## Data Quality Report:

The dataset contains a list of COVID-19 cases provided by the CDC, with 12 variables per case, including the target variable, death status.

There are 10,000 rows and 12 columns.

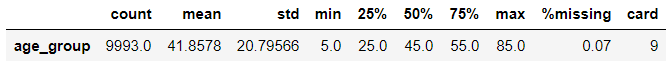
I changed the date columns (the first 4 columns) from strings to datetime objects. I treated these as categorical in the tables and plots, but I didn’t want to remove the exact dates from the dataset, so I kept the data as datetime objects but plotted a bar chart with months on the x axis.

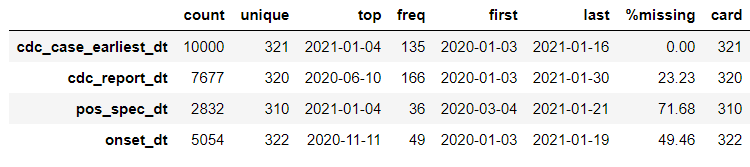
I changed age\_group to continuous, by replacing strings with ints, e.g. “20-29 Years” to ‘25’.

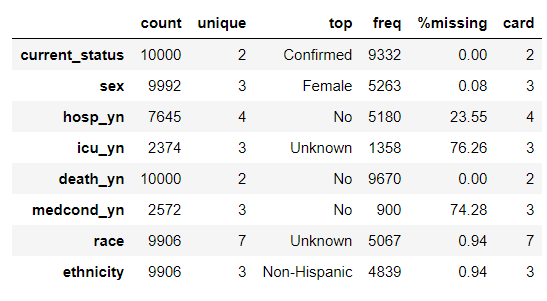
I split the race\_ethnicity\_combined column into two columns, race and ethnicity.

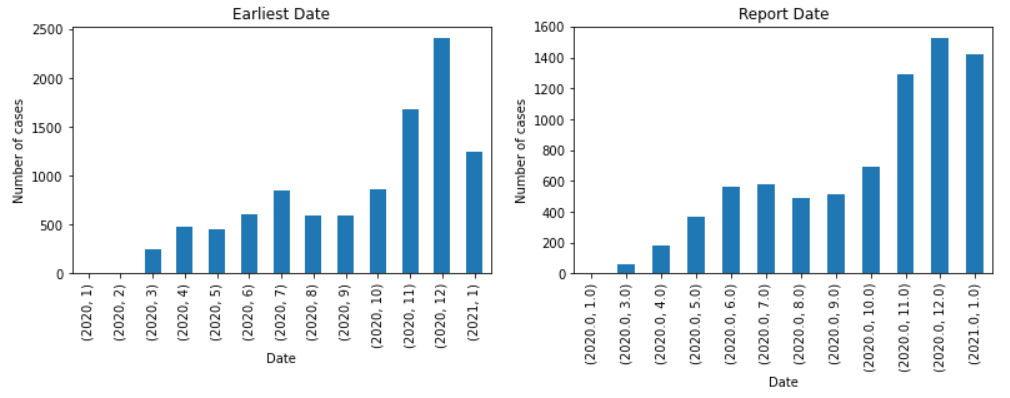
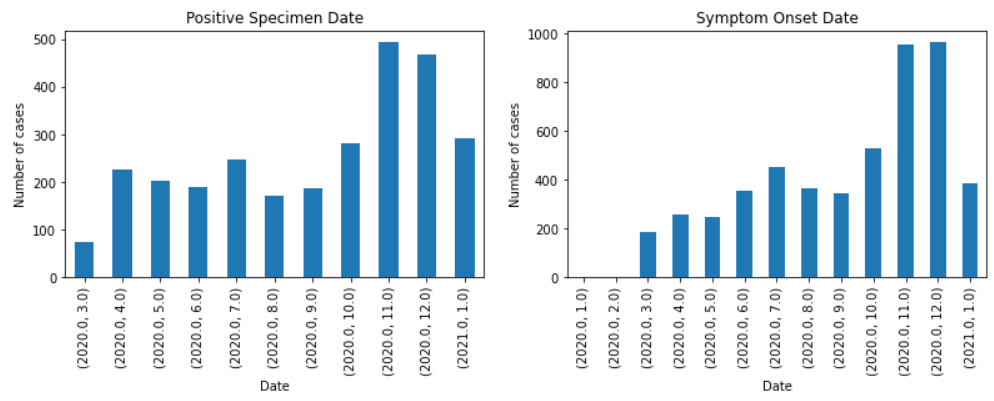
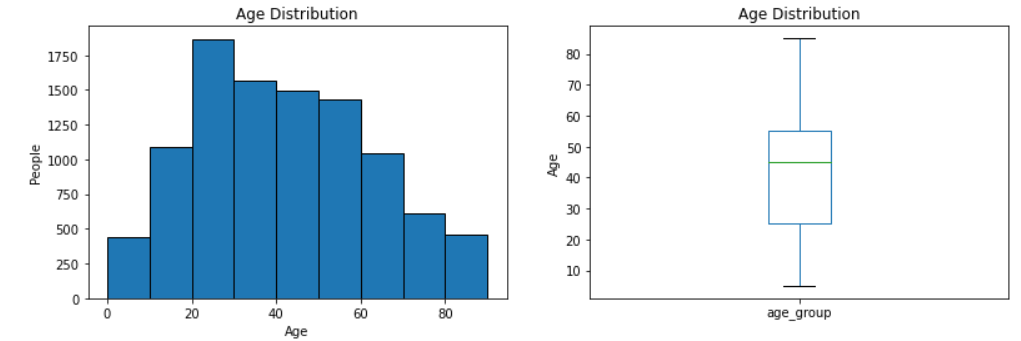
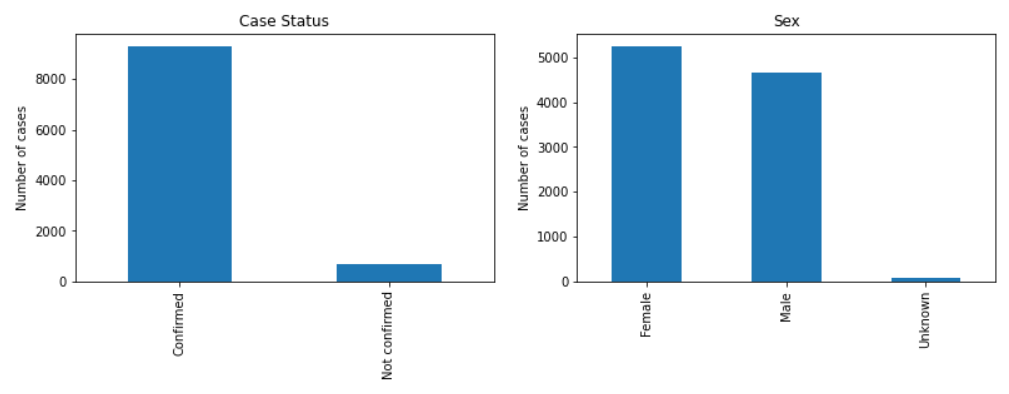
I replaced all missing values with None, but kept the ‘Unknown’ values as this is valid data.

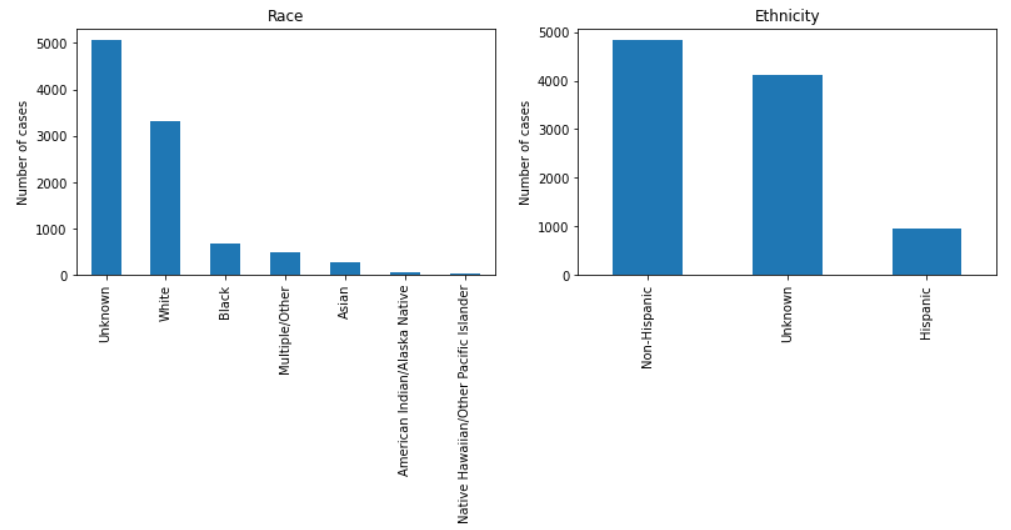
Then I produced the following tables and graphs.

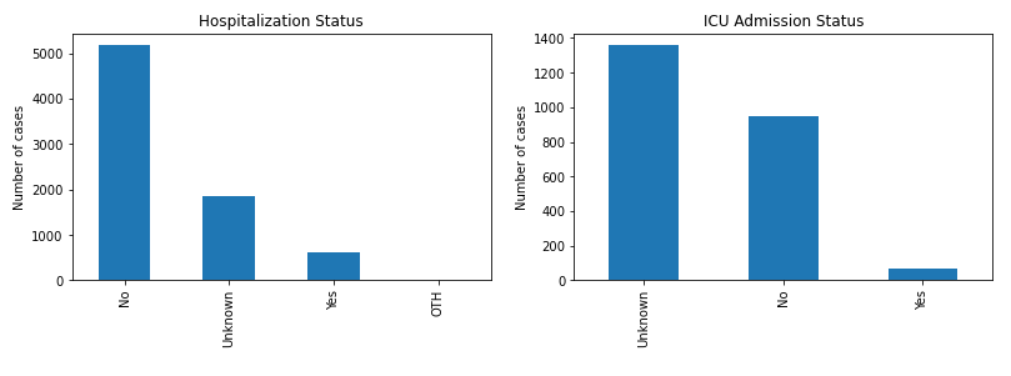


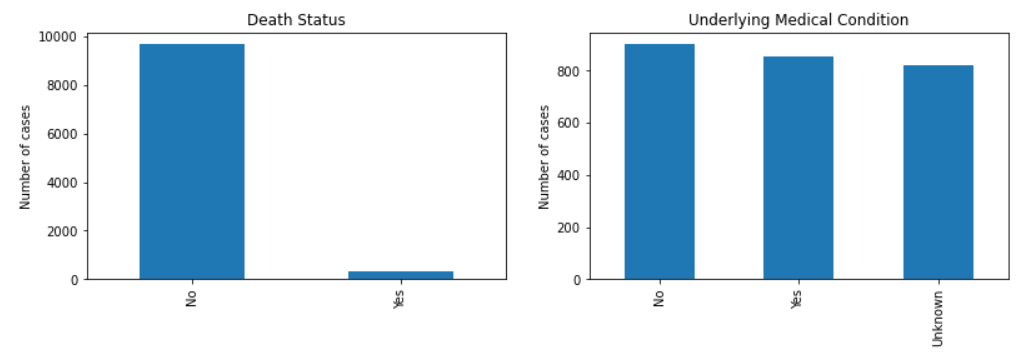










There were 3 features with no missing values, earliest date, current status and death. These features won’t require more handling.

There were 4 features with less than 1% missing values, age, sex, race and ethnicity. These features won’t require much more handling.

There were 3 features missing over 70% of their values, positive specimen date, ICU admission status and underlying medical condition. Onset date is missing 49%, and both hospitalization status and CDC report date are missing 23%. These features will need to be dealt with, particularly the ones with over 70%, which are likely to be dropped.

Hospitalization had 2 rows with the value “OTH”, which I think means “other hospital”, but it doesn’t really matter as it only appears twice.

There are 442 duplicate rows, all of them had ‘No’ for death status and 438 of them were laboratory confirmed cases. As there is no way to know for sure if these duplicates are errors or not, I decided to keep them.

There were no constant columns.

In my code, I originally set the correct data types at the start and then saved the dataframe as cleaned\_v1.csv, but when I imported it, the data types were reset back to object, so I had to set the data types after importing the cleaned\_1.csv instead.