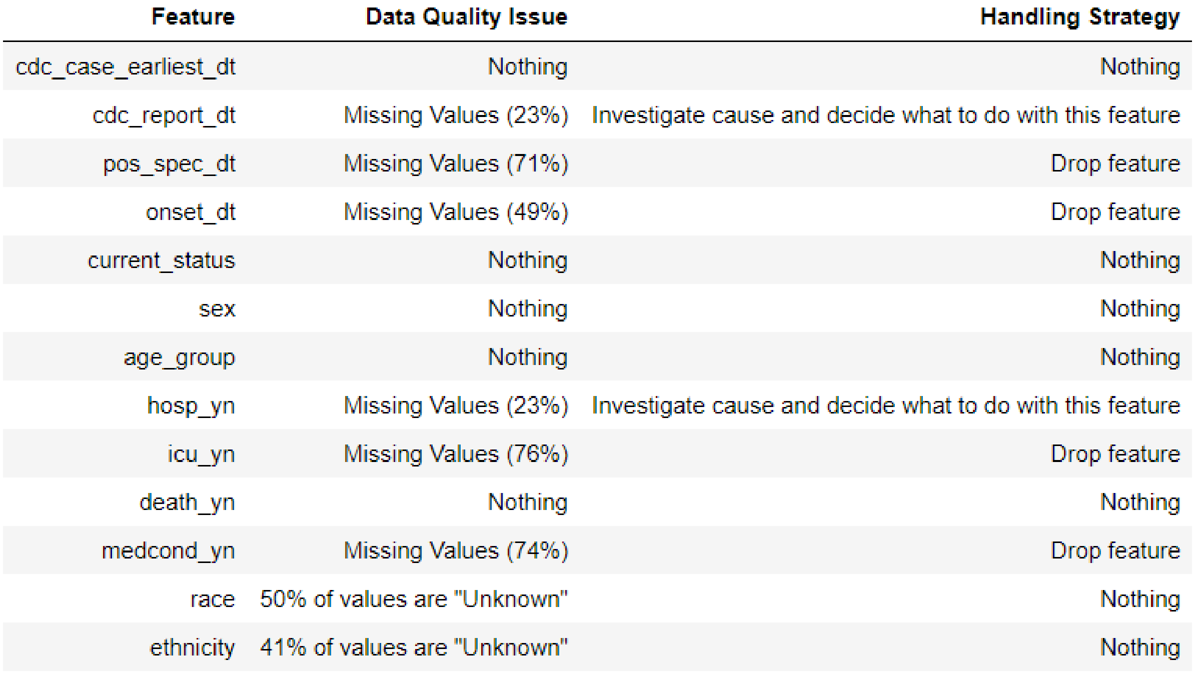
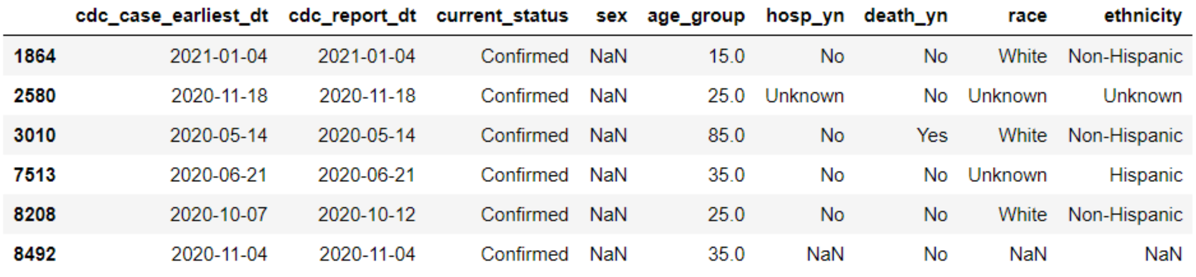
Data Quality Plan:

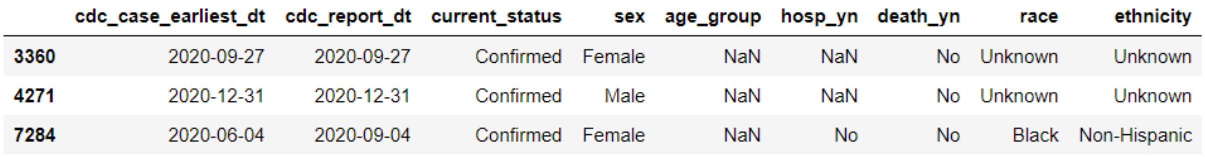


Feature	Data Quality Issue	Handling Strategy
cdc_case_earliest_dt	Nothing	Nothing
cdc_report_dt	Missing Values (23%)	Investigate cause and decide what to do with this feature
pos_spec_dt	Missing Values (71%)	Drop feature
onset_dt	Missing Values (49%)	Drop feature
current_status	Nothing	Nothing
sex	Nothing	Nothing
age_group	Nothing	Nothing
hosp_yn	Missing Values (23%)	Investigate cause and decide what to do with this feature
icu_yn	Missing Values (76%)	Drop feature
death_yn	Nothing	Nothing
medcond_yn	Missing Values (74%)	Drop feature
race	50% of values are "Unknown"	Nothing
ethnicity	41% of values are "Unknown"	Nothing

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.



	cdc_case_earliest_dt	cdc_report_dt	current_status	sex	age_group	hosp_yn	death_yn	race	ethnicity
1864	2021-01-04	2021-01-04	Confirmed	NaN	15.0	No	No	White	Non-Hispanic
2580	2020-11-18	2020-11-18	Confirmed	NaN	25.0	Unknown	No	Unknown	Unknown
3010	2020-05-14	2020-05-14	Confirmed	NaN	85.0	No	Yes	White	Non-Hispanic
7513	2020-06-21	2020-06-21	Confirmed	NaN	35.0	No	No	Unknown	Hispanic
8208	2020-10-07	2020-10-12	Confirmed	NaN	25.0	No	No	White	Non-Hispanic
8492	2020-11-04	2020-11-04	Confirmed	NaN	35.0	NaN	No	NaN	NaN



	cdc_case_earliest_dt	cdc_report_dt	current_status	sex	age_group	hosp_yn	death_yn	race	ethnicity
3360	2020-09-27	2020-09-27	Confirmed	Female	NaN	NaN	No	Unknown	Unknown
4271	2020-12-31	2020-12-31	Confirmed	Male	NaN	NaN	No	Unknown	Unknown
7284	2020-06-04	2020-09-04	Confirmed	Female	NaN	No	No	Black	Non-Hispanic

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.