## 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.

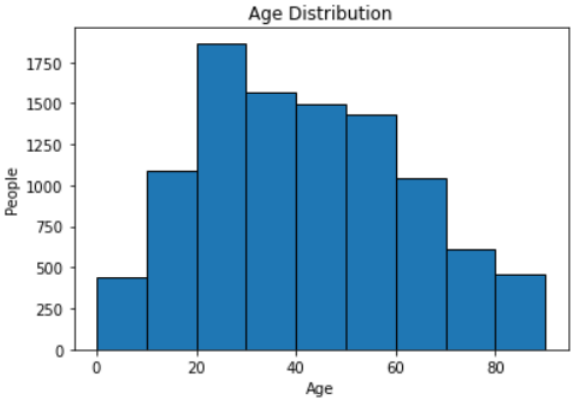
I replaced all missing values with None, but kept the 'Unknown' values as this is valid data.

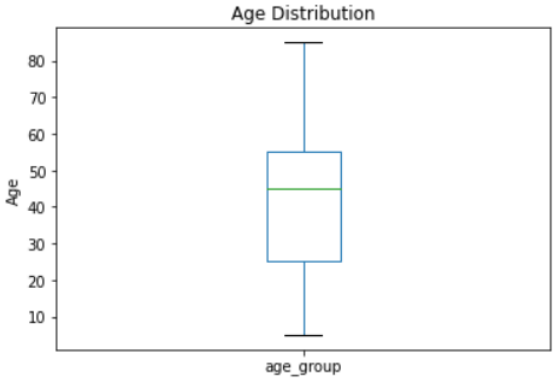
Then I produced the following tables and graphs.

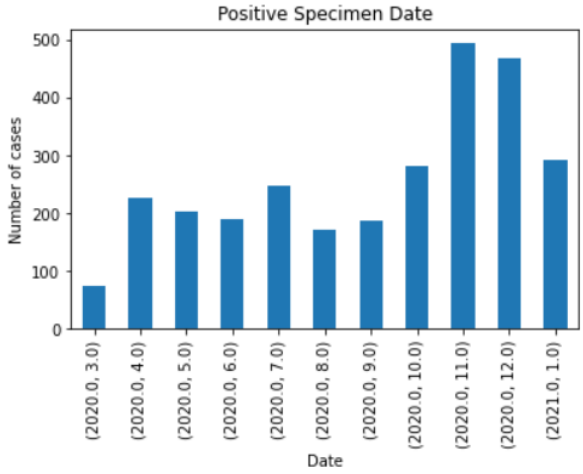
| | count | mean | std | min | 25% | 50% | 75% | max | %missing | card |
|---|---|---|---|---|---|---|---|---|---|---|
| **age_group** | 9993.0 | 41.8578 | 20.79566 | 5.0 | 25.0 | 45.0 | 55.0 | 85.0 | 0.07 | 9 |

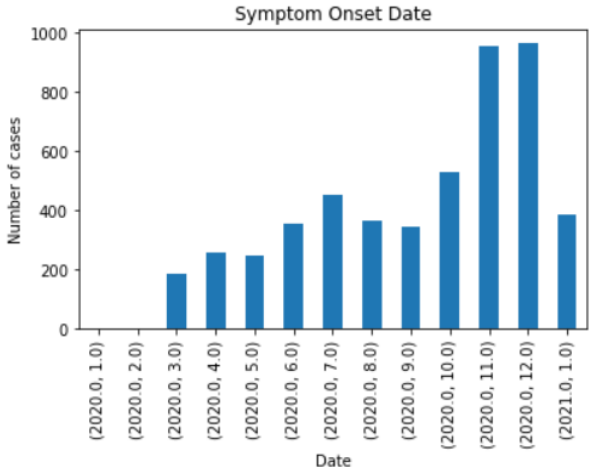
| | count | unique | top | freq | first | last | %missing | card |
|---|---|---|---|---|---|---|---|---|
| **cdc_case_earliest_dt** | 10000 | 321 | 2021-01-04 | 135 | 2020-01-03 | 2021-01-16 | 0.00 | 321 |
| **cdc_report_dt** | 7677 | 320 | 2020-06-10 | 166 | 2020-01-03 | 2021-01-30 | 23.23 | 320 |
| **pos_spec_dt** | 2832 | 310 | 2021-01-04 | 36 | 2020-03-04 | 2021-01-21 | 71.68 | 310 |
| **onset_dt** | 5054 | 322 | 2020-11-11 | 49 | 2020-01-03 | 2021-01-19 | 49.46 | 322 |

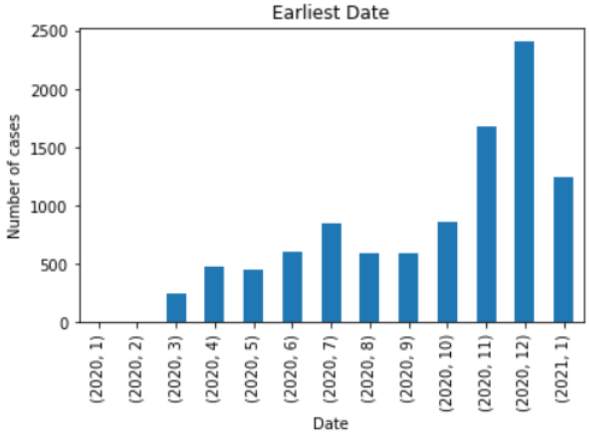
| | count | unique | top | freq | %missing | card |
|---|---|---|---|---|---|---|
| **current_status** | 10000 | 2 | Confirmed | 9332 | 0.00 | 2 |
| **sex** | 9992 | 3 | Female | 5263 | 0.08 | 3 |
| **hosp_yn** | 7645 | 4 | No | 5180 | 23.55 | 4 |
| **icu_yn** | 2374 | 3 | Unknown | 1358 | 76.26 | 3 |
| **death_yn** | 10000 | 2 | No | 9670 | 0.00 | 2 |
| **medcond_yn** | 2572 | 3 | No | 900 | 74.28 | 3 |
| **race** | 9906 | 7 | Unknown | 5067 | 0.94 | 7 |
| **ethnicity** | 9906 | 3 | Non-Hispanic | 4839 | 0.94 | 3 |

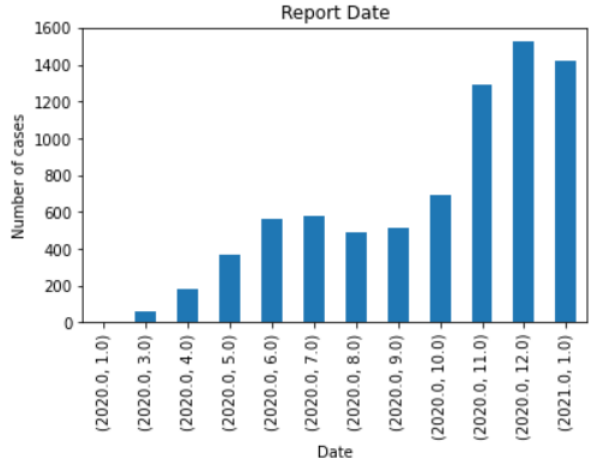

Age Distribution
People
Age
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80

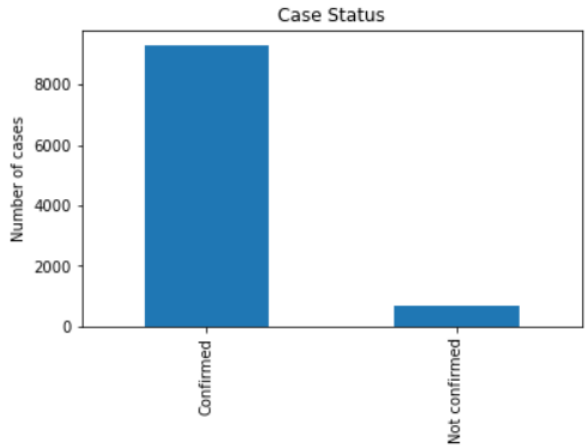

Age Distribution
Age
age_group
10
20
30
40
50
60
70
80

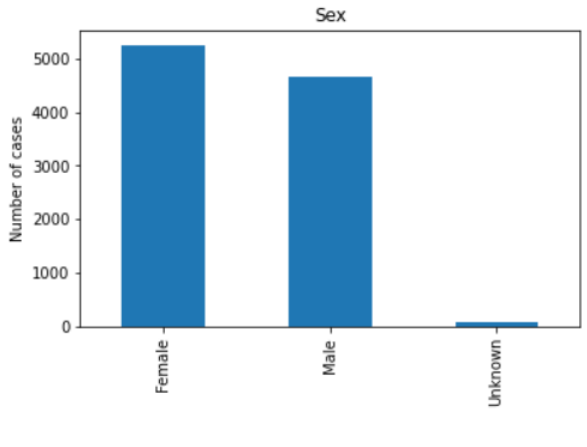

Positive Specimen Date
Number of cases
Date
0
100
200
300
400
500
(2020.0, 3.0)
(2020.0, 4.0)
(2020.0, 5.0)
(2020.0, 6.0)
(2020.0, 7.0)
(2020.0, 8.0)
(2020.0, 9.0)
(2020.0, 10.0)
(2020.0, 11.0)
(2020.0, 12.0)
(2021.0, 1.0)

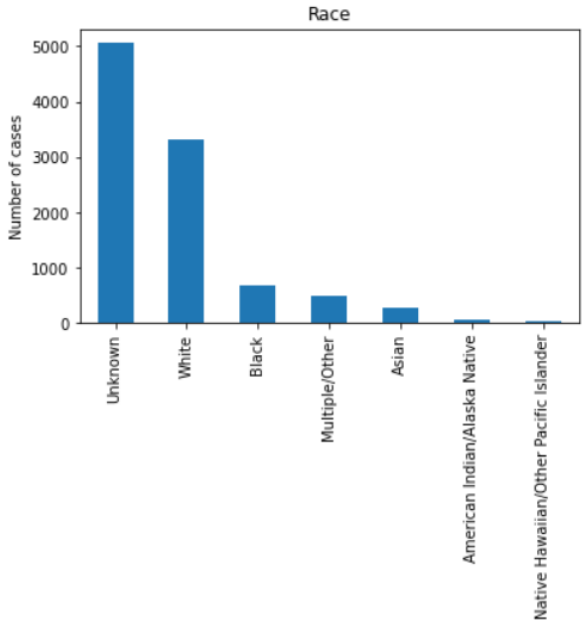

Symptom Onset Date
Number of cases
Date
0
200
400
600
800
1000
(2020.0, 1.0)
(2020.0, 2.0)
(2020.0, 3.0)
(2020.0, 4.0)
(2020.0, 5.0)
(2020.0, 6.0)
(2020.0, 7.0)
(2020.0, 8.0)
(2020.0, 9.0)
(2020.0, 10.0)
(2020.0, 11.0)
(2020.0, 12.0)
(2021.0, 1.0)

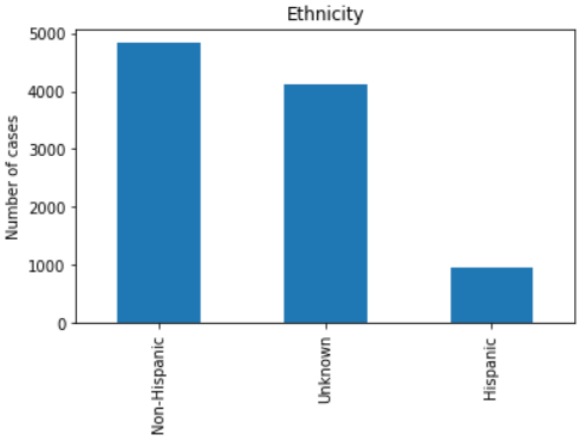

Earliest Date
Number of cases
Date
0
500
1000
1500
2000
2500
(2020, 1)
(2020, 2)
(2020, 3)
(2020, 4)
(2020, 5)
(2020, 6)
(2020, 7)
(2020, 8)
(2020, 9)
(2020, 10)
(2020, 11)
(2020, 12)
(2021, 1)

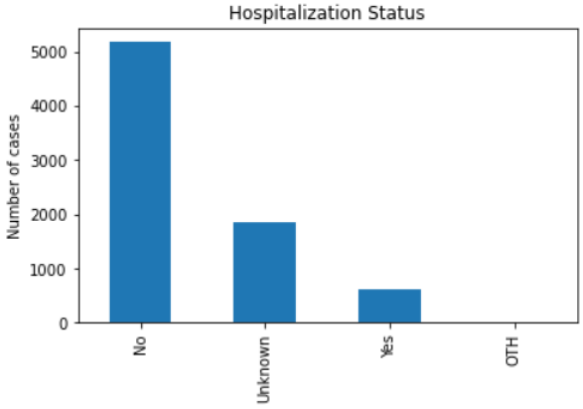

Report Date
Number of cases
Date
0
200
400
600
800
1000
1200
1400
1600
(2020.0, 1.0)
(2020.0, 3.0)
(2020.0, 4.0)
(2020.0, 5.0)
(2020.0, 6.0)
(2020.0, 7.0)
(2020.0, 8.0)
(2020.0, 9.0)
(2020.0, 10.0)
(2020.0, 11.0)
(2020.0, 12.0)
(2021.0, 1.0)

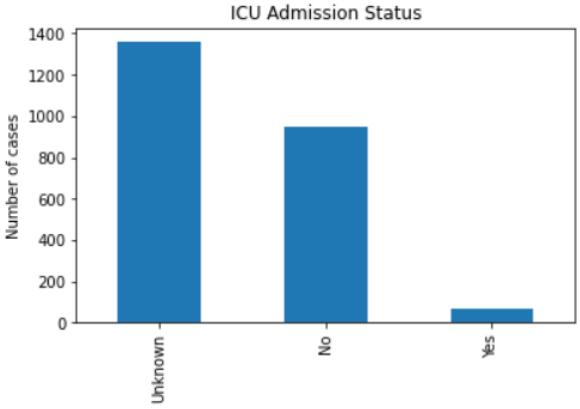

Case Status
Number of cases
0
2000
4000
6000
8000
Confirmed
Not confirmed

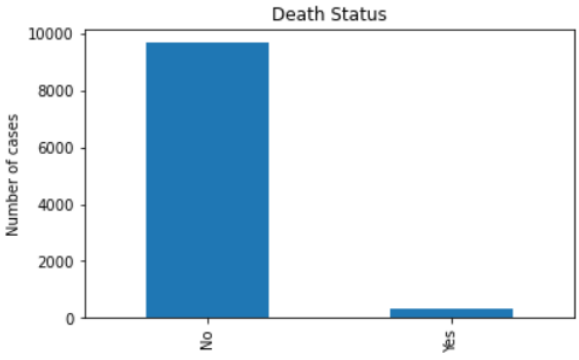

Sex
Number of cases
0
1000
2000
3000
4000
5000
Female
Male
Unknown

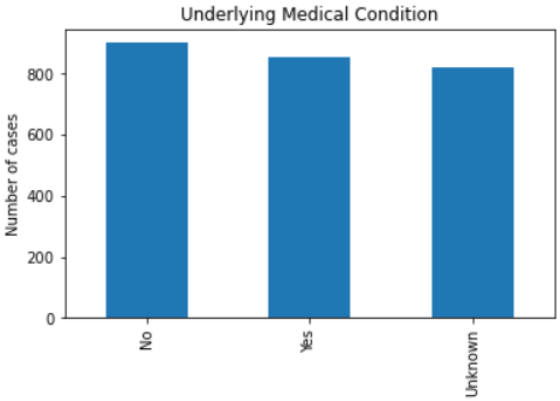

Race
Number of cases
0
1000
2000
3000
4000
5000
Unknown
White
Black
Multiple/Other
Asian
American Indian/Alaska Native
Native Hawaiian/Other Pacific Islander


Ethnicity
Number of cases
0
1000
2000
3000
4000
5000
Non-Hispanic
Unknown
Hispanic


Hospitalization Status
Number of cases
0
1000
2000
3000
4000
5000
No
Unknown
Yes
OTH


ICU Admission Status
Number of cases
0
200
400
600
800
1000
1200
1400
Unknown
No
Yes


Death Status
Number of cases
0
2000
4000
6000
8000
10000
No
Yes


Underlying Medical Condition
Number of cases
0
200
400
600
800
No
Yes
Unknown

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.