Data Quality Plan

Variable names	Data qulity issue	Handing Strategy
res_county	too many Sparse Data	delete data below 30
county_fips_code	too many Sparse Data	delete data below 30
age_group	too many ambiguous data	remove them when drawing plot
sex	too many ambiguous data	remove them when drawing plot
race	too many ambiguous data	remove them when drawing plot
ethnicity	too many ambiguous data	remove them when drawing plot
case_positive_specimen_interval	nagetive value(62 rows)	take absolute values
case_onset_interval	nagetive value(291 rows)	take absolute values
process	too many ambiguous data	remove them when drawing plot
exposure_yn	over 90% data are"Missing" while rest is "Yes"	delete it from dataset
current_status	unsuit data type	change data type to "category"
symptom_status	too many ambiguous data	remove them when drawing plot
hosp_yn	too many ambiguous data	remove them when drawing plot
icu_yn	too many ambiguous data	remove them when drawing plot
death_yn	unsuit data type	change data type to "category"
underlying_conditions_yn	over 90% data are null while rest is "Yes"	delete it from dataset
case_month	unsuit data type	change data type to "category"
res_state	unsuit data type	change data type to "category"
state_fips_code	unsuit data type	change data type to "category"

In [ ]: