Data Quality Report.

1. Overview

This report will outline the initial findings within the dataset (covid19-cdc-20202492.csv). It will summarise the data contained within. Describe said data, outline the data quality issues present within and how they will be addressed in the quality plan.

Initially the dataset contains 10000 entries with 12 Columns per entry. Initially all the datatypes for the dataset were object type. The majority of columns contained no empty values however cdc_report_dt, pos_spec_dt and onset_dt contained 2352, 7211 and 4941 missing entries respectively. The main issues present within the dataset were a significant amount of duplicate entries (841 including the first occurrence). Dates saved as an object type and a significant amount of rows with “Unknown” or “Missing” Data. No constant columns were discovered within the dataset

2. Summary

Several logical and data tests were carred out on the dataset. These raised some immediate concern. The dataset contained a large number of duplicate entries (841 including the first occurrence) and the values for many of the categorical columns were Missing or Unknown. One feature (race_ethnicity_combined) is almost 50% missing or unknown data.

For the categorical features several changes are recommended. Either the imputation of Missing to No values based on logical inferences or the dropping of data from the dataset on a feature by feature basis. It is also recommended to a drop a feature with an overwhelming amount of missing data.

The continuous data presents no immediate concern or issues however conversion from object to datetime is recommended

Review of logical consistency:

Test 1: Check if any entries have a positive icu status with a missing hospital status

- 0 cases found

Test 2: Check if any entries have a positive icu status with an Unknown hospital status

- 0 cases found

Test3: Check how many cases have Missing icu and positive hospital

- 301 cases found

Test4: check how many cases have Missing icu and negative hospital attendance:

- 3613 cases found

3 Review of Continuous features.

While the dataset does not include anytime of int65 or float64 traditional continous features the dataset does include columns for dates which after conversion to date-time can be examined much like continuous features . These are

cdc_case_earliest_dt:

The earliest available clinical date for the entry on record. This ranges from 2020-01-01 to 2021-01-16. Has no null values and a cardinality of 323

cdc_report_dt:
Date the care was initially reported to the CDC. Ranges from 2020-03-08 to 2021-01-29. 7648 entries and a cardinality of 324

pos_spec_dt:
Date of first positive specimen retrieval. Ranges from 2020-03-14 to 2021-01-23. 2789 entries and a cardinality of 314

onset_dt:
Symptom onset date. Ranges from 2020-02-01 to 2021-01-28. 5059 entries and a cardinality of 325

As shown above some of these features contain many missing entries however based on the features purpose and the included form these are not a cause for concern. As a missing entry in this case is not an error in data collection but an indication that the action was not taken (e.g not every case developed symptoms or had a positive specimen collected). Boxplots included in appendix at the end of this report show the frequency distribution of the dates. While there are some outliers in the data it is not expected for these to cause concern.

Based on the findings it is recommended to keep all entries and columns based on these features. However conversion from object to datetime is recommended to allow for more robust analysis

3.1 Histograms:

All histograms for continuous features can be found in the appendix at the end of this report

3.2 Box and whisker plots:

All boxplots for continuous features can be found in the appendix at the end of this report

4 Review of Categorical features.

There are 7 categorical features in the dataset. One of which is the target. These features are

- current status
- sex
- age_group
- race_ethnicity_combined
- hosp_yn
- death_yn
- medcond_yn

current_status:
This features uses the object datatype and idicates if the case is Laboratory confired or suspected There were no missing data or issues found for this feature

sex:
This feature indicates the sex of the entry. And contains 4 possible values (Male,Female,Unknown,Missing). As shown in the barchart in for sex in the appendix, the unknown and missing values make up a tiny percentage of the overall entries and so I would recommend removing these rows from the dataset as this would have little to no impact on the overall dataset

age_group:
This feature indicates the age group the entry falls into. There are 10 possible values within this feature:

- 0 - 9 Years
- 10 - 19 Years
- 20 - 29 Years
- 30 - 39 Years
- 50 - 59 Years
- 40 - 49 Years
- 60 - 69 Years
- 70 - 79 Years
- 80+ Years
- Missing

Issues exist within this feature, some entries for this feature have the value of Missing, however as shown in the barchart for age_group in the appendix, this makes up an insignificant amount of the overall dataset and It would be recommended to drop these rows as it would have little to no impact on the dataset

race_ethnicity_combined:
This feature indicates the race and/or ethnicity of the entry. This feature contains 9 possible values:

- 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

significant issues exist within this feature. As shown in barchart for race_ethnicity_combined in the appendix, a significant amount of the entries for this feature are either Unknown or Missing. As these values make up nearly half of the overall values it is recommended that this column be dropped from the dataset entirely as the amount of unsuitable data significantly hinders its ability to relate to the target feature in an accurate fashion. As there is no way to logically infer its correct value from the other data imputing the data would not be appropriate.

hosp_yn:
hosp_yn indicates if the entry has been admitted to hospital as a result of Covid-19. It contains 4 possible values:

- Yes
- No
- Missing
- Unknown

As seen in the appendix barchart for hosp_yn Missing and unknown values make up a significant amount of the entries for this feature. Based on the included form and the logical tests outlined and carried up above it is recommended to treat a "Missing" value as a no as a result of human error. However no such option exists for the Unknown values and as such, removal of the rows from the dataset is recommended.

icu_yn:
icu_yn indicates if the entry has been admitted to an intensive care unit (ICU) as a result of the Covid 19 virus. There are four values within the dataset:

- Yes
- No
- Missing
- Unknown

Based on the logical tests carried out conversion from "Missing" to No is appropriate as this most likely indicates a No as a result of human error (leaving the area blank instead of marking). No such inference can be made for the Unknown values and as such removal of those rows from the dataset is recommended

medcond_yn:
medcond_yn indicates if the entry has a pre-existing medical condition and has 4 possible values:

- Yes
- No
- Unknown
- Missing

Upon examination of the form and the trend with Missing values in the other categories it would be safe to infer that missing in this features is a result of human error, leaving the area blank as a no instead of ticking the No box. Therefore conversion of missing values to no would be appropriate.

No such inference can be made for the Unknown values and as these will no provide any predictive use towards the target feature (Unknown medical condition is an error in data collection not a valid status so any predictions or trends observed with it and the target condition are meaningless) it is recommended that these rows be dropped from the dataset

4.1 Bar charts:

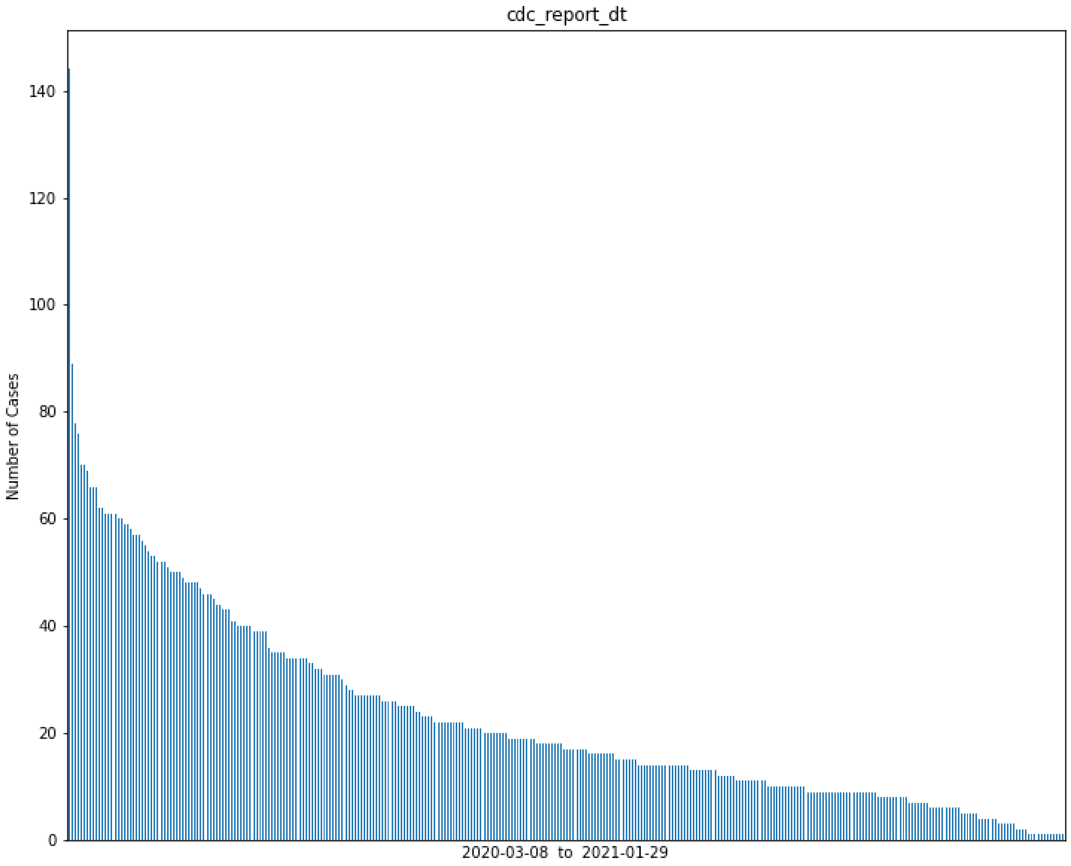
All bar charts for the categorical features can be found in the appendix at the end of this report

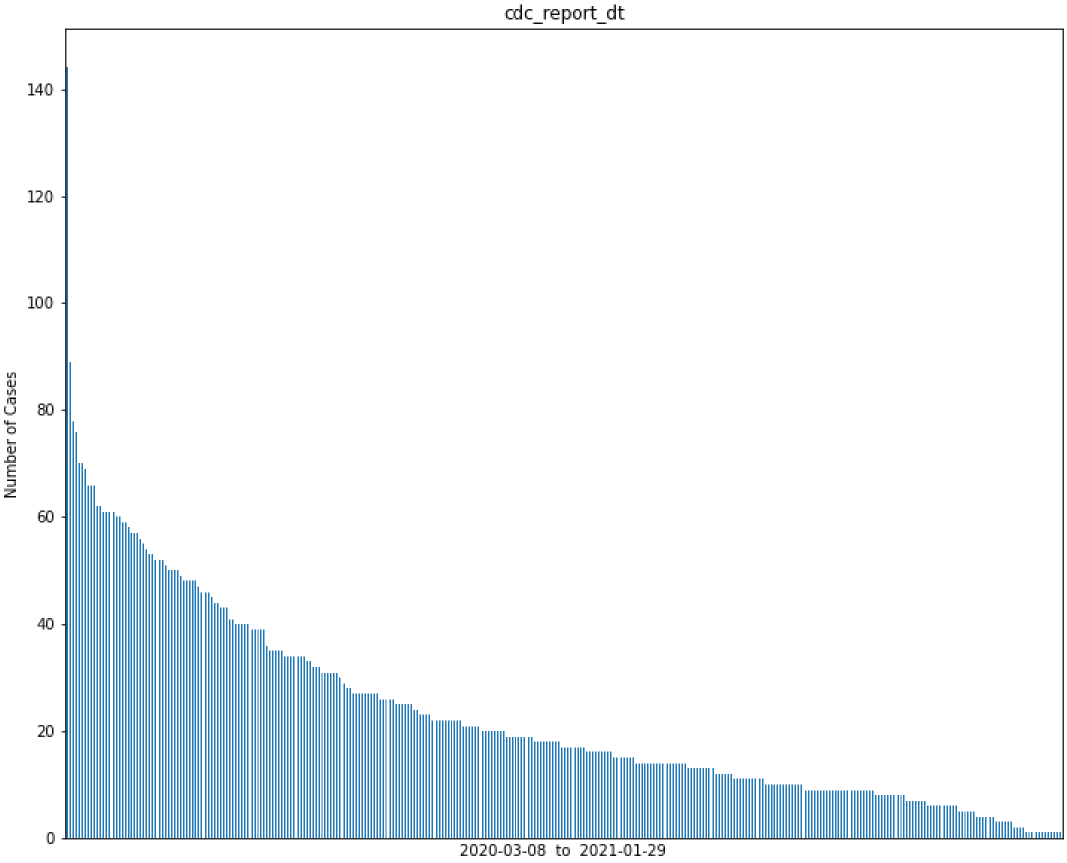
5 Accompanying Media:

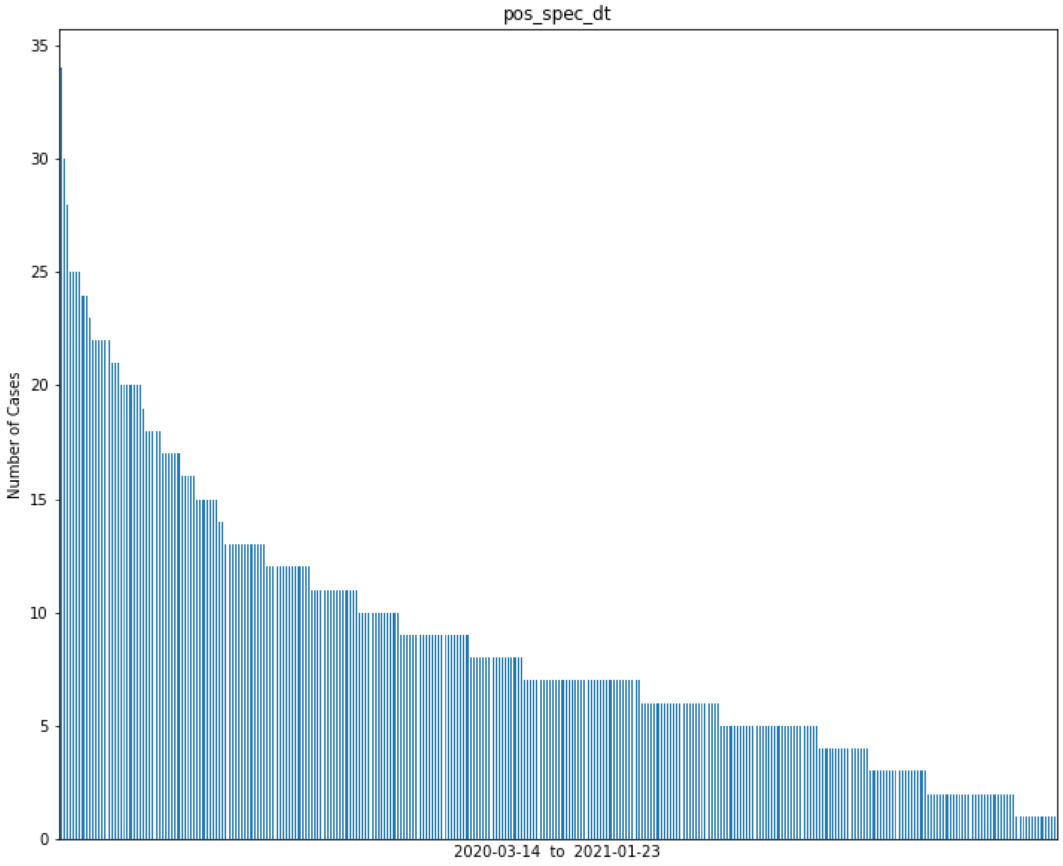
Form used for data collection is included in the folder (pui-form-1.pdf)

APPENDIX

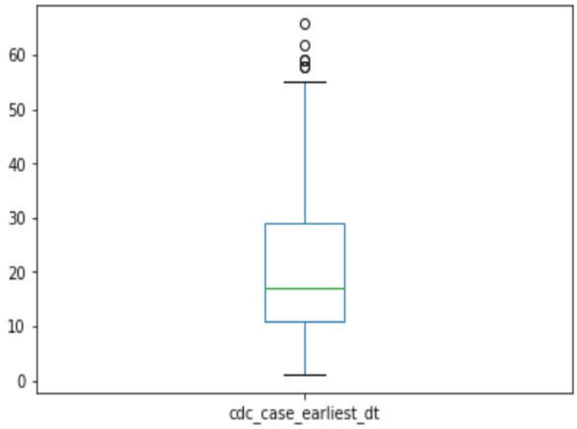
A) Continuous Histograms:

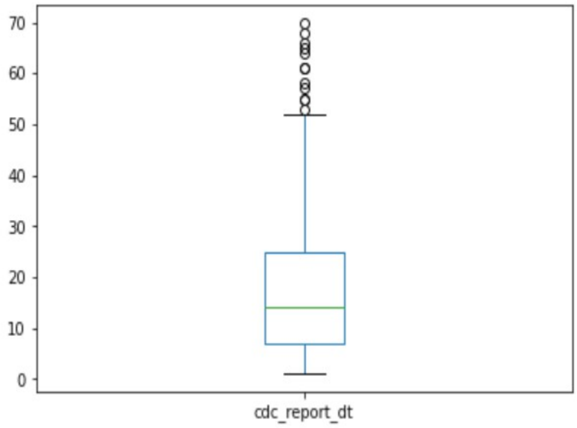


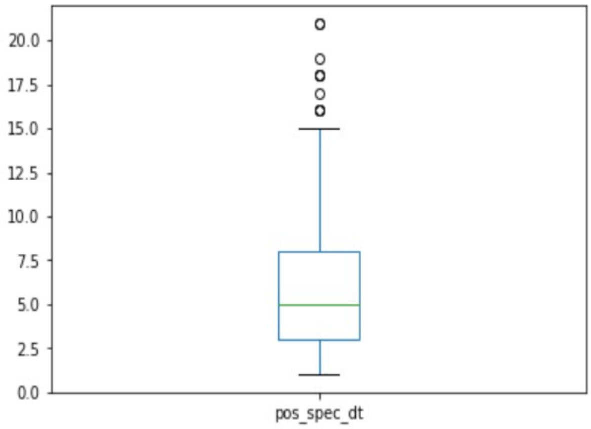

cdc_report_dt
Number of Cases
0
20
40
60
80
100
120
140
2020-03-08 to 2021-01-29

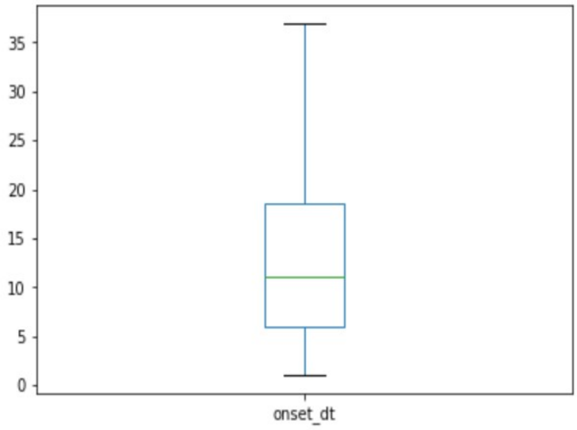

pos_spec_dt
Number of Cases
35
30
25
20
15
10
5
0
2020-03-14 to 2021-01-23

B) Continuous Boxplots:

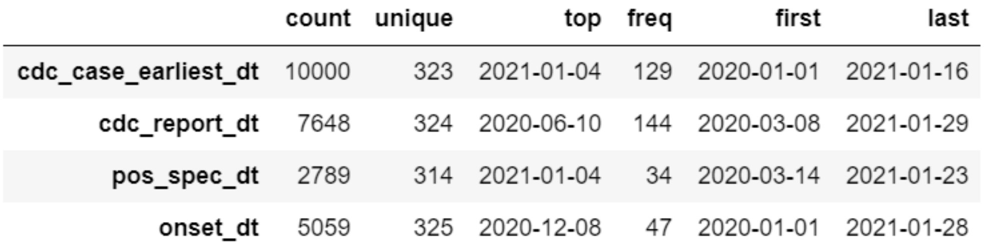






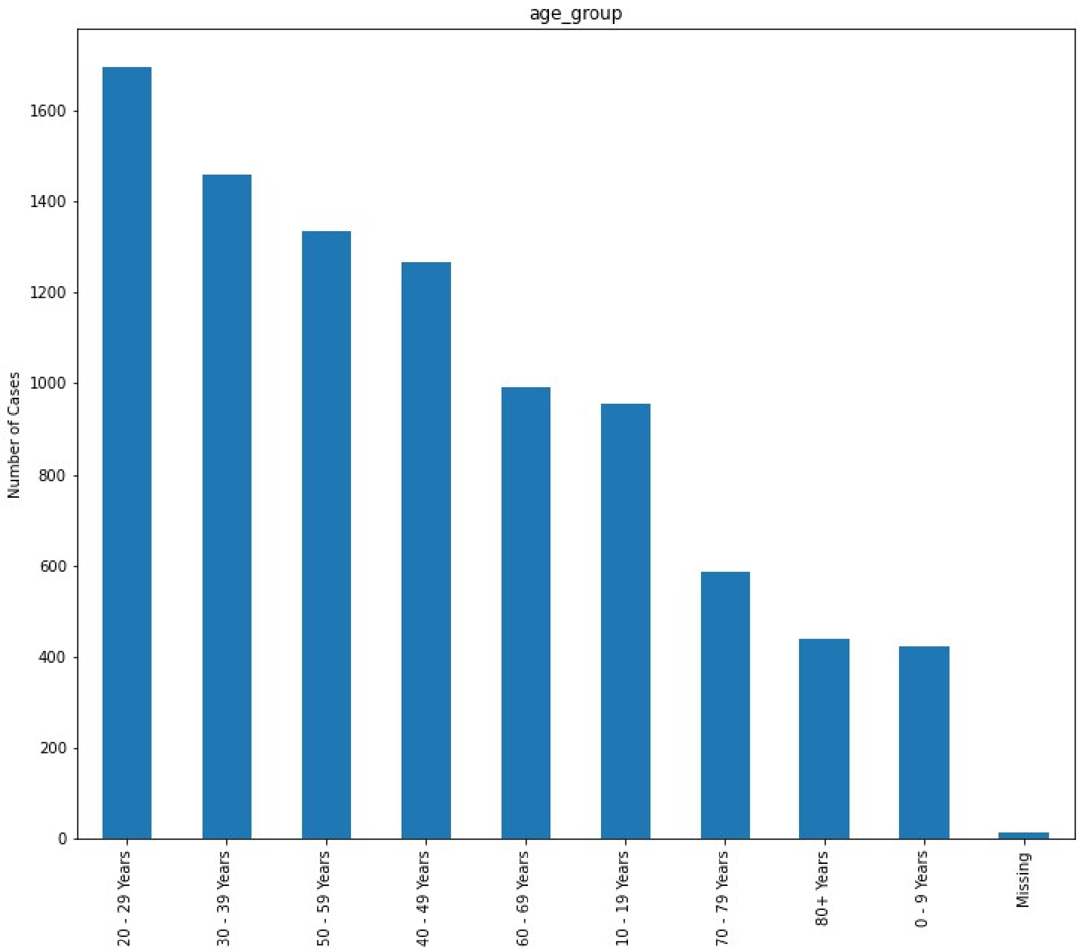


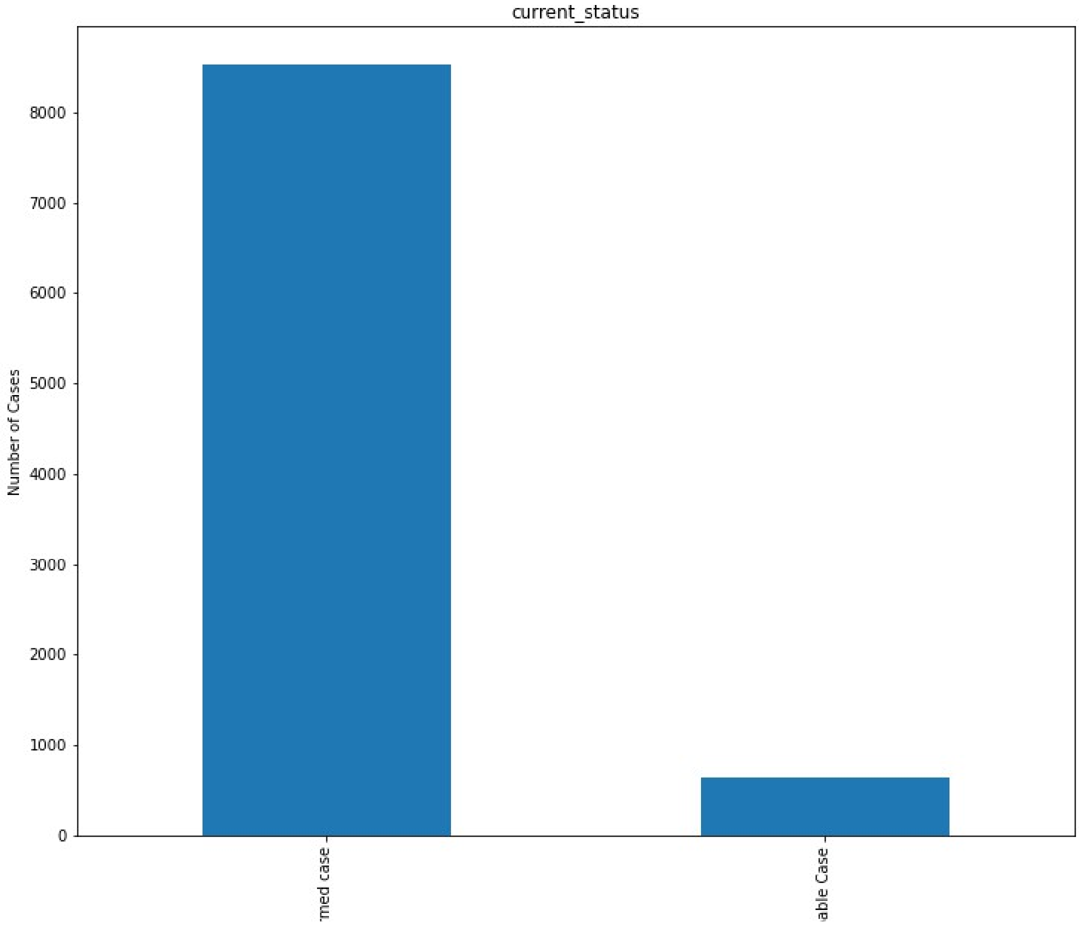
C) Descriptive statistics for Continuous

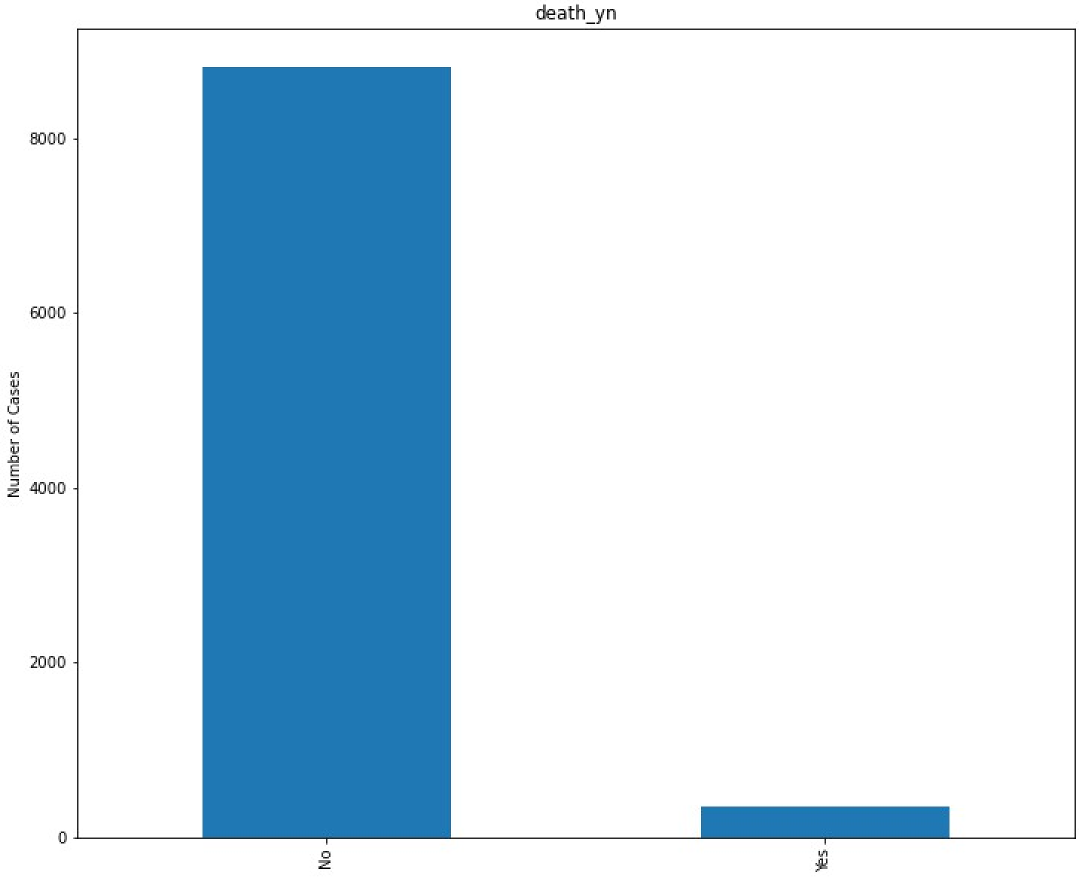


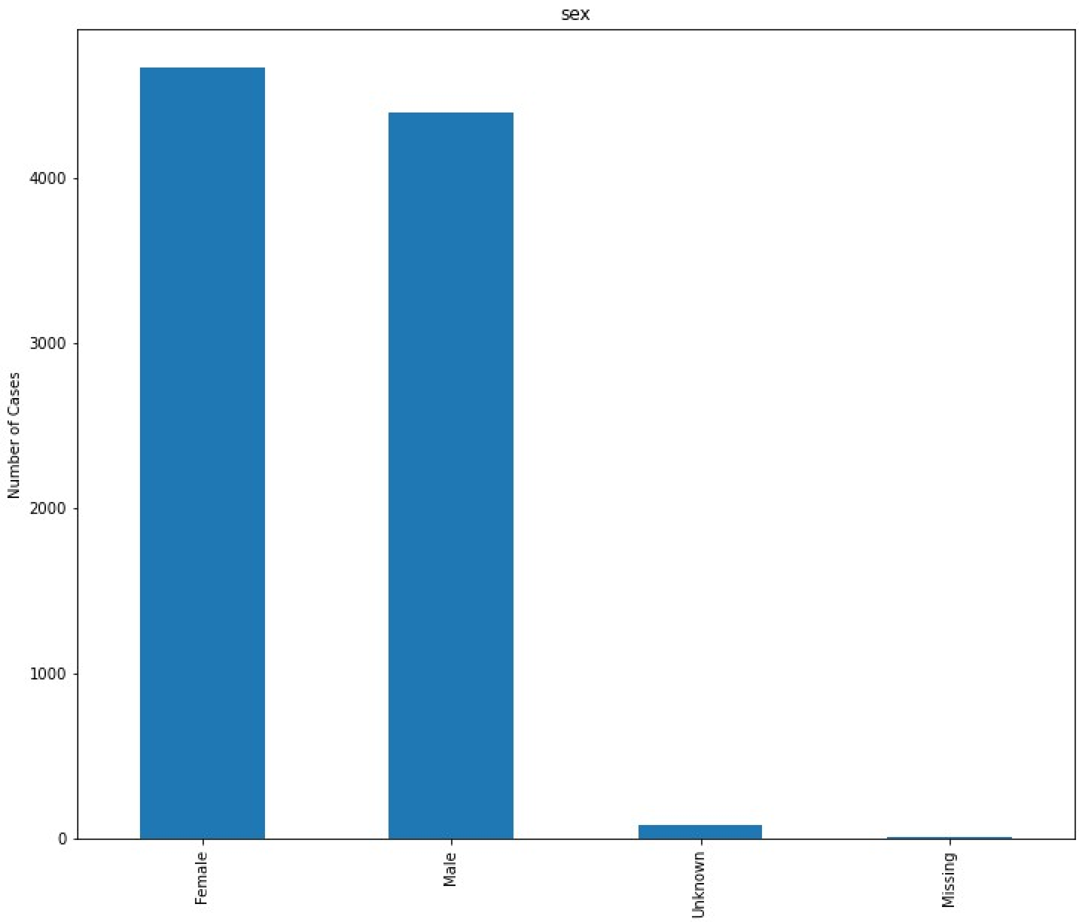
	count	unique	top	freq	first	last
cdc_case_earliest_dt	10000	323	2021-01-04	129	2020-01-01	2021-01-16
cdc_report_dt	7648	324	2020-06-10	144	2020-03-08	2021-01-29
pos_spec_dt	2789	314	2021-01-04	34	2020-03-14	2021-01-23
onset_dt	5059	325	2020-12-08	47	2020-01-01	2021-01-28

D) Categorical Barcharts:

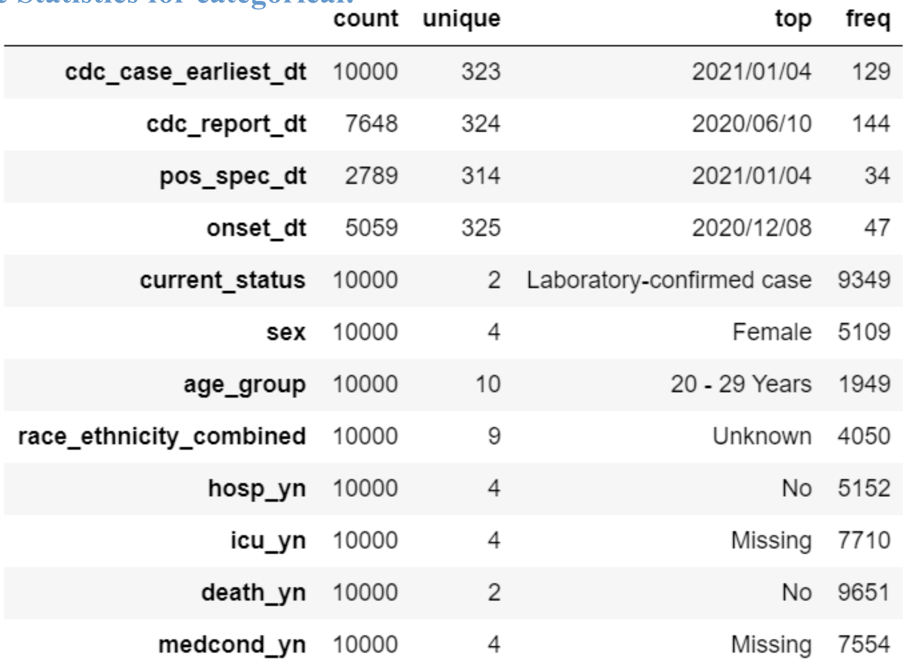



current_status
Number of Cases
0
1000
2000
3000
4000
5000
6000
7000
8000
'med case
able Case


death_yn
Number of Cases
0
2000
4000
6000
8000
No
Yes



E) Descriptive Statistics for categorical:



	count	unique	top	freq
cdc_case_earliest_dt	10000	323	2021/01/04	129
cdc_report_dt	7648	324	2020/06/10	144
pos_spec_dt	2789	314	2021/01/04	34
onset_dt	5059	325	2020/12/08	47
current_status	10000	2	Laboratory-confirmed case	9349
sex	10000	4	Female	5109
age_group	10000	10	20 - 29 Years	1949
race_ethnicity_combined	10000	9	Unknown	4050
hosp_yn	10000	4	No	5152
icu_yn	10000	4	Missing	7710
death_yn	10000	2	No	9651
medcond_yn	10000	4	Missing	7554