DATA QUALITY REPORT – INITIAL FINDINGS

Overview

This report will outline the initial findings based on the cleaned dataset (*covid19_cdc_cleaned.csv*). It will summarise the data, describe the various data quality issues observed and how they will be addressed. Please see appendix for some background to this dataset. Appendix includes terminology, assumptions, explanations and summary of changes made to the original dataset. This also includes feature summaries, histograms and boxplots used to visualise the data. On first indication the dataset appears have a great deal of missing data for particular features. On a surface level, there appears to be an absence of duplicate columns or columns, but this is difficult to decipher due to the lack of an index key. There is also an absence of irregular cardinalities or constant columns. There was a significant number of rows deemed to be duplicates that were removed (695 when using keep="first") due to their high degree of null attributes. The main issues observed were regarding special values for categorical data and a small number of logical errors.

Summary

Several tests were carried out to check the logical integrity of the data. This brought about a number of failures of the data. In total 90 instances of irrational data was observed. For example, in 1 instance an individual was committed to the ICU, but was not in hospital. The other case was 89 counts of the individual testing positive before the earliest records that the CDC have on them. This irrational data will need to be dealt with and should be checked with the domain expert. See logical integrity section for further details.

For the categorical features there was the inclusion of several special values i.e. "Unknown", and "Missing", which map to a specific meaning: "Unknown" is where individuals did not fill out the given field, and "Missing" is where the data was lost/not recorded. These values need to be reconciled into one field in order to clean up the uncategorised data.

Review of Logical Integrity

4 tests were carried out. The failures are below:

Test 1 – Check if any case has been admitted to ICU without being admitted to hospital. *1 case found*.

Test 2 – Check if any case has been reported before their earliest known date. *32 cases found*.

Test 3 – Check if any case has had a positive specimen test recorded before their earliest known date. *89 cases found*.

Test 4 – Check if any case had an onset of symptoms recorded before their earliest known date. *0 cases found*.

Review Continuous Features

There are 4 continuous features. All continuous features are in the form of "days since 1/1/2020". This is to ensure we can find average dates, standard deviations of dates, etc. Initial feedback suggested this is an interesting approach but it might make the data more difficult to interpret. However, I have decided to continue with this approach for two reasons. Firstly, if the data is to be used for machine learning it will be much easier to work with *integers* rather than *datetime*. Secondly, the dataset is lacking in solid continuous data. While there are other candidates to consider converting to continuous data, such as changing "*age*_group" categories to 25, 35, 45, etc., it has the possibility of skewing the data since there is no immediately obvious imputation for the 80+ category. Additionally, it might skew the data as by making incorrect assumptions/over-simplifications of ages in the dataset. Therefore, I would like to keep the data in *integer* format for demonstration purposes in order to satisfy all elements of the assignment. Regardless, it is unlikely that these dates will be used for further research as they are unlikely to have an impact on the target feature (*death_yn*).

cdc_case_earliest_dt – This feature appears to be clean, with no outliers or missing data. According to the CDC, it has been optimised for completeness and is recommended for time-based analyses. Therefore, it can function as a sort-of index for the dataset where we can base our other dates off of.

cdc_report_dt – According to the CDC, this feature has been depreciated. Thus, it lacks in informational quality and it significantly overlaps in its function with *cdc_case_earliest_dt*.

pos_spec_dt – This feature has a significant amount of missing data (70%).

onset_dt – This feature has large amount of missing data (46%).

Review of Categorical Features

There are 8 categorical features in the dataset, 1 of which is the target (*death_yn*) and will not be evaluated here. The 7 remaining are:

current_status – This feature appears to be clean, with no missing data.

sex - There appears to only be 1% of data missing from this feature.

age_group – This appears to only be 0.16% of missing data from this feature.

race_ethnicity_combined – There appears to be 38% of data missing from this feature.

hosp_yn – There appears to be 38% of data missing from this feature.

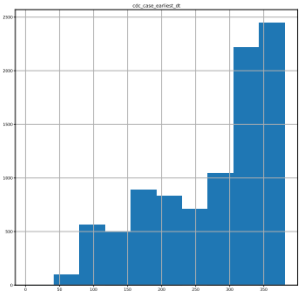
icu_yn – There appears to be a significant amount of data missing (88%) from this feature.

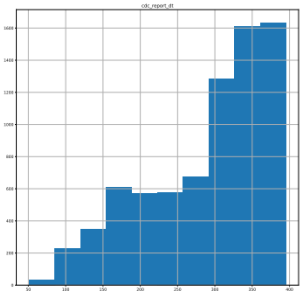
medond_yn – There appears to be a significant amount of data missing (81%) from this feature.

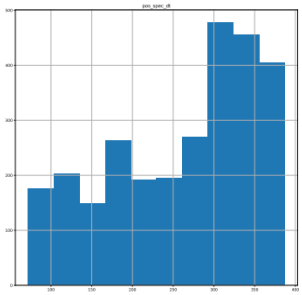
Note that every categorical feature besides the target feature and *current_status* had missing data in two formats: "*Unknown*" and "*Missing*". While "*Unknown*" is technically a valid field, it does not provide useful information, and is essentially the same as "*Missing*" for the purpose of data analysis. Thus, the two fields were combined and "*Unknown"* was imputed to "*Missing"*.

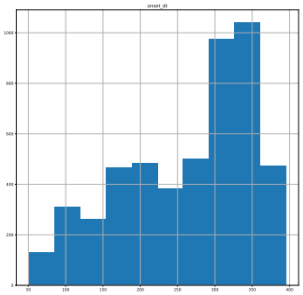
Charts

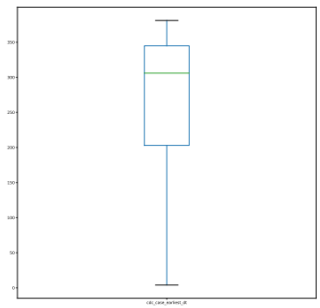
On initial inspection, there does not appear to be any problematic outliers. The data will be investigated further but no immediate action is expected in this regard.

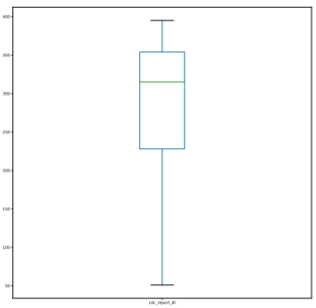


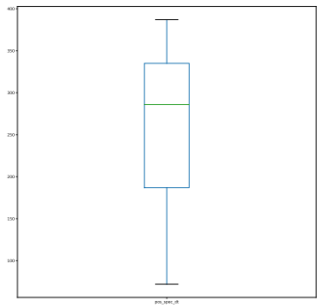


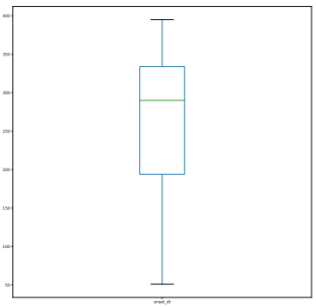


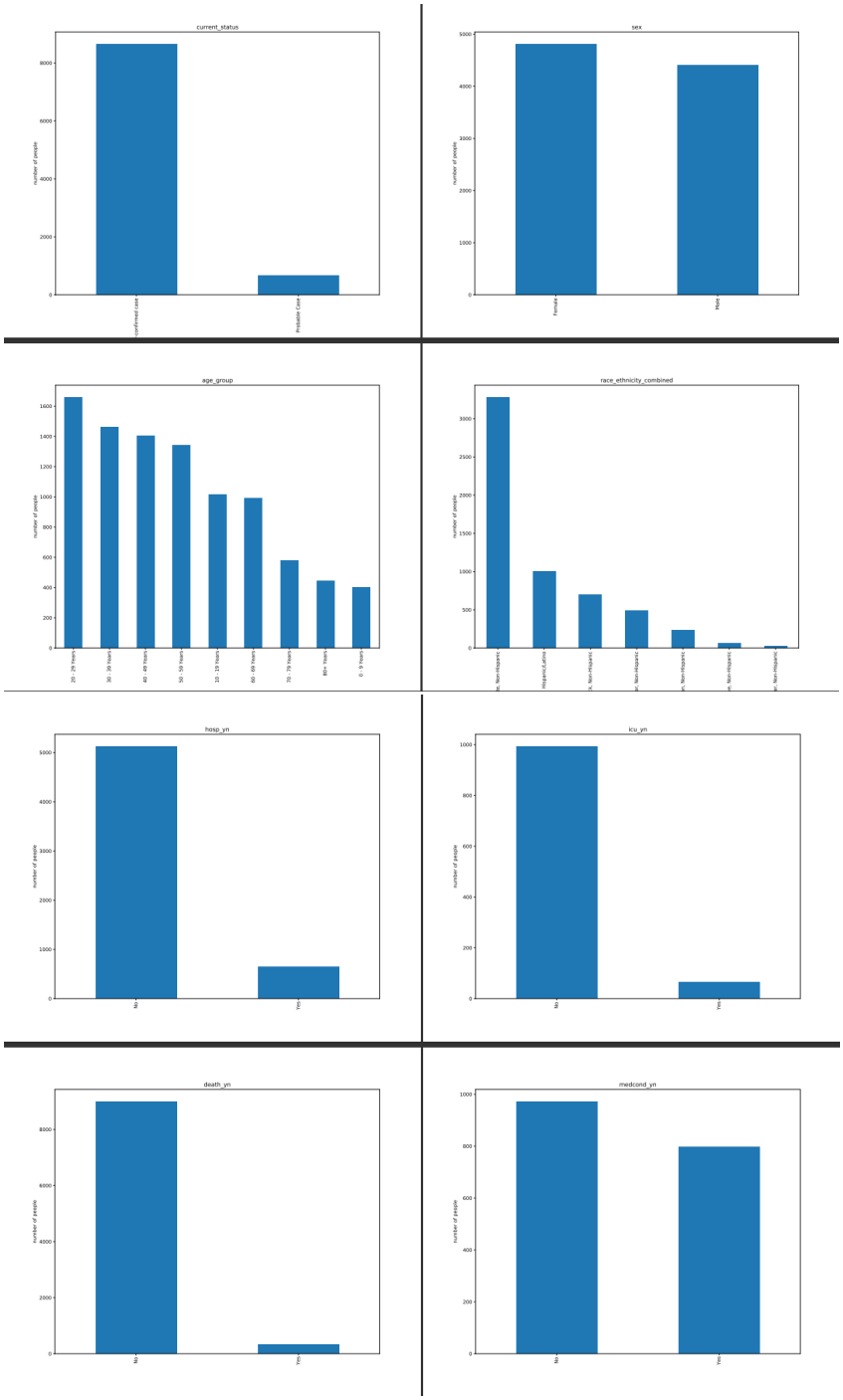












Action to Take

Missing data in continuous features will be imputed with their corresponding *cdc_case_earliest_dt*. This is because *cdc_earliest_date* is closely correlated with its peers. Additionally, it is the preferred date to use by the CDC.

Although the correlation of *cdc_report_dt* has a lower correlation to *cdc_case_earliest_dt*, which means it could provide some unique data, it will be dropped since the CDC said it is a depreciated feature, thus it provides no real value.

Drop row where an individual was in ICU *icu_yn* without being in hospital *hosp_yn*. It is a logical error.

Where other dates precede *cdc_case_earliest_dt*, they are imputed to be equal to *cdc_case_earliest_dt* in order to preserve the value of the other features in the row, while remaining logically consistent.

Appendix

cdc_case_earliest_dt – Earliest available date in the CDC's record, taken from either the available data set of clinical dates (date related to the illness or specimen collection) or the calculated date representing initial date case was received by CDC (calculated and optimised).

cdc_report_dt – Calculated date represented initial date case was reported to CDC (depreciated).

pos_spec_dt – Date of first positive specimen collection.

onset_dt – Symptom onset date, if symptomatic.

current_status – Case status: Laboratory-confirmed case; Probable case; Missing.

sex – Male; Female; Other; Missing.

age_group – 0-9 Years; 10-19 Years; 20-39 Years; 40-49 Years; 50-59 Years; 60-69 Years; 70-79 Years; 80+ Years; Missing.

race_ethnicity_combined – Race and ethnicity (combined). Hispanic/Latino; American Indian/Alaska Native, Non-Hispanic; Native Hawaiian/Other Pacific Islander, Non-Hispanic; White, non-Hispanic; Multiple/Other, Non-Hispanic; Missing.

hosp_yn – Hospitalisation status.

icu_yn – Death status.

medond_yn – Presence of underlying comorbidity or disease

Continuous Features – Descriptive Statistics

Prior to full cleaning (NaN data, logical errors, etc.)

	count	mean	std	min	25%	50%	75%	max	%missing	card
cdc_case_earliest_dt	8845.0	270.849406	86.477714	4.0	201.0	301.0	343.0	381.0	0.000000	320
cdc_report_dt	7478.0	288.258492	81.612543	51.0	227.0	315.0	354.0	395.0	15.455059	327
pos_spec_dt	2751.0	260.197383	88.994288	72.0	187.0	285.0	335.0	387.0	68.897682	308
onset_dt	4971.0	262.706699	85.986146	51.0	194.0	290.0	334.0	395.0	43.798756	321

Categorical Features – Descriptive Statistics

Prior to full cleaning (NaN data, logical errors, etc.)

	count	unique	top	freq	%missing	card
current_status	8845	2	Laboratory-confirmed case	8179	0.000000	2
sex	8751	2	Female	4574	1.062747	2
age_group	8830	9	20 - 29 Years	1544	0.169587	9
race_ethnicity_combined	5713	7	White, Non-Hispanic	3233	35.409836	7
hosp_yn	5682	2	No	5043	35.760317	2
icu_yn	1055	2	No	990	88.072357	2
death_yn	8845	2	No	8518	0.000000	2
medcond_yn	1766	2	No	969	80.033917	2