Data Quality Report

Initial Findings

Brian Manning

17324576

COMP47350: Data Analytics

UCD School of Computer Science

12/03/2020

1. Overview

This report will outline my initial findings based on the cleaned dataset (data/covid19-cdc-17324576_final_df_orig.csv). It will summarise the data, describe the various data quality issues observed and how they will be addressed. Please see the appendix for some background on 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 have relatively few poor quality features – this is because only 3 columns (which are datetime) have null values in them. Upon further inspection it is clear that the other categorical features also contain null values which appear as Missing/Unknown. The main issues observed below were inconsistencies between the date columns identified through integrity testing, deprecated report date column, a date column with a large amount of null values and also the large amount of missing values found in the categorical columns.

2. Summary

Several checks have been conducted on the dataset to check the logical integrity of the data. Many failures of the dataset have been found and discussed both below and in the notebook. In total there were 7,038 failures below, although most of these fall under test 4. Test 4 checked the columns which have no positive specimen date with a current status of laboratory which cannot be the case. The irrational data will need to be verified by the domain expert to ensure the data is safe to be analysed.

For the datetime features, a number of issues were found during the logical integrity testing. These were mainly time interval which should not have been possible i.e. date reported before earliest reported date. The rows which failed these tests have been dropped but this would need to be verified by the domain expert as to whether there is a valid reason for these occurrences.

For several categorical features there were many special values. These values were Missing and Unknown. The meaning of these values is as follows:

- Missing: These are values which were left unanswered when the initial form was being filled out
- Unknown: This is a choice on the form i.e., for sex the person answering the form could answer Male, Female, Other or Unknown.
- OTH: One value in hosp_yn contained this value which was not specified in the data dictionary

These values will need to be dealt with although I have chosen not the drop these features because of the high correlation between them and the target outcome.

3. Review Logical Integrity

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

1. Check the time difference between the date reported to the CDC and the earlier reported test
    a. 43 rows failed this test.
    b. The maximum negative time difference found was -40 days

2. Check the time difference between the pos_spec_dt and cdc_case_earliest_dt
    a. 96 rows failed this test
    b. The maximum negative time difference found was -40 days

3. Check the time difference between the onset_dt and cdc_case_earliest_dt
    a. 0 rows failed this test
    b. The smallest time difference was 0 days as expected

4. Check the columns which have no positive specimen date vs current_status
    a. 6657 rows failed this test
    b. Possibly due to data never being inputted

5. Check rows to see if any are probable case (ie not laboratory confirmed) but do have positive specimen collection date
    a. 242 rows fail this test

6. Check if there are cases which were in the ICU but have a no value for hospitalization status
    a. 0 rows failed this test

7. Check if there are cases which were in the ICU but have value other than yes for hospitalization
    a. 0 rows failed this test

8. Compare case-fatality rate of the dataset sample with the actual value in the United States
    a. Case fatality rate of the dataset is over double that of the current United States case fatality rate

4. Review DateTime Features

4.1 Descriptive Statistics

There are 4 datetime features in the dataset. These are as follows:

- cdc_case_earliest_dt:
    - This has values for all 10,000 rows and will therefore be the main datetime feature used to analysis the dataset over time.
    - Several logical integrity test are failed when checking this feature against the below datetime features.

- cdc_report_dt
    - This feature had a number of null values (275/ 2.8%) and also failed logical integrity tests.
    - I have dropped this column completely as it is described as deprecated in the data dictionary. The above cdc_case_earliest_dt column will be used for the case report date.

- pos_spec_dt
    - This feature had a majority of null rows (7100/71% null).
    - I have decided to drop this feature for a number of reasons:
        - Very large number of null values
        - Specimen collection date would have no effect on the target outcome (death), report date is more useful here
        - Failure of integrity tests

- onset_dt
    - This feature had quite a lot of null values (4023/40%)
    - I have chosen not to drop this feature as it contains new useful information about the case, i.e. I will be able to analyse the time between the case being reported and the onset of symptoms which may have some relation to the target outcome.
    - I will deal with the large amount of null values by excluding them for the analysis.

4.2 Charts

Line plots and box plots can be seen of the above features below.

5. Review Categorical Features

5.1 Descriptive Statistics

There are 9 categorical features in this dataset. These are as follows:

- current_status:
    - This feature has 10,000 values and no null values

  - A large number of these values fail test 4 above, which checks if a case with a positive specimen date is a laboratory-confirmed case and not a probable case. 6657 fail this test – this will need to be checked with the domain expert as I can see no logical reason for this other than omission/non-collection of positive specimen collected date data.
  - I have chosen not to drop the rows failing the test as it represents 67% of the dataset. The positive specimen collection date column has been dropped

- Sex
  - No issues identified with this dataset, the majority (>99%) of rows contain valid data

- age_group
  - No issues identified with this dataset, the majority (>99%) of rows contain valid data

- race_ethnicity_combined
  - Large amount of unknown values (40.4%).
  - This feature has not been dropped as it will be useful for our analysis
  - When it is being analysed the unknown values will be both included and excluded to see if there is any relationship between any of the valid values and the target outcome and the unknown value and the target outcome

- hosp_yn
  - This feature has 41% missing/unknown/OTH values
  - The row which contained the OTH value as no documentation could be found in regard to this value
  - The missing/unknown values have not been dropped as they represent a large amount of the dataset

- icu_yn
  - Very large amount of missing/unknown values (89%)
  - I have chosen not to drop this feature even though there are so many missing values as this feature could be a strong predictor of the target outcome.
  - When analysing this feature the missing/unknown values can be removed and only the yes/no answers can be studied

- death_yn
  - No issues found with this feature

- medcond_yn
  - Large amount of missing/unknown values (84%)
  - This feature will also not be dropped because the feature could be a strong predictor of the target outcome.

5.2 Charts

Line plots and box plots can be seen of the above features below.

6. Action to take

The following main actions will be taken:

- pos_spec_dt
  - This feature will be dropped for the reason specified above
- cdc_report_dt
  - This feature will be dropped for the reasons specified above
- Values failing logical integrity tests:
  - Cases that failed logical integrity tests 1, 2 & 3 have been dropped from the dataset.
- Unknown/Missing values
  - I have chosen for all of the features with large amounts of missing/unknown values not to drop the features/rows containing the missing values as they contain valid information for other features (e.g. age_group, sex).
  - All Missing values will be replaced with Unknown to represent one value
  - Unknown values will be removed individually when studying individual features rather than remove all rows that contain Missing/Unknown. I have chosen to do this as it gives us the maximum amount of information about each feature while removing the Unknown values individually rather than all at once. (This technique of analysis was also used here: https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm)
- OTH value
  - One row contained OTH as the value of hosp_yn. I have chosen to drop this row due to the lack of information surrounding this value.
- Imputation:
  - Imputation will not be carried out on the features with missing values above, this is due to the difficulty of imputation with categorical features.
  - The features will be analysed both with and without Unknown/Missing values to show relationships between them and the target outcome.

7. References

[1] CDC Descriptions of data

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

8. Appendix

a. Data Dictionary

- cdc_case_earliest_dt: The earlier of the Clinical Date (date related to the illness or specimen collection) or the Date Received by CDC
- cdc_report_dt: Initial case report date to CDC. Deprecated, use new cdc_case_earliest_dt
- 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

b. DateTime Features

Feature	cdc_case_earliest_dt	cdc_report_dt	pos_spec_dt	onset_dt
count	10000	7725	2900	5077
mean	35:42.7	41:43.9	51:28.6	21:32.7
min	04/01/2020 00:00	02/03/2020 00:00	08/03/2020 00:00	09/01/2020 00:00
25%	23/07/2020 18:00	13/08/2020 00:00	07/07/2020 00:00	11/07/2020 00:00
50%	07/11/2020 00:00	12/11/2020 00:00	19/10/2020 00:00	19/10/2020 00:00
75%	15/12/2020 00:00	21/12/2020 00:00	02/12/2020 00:00	03/12/2020 00:00
max	16/01/2021 00:00	29/01/2021 00:00	20/01/2021 00:00	24/01/2021 00:00

c. Categorical Features

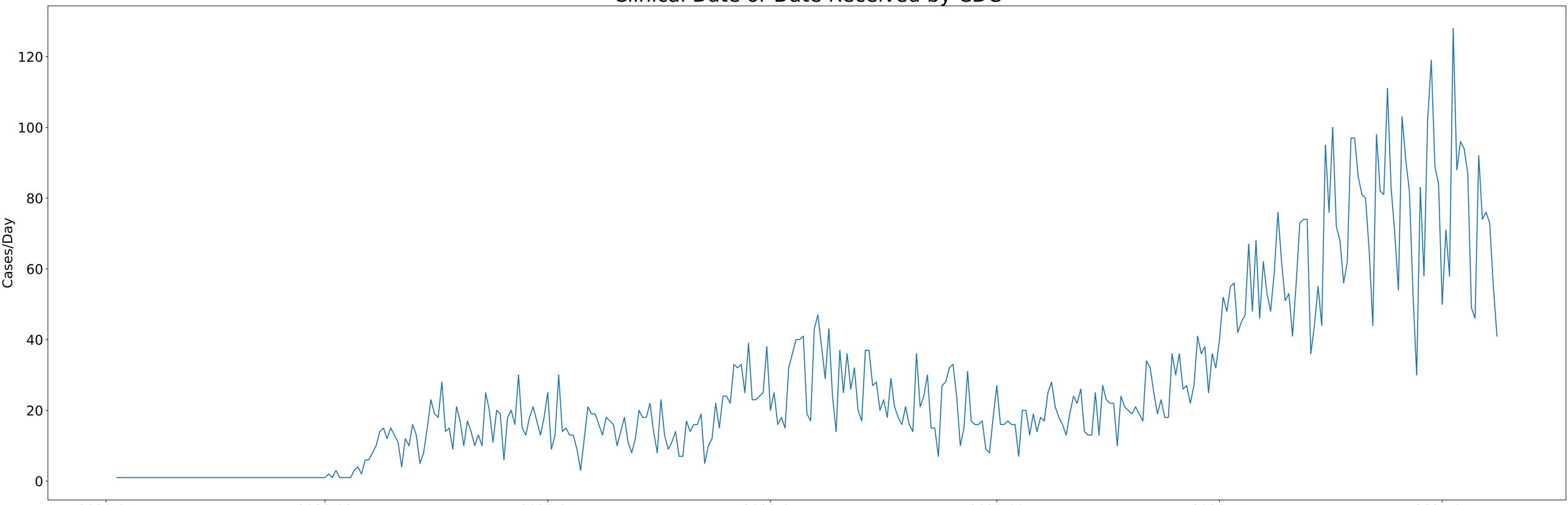
Feature	count	unique	top	freq
current_status	10000	2	Laboratory-confirmed case	9315
sex	10000	4	Female	5234
age_group	10000	10	20 - 29 Years	1921
race_ethnicity_combined	10000	9	Unknown	4039
hosp_yn	10000	5	No	5181
icu_yn	10000	4	Missing	7618
death_yn	10000	2	No	9637
medcond_yn	10000	4	Missing	7448

d. Box Plots & Histograms

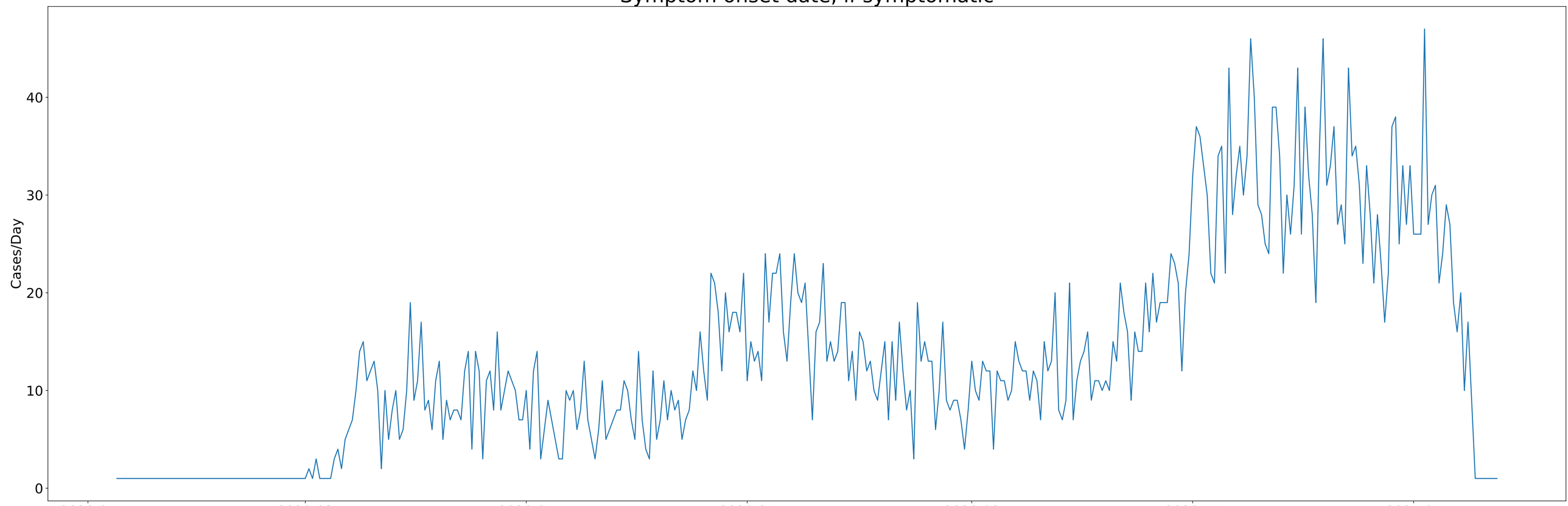
See below for summary plots and histograms. Accompanying PDF files will show larger plots.

Line Charts of Cases per Day

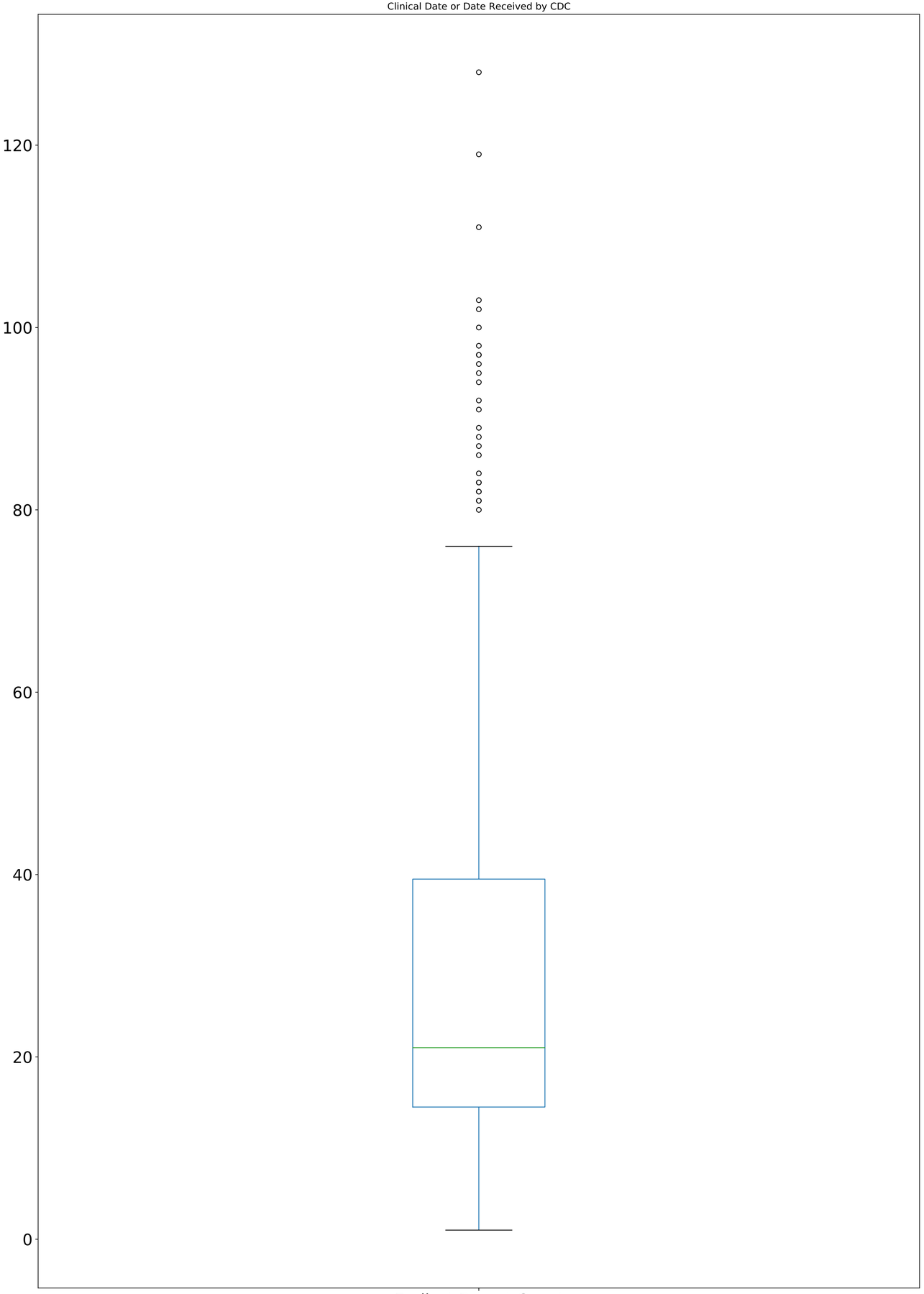
Clinical Date or Date Received by CDC



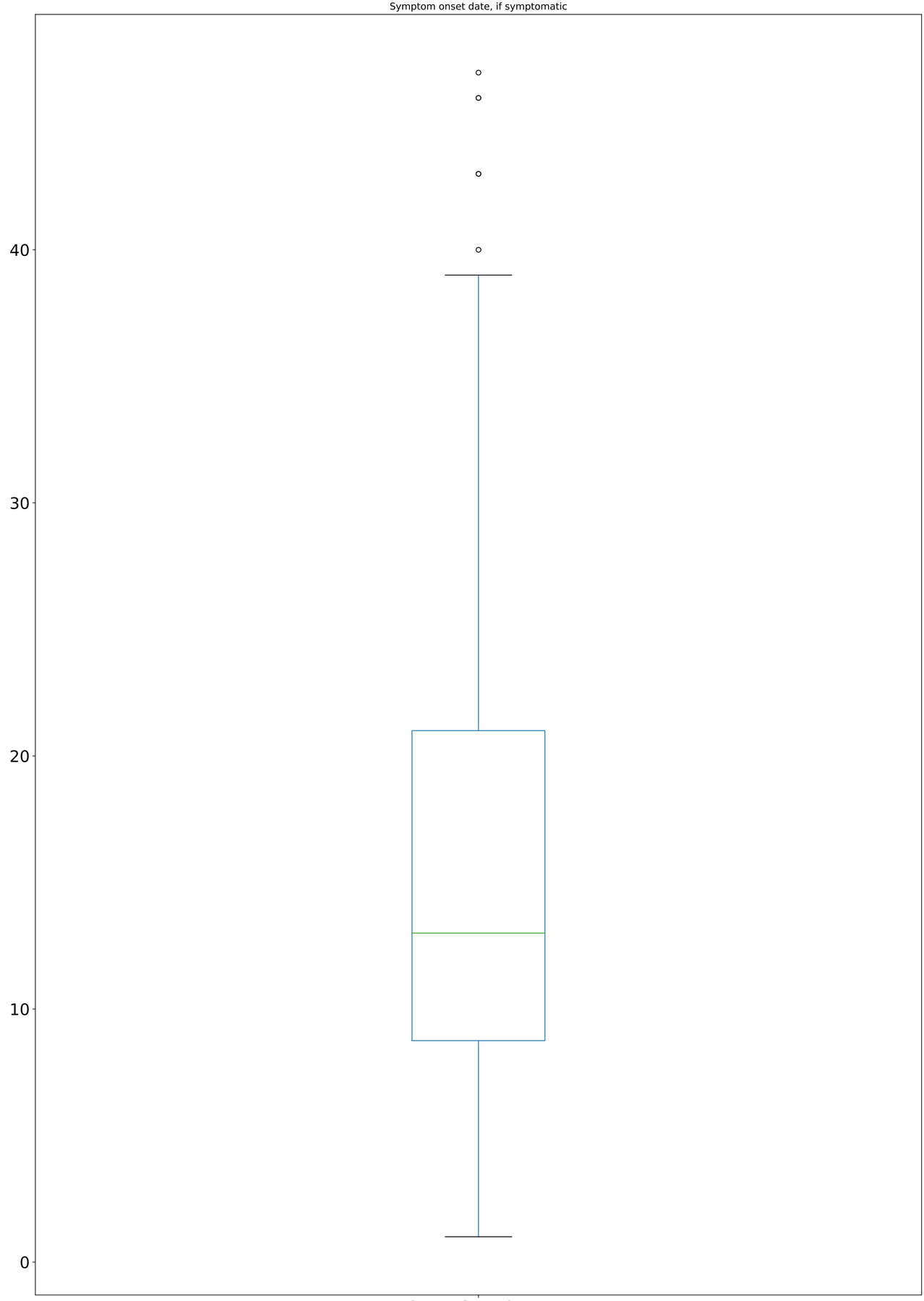
Symptom onset date, if symptomatic



Clinical Date or Date Received by CDC

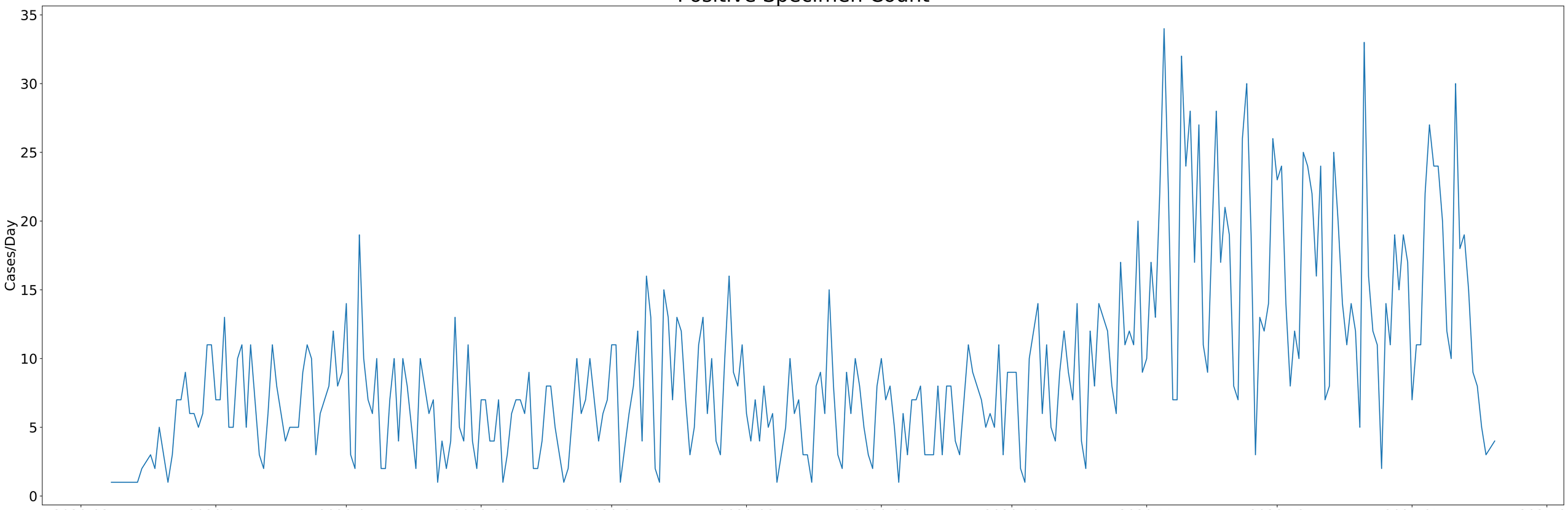


Symptom onset date, if symptomatic

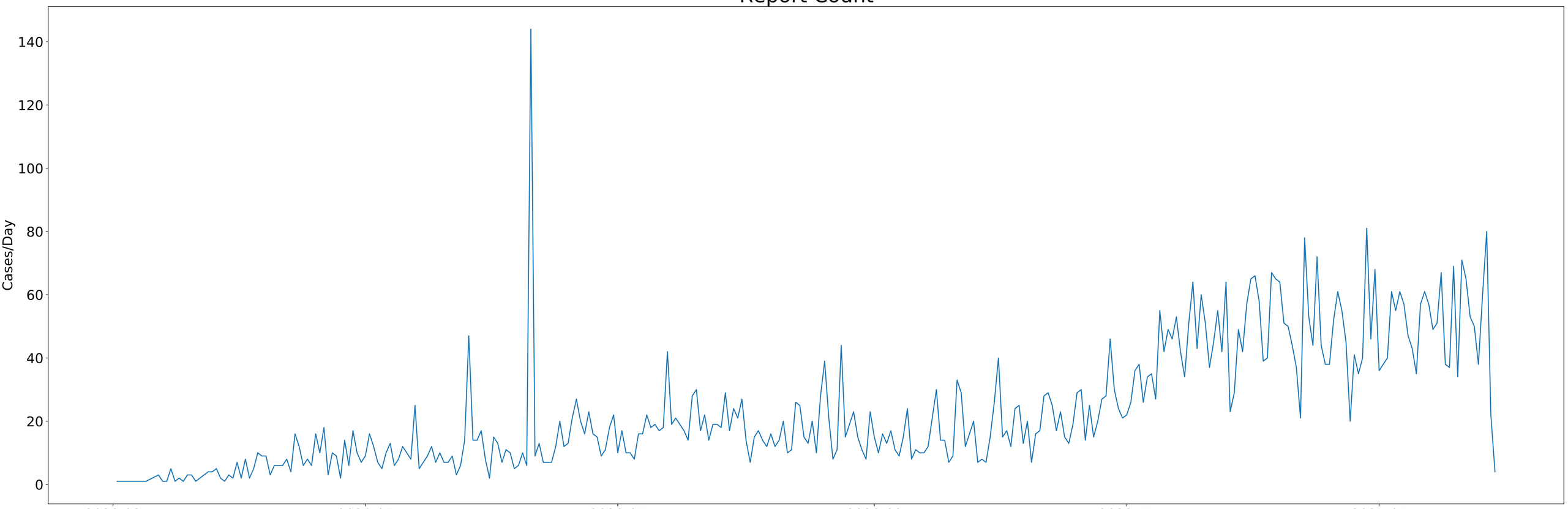


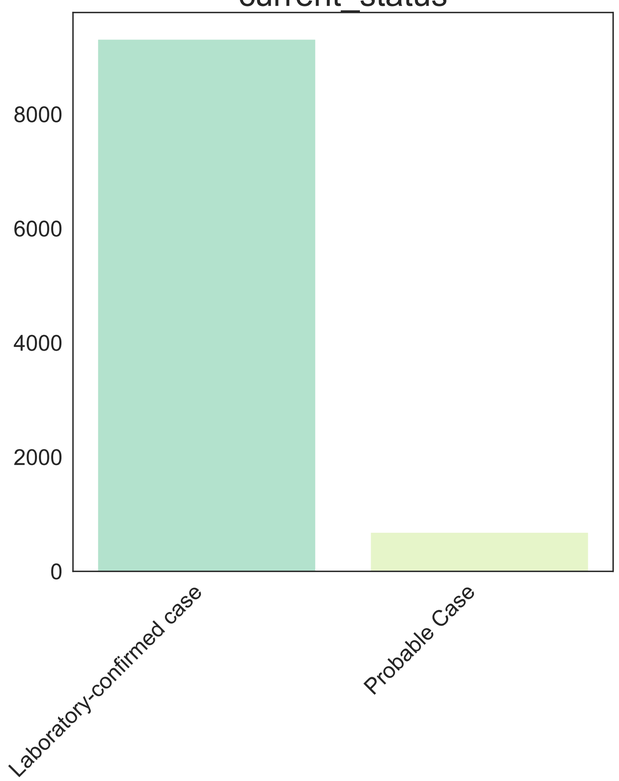
Line Charts of Cases per Day (Dropped Features)

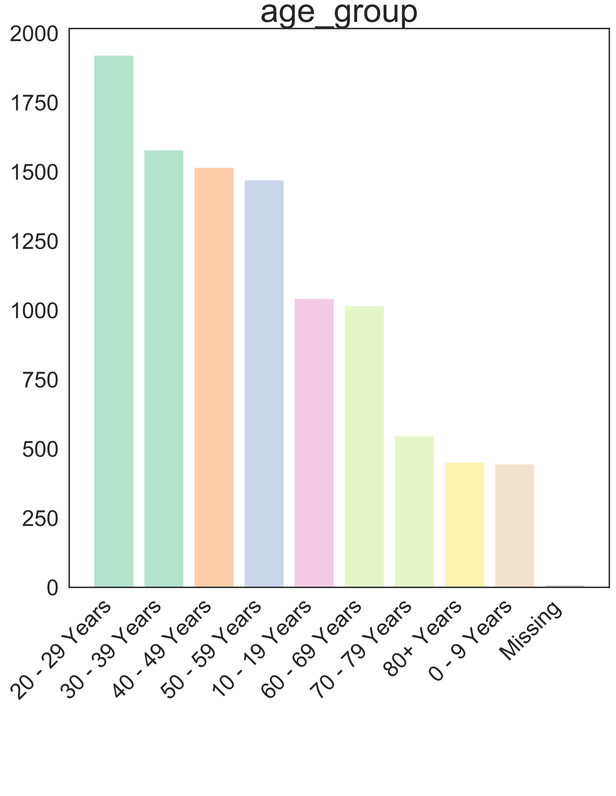
Positive Specimen Count

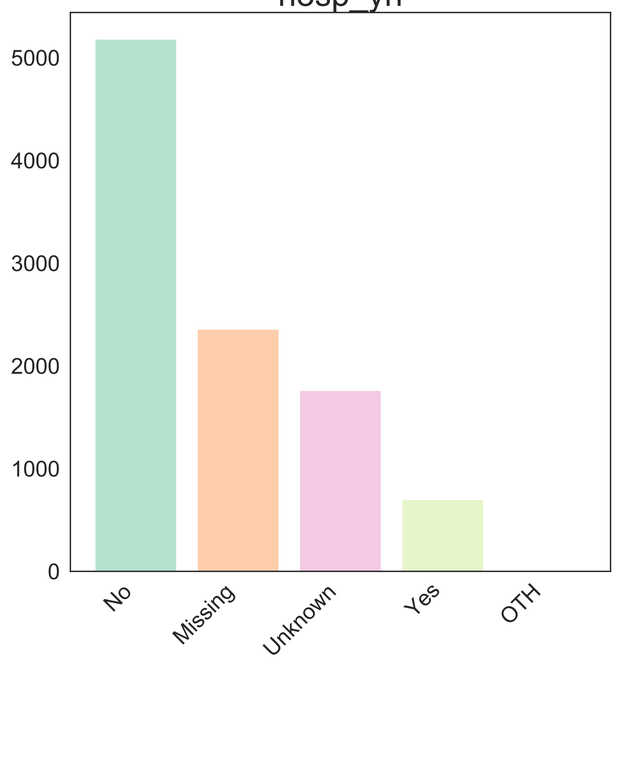


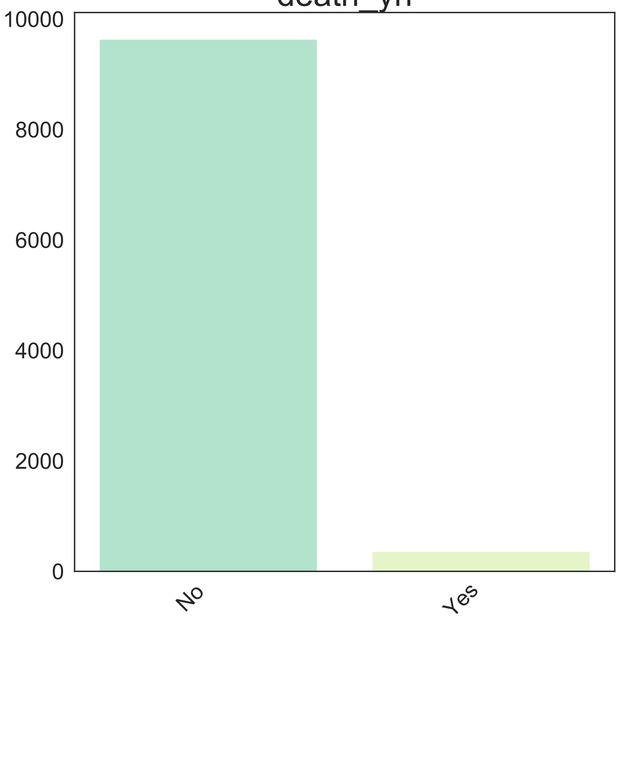
Report Count

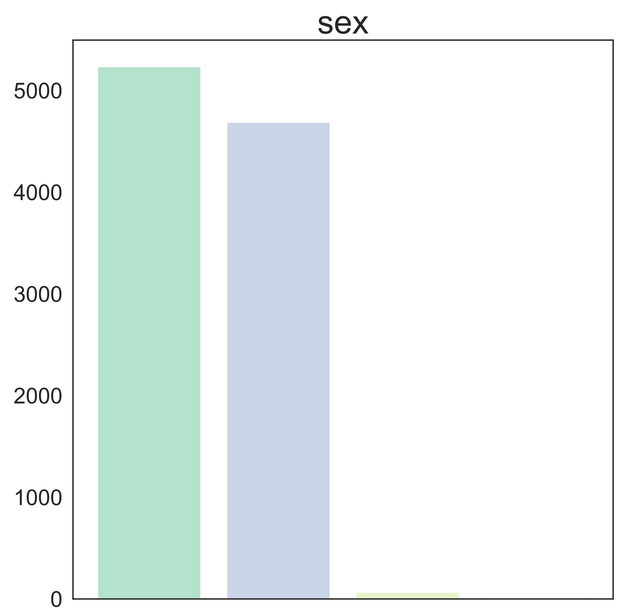


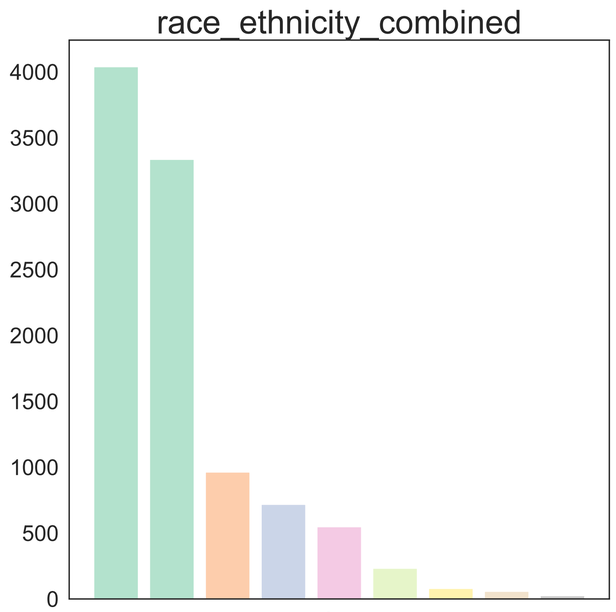

current_status
8000
6000
4000
2000
0
Laboratory-confirmed case
Probable Case

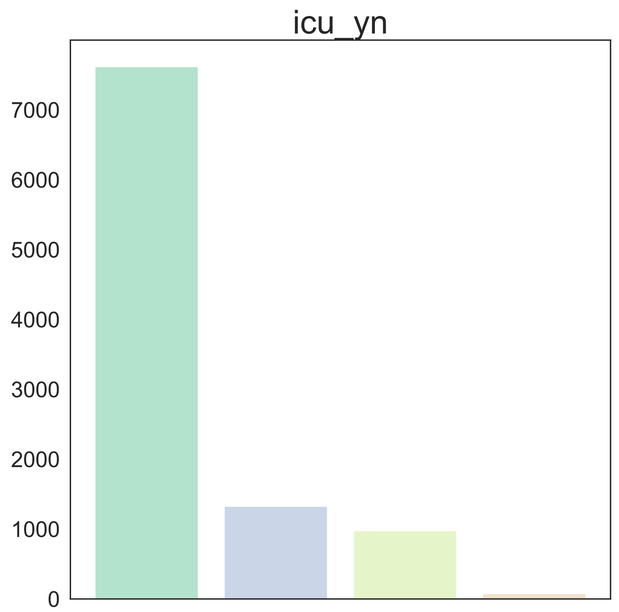

age_group
2000
1750
1500
1250
1000
750
500
250
0
20 - 29 Years
30 - 39 Years
40 - 49 Years
50 - 59 Years
10 - 19 Years
60 - 69 Years
70 - 79 Years
80+ Years
0 - 9 Years
Missing

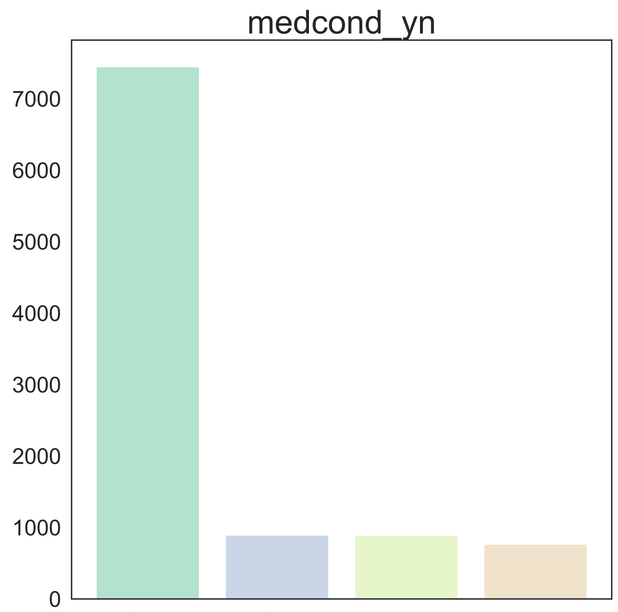

hosp_yn
5000
4000
3000
2000
1000
0
No
Missing
Unknown
Yes
OTH


death_yn
10000
8000
6000
4000
2000
0
No
Yes


sex
5000
4000
3000
2000
1000
0
Female
Male
Unknown
Missing

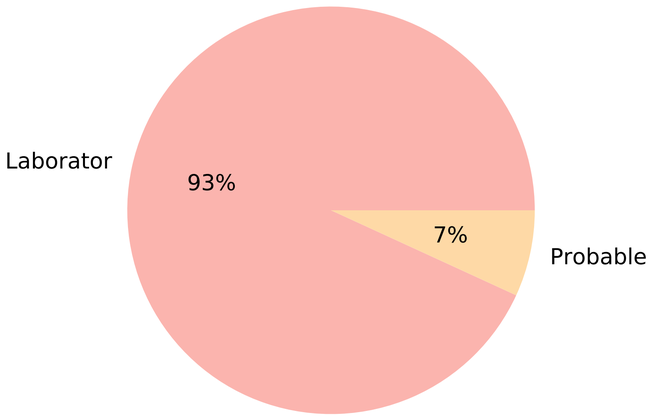

race_ethnicity_combined
4000
3500
3000
2500
2000
1500
1000
500
0
Unknown
White, Non-Hispanic
Hispanic/Latino
Black, Non-Hispanic
Multiple/Other, Non-Hispanic
Asian, Non-Hispanic
Missing
American Indian/Alaska Native, Non-Hispanic
Native Hawaiian/Other Pacific Islander, Non-Hispanic


icu_yn
7000
6000
5000
4000
3000
2000
1000
0
Missing
Unknown
No
Yes

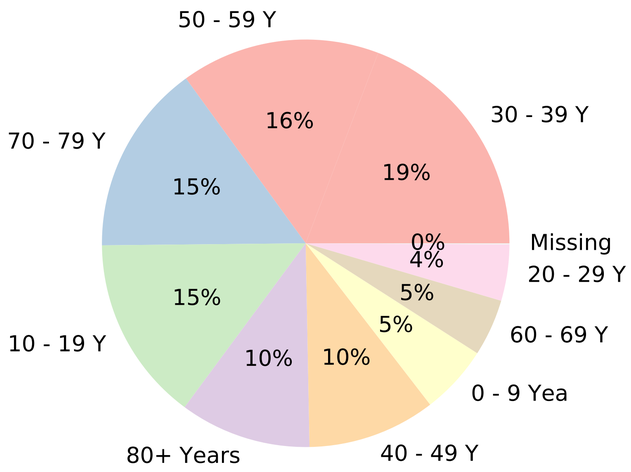

medcond_yn
7000
6000
5000
4000
3000
2000
1000
0
Missing
Yes
No
Unknown

Pie Charts of Categorical Features

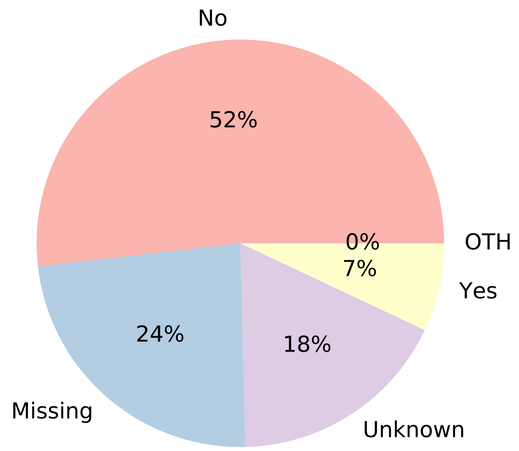
current_status



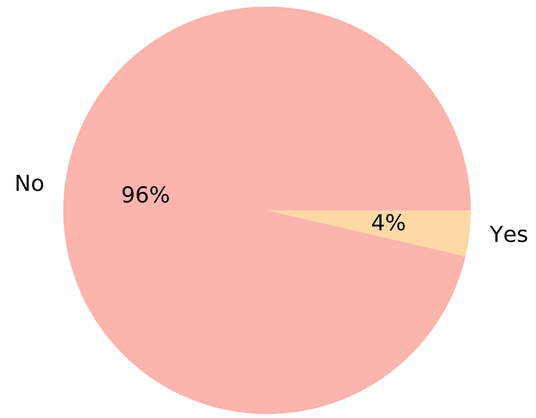
age_group



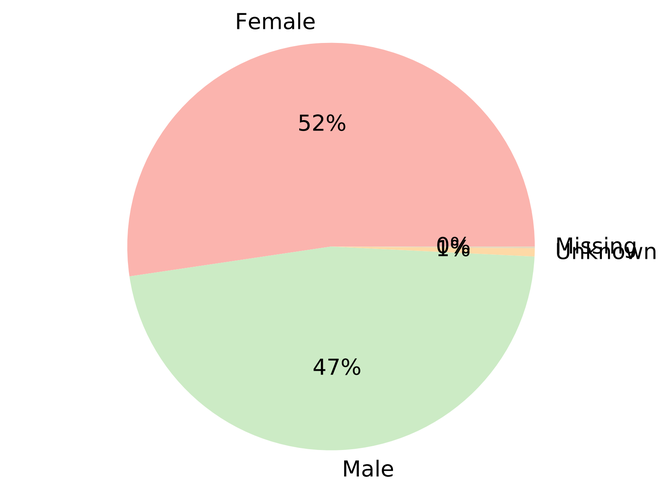
hosp_yn



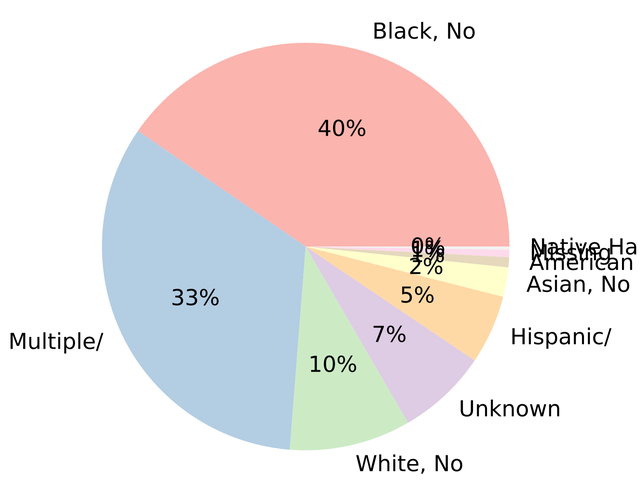
death_yn



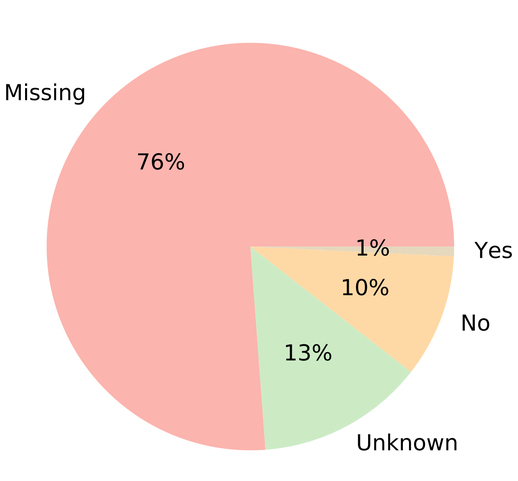
sex



race_ethnicity_combined



icu_yn



medcond_yn