Data Quality Report Initial Findings

1 Overview

The aim of the data quality report is to outline the condition of the raw data and how the data needed to be cleaned for in order for information to be extrapolated. The report will summarise the data, discuss where the issues with the data lie, detail how the data was cleaned and present the results of the data cleaning in the forms of graphs and tables. Additional information in regard to the terminology of the dataset and the additional information on the dataset can be found in the appendix.

On first look at the dataset, it was quite apparent that the data would not be easy to work with and would need a decent amount of work to clean it up. There were missing values scattered throughout the data. Due to the missing information, rows could appear as possible duplicates, due to the lack of identifying data and the missing data. The most difficult part of the dataset was the lack on numerical data in the dataset. Additionally, some columns had outlier information in regard to their dates, however, these dates appear at the end of the dataset and could be attributed to a lack of updated data.

Upon investigation it was determined that the dataset contained no duplicate columns, or columns with irregular cardinalities. A number of logic tests were carried out on the datasets which did showed inaccuracies with the *cdc_case_earliest_dt* column.

2 Summary

Before any cleaning of the data the dataset had a size of 10,000 rows and 12 columns of data. The column names were *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,* and *medcond_yn.* Descriptions of the columns can be found in the appendix.

Of the 12 columns, four had date/time data and the remaining eight were text data. Initially all the data was imported into the data frame as objects. Hence, the four columns of dates were transformed into datetime64[ns] data types and the eight text-based data were recategorized as categorical data. Initially it was difficult to determine if the dates should be categorical or continuous, however, as the dates were continual time ranges, it was determined that the date columns should be treated as continuous data.

Once the data had been categorised into the relevant date types the data was checked for duplications, percentage missing, cardinality and constant columns. As discussed in the overview, no constant columns, duplicate columns or cardinality irregularities were discovered.

There were 969 duplicate rows in the table, which was reduced down to 540, when the first instance of the row was skipped. Upon looking at the data being presented the decision to keep the duplicate entries was made. The decision was made due to the type of data being discussed. The dataset had been anonymised before issue of the data and as such any identifying data had been removed. The personal data left in the dataset was very ambiguous and could very easily apply to more than one person as is the point of anonymising the data. For example there is a duplicate in row 18 and 5466, however, it would be difficult to determine if they are the same person or not as the only personal features commented on is a range of ages, their sex and there race and ethnicity. It would be possible that two white females in their 20s could have tested yes on the same day and developed symptoms on tsshe same day.

Next the data was checked for the missing value rate of the information. As the database owners fill in 'Missing' for some columns where information is missing the dataset needed to be reformatted so that the missing values were uniform. All cells which contained 'Missing' were replaced with 'NaN' and the missing dates were replaced with 'NaT'. This meant the dataset could easily be searched to determine the number of missing values in the data set for each column.

	Num of miss	total rows	%miss
cdc_case_earliest_dt	0	10000	0
cdc_report_dt	2311	10000	23.11
pos_spec_dt	7195	10000	71.95
onset_dt	4966	10000	49.66
current_status	0	10000	0
sex	10	10000	0.1
age_group	15	10000	0.15
race_ethnicity_combined	89	10000	0.89
hosp_yn	2366	10000	23.66
icu_yn	7649	10000	76.49
death_yn	0	10000	0
medcond_yn	7448	10000	74.48

As indicated in the table above, the miss rate of *pos_spec_dt, icu_yn,* and *medcond_yn* was over 70%, and as such they were dropped from the table. The data could have been kept, however, with the level of missing data, any attempt at filling in the data could have skewed the results of a data analysis performed on the cleaned data at a later date.

Additionally, *onset_dt* was looked at for potentially removing from the dataset, however as the data had an approximate 50% miss rate, it was decided to keep the column and perform some logical tests on it to see if it was worth keeping. It was determined that the *onset_dt* could be beneficial for determining spikes in the transmission of the virus.

After the columns had been dropped a check was performed on the remaining data to determine how many rows contained data in all of the columns. Unfortunately, it was determined that of the 10000 rows of data, 54.46% of the rows were missing data from one or more column. At risk of losing over half of the data set, the discission was made to keep all the rows and work with the NaN values where they arose. As the NaN values occur in the categorical and continuous data, it was difficult to mask them, as such they were left as NaN due to the type of information we were dealing with. The NaN values can be useful with this data set in determining what type of data a person is willing to give.

Imputation, i.e. the mapping of the mean to a missing continuous number could not work in this instance as the data was in date format and could throw of analysis or cause inconsistencies in the data.

3 Review Logical Integrity

A total of eight tests were performed on the data to determine if the data was logical. Most of the tests focused on the deaths of individuals. Originally more tests were scheduled, however due to the decision to drop columns due to the high levels, some tests couldn't be run.

From the results there are inconsistencies with the cdc_case_earliest_dt being the earliest date in the data set. As this is more of an indicator data, it was left in as one of the most complete columns in the data set. Additionaly, it is noted that there are quite a few instances of the cdc_report_dt being before the onset_dt. This may be an inconsistency or additional data could show that the patient was a close contact and so was tested before the onset of symptoms. As such the data was kept.

Test 1a: Test if *cdc_case_earliest_dt* contains dates before or equal to the *cdc_report_dt.*

26 instances were detected as being set later than the *cdc_report_dt*

Test 1b: Of the 26 instances check if *cdc_case_earliest_dt* contains dates later than *onset_dt*.

0 instances of *cdc_case_earliest_dt* containing a date later than *onset_dt.*

Test 1c: Of the 26 instances check that cdc_report does not have a date later than *onset_dt*

O instances of *cdc_report_dt* containing a date later than *onset_dt*

Test 1d: tests if cdc_case_earliest_dt occurs after onset_dt

1 occurrence of this in the dataset

Test 1e: of the 1 instance check if cdc_report_dt is later than onset_dt

1 occurrence of this failing

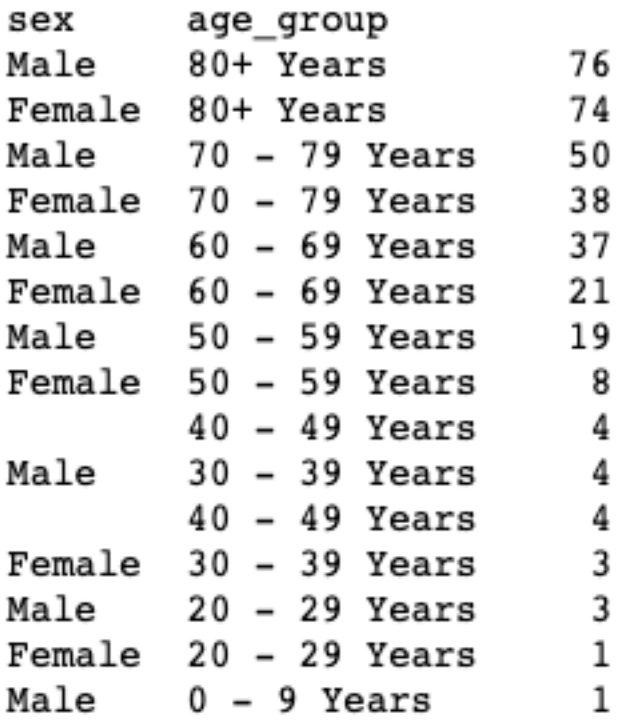
Test 1f: test if onset_dt occurs always occurs before cdc_report_dt

3500 instances of this failing

Test 1g: Test if cdc_report_dt occurs before onset_dt

39 instances of this failing

Test 2a: Determine if the deaths per age_group and sex conform to the repo rates.



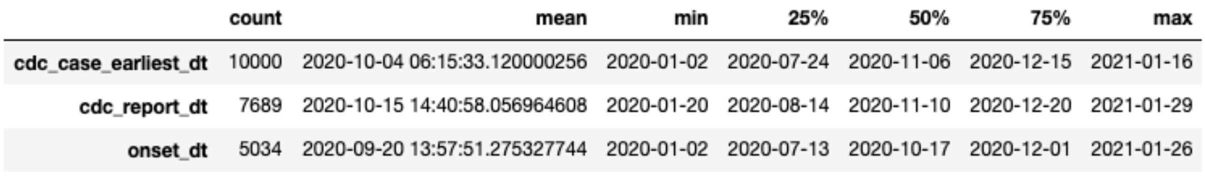
```
sex     age_group
Male    80+ Years        76
Female  80+ Years        74
Male    70 - 79 Years    50
Female  70 - 79 Years    38
Male    60 - 69 Years    37
Female  60 - 69 Years    21
Male    50 - 59 Years    19
Female  50 - 59 Years     8
        40 - 49 Years     4
Male    30 - 39 Years     4
        40 - 49 Years     4
Female  30 - 39 Years     3
Male    20 - 29 Years     3
Female  20 - 29 Years     1
Male    0 - 9 Years       1
```

The counts detail the trend that older people are more susceptible to the virus than younger people

4 Review Continuous Features

4.1 Descriptive Statistics

There are three continuous features in the data set. All three continuous features are date information and as such can be difficult to work with as continuous data. In the figure below we can see the description of the continuous data. Note that the time part of the mean date is used to signify a fraction of a day.



	count	mean	min	25%	50%	75%	max
cdc_case_earliest_dt	10000	2020-10-04 06:15:33.120000256	2020-01-02	2020-07-24	2020-11-06	2020-12-15	2021-01-16
cdc_report_dt	7689	2020-10-15 14:40:58.056964608	2020-01-20	2020-08-14	2020-11-10	2020-12-20	2021-01-29
onset_dt	5034	2020-09-20 13:57:51.275327744	2020-01-02	2020-07-13	2020-10-17	2020-12-01	2021-01-26

4.2 Histograms

A set of histograms for the data is available in the appendix.

The histograms indicate a right-skewed dataset whereby the number of cases being reported per day are increasing every day. It is interesting to see that although the number of reports is still high in January 2021, there is a reduction in the number of people experiencing the onset of symptoms. Unfortunately, as onset_dt had a 50% miss rate, this analysis might be out slightly considering onset tracks with cdc_reports_dt quite well for the previous bins.

4.3 Box Plots

Due to the datetime types of the continuous data, conventional box plots could not be relied upon to visualise the data effectively and as such were removed from the discussion.

5 Review Categorical Features

5.1 Descriptive Statistics

Due to the type of data included in the data set, the majority of the columns were categorical features. In the original dataset there were eight categorical columns, due to a large number of missing values in two of the columns as mentioned above. Hence, six features remained to be reviewed.

Of the six remaining features hosp_yn still had quite a large miss percentage of 23.66%. To facilitate this the NaN values were marked as 'Missing' and included in the analysis. This could be combined with the unknown category in hosp_yn at a later date or treated separately.

The cardinality and permitted values of the hosp_yn feature showed an inconsistency with the pui form in so far as it showed 2 instances of OTH. OTH is not an input field on the form nor the excepted missing value number.

The cardinality and permitted values for the other features are correct and do not need intervention.

5.2 Bar Plots

A set of Bar plots for the categorical data can be found below in the appendix.

Current Status:

This featured measured whether the patient had received laboratory confirmation or if they were awaiting results, marked as a probable case. From the graphs it is clear that the majority of patients have received a laboratory confirmation of a positive test result.

It is noted that there may be issues with the reporting of information as approximately 93% of instances have a laboratory confirmation, however, pos_spec_dt was removed due to having a 72% miss rate.

Sex:

The graph indicates that the data set contains more female than male patients with a miniscule number of unknown or missing entries. This information could be useful in determining whether the virus is worse, better or the same for the different sexes.

Age Group:

The bar plot for the age group tells a very telling story. From an initial review of the data, it is clear that the employable age group are at most likely to contract the virus. The majority of patients in this data set range between 20 and 59. The largest spike is in the 20-29.

It can also be noted that the 70+ age group are relatively low risk for contracting the virus, however this could be due to them not being in contact with a large number of people.

Race & Ethnicity Combined

Race and ethnicity combined is a difficult column to analyse as approximately 40% of instances were reported as unknown and 34.5% were reported as White, not Hispanic. From the data you could hypothesise that White people are more at risk at contracting the virus than non-White people. However, due to the large number of unknown results in this section it makes it difficult to make any decisions.

These results could be due to people not wanting to indicate their race or ethnicity on the form for personal reasons.

Attended Hospital

This graph is a tad confusing, when looked at knowing the 93% positivity rate of the information. Over 50% of people did were not hospitalized for the virus and only 6.85% admitted to being hospitalized. This could be a result of the cost of healthcare in the United States, or it could be a result of mild symptoms in those who contracted the virus.

Death_yn

96% of instances were marked as not deceased. This could corroborate the hypothesis that the instances of the virus in this dataset were patients with mild symptoms. As we had to drop the medical conditions column due to lack of data, it would be difficult to determine whether the dataset included mainly healthy people or not.

6 Actions to Take

1. Remove the rows containing OTH in the hosp_yn feature
2. Remove the rows that failed the logic tests for the dates
3. Remove the column for race and ethnicity combined due to the high level of missing and unknown data.
4. Remove the rows containing missing age data
5. Remove the rows containing missing and unknown hosp_yn data
6. Remove the rows containing missing and unknown sex data
7. Remove the probable cases

7 References

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

8 Appendix

8.1 Terminology & Assumptions

Column Name: cdc_case_earliest_dt
Description: 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.
Data Type: Date & Time

Column Name: cdc_report_dt
Description: 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.
Data Type Date & Time

Column Name: pos_spec_dt
Description: Date of first positive specimen collection
Data Type Date & Time

Column Name: onset_dt
Description: Symptom onset date, if symptomatic
Data Type Date & Time

Column Name: current_status
Description: Case Status: Laboratory-confirmed case; Probable case
Data Type Plain Text

Column Name: sex
Description: Sex: Male; Female; Unknown; Other
Data Type Plain Text

Column Name: age_group
Description: 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
Data Type Plain Text

Column Name: race_ethnicity_combined
Description: 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
Data Type Plain Text

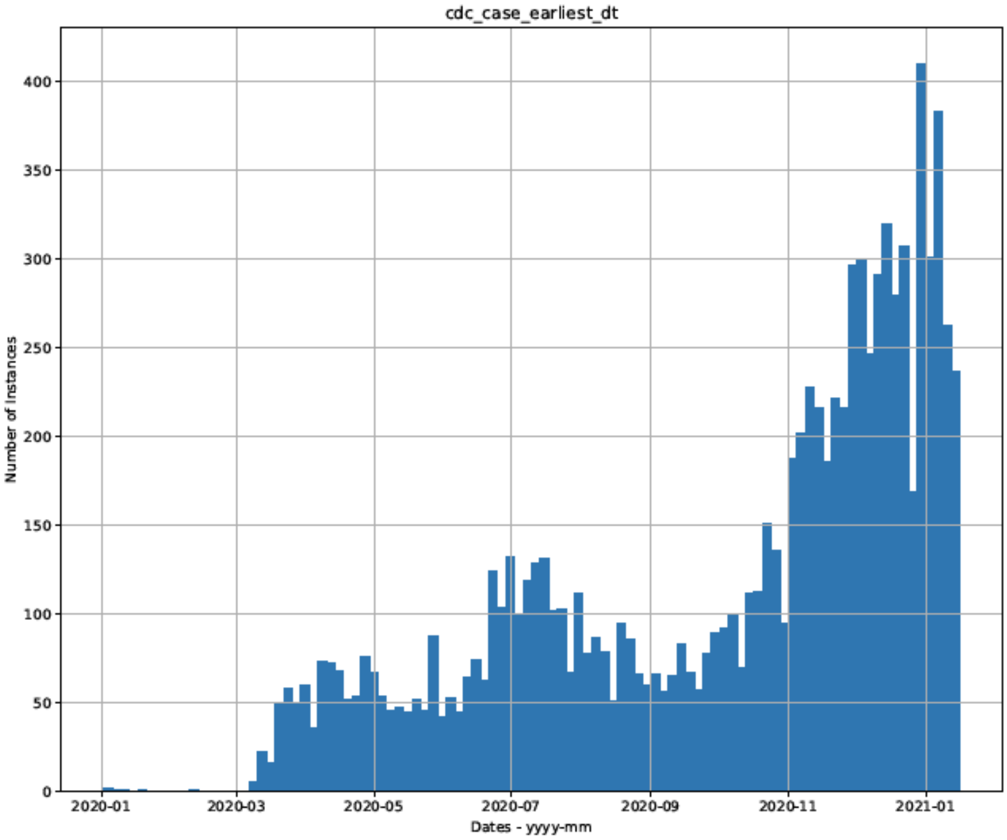
Column Name: hosp_yn
Description: Hospitalization status
Data Type Plain Text

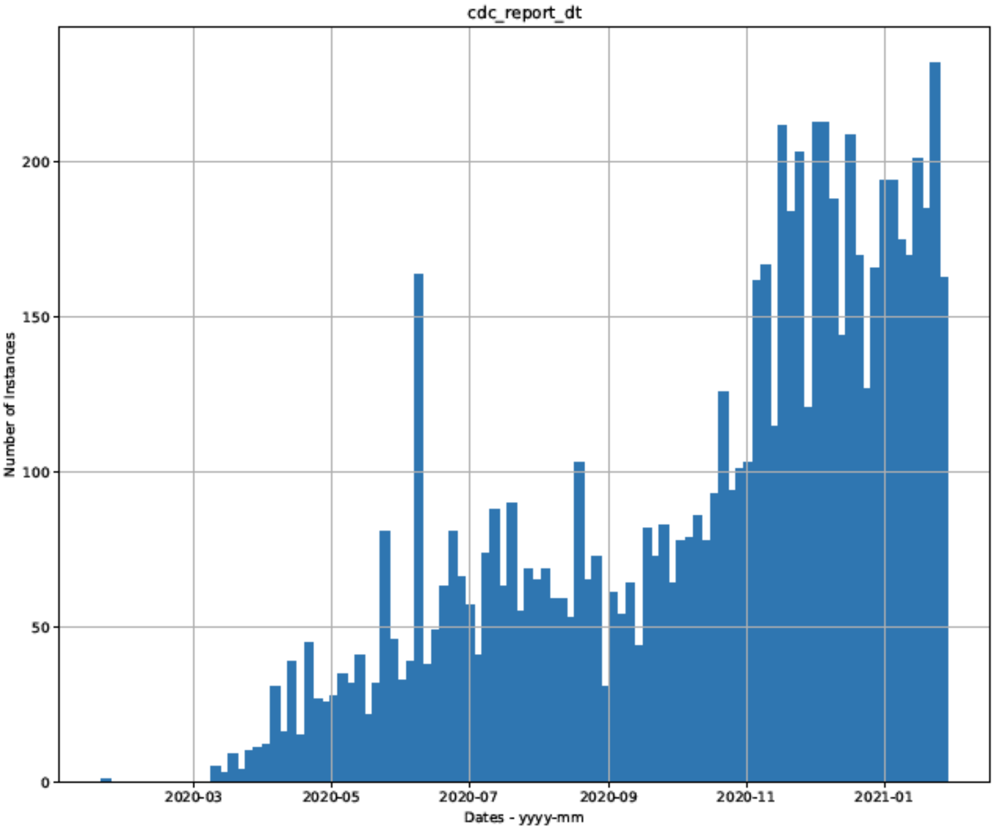
Column Name: icu_yn
Description: ICU admission status
Data Type Plain Text

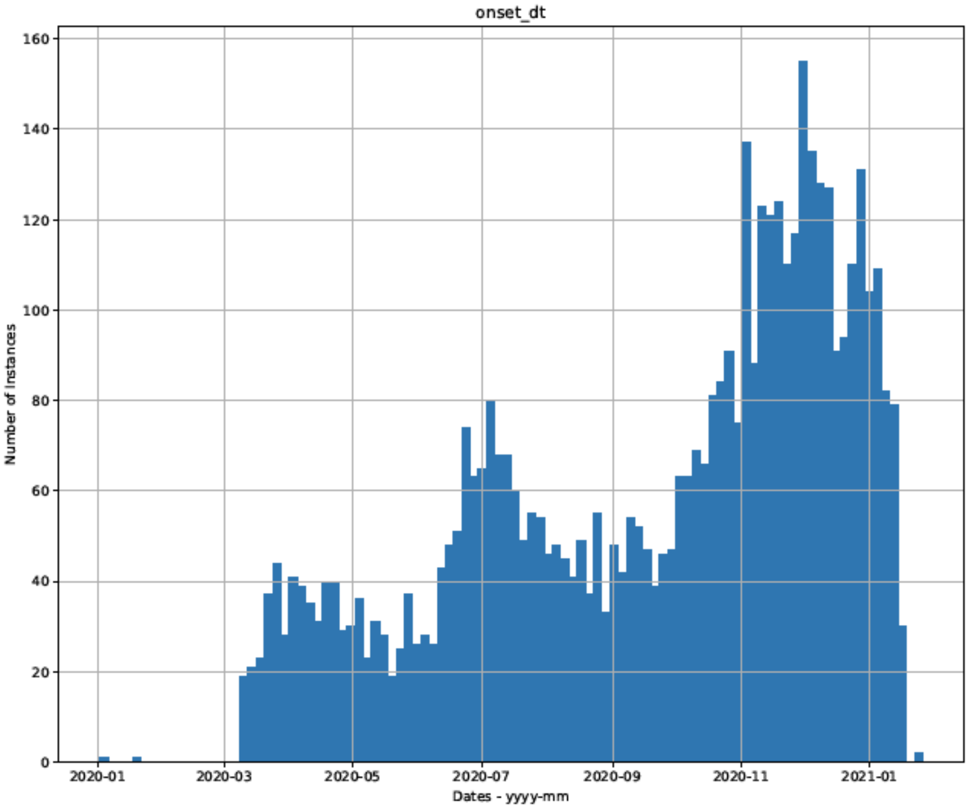
Column Name: death_yn
Description: Death status
Data Type Plain Text

Column Name: medcond_yn
Description: Presence of underlying comorbidity or disease
Data Type Plain Text

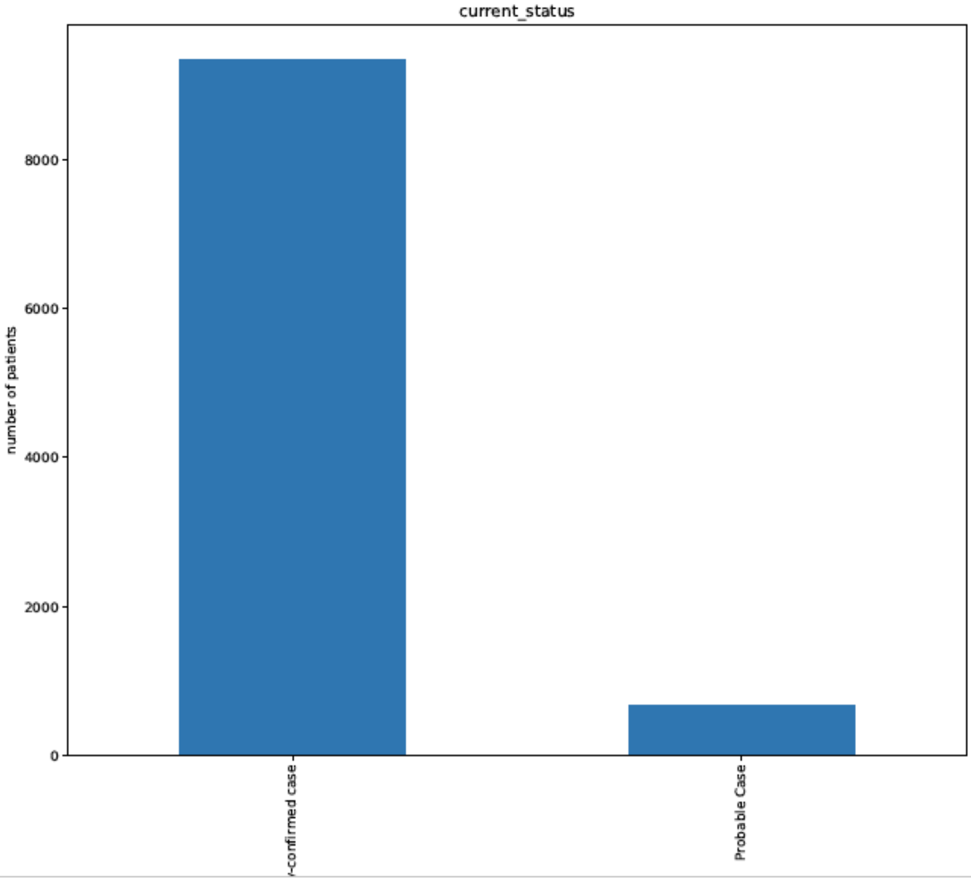
8.2 Histograms

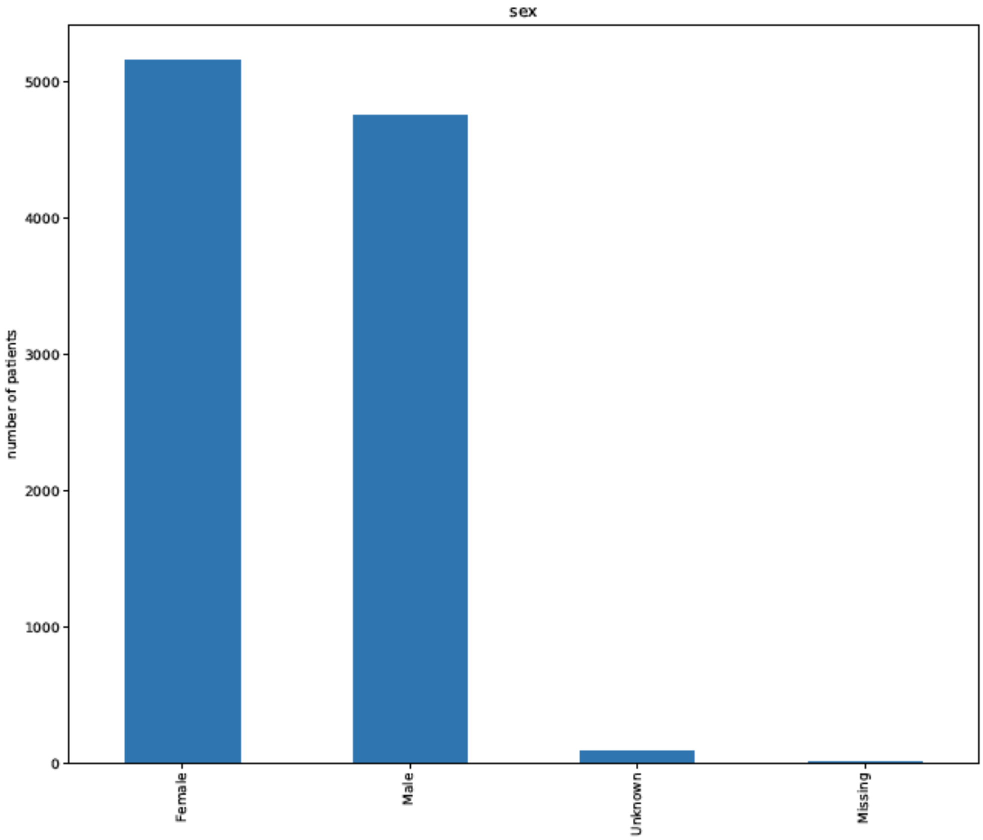


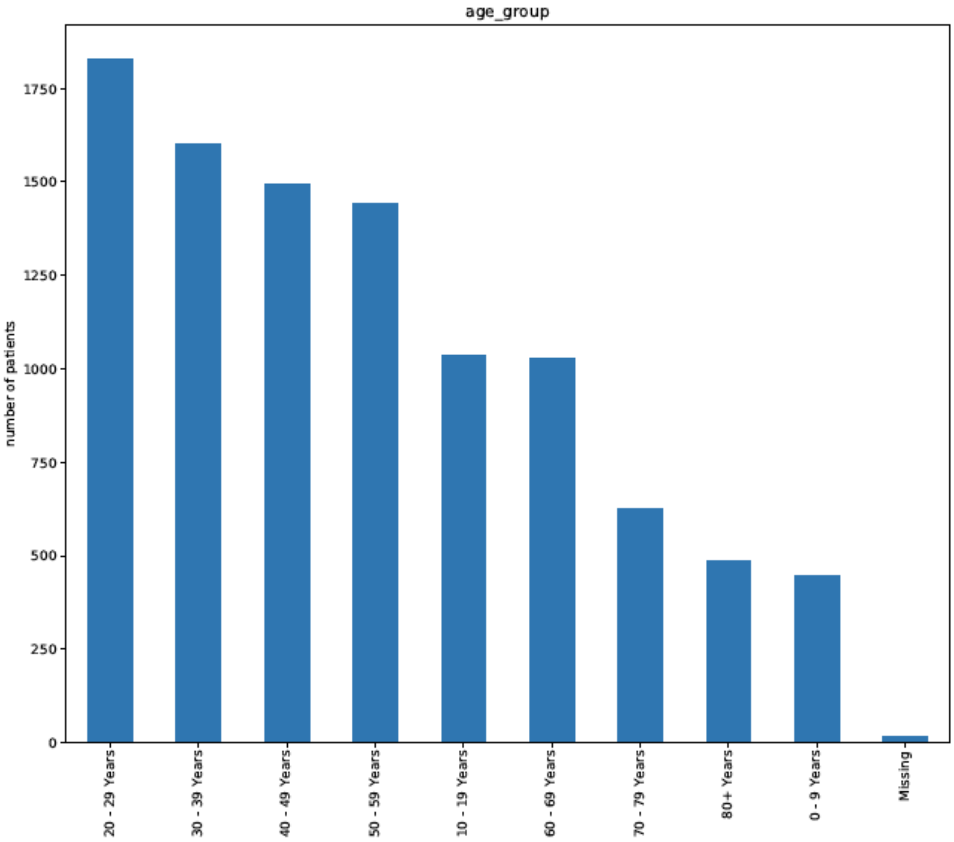

cdc_report_dt
Number of Instances
0
50
100
150
200
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
Dates - yyyy-mm

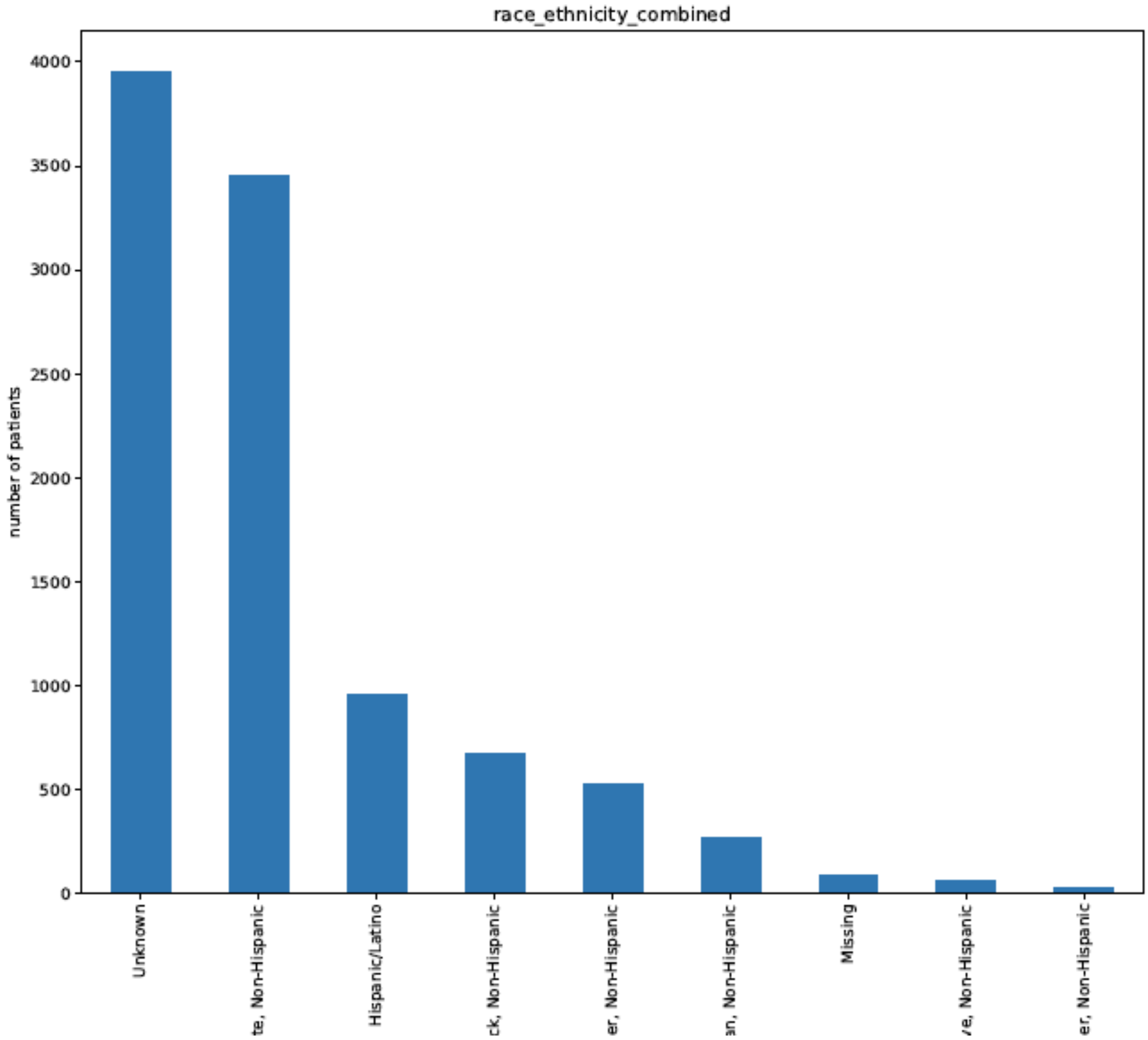

onset_dt
Number of Instances
0
20
40
60
80
100
120
140
160
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
Dates - yyyy-mm

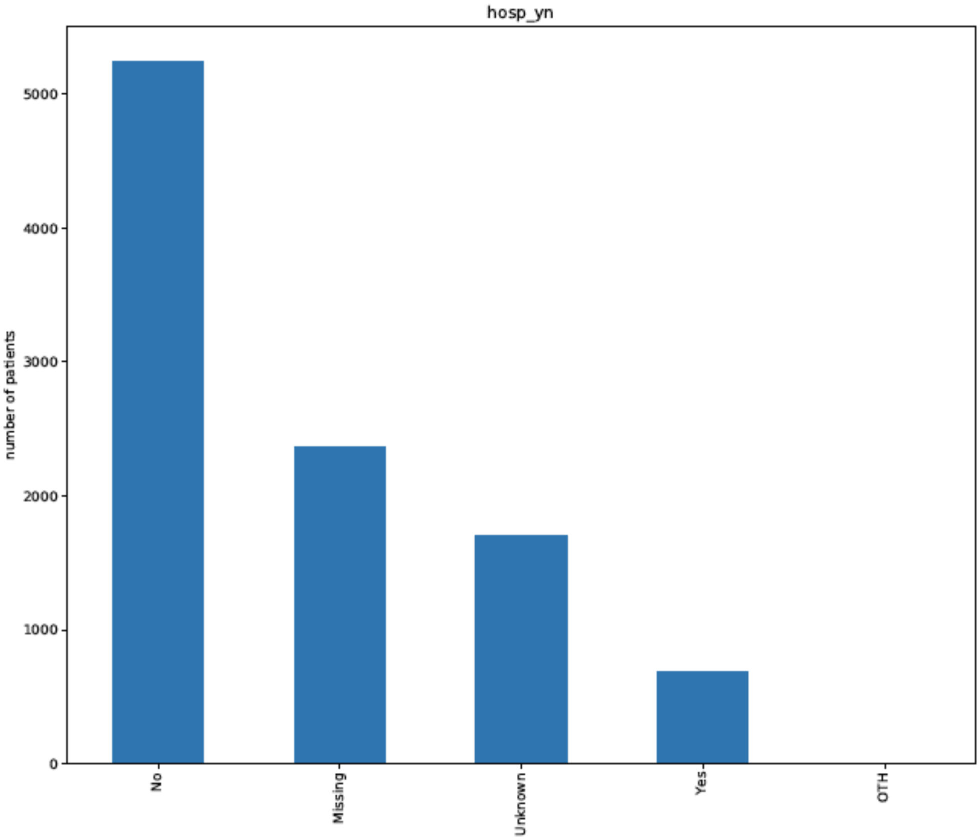
8.3 Bar plots

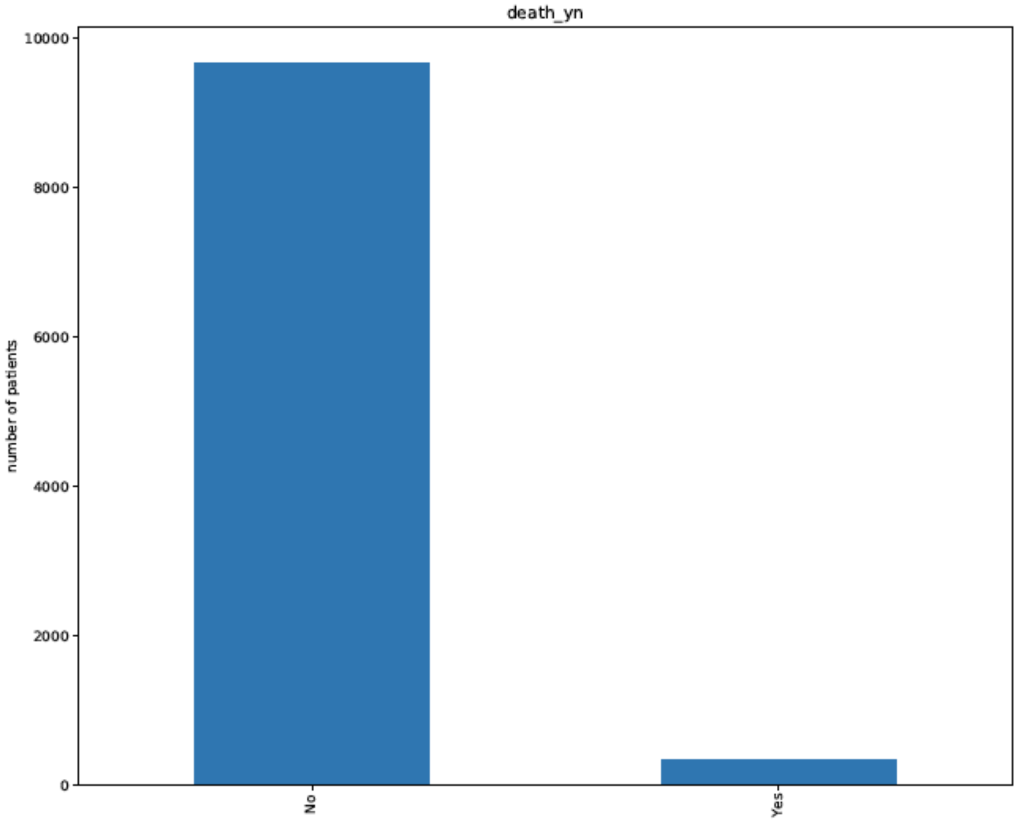

current_status
number of patients
0
2000
4000
6000
8000
-confirmed case
Probable Case


sex
number of patients
0
1000
2000
3000
4000
5000
Female
Male
Unknown
Missing


age_group
number of patients
0
250
500
750
1000
1250
1500
1750
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


race_ethnicity_combined
number of patients
4000
3500
3000
2500
2000
1500
1000
500
0
Unknown
te, Non-Hispanic
Hispanic/Latino
ck, Non-Hispanic
er, Non-Hispanic
an, Non-Hispanic
Missing
ve, Non-Hispanic
er, Non-Hispanic


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


death_yn
number of patients
0
2000
4000
6000
8000
10000
No
Yes