Data Quality Plan

Introduction

Following on from my data quality report I will know summarise the data quality issues I believe the data sample has and try to implement an effective handling strategy that will keep as much of the data intact as possible and provide us with the most relevant data. I will also provide some suggestions on data collection as it may help to alleviate some of the data cleaning required in the future and provide more accurate complete data.

Feature	Data Quality Issue	Handling Strategy
Pos_spec_dt	Missing values (71.34%)	Drop feature. I choose to drop the feature due to the large amount of missing data and we have 3 other sources of dates that could provide more information.
Icu_yn	Missing values (89.05%) (Calculated with missing and unknown values)	Possibly map missing values to no if the figures we have are in line with national figures on ICU admissions. Based on figures from (ourworldindata.org, accessed 12/03/2020) the covid ICU rate is 25% so it is impossible to map the missing data to a no as our data only shows the ICU rate being approximately 9%. Drop feature
Medcond_yn	Missing values (82.26%) (Calculated with missing and unknown values)	Drop feature. I decided to drop this feature as it could not give us information that we could confidently extrapolate to a wider data sample due to the large number of missing values. It is an important feature to track to see if trends can be established between Covid-19 and other medical conditions. I would encourage a revisit to how this data is collected and try to get more accurate information.
All features	437 Duplicate Rows	Keep these rows, as it is plausible due to the nature of the information gathered that duplicate rows are possible.
Race/Ethnicity	1. 4020 unknown values and a further 91 missing. 2. The data that has been collected a large proportion is White, non-Hispanic race/ethnicity.	1. Investigate as to why so many are unknown see if it is possible to get that data. If not perhaps giving a general location may allow mapping to a particular race/ethnicity based on geolocation.

		2. This could skew the data that it is more prevalent in those communities when it is more likely due to White communities having better healthcare and access to medical care. Do not treat the data as representative of how prevalent Covid-19 is in different communities. 3. We could try to re sample the full data set with a more balanced data sample when it comes to this feature while ensuring all other features remain evenly distributed amongst age and gender.
CDC case earliest date	The data is skewed right starting around November 2020.	Investigate why the data collection has been quite erratic and see how we can improve on this to collect more consistent data.
CDC report date	1. Outlier 10/06/2020. 140 reports to the CDC. 2. Missing 23.28% of data	1. Investigate why on this day there were approximately 40 more cases reported than the next highest day. Maybe there was a delay in the reporting, or perhaps a batch had no date and were given the date this was discovered. 2. Map the missing dates to the cdc_case_ earliest_dt
Onset_dt	Missing 49.58% of data.	Map the missing dates to the cdc_case_ earliest_dt. I mapped to this date as I noticed many of the entries with all three dates completed had the same date and it is the most accurate date, we have access to.
CDC report date **Onset date**	Both graphs end at 0.	Check with data domain expert perhaps these are no longer being collected and should be removed from the dataset going forward.
Hosp_YN	41.54% missing values (Calculated with missing and unknown values)	Perhaps check where the data documents were sent from if it was a hospital you could assume the patient was admitted if it was a testing centre or other facility you could assume, they were not admitted. Impossible to implement without further data.
Death_yn	3.45% significantly higher than the national average of 1.81%	We are working with a small subset of data and as such it can throw figures of

	(calculated Numberofdeath/total (345/10000)*100=3.45%) (521,625/28,700,966)*100=1.81%)	like this. It must be noted clearly that findings may not be indicative of the national picture.

Conclusion

There are many avenues to pursue in ensuring a fair and equally distributed data sample. The data quality is heavily impacted by missing or unknown data and we should revisit the collection points to see if we can streamline and improve this collection process. The form requires a lot of data to be input and I can imagine it is a strain on an already overwhelmed health service. A review of the form and keeping of only the most essential data points should be considered.

Bibliography

Our world in data. Available at: https://ourworldindata.org/covid-hospitalizations (Accessed:12/03/2020)