Data Quality Plan_22200374

Summary of data quality plan:

Variable Names	Data Quality Issue	Handling Strategy
case_positive_specimen_interval	Outliers	Do Nothing
case_positive_specimen_interval	Negative Value (56 rows)	Take its absolute value
case_positive_specimen_interval	When 'current_status' = 'Laboratory-confirmed case','case_positive_specimen_interval' = 'NaN'(7697 rows)	Replace missing values with median value 0
case_onset_interval	Outliers	Do Nothing
case_onset_interval	Negative Value (285 rows)	Take its absolute value
case_onset_interval	When 'symptom_status' = 'Symptomatic', 'case_onset_interval' = 'NaN'(665 rows)	Replace missing values with median value 0
res_county	'NaN' Value	Replace with 'Unknown'
county_fips_code	'NaN' Value	Replace with 'Unknown'
age_group	'NaN' Value and 'Missing' Value	Replace with 'Unknown'
sex	'NaN' Value and 'Missing' Value	Replace with 'Unknown'
race	'NaN' Value and 'Missing' Value	Replace with 'Unknown'
ethnicity	'NaN' Value and 'Missing' Value	Replace with 'Unknown'
process	The missing rate is too high	Drop this column
exposure_yn	The missing rate is too high	Drop this column
symptom_status	'Missing' Value	Replace with 'Unknown'
symptom_status	When 'death_yn' = 'Yes', 'symptom_status' = 'NaN' or 'Missing' or 'Unknown'	Replace with 'Symptomatic'
hosp_yn	'Missing' Value	Replace with 'Unknown'
icu_yn	'Missing' Value	Replace with 'Unknown'
icu_yn	When 'death_yn' = 'No' and 'hosp_yn'= 'No', 'icu_yn' = 'NaN'	Replace missing values with 'No'
underlying_conditions_yn	The missing rate is too high	Drop this column
Failing Rows	Duplicates(1052rows)	Drop duplicated rows