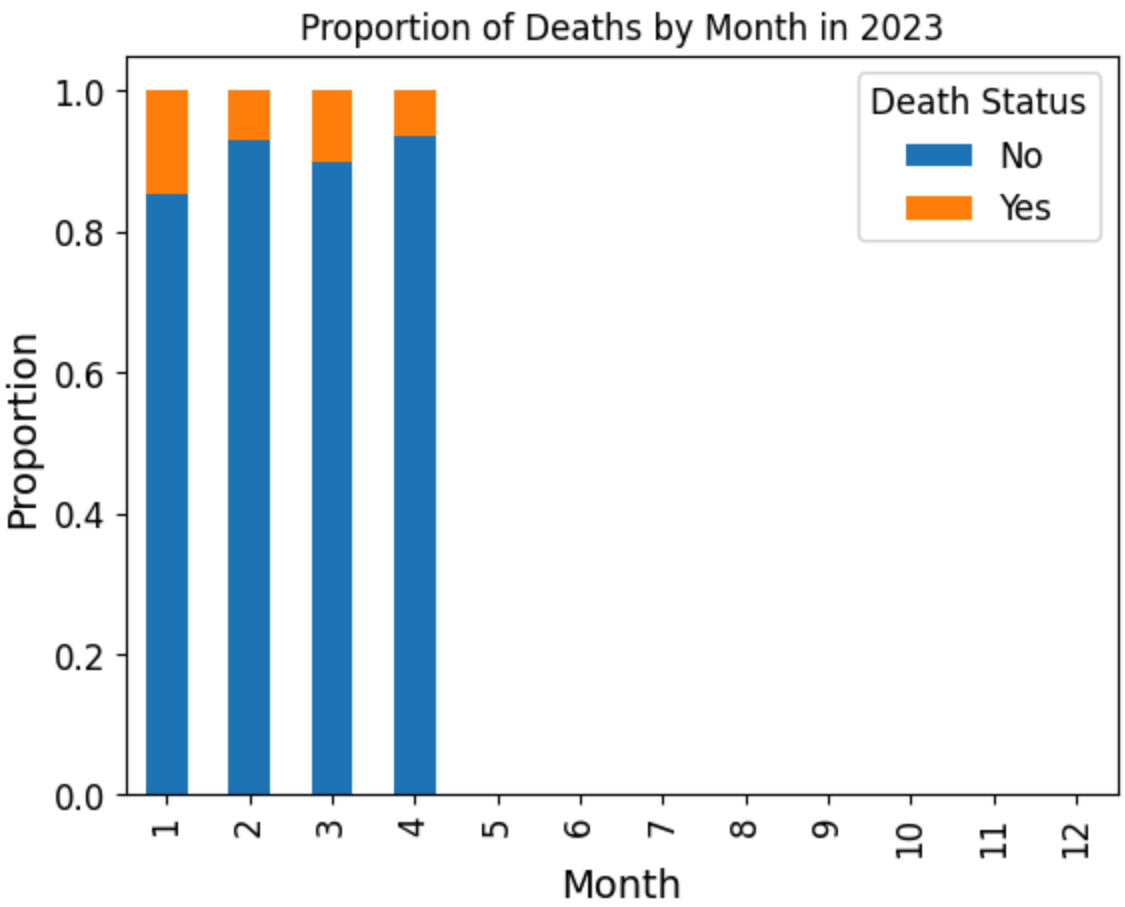

Proportion of Deaths by year
Death Status
No
Yes
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
2020
2021
2022
2023
year

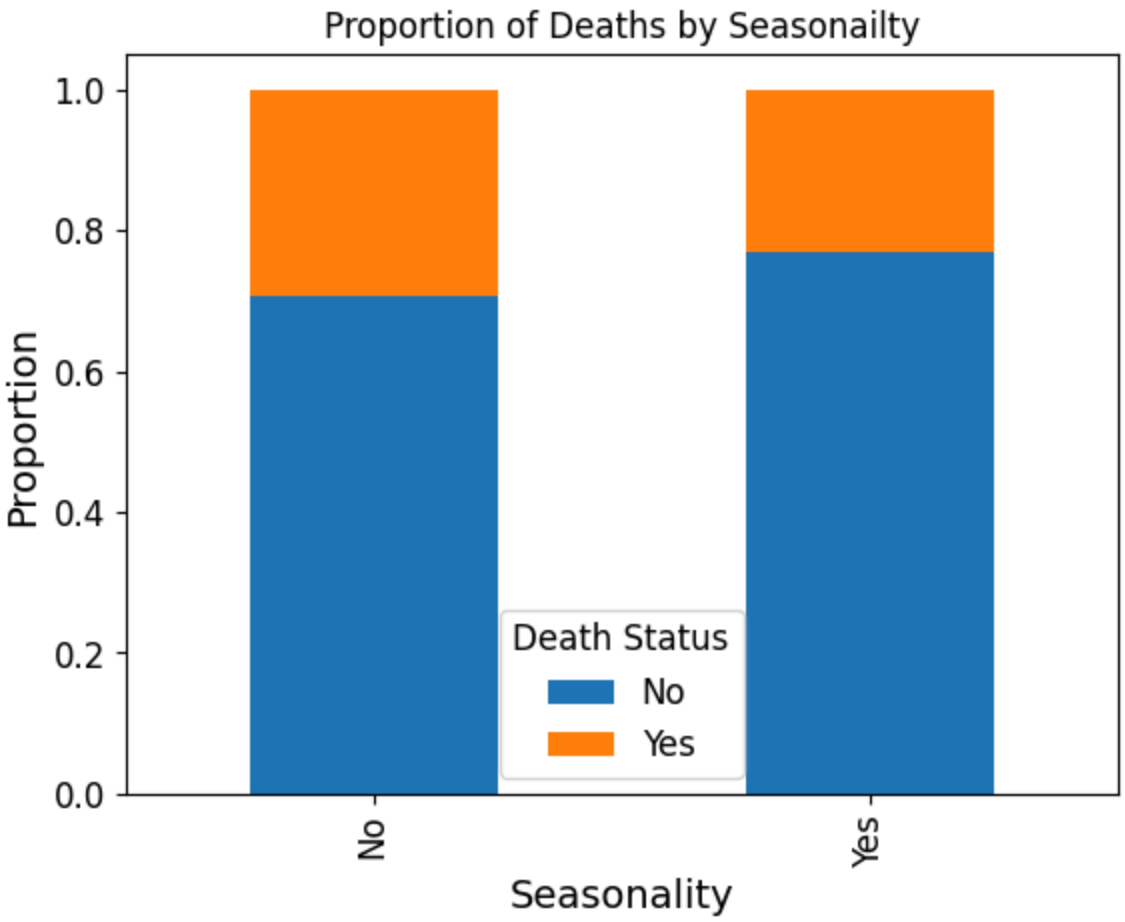

Proportion of Deaths by month
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Death Status
No
Yes
1
2
3
4
5
6
7
8
9
10
11
12
month

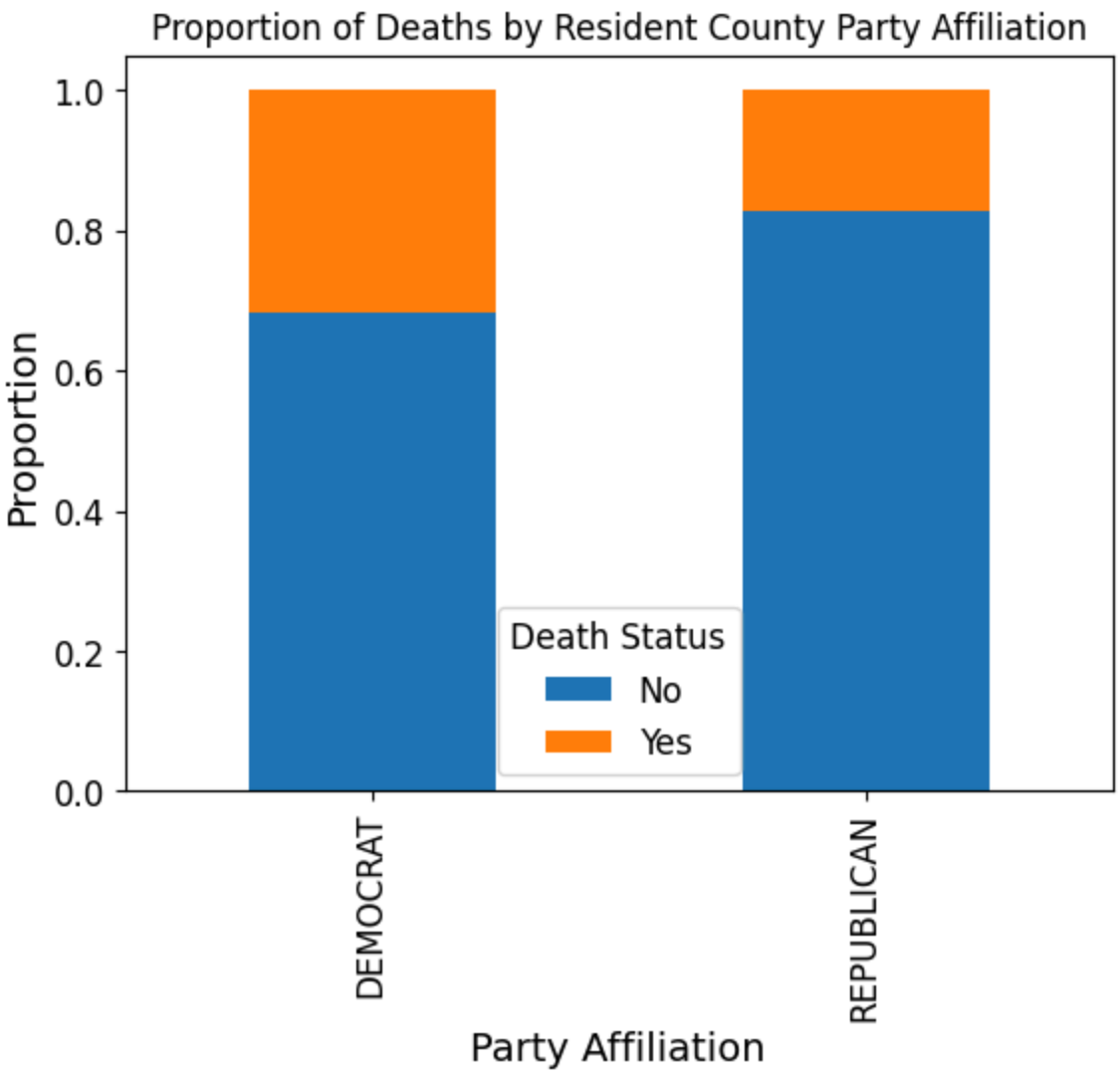

Proportion of Deaths by Month in 2020
Death Status
No
Yes
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
11
12
Month

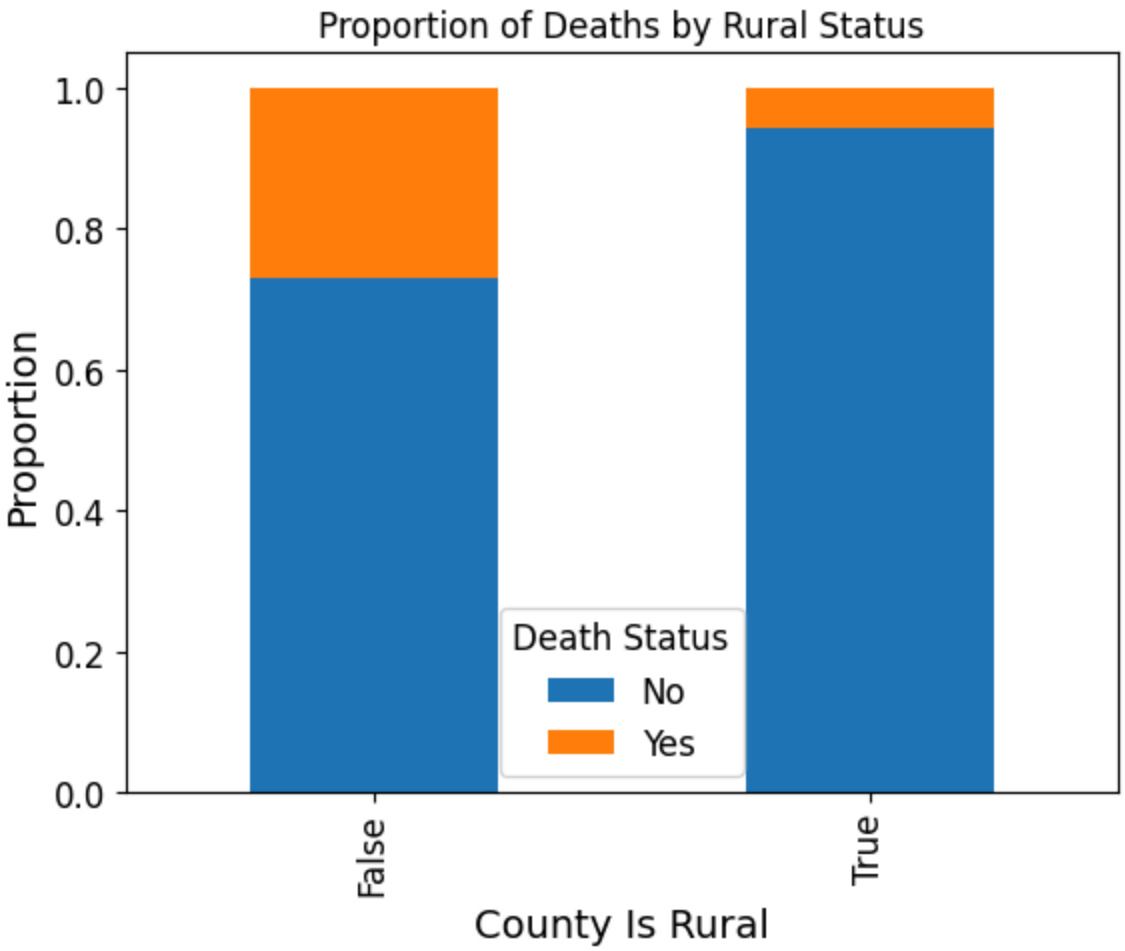

Proportion of Deaths by Month in 2021
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Death Status
No
Yes
1
2
3
4
5
6
7
8
9
10
11
12
Month


Proportion of Deaths by Month in 2022
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Death Status
No
Yes
1
2
3
4
5
6
7
8
9
10
11
12
Month


Proportion of Deaths by Month in 2023
Death Status
No
Yes
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
11
12
Month


Proportion of Deaths by Seasonailty
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
Death Status
No
Yes
No
Yes
Seasonality


Proportion of Deaths by Resident County Party Affiliation
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
Death Status
No
Yes
DEMOCRAT
REPUBLICAN
Party Affiliation


Proportion of Deaths by Rural Status
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
Death Status
No
Yes
False
True
County Is Rural