COMP 47350

Assignment 1
Data Quality Report (DQR)

Brian McMahon

15463152

UCD School of Computer Science

University College Dublin

March 22, 2024

1 Overview

This report outlines the initial findings of the data quality report regarding (covid19-cdc-15463152-clean.csv), providing an in-depth analysis of its various features. The following points are addressed: Each data feature is examined and highlighted where NaN, missing, or unknown values are present. A Series of logical tests are performed to identify data where it is not valid, especially concerning the continuous features in the dataset. The analysis of continuous features reveals skewed distributions and outliers, particularly in case positive specimen interval and case onset interval, indicating potential data inconsistencies. Categorical features are examined for missing or unknown values, with a focus on unifying these under the category of "missing" to remove ambiguity.

2 Review Logical Integrity

The data is mostly categorical so the logical integrity section is small. However two features are of consideration when examining logical integrity of data in this context: case onset interval and case positive specimen interval. These categories are calculated categories. How they are calculated can be see in the appendix.

2.1 case positive specimen interval

The case positive specimen interval is calculated in weeks and should have a strictly positive value, discussed in appendix. This means that any negative value is erroneous and should be removed or set to a special value.

2.2 case onset interval

The case onset interval is calculated in weeks and should have a strictly positive value, discussed in appendix. This means that any negative value is erroneous and should be removed or set to a special value.

3 Review Continuous Features

There are 2 continuous features present in this data set: case positive specimen interval and case onset interval. see fig 3 and fig 2 to see the frequency of missing data and breakdown of data issues. Observing the histogram fig 1 we can see that the distribution is massively skewed towards 0 (notice log y scale). It would be expected that if a test was performed with a positive result a CDC entry would be made within the week. So the results are not surprising, however it is hard to see the value in using this data as almost all of the data has the same value. Due to the high concentration near 0 it also means we have quite a few outliers, and some outliers that seem to be erroneous.

3.1 case positive specimen interval

Case positive specimen interval is calculated in weeks by computing the number of days between the earliest positive specimen date and the the earliest CDC case date and dividing by 7. There are several issues associated with this feature. Firstly, negative values are present, As it is impossible by definition of the feature to have a negative interval this must be erroneous data caused by some previous CDC case date, relating to a different instance of infectious disease or instance of COVID. Secondly, outliers exist which are inconsistent with the infectious nature of COVID-19. In one instance a positive interval of over 100 weeks exists, it seems likely that this is an erroneous data point relating to a previous CDC case date. As the calculation uses the earliest CDC case date available this could be the case. Thirdly the scale of weeks in relation to this feature makes it difficult to see any useful distribution as an extremely large portion of results have the same value, ie it is not very useful if almost all of the results have the same value.

3.2 case onset interval

Case positive specimen interval is calculated in weeks by taking the difference between the earliest CDC case date and the date symptoms are first observed. similar as to above the following points are observed: There are several issues associated with this feature. Firstly, negative values are present, As it is impossible by definition of the feature to have a negative interval this must be erroneous data caused by some previous CDC case date, relating to a different instance of infectious disease or instance of COVID. Secondly, outliers exist which are inconsistent with the infectious nature of COVID-19. In one instance a positive interval of over 100 weeks exists, it seems likely that this is an erroneous data point relating to a previous CDC case date. As the calculation uses the earliest CDC case date available this could be the case. Thirdly the scale of weeks in relation to this feature makes it difficult to see any useful distribution as an extremely large portion of results have the same value, ie it is not very useful if almost all of the results have the same value.

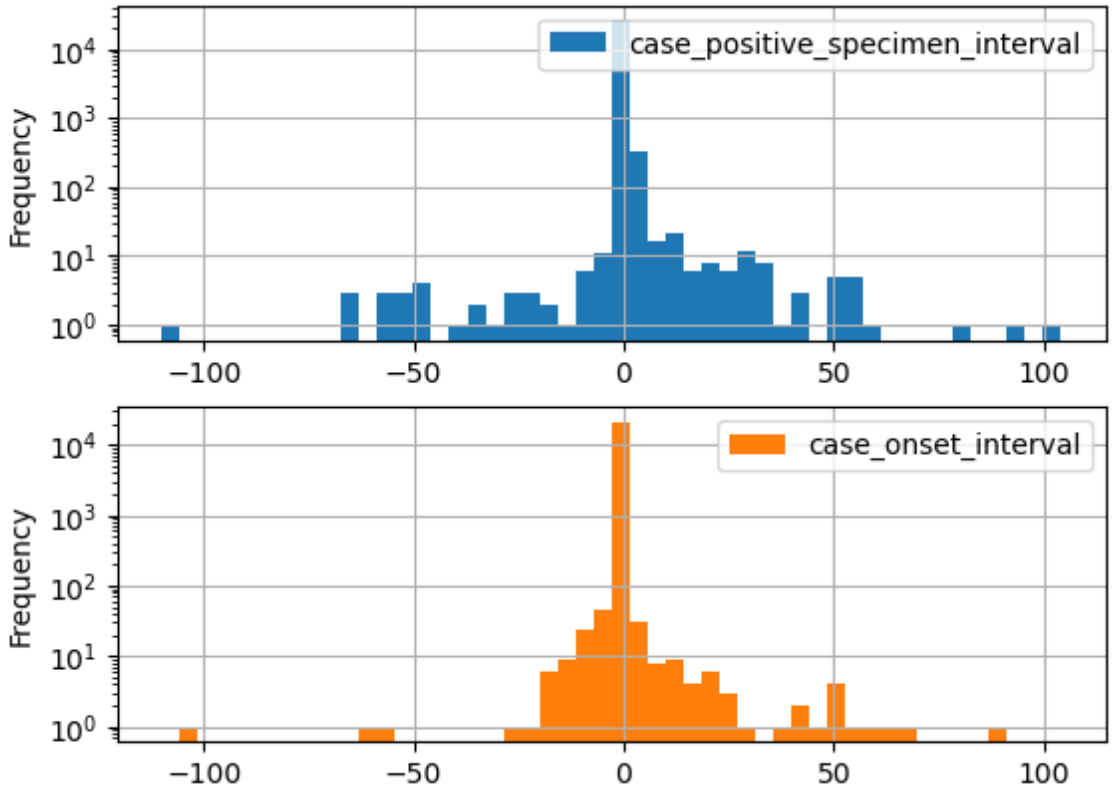

Figure 1: shows the histogram of the two continuous features: 'case positive specimen interval' and 'case onset interval'. Y axis is using a log scale to show data even though the mean dominated the distribution. negative values and outliers are discussed in the continuous feature section.

4 Review Categorical Features

There are 17 categorical features present in the dataset. The general problem with the categorical features present is the presence of three values for unknown or missing data. The value of this missing data is represented as either "Missing", "Unknown", or NaN. In most cases the data can be unified. For these procedures we will consider "Missing" as the correct value to be associated with blank unknown or missing data. See fig. 4 to see the breakdown of the missing and unknown data vs NaN, and see fig 5 to see the frequency of missing data and cardinality per feature.

4.1 case month

We consider case month as a categorical feature because it groups together data from different time periods. However it could be considered as a continuous feature if we applied some mapping to integer values. This might be beneficial as it would allow some sort of time series analysis to be applied to this feature. For the time being we are leaving it as a categorical data type. The dataset is complete with no blank or missing values.

Feature	Issue	Description
case_positive_specimen_interval	blank rows	23634 blank rows
case_positive_specimen_interval	less 0	138 rows less than 0
case_onset_interval	blank rows	28788 blank rows
case_onset_interval	less 0	685 rows less than 0

Figure 2: There are numerous blank values in both the 'case positive specimen interval' and 'case onset interval' indicating that the feature is somewhat unreliable. Furthermore we highlight that some of the rows have negative values which is not allowed based on the description of the feature.

	count	mean	std	min	25%	50%	75%	max	%missing	card
case_positive_specimen_interval	26366.0	0.152886	2.472582	-110.0	0.0	0.0	0.0	104.0	47.544	81
case_onset_interval	21212.0	-0.022252	1.900717	-105.0	0.0	0.0	0.0	91.0	58.946	54

Figure 3: Observing the descriptive statistics of the 'case positive specimen interval' and 'case onset interval' features we see that due to the massive density of 0 values in the data set means that almost anything outside of 0 is an outlier. This being said there are min and max values many standard deviations from the mean >5 std hinting that there might be some erroneous data in these features.

4.2 res state and state fips code

Residential state has been changed from a float value to a categorical value as the number chosen to represent the state has no particular meaning. In 2 cases the state has been suppressed leading to blank values being present. The state fips code is consistent with the residential state so these features have been grouped together. Considerations can be made to remove one of these features as they repeat the same data.

4.3 res county and county fips code

Residential state has been changed from a float value to a categorical value as the number chosen to represent the state has no particular meaning. In 2829 cases the state has been suppressed leading to blank values being present. The state fips code is consistent with the residential state so these features have been grouped together. Considerations can be made to remove one of these features as they repeat the same data.

4.4 age group

Age group contains 368 rows with null data and 73 with missing or unknown data. For our purposes these categories are analogous so 368+73 gives 441 rows with no age data.

4.5 sex

Sex feature contains 1095 rows with no data, which have been suppressed, and 207 rows containing missing or unknown in the data field. For our purposes these can be represented as the same, giving a total of 1302 rows with no data.

4.6 race

Race feature contains 6145 rows with no data, which have been suppressed, and 6405 rows containing missing or unknown in the data field. For our purposes these can be represented as the same, giving a total of 12550 rows with no data, totaling 25.1% of data.

4.7 ethnicity

Ethnicity feature contains 6733 rows with no data, where the data has been suppressed, and 9558 missing or unknown rows. For our purposes these can be represented as the same, giving a total of 16291 rows with no data, totaling 32.6% of data.

4.8 process

Process contains 45524 missing or unknown values, totaling 91% of data. This column should be considered for dropping as the data is very sparse.

4.9 exposure yn

Exposure_yn contains 45237 missing or unknown values, totaling 90% of data. This column should be considered for dropping as the data is very sparse.

4.10 current status

Current_status has 0 missing or unknown data.

4.11 symptom status

Symptom status feature contains 27011 rows with missing or unknown data totaling 54% of data.

4.12 hosp yn

hosp_yn status feature contains 17645 rows with missing or unknown data totaling 35.3% of data.

4.13 icu yn

icu_yn status feature contains 46041 rows with missing or unknown data totaling 92.1% of data, making it a good candidate to be dropped. However any positive ICU case will assumedly have high success in predicting likelihood of death as an outcome. So we will investigate more before dropping.

4.14 death yn

Feature is complete and has no percieved issues.

4.15 underlying conditions yn

underlying_condition_yn status feature contains 45896 rows with NaN data totaling 91.7% of data, making it a good candidate to be dropped. However underlying conditions might have high success in predicting likelihood of death as an outcome. So will investigate more before dropping.

5 Actions to take

Review missing/ unknown data Review all cases of missing, unknown, and NaN data from the above features and unify them under the heading of "missing". This is a data cleaning step to remove ambiguity.

Remove negative data from continuous features The two continuous features in our dataset should not contain negative values. This is due to the definition of the features. Logical inconsistencies in the data should be removed.

Feature	Issue	Description
case_month	no issues	all data correct
res_state	contains null data	2 rows contain null values state fips code aswell
state_fips_code	contains null data	2 rows contain null values res state aswell
res_county	contains null data	2829 rows contain no county
county_fips_code	contains null data	2829 rows contain no county fips code
age_group	contains null data	368 rows contains no age group, 73 contain missing or unknown
sex	contains null data, missing, or unknown	1095 rows contains no sex data, 32 contain missing, and 175 contain unknown
race	contains null data, missing, or unknown	6145 no data, 6405 missing or unknown
ethnicity	contains null data, missing, or unknown	6733 no data, 9558 missing or unknown
process	contains missing or unknown	45524 missing or unknown
exposure_yn	contains missing or unknown	45237 missing or unknown
current_status	2 values	maybe rename one value
symptom_status	contains missing or unknown	27011 missing or unknown
hosp_yn	contains missing or unknown	17645 missing or unknown
icu_yn	contains missing or unknown	46041 missing or unknown
death_yn	no issues	all data correct
underlying_conditions_yn	contains null data	45896 rows contain no information

Figure 4: Table highlighting the data quality issues present in the categorical features in the data set

	count	unique	top	freq	%missing	card
case_month	50000	40	2022-01	6397	0.000	40
res_state	49998	51	NY	5569	0.004	51
state_fips_code	49998	51	36	5569	0.004	51
res_county	47171	949	MIAMI-DADE	989	5.658	949
county_fips_code	47171	1360	12086	989	5.658	1360
age_group	49632	5	18 to 49 years	20336	0.882	5
sex	48905	4	Female	25832	2.604	4
race	43855	8	White	30432	25.100	8
ethnicity	43267	4	Non-Hispanic/Latino	29703	32.582	4
process	50000	10	Missing	45370	91.048	10
exposure_yn	50000	3	Missing	43111	90.474	3
current_status	50000	2	Laboratory-confirmed case	42324	0.000	2
symptom_status	50000	4	Symptomatic	22259	54.022	4
hosp_yn	50000	4	No	25178	35.290	4
icu_yn	50000	4	Missing	39100	92.082	4
death_yn	50000	2	No	40000	0.000	2
underlying_conditions_yn	4172	2	Yes	4104	91.656	2

Figure 5: descriptive statistics including the cardinality and % missing data present in each of the features