Data Quality Report

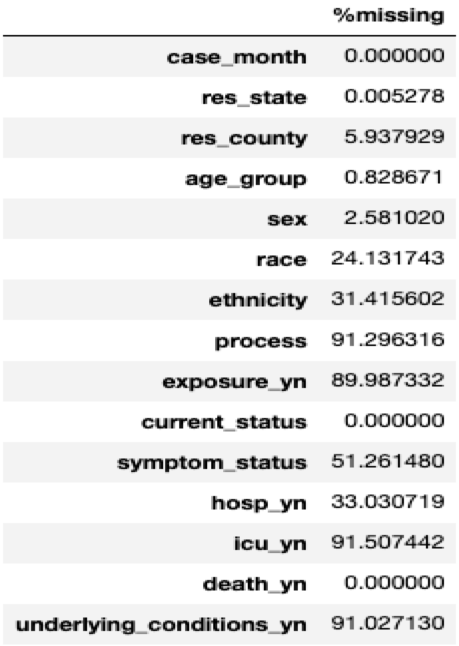
Overview

This report aims to give an initial discussion of my findings of the cleaned dataset (cleaned_covid.csv). It will also summarise the data, describe the various data quality issues associated with the file and how they will be addressed.
Please see the 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.

Summary

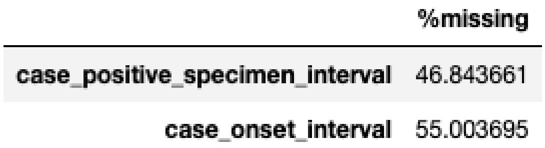
Null Values

This dataset has quite a few null values under each column along with a lot of 'Missing' or 'Unknown' values in a string format. These were not recognised as unknown values, therefore, I converted these values to NaN values so that the amount of null values was accurate.



	%missing
case_month	0.000000
res_state	0.005278
res_county	5.937929
age_group	0.828671
sex	2.581020
race	24.131743
ethnicity	31.415602
process	91.296316
exposure_yn	89.987332
current_status	0.000000
symptom_status	51.261480
hosp_yn	33.030719
icu_yn	91.507442
death_yn	0.000000
underlying_conditions_yn	91.027130

Process, exposure_yn, icu_yn & underlying_conditions all have 90+% missing values in this dataset. This is obviously a huge proportion of the data. Process was how the case was first identified. Exposure_yn was if this person was exposed to covid so mostly unknown values are expected.



	%missing
case_positive_specimen_interval	46.843661
case_onset_interval	55.003695

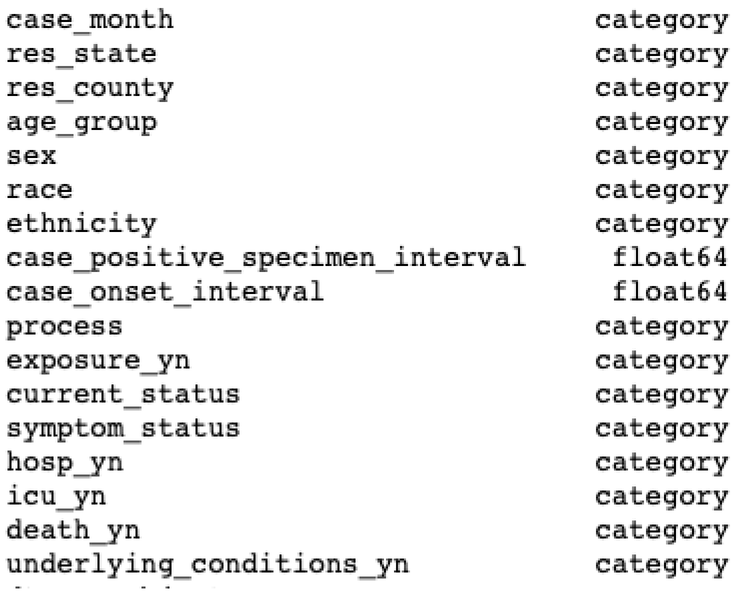
Dropped Values

There are duplicate rows which I have decided to drop as we don't want to count these duplicate rows in our dataset (count 1054) . There are no duplicate columns in the dataset, however I decided to drop two columns, the reason I decided to drop these columns was because they were effectively duplicates of the state and county columns which were already asked in the dataset (state_fips_code and county_fips_code). I also decided to change the negative values to the absolute value as I believe these negatives were inserted in error as these values would not be possible to be negative.

Types

I decided to convert all the object columns to 'category' because converting categorical variables from object to category type can help reduce memory usage and improve the performance of data manipulation and modelling tasks.

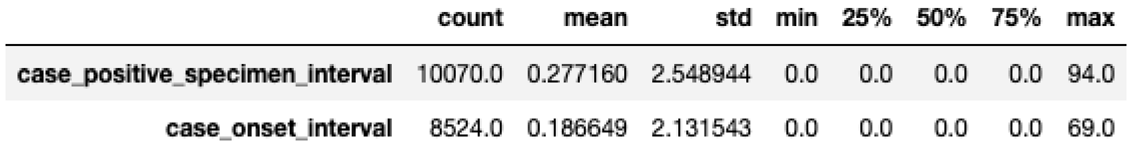
I also decided to convert all numerical columns to floats as converting to floats in data is necessary when working with decimal values or datasets that contain values with a high degree of precision. It can also be useful for data normalisation or standardisation tasks in data analysis. There were also a number of outliers in the case_positive_specimen_interval and case_onset_interval columns in the numerical columns.



```
case_month                              category
res_state                               category
res_county                              category
age_group                               category
sex                                     category
race                                    category
ethnicity                               category
case_positive_specimen_interval          float64
case_onset_interval                      float64
process                                 category
exposure_yn                             category
current_status                          category
symptom_status                          category
hosp_yn                                 category
icu_yn                                  category
death_yn                                category
underlying_conditions_yn                category
```

Continuous Features

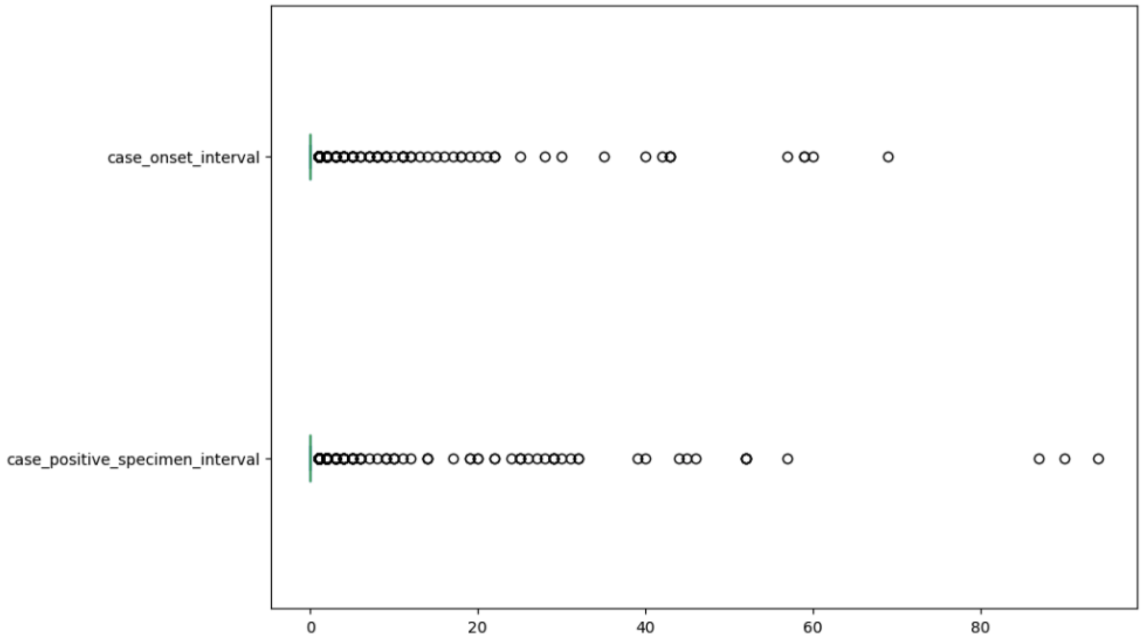
There are two continuous features in the dataset which I decided to proceed with, case_positive_specimen_interval and case_onset_interval.

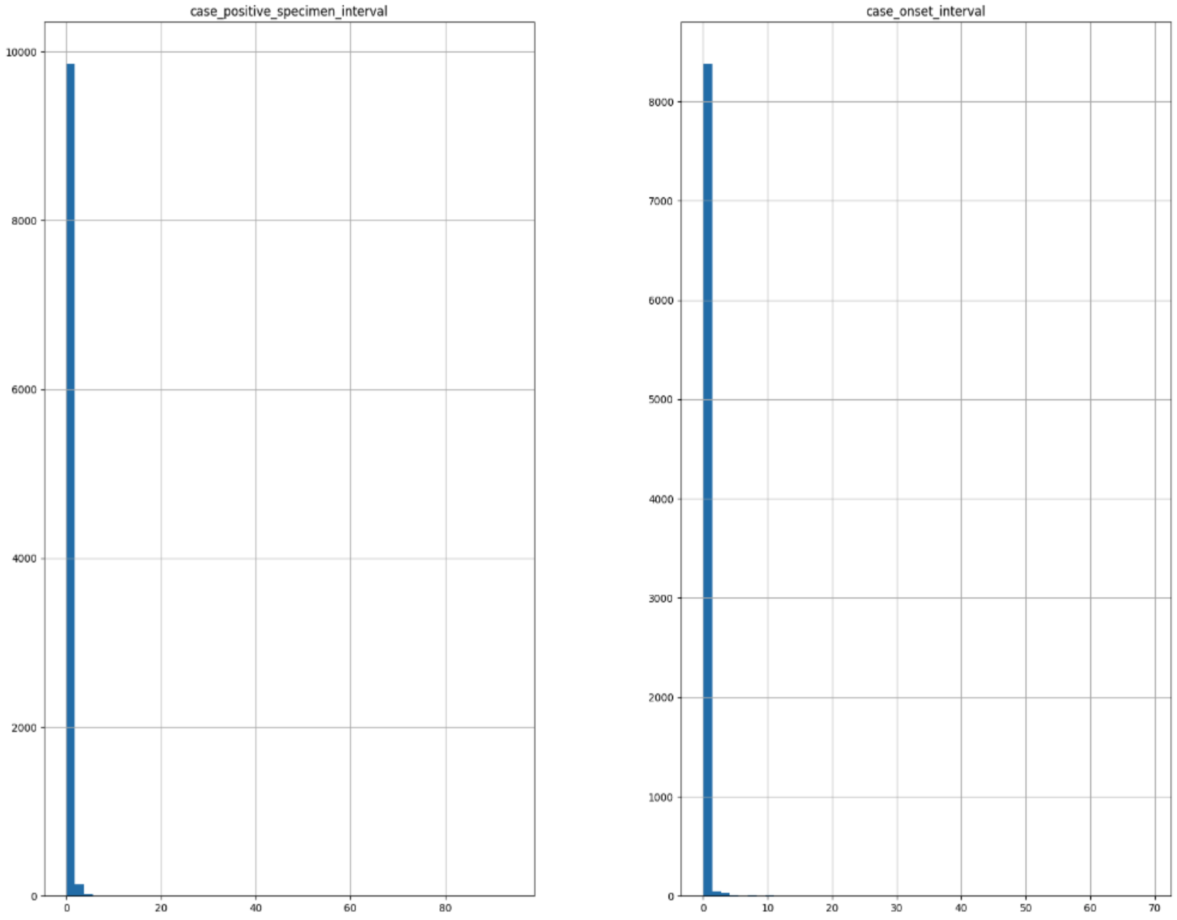


	count	mean	std	min	25%	50%	75%	max
case_positive_specimen_interval	10070.0	0.277160	2.548944	0.0	0.0	0.0	0.0	94.0
case_onset_interval	8524.0	0.186649	2.131543	0.0	0.0	0.0	0.0	69.0

Observations:

- The majority of the values in this column are equal to 0. This means that the days between pos_spec_dt and cdc_case_earliest_dt are equal to zero the majority of the time. There are huge outliers in case_positive_specimen_interval of values 94 as the std is only 2.5489 and the mean is 0.277. It can be seen that the majority of the values in this were 0 as the interquartile numbers are all equal to 0.
- The majority of the values in this column are equal to 0. This means that the cdc_case_earliest_dt and onset_dt are the same and so the interval is 0 and have the same value as cases with 0 week intervals. There are also huge outliers in case_onset_interval of value a 69 as the std is only 2.13 and the mean is 0.18. It can also be seen that the majority of the values in this were 0 as the interquartile numbers are all equal to 0.

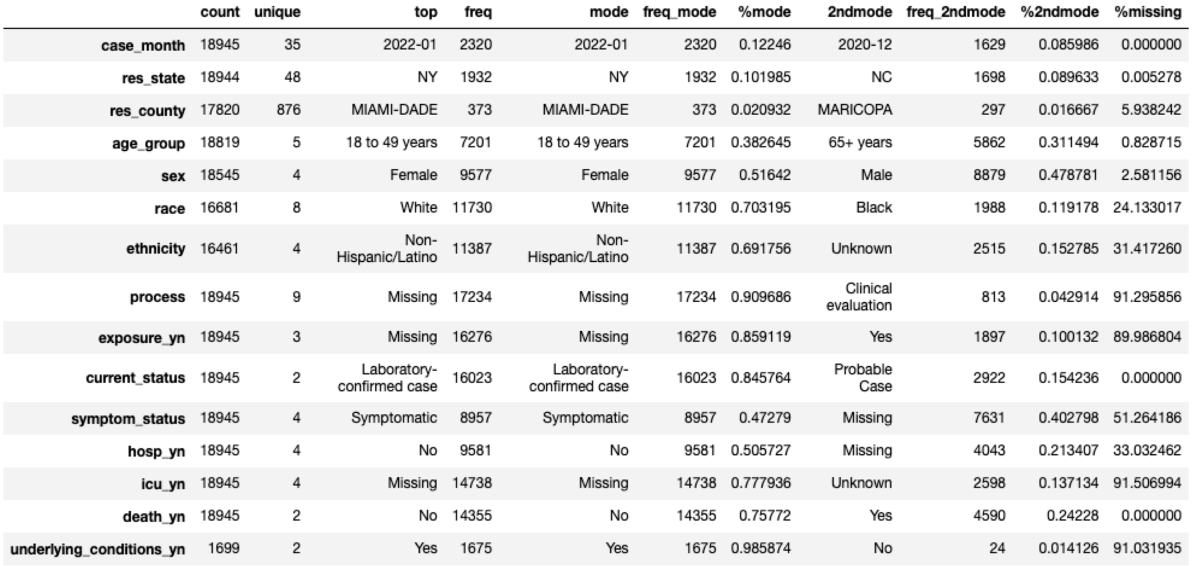

case_onset_interval
case_positive_specimen_interval
0
20
40
60
80


case_positive_specimen_interval
10000
8000
6000
4000
2000
0
0
20
40
60
80
case_onset_interval
8000
7000
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
60
70

Review Logical Integrity

- Test 1: Validity check for case_positive_specimen_interval: False (9512 times False)
- Test 2: Validity check for case_onset_interval: False (11197 times False)
- Test 3: Completeness check for county_fips_code: False
- Test 4: Completeness check for state_fips_code: False
- Test 5: Uniqueness check for county_fips_code and state_fips_code: False

Review Categorical Features



	count	unique	top	freq	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode	%missing
case_month	18945	35	2022-01	2320	2022-01	2320	0.12246	2020-12	1629	0.085986	0.000000
res_state	18944	48	NY	1932	NY	1932	0.101985	NC	1698	0.089633	0.005278
res_county	17820	876	MIAMI-DADE	373	MIAMI-DADE	373	0.020932	MARICOPA	297	0.016667	5.938242
age_group	18819	5	18 to 49 years	7201	18 to 49 years	7201	0.382645	65+ years	5862	0.311494	0.828715
sex	18545	4	Female	9577	Female	9577	0.51642	Male	8879	0.478781	2.581156
race	16681	8	White	11730	White	11730	0.703195	Black	1988	0.119178	24.133017
ethnicity	16461	4	Non-Hispanic/Latino	11387	Non-Hispanic/Latino	11387	0.691756	Unknown	2515	0.152785	31.417260
process	18945	9	Missing	17234	Missing	17234	0.909686	Clinical evaluation	813	0.042914	91.295856
exposure_yn	18945	3	Missing	16276	Missing	16276	0.859119	Yes	1897	0.100132	89.986804
current_status	18945	2	Laboratory-confirmed case	16023	Laboratory-confirmed case	16023	0.845764	Probable Case	2922	0.154236	0.000000
symptom_status	18945	4	Symptomatic	8957	Symptomatic	8957	0.47279	Missing	7631	0.402798	51.264186
hosp_yn	18945	4	No	9581	No	9581	0.505727	Missing	4043	0.213407	33.032462
icu_yn	18945	4	Missing	14738	Missing	14738	0.777936	Unknown	2598	0.137134	91.506994
death_yn	18945	2	No	14355	No	14355	0.75772	Yes	4590	0.24228	0.000000
underlying_conditions_yn	1699	2	Yes	1675	Yes	1675	0.985874	No	24	0.014126	91.031935

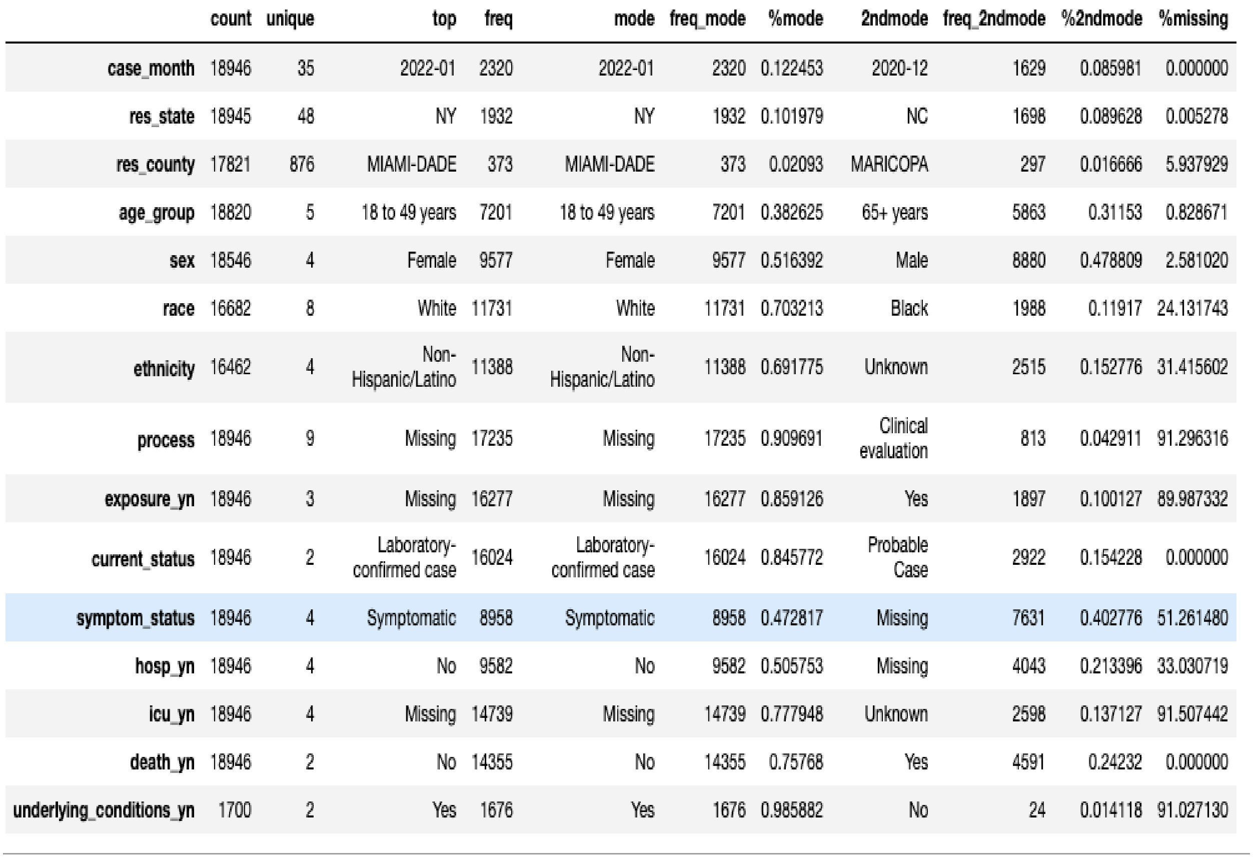
Descriptive Statistics

- There are 14 Categorical features in this dataset. These categorical features were all objects which I converted into categorical.
- The number of unique values is ok for each feature with the only large amount being seen in the county which is expected.
- I decided to drop the rows of 3 columns; process, exposure_yn, & icu_yn as they all had 89.9% and above missing values.
- I decided to keep underlying_conditions_yn as I thought it would be a particularly interesting column.

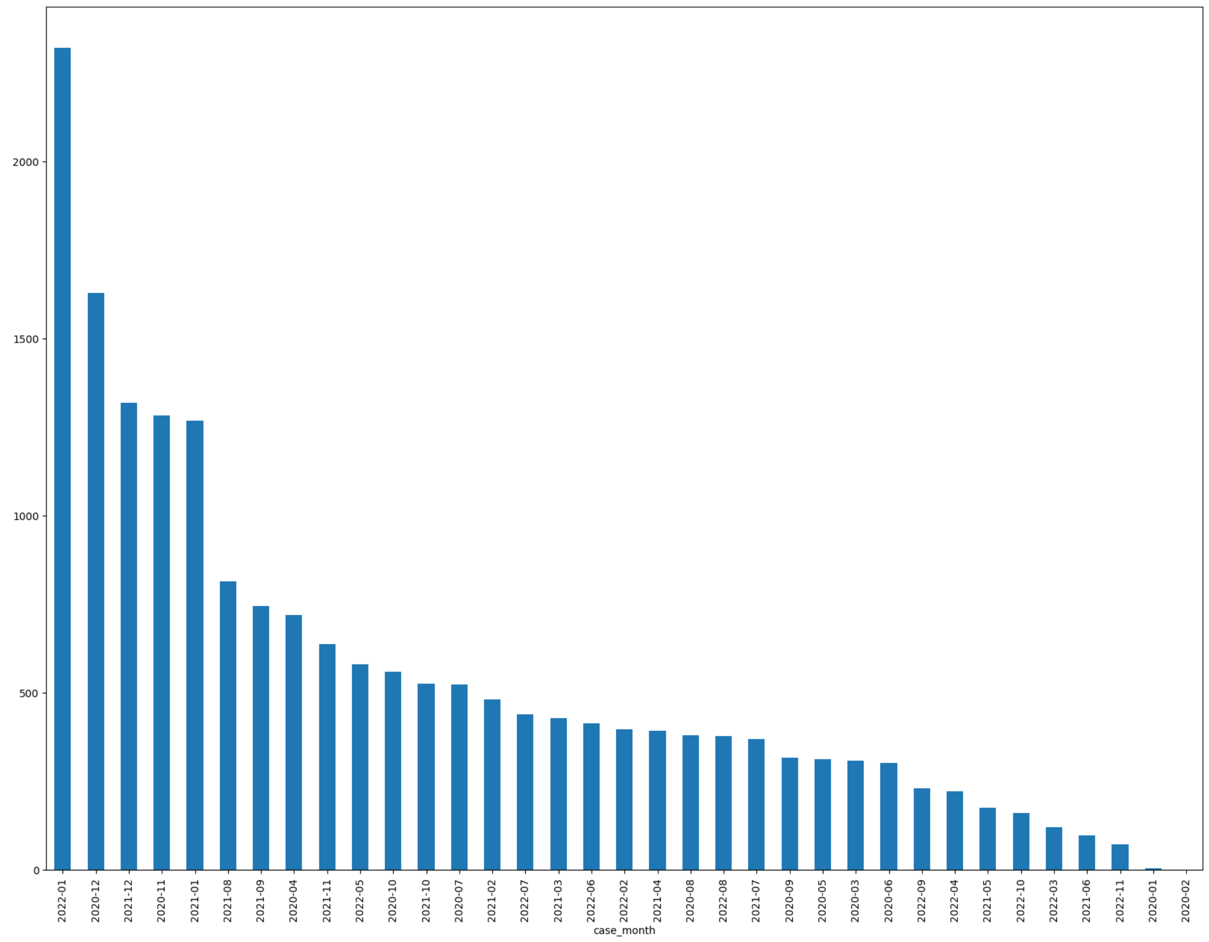
Bar Plots

The bar plots can be found in the appendix below.

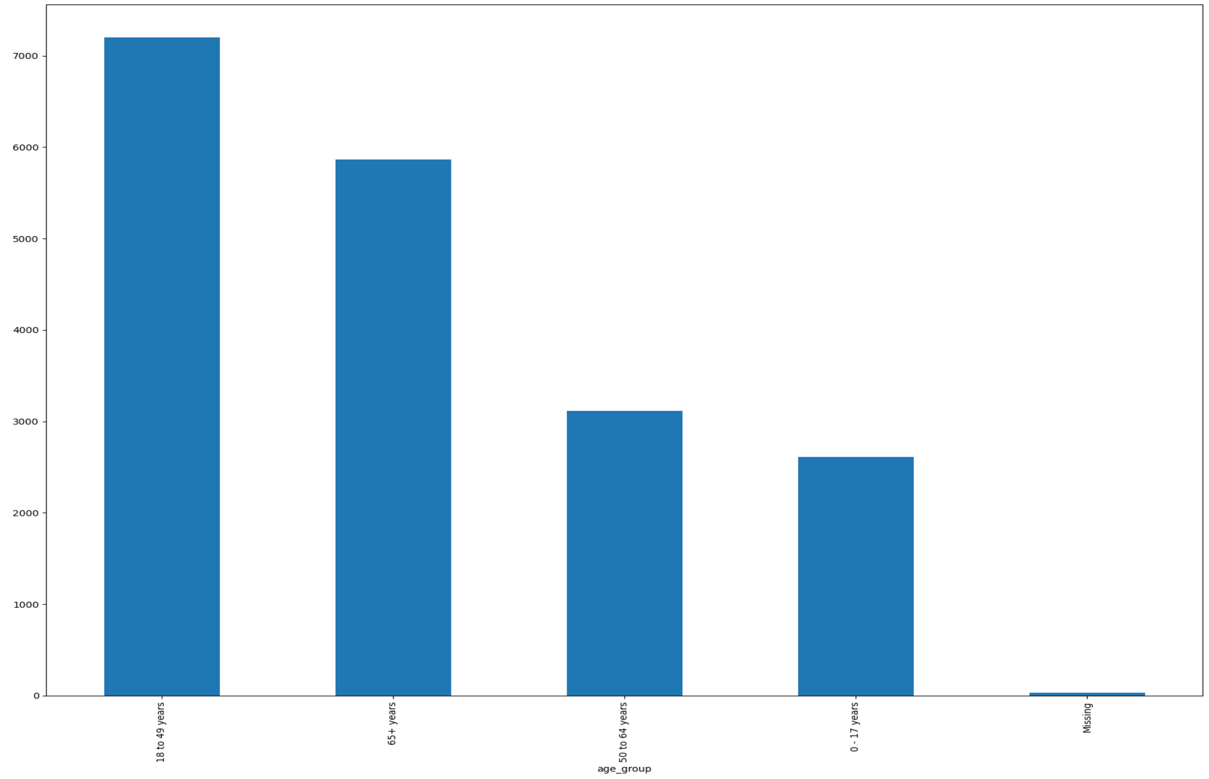
Appendix A



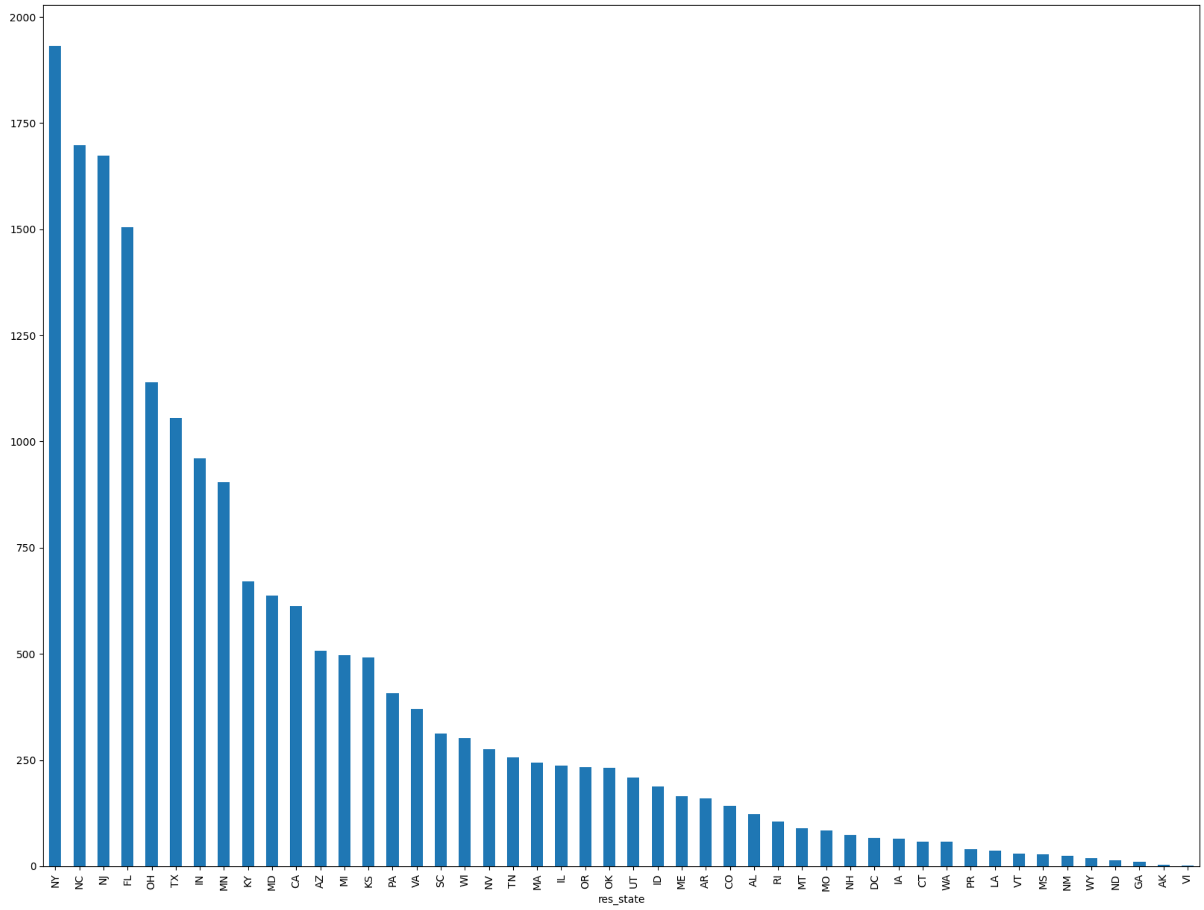
	count	unique	top	freq	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode	%missing
case_month	18946	35	2022-01	2320	2022-01	2320	0.122453	2020-12	1629	0.085981	0.000000
res_state	18945	48	NY	1932	NY	1932	0.101979	NC	1698	0.089628	0.005278
res_county	17821	876	MIAMI-DADE	373	MIAMI-DADE	373	0.02093	MARICOPA	297	0.016666	5.937929
age_group	18820	5	18 to 49 years	7201	18 to 49 years	7201	0.382625	65+ years	5863	0.31153	0.828671
sex	18546	4	Female	9577	Female	9577	0.516392	Male	8880	0.478809	2.581020
race	16682	8	White	11731	White	11731	0.703213	Black	1988	0.11917	24.131743
ethnicity	16462	4	Non-Hispanic/Latino	11388	Non-Hispanic/Latino	11388	0.691775	Unknown	2515	0.152776	31.415602
process	18946	9	Missing	17235	Missing	17235	0.909691	Clinical evaluation	813	0.042911	91.296316
exposure_yn	18946	3	Missing	16277	Missing	16277	0.859126	Yes	1897	0.100127	89.987332
current_status	18946	2	Laboratory-confirmed case	16024	Laboratory-confirmed case	16024	0.845772	Probable Case	2922	0.154228	0.000000
symptom_status	18946	4	Symptomatic	8958	Symptomatic	8958	0.472817	Missing	7631	0.402776	51.261480
hosp_yn	18946	4	No	9582	No	9582	0.505753	Missing	4043	0.213396	33.030719
icu_yn	18946	4	Missing	14739	Missing	14739	0.777948	Unknown	2598	0.137127	91.507442
death_yn	18946	2	No	14355	No	14355	0.75768	Yes	4591	0.24232	0.000000
underlying_conditions_yn	1700	2	Yes	1676	Yes	1676	0.985882	No	24	0.014118	91.027130



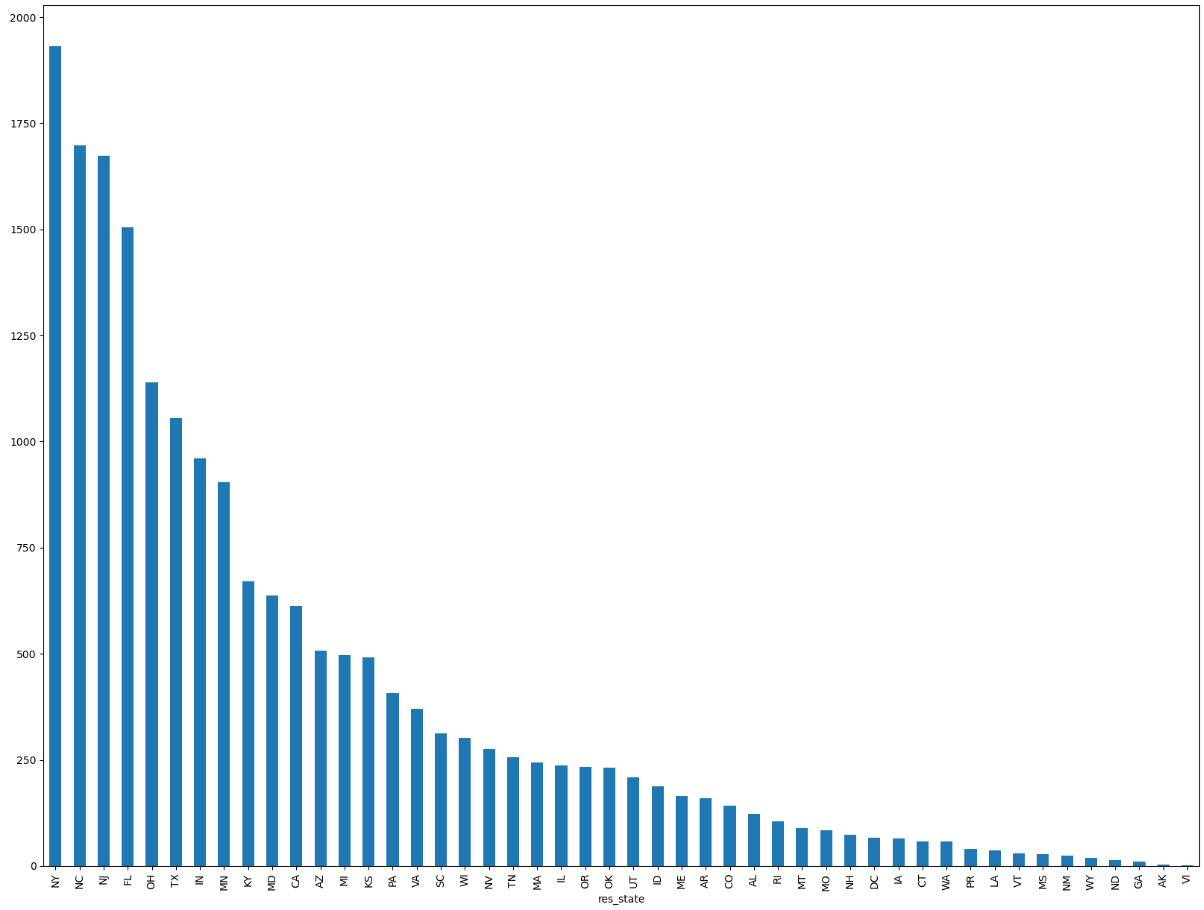
2000
1500
1000
500
0
2022-01
2020-12
2021-12
2020-11
2021-01
2021-08
2021-09
2020-04
2021-11
2022-05
2020-10
2021-10
2020-07
2021-02
2022-07
2021-03
2022-06
2022-02
2021-04
2020-08
2022-08
2021-07
2020-09
2020-05
2020-03
2020-06
2022-09
2022-04
2021-05
2022-10
2022-03
2021-06
2022-11
2020-01
2020-02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



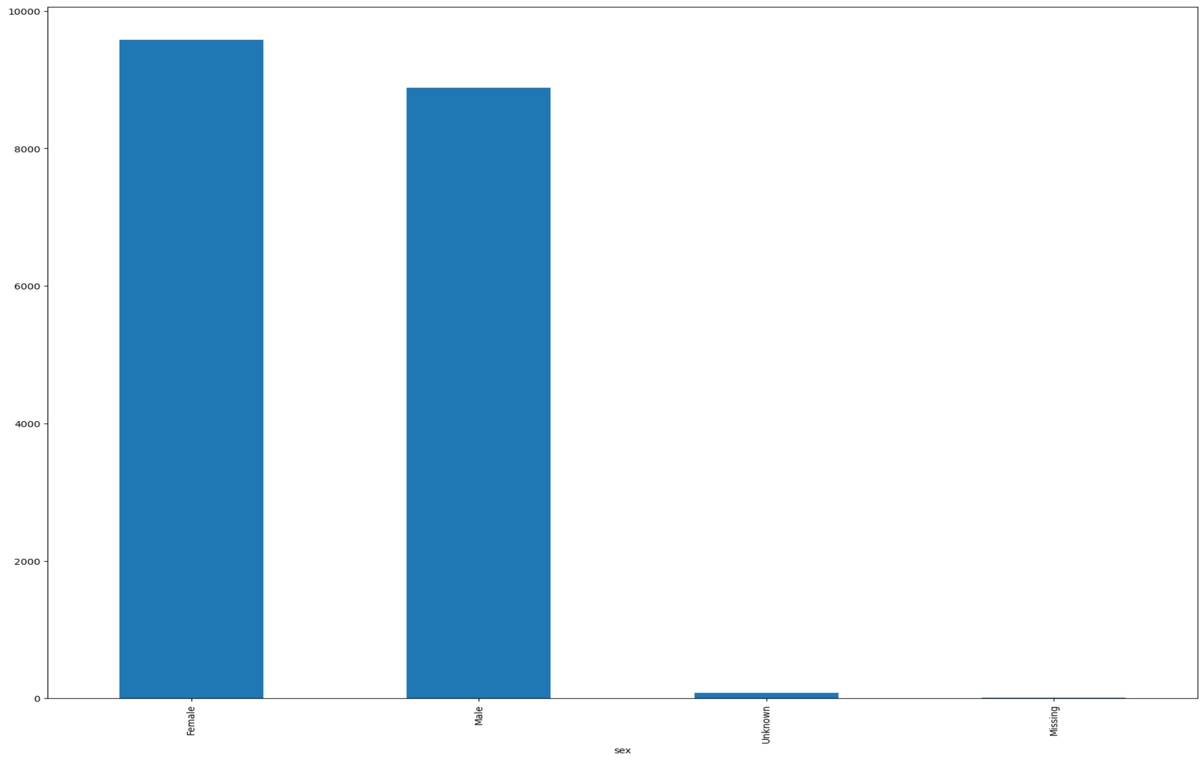
7000
6000
5000
4000
3000
2000
1000
0
18 to 49 years
65+ years
50 to 64 years
0 - 17 years
Missing
age_group

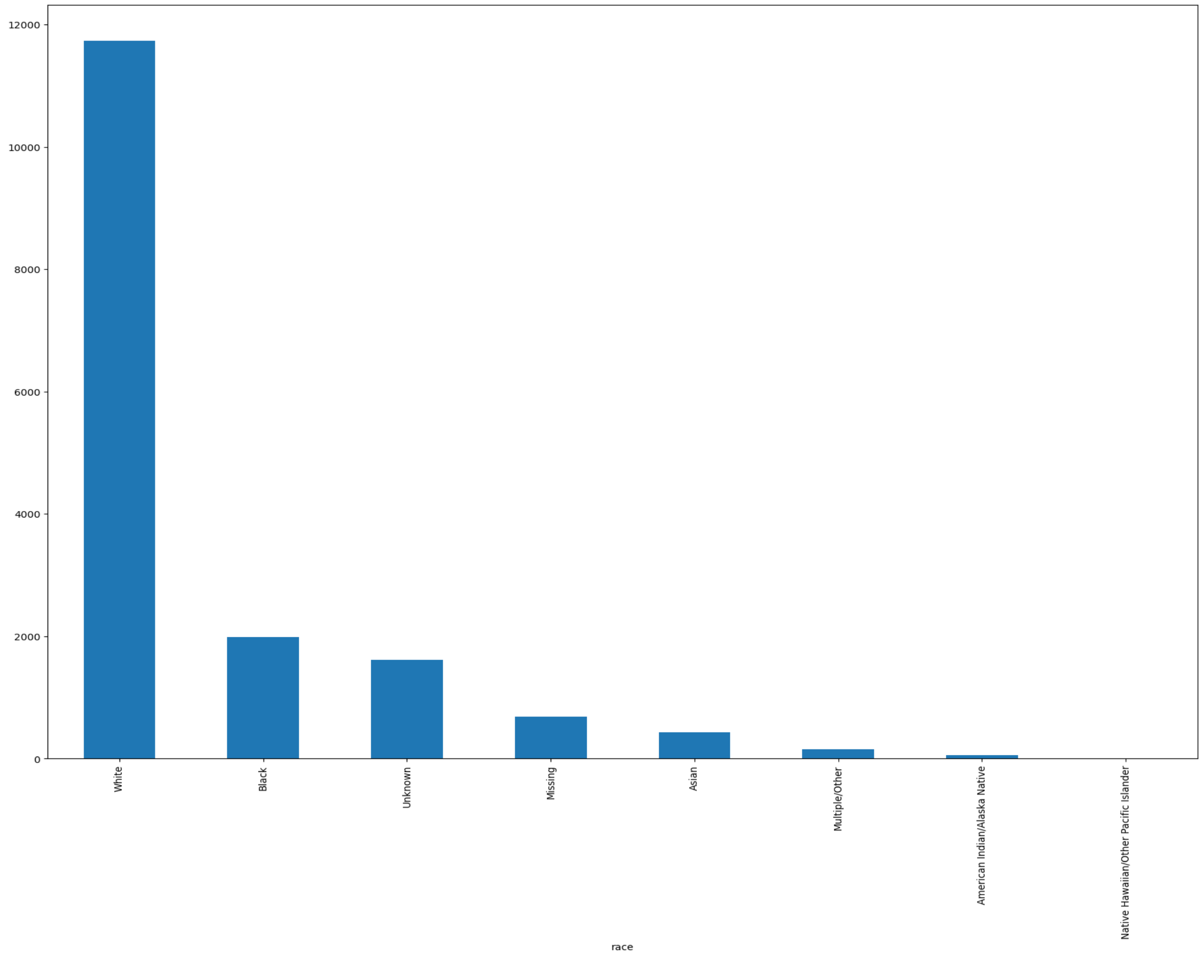


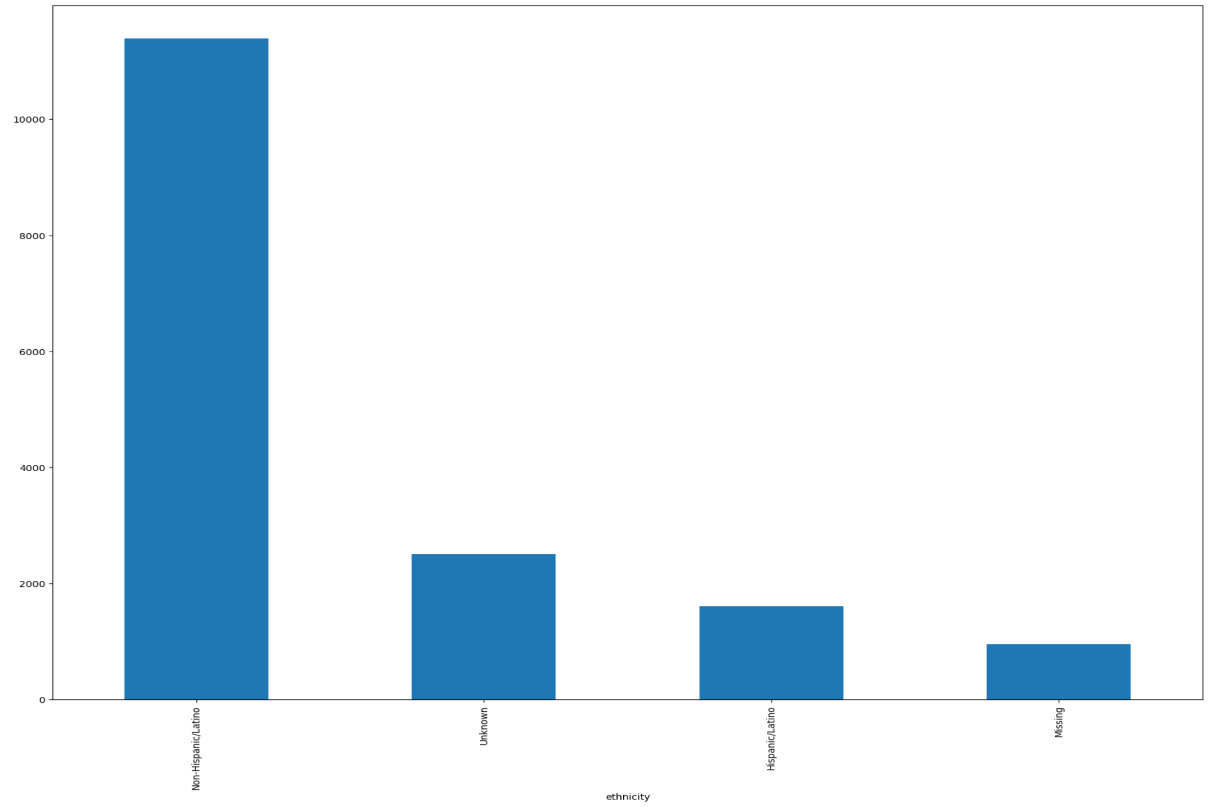
0
250
500
750
1000
1250
1500
1750
2000
NY
NC
NJ
FL
OH
TX
IN
MN
KY
MD
CA
AZ
MI
KS
PA
VA
SC
WI
NV
TN
MA
IL
OR
OK
UT
ID
ME
AR
CO
AL
RI
MT
MO
NH
DC
IA
CT
WA
PR
LA
VT
MS
NM
WY
ND
GA
AK
VI
res_state

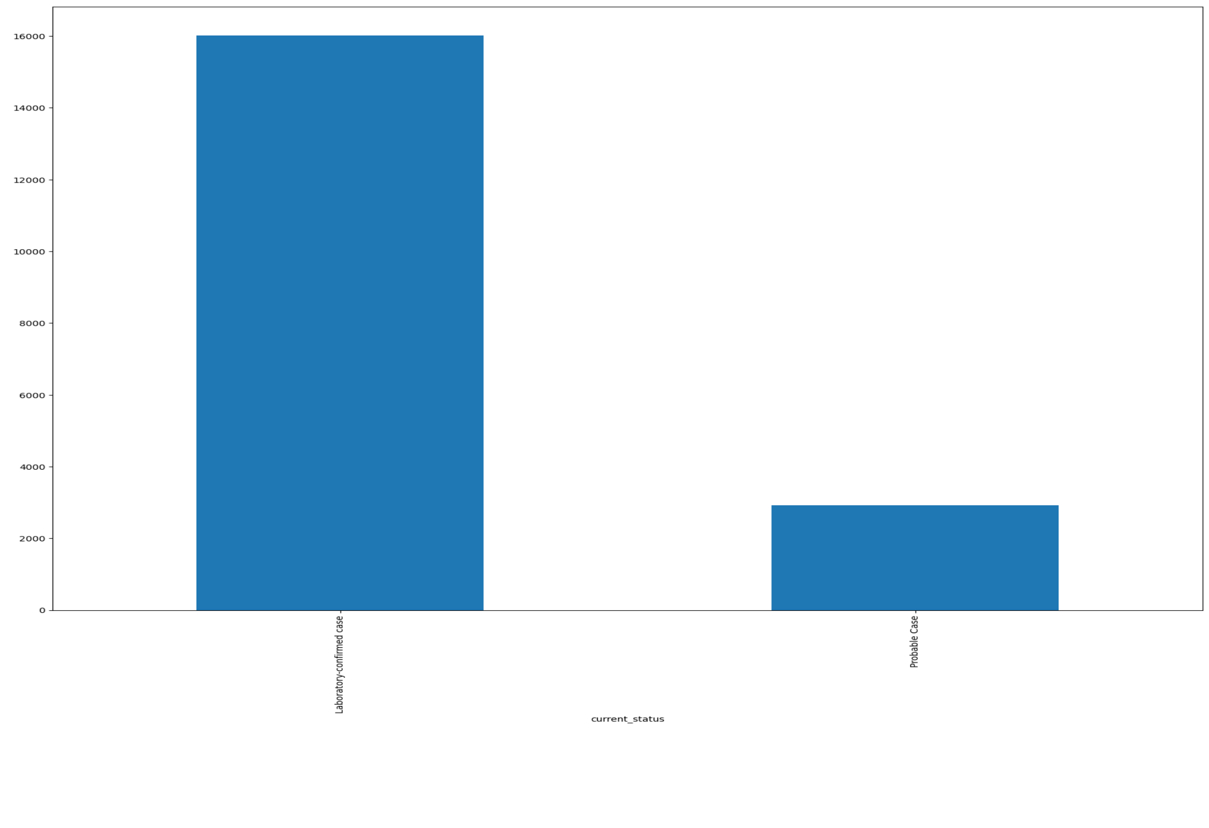


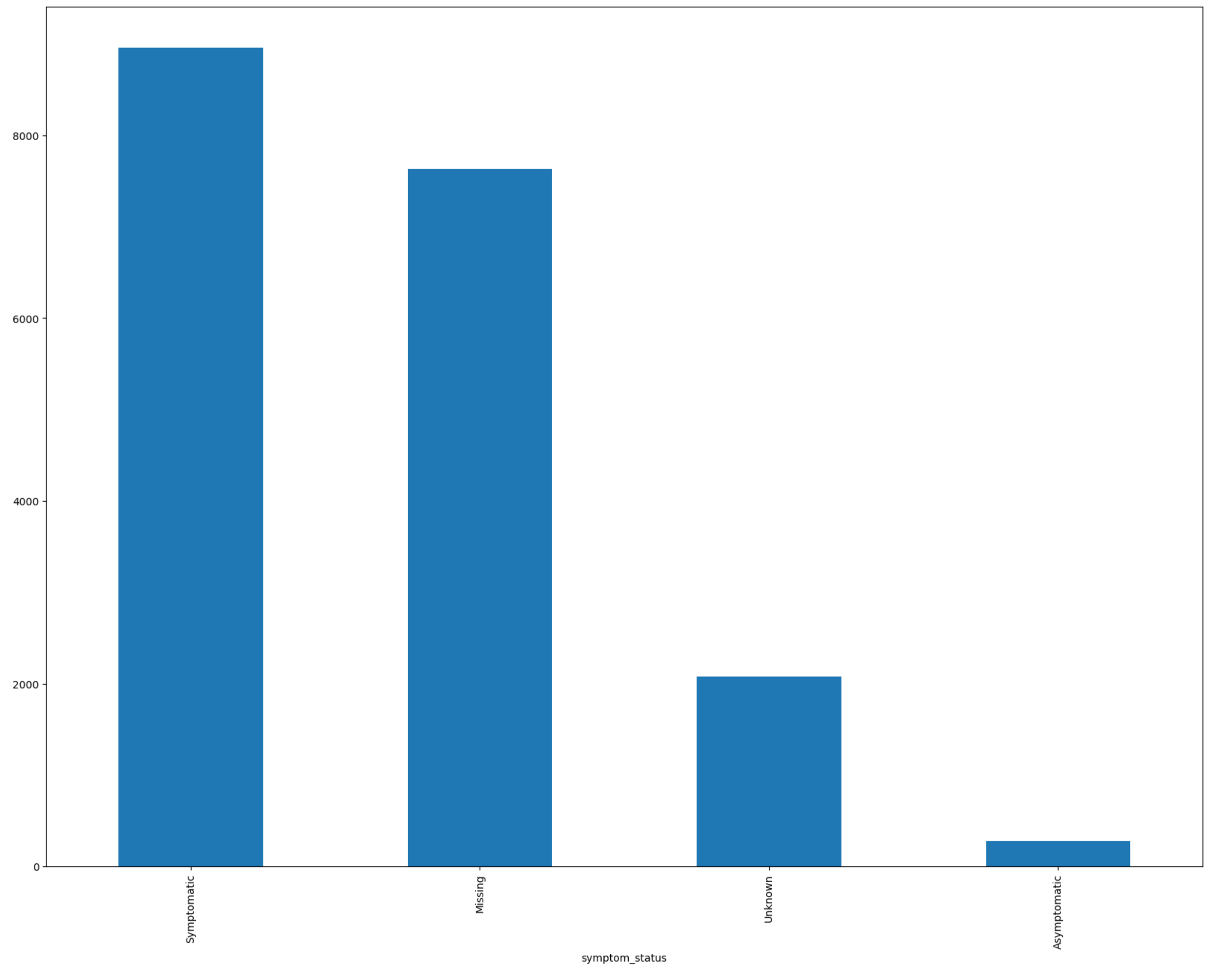
0
250
500
750
1000
1250
1500
1750
2000
NY
NC
NJ
FL
OH
TX
IN
MN
KY
MD
CA
AZ
MI
KS
PA
VA
SC
WI
NV
TN
MA
IL
OR
OK
UT
ID
ME
AR
CO
AL
RI
MT
MO
NH
DC
IA
CT
WA
PR
LA
VT
MS
NM
WY
ND
GA
AK
VI
res_state

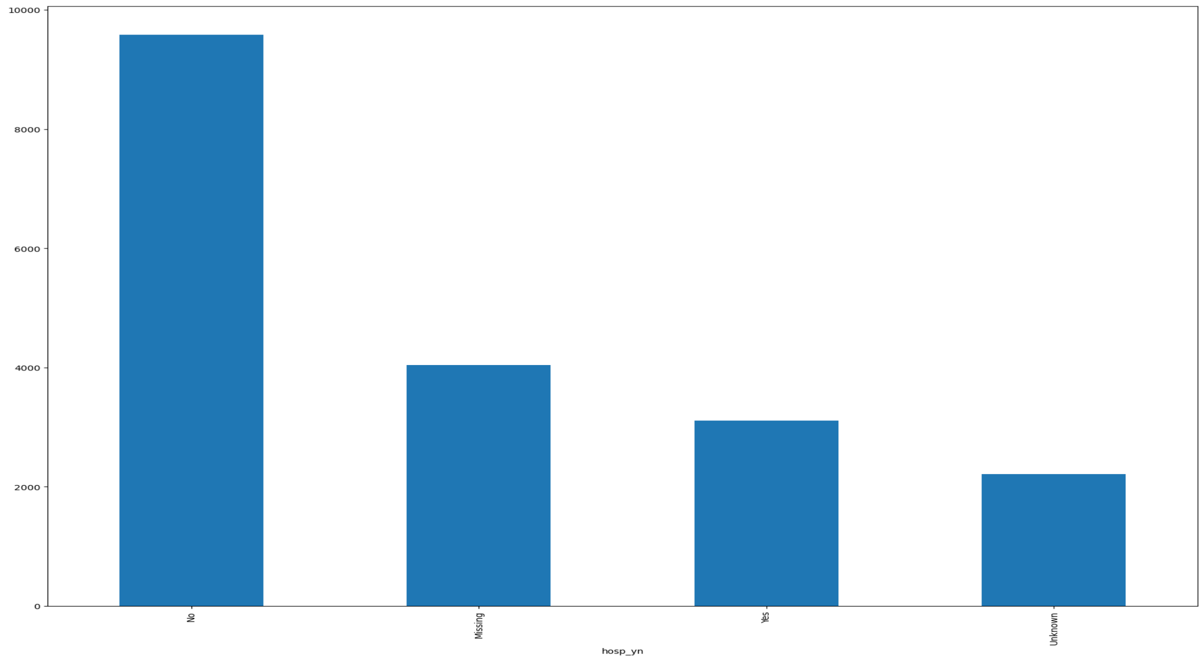

10000
8000
6000
4000
2000
0
Female
Male
Unknown
Missing
sex

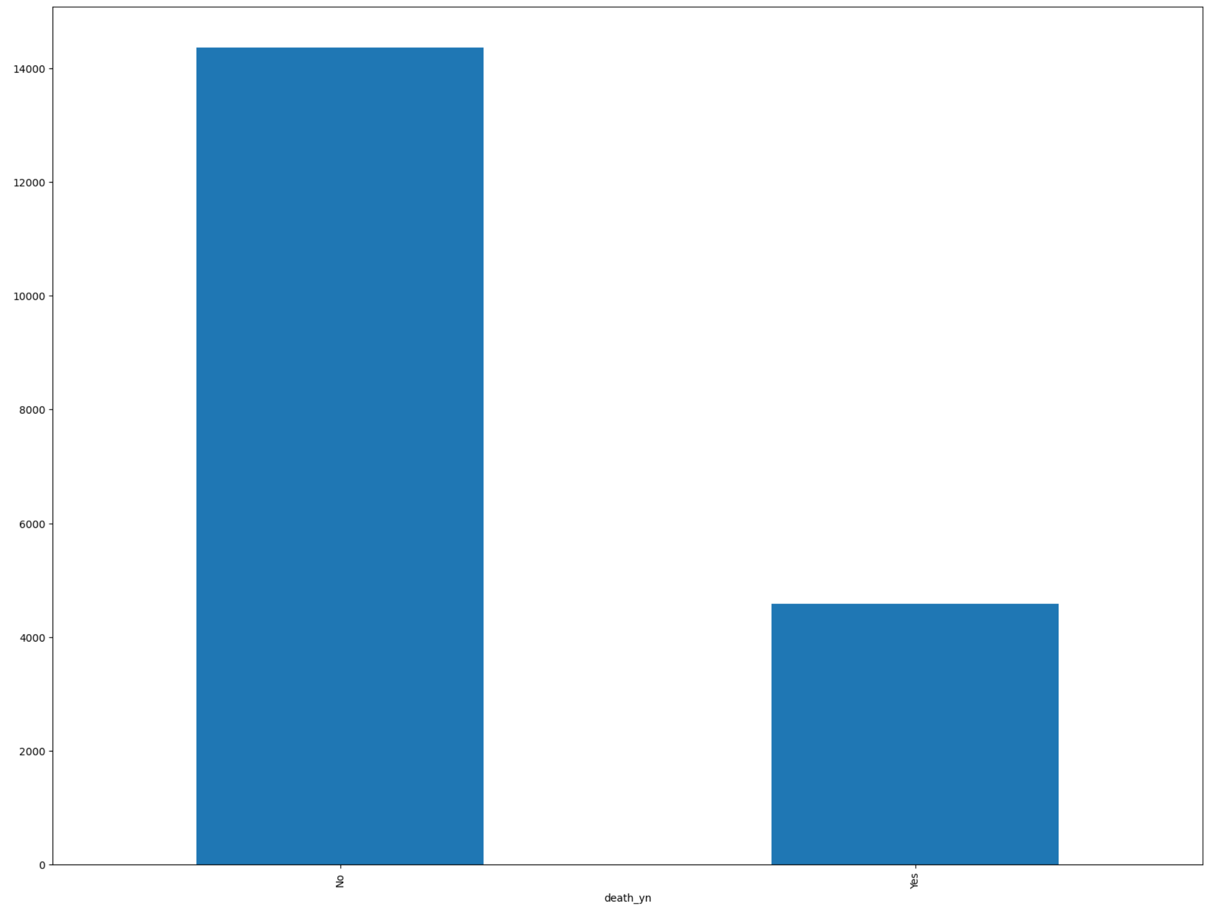

12000
10000
8000
6000
4000
2000
0
White
Black
Unknown
Missing
Asian
Multiple/Other
American Indian/Alaska Native
Native Hawaiian/Other Pacific Islander
race

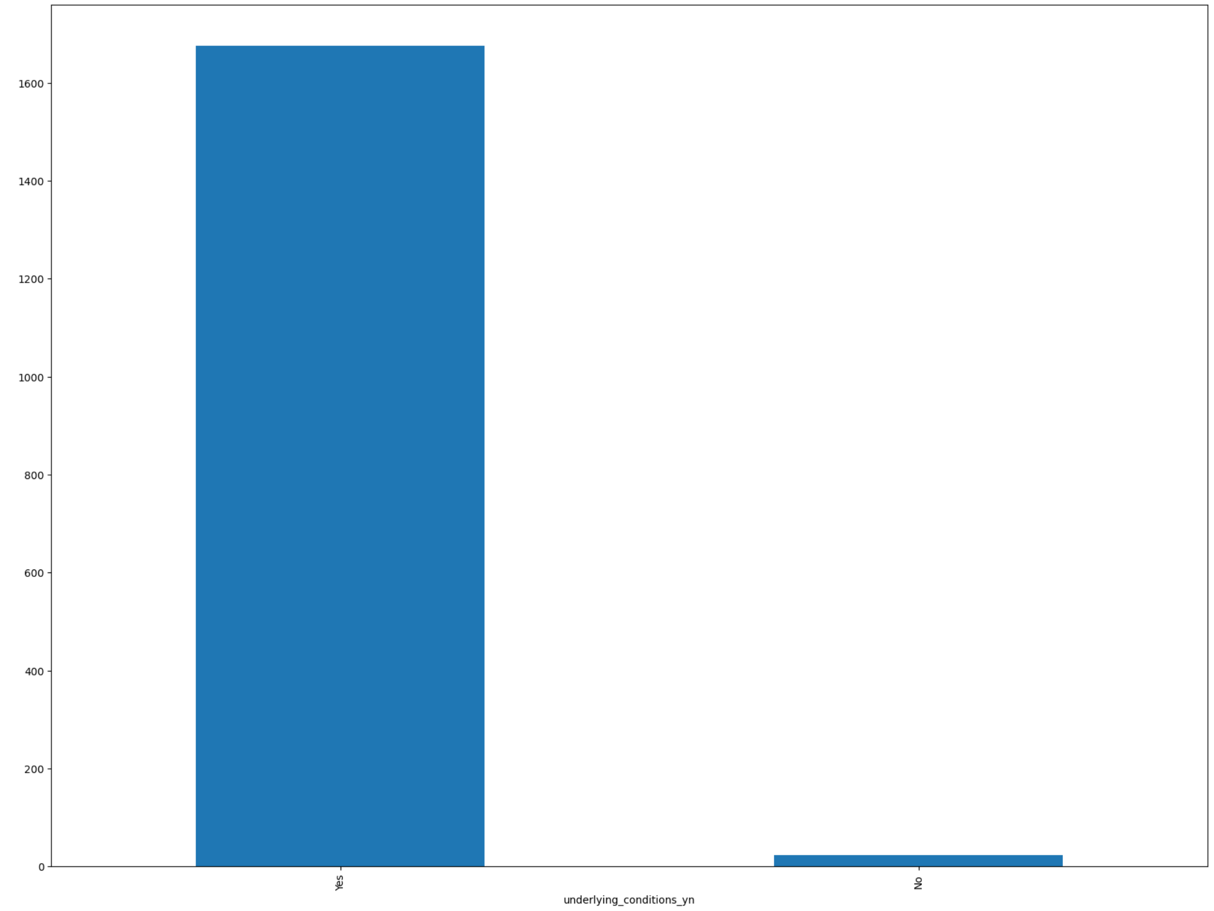

10000
8000
6000
4000
2000
0
Non-Hispanic/Latino
Unknown
Hispanic/Latino
Missing
ethnicity

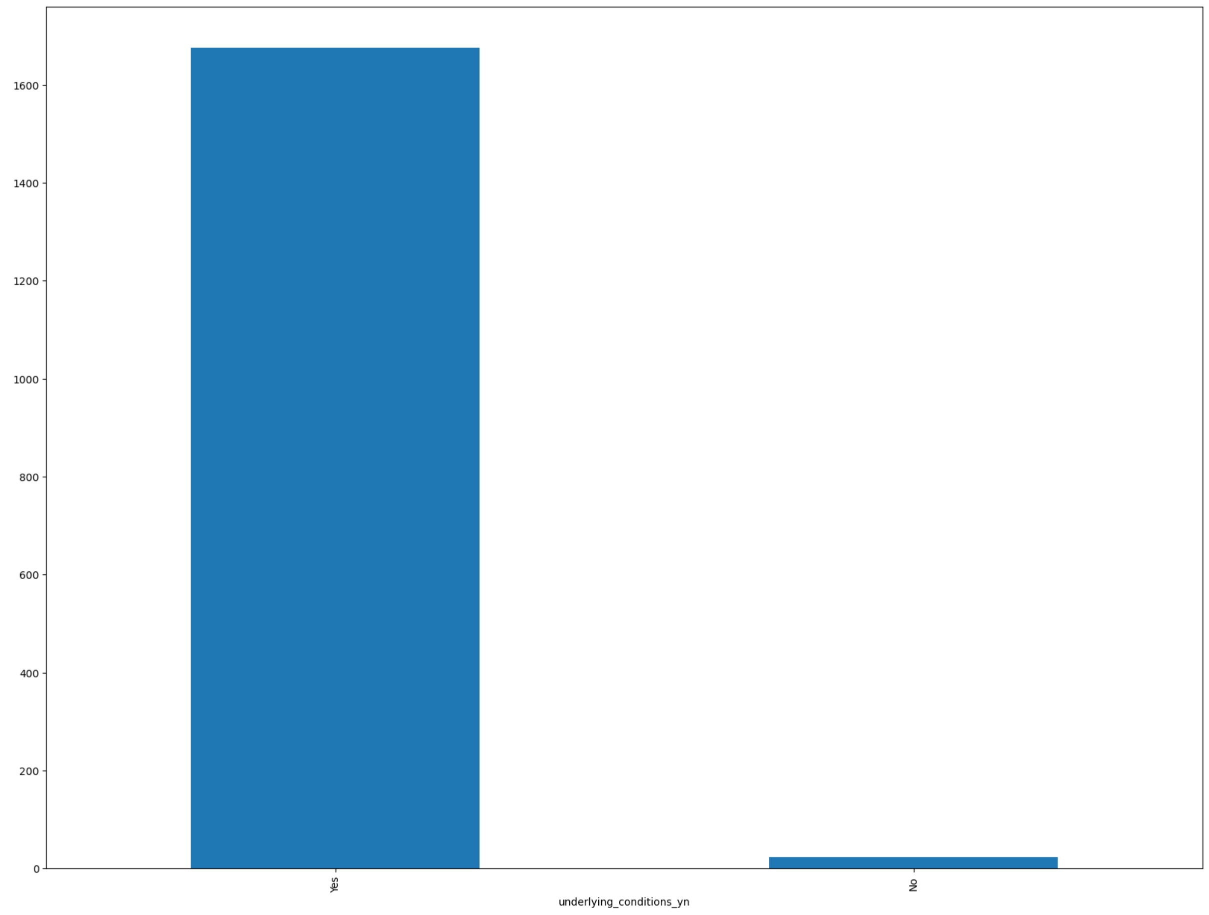

16000
14000
12000
10000
8000
6000
4000
2000
0
Laboratory-confirmed case
Probable Case
current_status


8000
6000
4000
2000
0
Symptomatic
Missing
Unknown
Asymptomatic
symptom_status


10000
8000
6000
4000
2000
0
No
Missing
Yes
Unknown
hosp_yn


14000
12000
10000
8000
6000
4000
2000
0
No
Yes
death_yn


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


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