Data Understanding: Data Quality Plan

A potential data quality plan lists all features and actions involving those features:

Feature	Data Quality Issue	Handling Strategy
case_month (category)	None	Keep as is
res_state (category)	Missing values(0.005%)	Investigate rows where this value appears. If invalid data and only few rows affected, drop rows
state_fips_code (category)	Missing values(0.005%)	Drop this feature because the state code can be replaced by state
res_county (category)	Missing Values (6.42%)	Investigate rows where this value appears, impute as needed
county_fips_code (category)	Missing Values (6.42%)	Investigate rows where this value appears, consider imputation.
county_fips_code (category)	Large cardinality(1208.0)	Leave it as an identifier of small towns of the counties.
age_group (category)	Missing Values (0.99%)	Investigate rows where this value appears, consider imputation.
sex (category)	Missing Values (2.84%)	Investigate rows where this value appears, derive a missing indicator feature from this feature with missing values
race (category)	Missing Values (24.09%)	Investigate rows where this value appears, would derive an new feature with missing values.
ethnicity (category)	Missing Values (31.46%)	Investigate rows where this value appears, impute as needed.
case_positive_specimen_interval (int64)	Missing Values （47.07%）	Although with many missing values, the feature is very important and may affect the following analysis. Investigate rows affected, consider imputation
case_positive_specimen_interval (int64)	Outlier, irregular negative values	Convert negative values to postive one and keep as is. Given that the outliers should be normal case, drop it may distrub the following analysis
case_onset_interval (int64)	Missing values (55.22%)	Although with many missing values, the feature is very important and may affect the following analysis. Investigate rows affected, consider imputation
case_onset_interval (int64)	Outlier, irregular negative values	Convert negative values to postive one and keep as is. Since the outliers should be normal case, drop it may distrub the following analysis
process (category)	Missing values (91.16%)	Too many missing values, drop feature
exposure_yn (category)	Irregular Cardinality (1)	Constant column, drop feature
current_status (category)	None	Keep as is
symptom_status (category)	Missing values (51.26%)	Although with many missing values, the feature is very important and may affect the following analysis. Would derive an new feature with missing values.
hosp_yn (category)	Missing values (33.13%)	Investigate rows where this value appears, would derive an new feature with missing values.
icu_yn (category)	Missing values (91.31%)	Too many missing values, drop feature
death_yn (category)	None	Keep as is
underlying_conditions_yn (category)	Missing values (90.88%)	Too many missing values, but won't drop this feature because the underlying condition has a big impact on the analysis. Would derive an new feature with missing values.