## Summary of data quality plan:

| Variable Names | Data Quality Issue | Handling Strategy |
|---|---|---|
| cdc_case_earliest_dt | Some dates earlier than earliest date, including cdc_report_dt (32 rows) and pos_spec_dt (46 rows) | Carry out general integrity tests and imputted data when cdc_case_earliest_dt was not the earliest date |
| cdc_report_dt | Missing Values (22.61%) and Depreciated column | We used this column for comparison in integrity checks and then dropped as depreciated column according to CDC guidelines |
| pos_spec_dt | Missing Values (71.74%) | We used this column for comparison in integrity checks and then dropped due to large missing values |
| onset_dt | Missing Values(48.92%) | Used column for general integrity checks. Not useful for target class, dropped due to large missing values |
| current_status | None | Do Nothing |
| sex | Missing (0.12%) Unknown (0.84%) | Since very few Missing & Unknowns, rows were removed |
| age_group | Missing (0.13%) | Since very few Missing, rows were removed |
| race_ethnicity_combined | Missing (0.94%) Unknown (39.91%) | Combined into single Unknown value and keep |
| hosp_yn | Missing(23.5%) Unknown (17.3%) | Combine into single Unknown value and keep |
| icu_yn | Missing (76.35%) Unknown (13.14%) | Infer values from hosp_yn where possible (1 row), combine into single Unknown value and keep |
| medcond_yn | Missing (74.64%) Unknown (7.8%) | Combine into single Unknown value and keep |
| death_yn | None | Do Nothing |