COMP 47350

Assignment 1
Data Quality Plan (DQP)

Brian McMahon

15463152

UCD School of Computer Science

University College Dublin

March 22, 2024

1 Overview

The Data Quality Plan will outline the steps taken to clean the data after the data quality report. The steps outlined below address the problems highlighted in the data quality report and propose a method for addressing these issues. The new data set is named "covid19-cdc-15463152-cleaned-data-dqp.csv". There is one more data set present as a csv which is the data set with the new features from section 4, this is called "covid19-cdc-15463152-cleaned-new-features.csv".

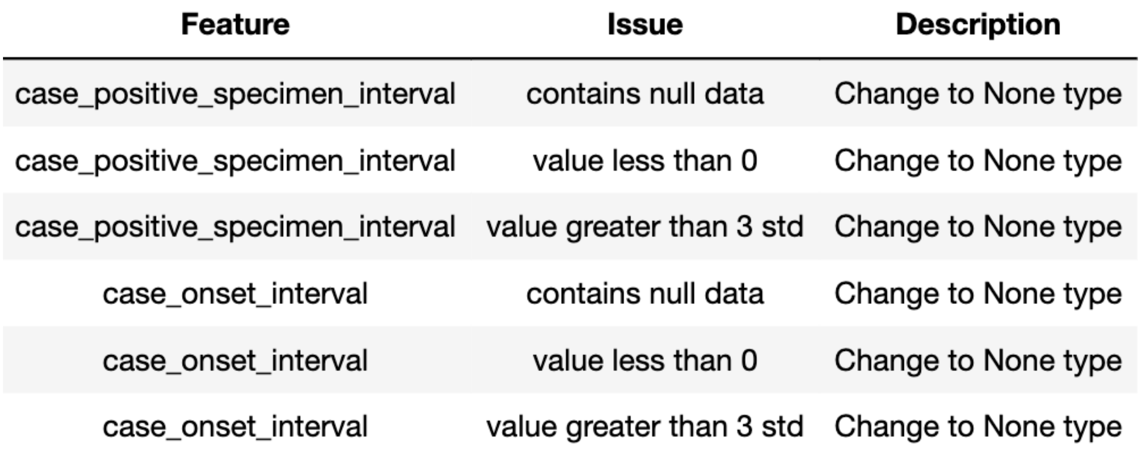
2 Continuous Features

The continuous features are the case positive specimen interval and the case onset interval. Both of these features contain outliers far from the mean and posses logical inconsistencies in the form of negative values. The description given for these features means that the value should be strictly positive. Under the description it is possible to have outliers at high positive values, but it does not make sense in regards to how the data was reported. For example it should not take 104 weeks to identify if a sample was positive.

Take case_positive_specimen_interval for example. Below we can see two plots. Consider the description of this feature. It is the number of weeks between a case creation date and the positive specimen date. Naturally we would expect most cases to receive a specimen test result in a short amount of time with a smaller amount of cases taking longer and longer. We remove anything outside 3 standard deviation due to the nature of a COVID test and how long it should take. Reasonably there's no reason for a COVID test to take 15 weeks or more to receive a result. Therefore it is assumed that anything above 3 std is erroneous and perhaps is a mix of multiple testing dates from repeat tests. Nan is the correct data type

Therefore for both continuous features the suggested data quality plan is as follows:

1. Set all missing data as NaN
2. Set all negative values to NaN
3. Set all positive values above 3 standard distributions from the mean to NaN



Feature	Issue	Description
case_positive_specimen_interval	contains null data	Change to None type
case_positive_specimen_interval	value less than 0	Change to None type
case_positive_specimen_interval	value greater than 3 std	Change to None type
case_onset_interval	contains null data	Change to None type
case_onset_interval	value less than 0	Change to None type
case_onset_interval	value greater than 3 std	Change to None type

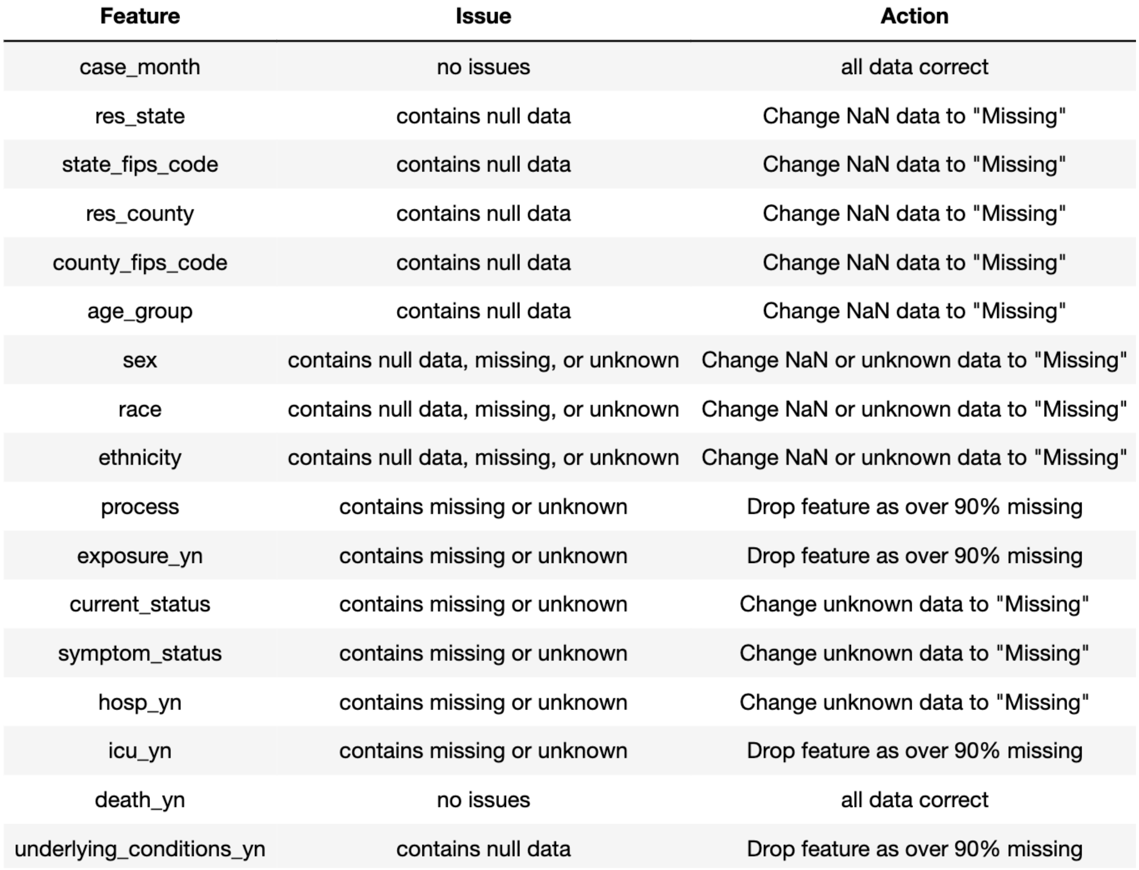
Figure 1: While the table shows None type it should actually read Numpy NaN type. My overleaf is not taking any more images for some reason.

3 Categorical Features

The categorical features present in our dataset appear mostly clean however there is a variety of titles given when dealing with missing data. These titles might refer to whether the data is missing or suppressed but for our purposes the categories should be the same. Also features which have a large percentage of missing data >90% should be dropped as they do not add any information.

Therefore the plan for categorical features is as follows:

1. Set all NaN or unknown data to "Missing" to unify the name for missing data
2. Remove features whose missing data is >90%



Feature	Issue	Action
case_month	no issues	all data correct
res_state	contains null data	Change NaN data to "Missing"
state_fips_code	contains null data	Change NaN data to "Missing"
res_county	contains null data	Change NaN data to "Missing"
county_fips_code	contains null data	Change NaN data to "Missing"
age_group	contains null data	Change NaN data to "Missing"
sex	contains null data, missing, or unknown	Change NaN or unknown data to "Missing"
race	contains null data, missing, or unknown	Change NaN or unknown data to "Missing"
ethnicity	contains null data, missing, or unknown	Change NaN or unknown data to "Missing"
process	contains missing or unknown	Drop feature as over 90% missing
exposure_yn	contains missing or unknown	Drop feature as over 90% missing
current_status	contains missing or unknown	Change unknown data to "Missing"
symptom_status	contains missing or unknown	Change unknown data to "Missing"
hosp_yn	contains missing or unknown	Change unknown data to "Missing"
icu_yn	contains missing or unknown	Drop feature as over 90% missing
death_yn	no issues	all data correct
underlying_conditions_yn	contains null data	Drop feature as over 90% missing

Figure 2: Unify missing data as "Missing" string.