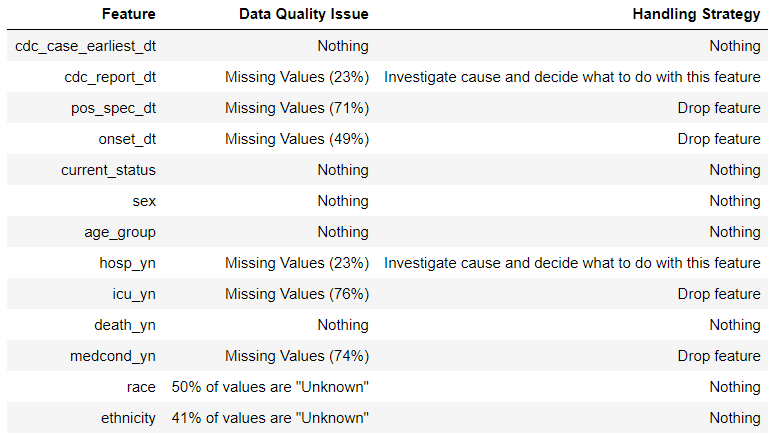
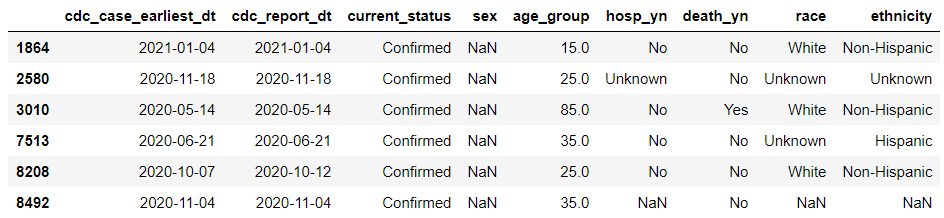
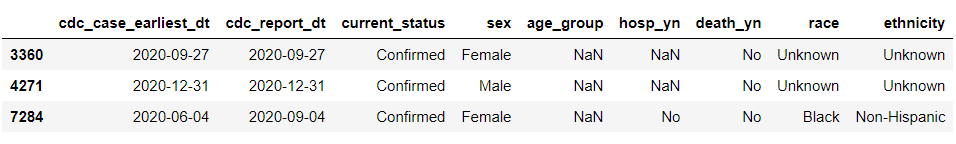
## Data Quality Plan:



I didn’t change anything for the features, earliest date, current status, death, age, sex, race or ethnicity. For race and ethnicity, the high percent of “Unknown” values is of concern, but I didn’t see any way to improve this. The following tables show the rows with missing values for sex and age.





I decided to keep these rows, as it wouldn’t make any significant difference if I removed them, and they will be used in plots involving age or sex.

I decided to drop the features with over 70% missing values (positive specimen date, ICU admission, underlying medical condition) and the one with 49% missing values (symptom onset date) as these percentages are just too high for these features to be of any real use. Around 50% of the missing ICU values could have been inferred from the hospitalization values, assuming a “No” for hospitalization implies a “No” for ICU admission, but this would be pointless as the information is already contained in the hospitalization data.

I kept the CDC report date and hospitalization status features, as I think even with 23% missing values, they can still be useful.

I dropped the rows where the earliest date was greater than the report date, as this is a logical error.

I dropped the 668 rows with “Not confirmed” for current status, as I thought it would be better to work with only confirmed cases.