Data Quality Plan

Variable Name	Data Quality Issue	Handling Strategy
cdc_report_dt	Deprecated	Drop Feature
pos_spec_dt	Missing Values	Drop Feature
onset_dt	Missing values	Do Nothing
current_status	Primarily 'Laboratory-confirmed case'	Remove 'Probable case rows'
sex	Missing Values	Replace with 'Unknown'
age_group	Missing Values	Drop Entries
race_ethnicity_combined	Missing Values	Replace with 'Unknown'
hosp_yn	Missing Values	Replace with 'Unknown'
icu_yn	Missing Values	Replace with 'Unknown'
medcond_yn	Missing Values	Replace with 'Unknown

I chose to drop the two continuous features as I believe that for building a solution for death risk prediction only the earliest case date and the date on which symptoms occur will be needed.

I chose to replace missing values with 'Unknown' as in most cases the missing values made up a larger proportion of the dataset.

I chose to drop the entries where age_group was missing as they were less than 1% of the dataset and it would be impractical for future analysis to replace them with 'Unknown'.

I chose to drop entries which were probable cases as I believe for building a reliable death risk prediction model only the data from genuine confirmed cases should be used so as not to skew any results.