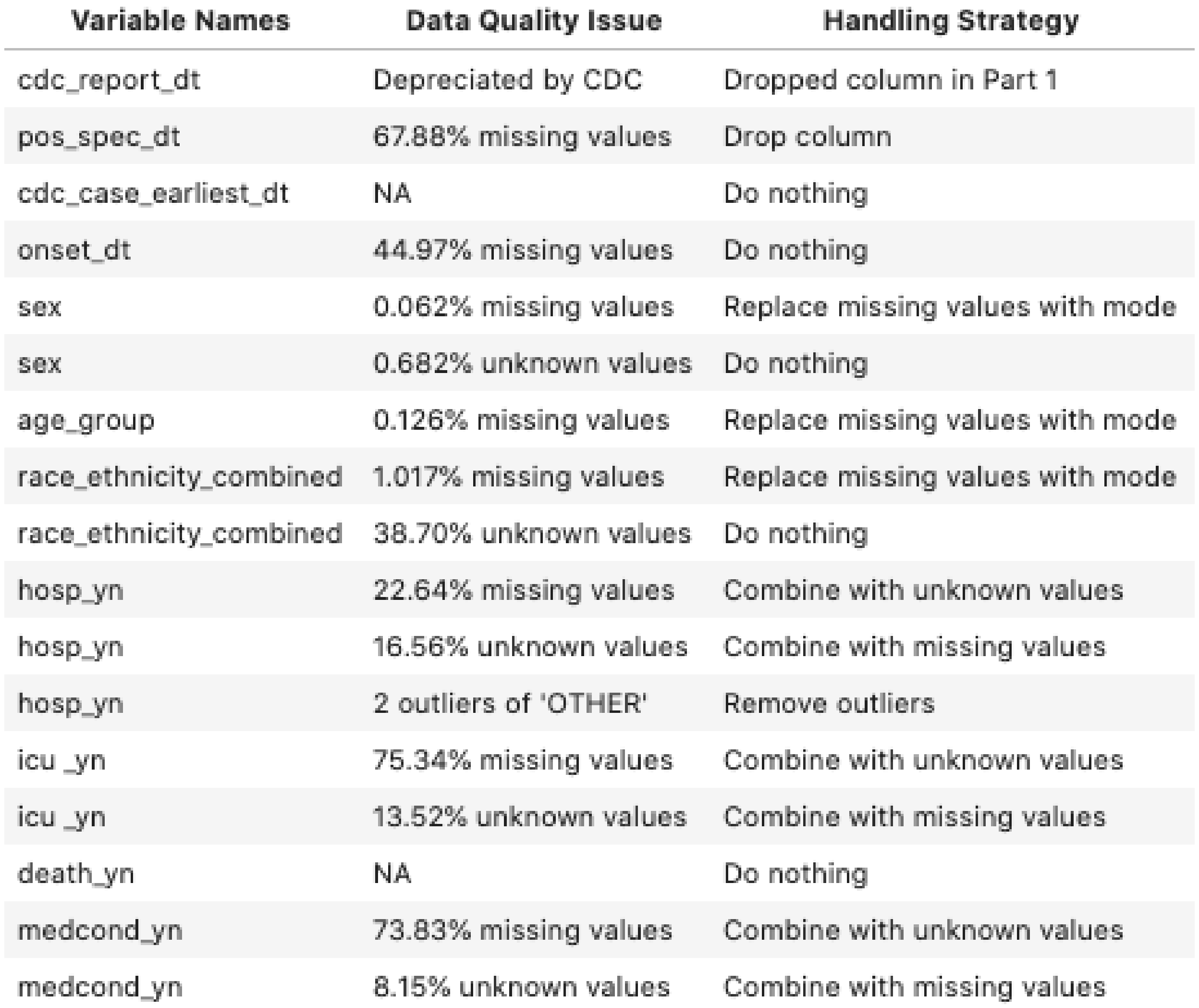
Summary of Data Quality Plan:



Variable Names	Data Quality Issue	Handling Strategy
cdc_report_dt	Depreciated by CDC	Dropped column in Part 1
pos_spec_dt	67.88% missing values	Drop column
cdc_case_earliest_dt	NA	Do nothing
onset_dt	44.97% missing values	Do nothing
sex	0.062% missing values	Replace missing values with mode
sex	0.682% unknown values	Do nothing
age_group	0.126% missing values	Replace missing values with mode
race_ethnicity_combined	1.017% missing values	Replace missing values with mode
race_ethnicity_combined	38.70% unknown values	Do nothing
hosp_yn	22.64% missing values	Combine with unknown values
hosp_yn	16.56% unknown values	Combine with missing values
hosp_yn	2 outliers of 'OTHER'	Remove outliers
icu _yn	75.34% missing values	Combine with unknown values
icu _yn	13.52% unknown values	Combine with missing values
death_yn	NA	Do nothing
medcond_yn	73.83% missing values	Combine with unknown values
medcond_yn	8.15% unknown values	Combine with missing values

Variable Name	Data Quality Issue	Justification
Cdc_report_dt	Depreciated by CDC	Dropped due to depreciation in value by CDC
Pos_spec_dt	67.88% missing values	Dropped due to 67.88% missing values and the pos_spec_dt did not give much insight into the target feature, which is death, so it was determined to be a low value feature.
Onset_dt	44.97% missing values	Decided to keep this feature in the dataset as the feature may bring insight into incubation period of the disease, from test to onset of symptoms. Even though there was a high % of missing values, the onset_dt may give valuable insight into the target feature. Imputation was not an option as a strong bias could be introduced into the dataset with that high a percentage.
sex	0.062% missing values	Replaced values with mode due to low % missing values, that would not severely skew bias in the dataset.
sex	0.682% unknown values	Decided to keep the unknown values as these values may be due to valid data, if a person did not feel they fit into the available categories.
Age_group	0.126% missing values	Replaced values with mode due to low % missing values, that would not severely skew bias in the dataset.
Race_ethnicity_combined	1.017% missing values	Replaced values with mode due to low % missing values, that would not severely skew bias in the dataset.
Race_ethnicity_combined	38.70% unknown values	Decided to keep the unknown values as these values may be due to valid data, if a

		person did not feel they fit into the available categories. Imputation was not an option as a strong bias could be introduced into the dataset with that high a percentage.
Hosp_yn	22.64% missing values and 16.56% unknown values	Decided to combine missing and unknown data into one missing data feature due to the fact that they provide the same information and I decided to keep the missing data in the dataset as this feature is a high value feature, as hospitalisations could be highly correlated to the target feature. Imputation was not an option as a strong bias could be introduced into the dataset with that high a percentage. Removing the missing data could also introduce a strong bias.
Icu_yn	75.34% missing values and 13.52% missing values	Decided to combine missing and unknown data into one missing data feature due to the fact that they provide the same information and I decided to keep the missing data in the dataset as this feature is a high value feature, as ICU admittance could be highly correlated to the target feature. Imputation was not an option as a strong bias could be introduced into the dataset with that high a percentage. Removing the missing data could also introduce a strong bias. This data is also sensitive data and therefore, I believe valid data. I believe that this missing data is missing at random, as I believe there is a correlation between missing ICU data and

		people who were not hospitalised. The PIC form shows a box to the right of the hospitalisation box where extra information is inputted about ICU admittance. It is my belief that the position of the ICU box beside the hospital data meant that people who ticked no for hospitalisation, did not feel the need to tick ICU information and as a result data was missing.
Medcond_yn	73.83% missing values and 8.15% unknown values.	Decided to combine missing and unknown data into one missing data feature due to the fact that they provide the same information and I decided to keep the missing data in the dataset as this feature is a high value feature, as having underlying medical conditions could be highly correlated to the target feature. Imputation was not an option as a strong bias could be introduced into the dataset with that high a percentage. Removing the missing data could also introduce a strong bias. This data is also sensitive data and therefore, I believe valid data. Missing data may be as a result of the fact that people were not comfortable sharing that information and therefore, a high % of missing data was observed.