Data Quality Report – Initial Findings

1. Overview

This report will list and describe the findings about the dataset covid19-cdc-19200690_CLEANED.csv. The features in the dataset are divided into continuous and categorical types. Each entry in the dataset is tested to make sure it makes logical sense: for example, an entry in the data set does not make sense if a patient is marked as being in the ICU in hospital, but is also marked as not being in hospital. Each feature will be described: what the feature describes, if there any issues with the data, if any action needs to be taken. There are tables, histograms and bar charts in the appendix section of this report to illustrate the data being described.

2. Summary

2.1. Logical Integrity

Three tests for the logical integrity of the data were carried out. The tests returned a significant number of illogical rows. These tests will be discussed in the Logical Integrity section of the report.

2.2. Continuous Features

For the continuous features, there were no obvious errors, aside from missing data. However, the 'earliest_date' column had no missing values at all. No special values were found in the continuous features, as each value corresponded to a date, or was a null value.

2.3. Categorical Features

Upon first viewing, the categorical features had no null values. But on closer inspection, it could be seen that large numbers of the entries had the entry 'Unknown', or 'Missing'. Other than missing values, and logical errors, there were no obvious errors with the categorical features.

3. Logical Integrity

3 tests were carried out to make sure the data made sense. The results can be seen below:

1. Check if there are any rows where the 'earliest_date' column is a later date than any of the other columns for that row. This should not be possible, as the 'earliest_date' column should be earlier than or equal to all other continuous columns in a row.
    - 68 rows were found that failed this test.

2. Check if there are any rows that have a value of 'Probable Case' for the 'status' column, and a not-null value for the 'posSpec_date column'. This should not be possible, as any entries that are not 'Laboratory-confirmed case' should not have a positive specimen collected for coronavirus, therefore they would not have a date for when the positive specimen was collected.
    - 215 rows were found that failed this test.

3. Check if there are any rows that have a 'yes' value for the 'icu' column, and 'no', 'unknown', or 'missing' for the 'hosp' column. This should be impossible if the 'hosp' value is 'No'. It should indicate the value of the 'hosp' column if the value is 'Unknown' or 'Missing'.
    - 0 rows were found that failed this test.

4. Review Continuous Features

4.1. Descriptive Statistics

There are just four continuous features in this dataframe – all of type datetime. Each one is summarised below.

- earliest_date
  - The range of dates of entries for this column is 377 days: from 5 January, 2020 to 16 January, 2021.
    - The first 25% of these entries – 2500 entries – take place over 203 days from the first date.
    - The last 50% of the entries comes in the final 71 days.
  - This seems plausible, as the number of COVID-19 cases has continued to rise over time since the pandemic began. Therefore the higher concentration of cases as the year progresses makes sense.
  - There are no special or null values.

- report_date
  - 24.08% of the values for this column are marked null.
  - The range of dates for this column spans 390 days: from 5 January, 2020 to 29 January 2021.
  - There is a similar distribution of entries here as the earliest_date column: the entries occur at a higher rate towards the end of the range of dates.
  - This is again consistent with more cases COVID-19 occuring as the year progresses.
  - No special values were found.

- posSpec_date
  - 72.74% of the values for this column are marked null.
  - The range of dates for this column is just 325 days: from 6 March, 2020 to 25 January, 2021.
  - The distribution of entries is less extreme here, with the first 25% of entries coming in 118 days, and the final 50% of entries coming in 106 days.
    - This is most likely due to the much smaller sample size, and the fact that entries only started being recorded for this column in March, rather than January.
    - However, the larger number of entries is stil skewed towards the later range of dates, which is still consistent with more COVID-19 cases happening as the year progresses.
  - No special values were found.

- Onset
  - 49.16% of the values for this column are marked null.
  - The range of dates for this column is just 388 days: from 1 January, 2020 to 27 January, 2021.

- There is a similar distribution of entries here as the earliest_date column and the report_date column: the entries occur at a higher rate towards the end of the range of dates.
- This is again consistent with more cases COVID-19 occuring as the year progresses.
- No special values were found.

4.2. Histograms

All histograms can be found in the appendix section of this document. Some individual plots can be found in the accompanying notebook. Overall the distribution of entries was consistent and no outliers were found: which is to say that there were no days where a large number of cases was found in the earlier dates.

5. Review Categorical Features

5.1. Descriptive Statistics

There are 8 categorical features: status, sex, age_group, race_ethnicity, hosp, icu, death and medcond. None of these columns contain any missing or null values, but some contanis values that say 'Missing' or 'Unknown', which amounts to the same thing. Columns that have high percentages of these values may need to be dropped.

- status: describes whether the entry is a confirmed COVID-19 case or not.
    - It has just 2 values: 'Laboratory-confirmed case' and 'Probable Case'.
    - Apart from the logical errors found, the results here seem plausible: 93.4% of entries are 'Laboratory-confirmed case' and the remaining 6.6% are 'Probable case'.
    - There are no missing values.

- sex: the sex of the person in the entry.
    - It has 4 values: 'Male', 'Female', 'Missing' and 'Unknown'.
    - The distribution of results makes sense: 52.28% are 'Female' and 46.8% are 'Male'.
    - 0.82% of the entries have a 'Missing' or 'Unknown' value.

- age_group: ranges of ages for the person in the entry, grouped in decades, from 0 years up to 80 plus years.
    - It has 10 values – '0 – 9 Years', '10 – 20 Years' and so on, up to '80+ Years'.
    - There is no siginifcant outlier: no age range has significantly more entries than any other.
    - Just 0.1% of entries are missing.

- race_ethnicity: the racial and ethnic background of the person in the entry.
    - It has 9 values.
    - 42.54% of these are 'Missing' or 'Unknown'.
    - Another 32.84% are 'White, Non-Hispanic'.
    - The next-most common value is 'Hispanic/Latino' at 9.09%.

- hosp: was the person in the entry admitted to hospital.
    - Has 4 values: 'Yes', 'No', 'Missing', 'Unknown'.
    - 52.23% of the results for this column are 'No', and 23.52% are 'Missing'.

- icu: was the person in the entry admitted to intensive care.
    - 89.35% of the cases for this column are either 'Missing' or 'Unknown', so it will not be used.

- death: did the person in the entry die.
    - Has 2 values: 'No' and 'Yes'.
    - 96.8% of the entries were 'No' and the remainder was 'Yes'.

    - There are no missing values for this entry.

- medcond: did the person in the entry have an underlying medical condition.
    - 82.83% of the entries for this column were either ‘Missing’ or ‘Unknown’, so it will not be used.

5.2. Bar Charts

Bar charts for he categorical features can be found in the appendix section of this report.

6. Action to take

8 main actions will be taken, which are summarised below:

- Logical Integrity
    - For the rows that failed Test 1, the 'earliest_date' column for this row should have its value changed to whatever the minimum value of the continuous features of that row is.

    - The rows that failed Test 2 should be dropped. There is no way of knowing for these rows, which entry was the mistake: the entry in the 'status' column or the entry in the 'posSpec_date' column.
        - Thus it is not possible to determine whether the entry has definitely collected a positive specimen sample – meaning its status should be changed to 'Laboratory-confirmed case' – or not – meaning the entry was mislabeled.

- Continuous Features
    - The posSpec_date column should be dropped.
        - 72.74% of its values are missing. The high percentage of null entries and the smaller range of dates for this column being recorded causes this column's distribution differs from that of the other continuous columns.
        - There are also no rows that have a value only for this row in terms of its continuous features, so no essential data will be lost in dropping this row.

    - The onset column should be dropped.
        - 49.16% of its values are missing.
        - There are no rows that have only a value for this row in terms of its continuous feature, so no essential data will be lost.
        - The reporting of symptoms by patients is imperfect. It is possible that patients will misremember when their symptoms came on. It is also possible that patients may think they are symtomatic, when in fact they are not, or vice versa.

- Categorical Features
    - The icu column should be dropped.
        - 89.35% of its values are ‘Missing’ or ‘Unknown’.

    - The medcond column should be dropped.
        - 82.83% of its values are ‘Missing’ or ‘Unknown’.

    - Any rows that have a value of ‘Missing’ or ‘Unknown’ for the sex column should be dropped.
        - In terms of identifying trends, knowing the sex of the person in the entry is useful information.
        - 99.18% of the data for this column is available, so not a lot of data will be lost in dropping these rows.

    - Any rows that have a value of ‘Missing’ or ‘Unknown’ for the age_group column will be dropped.
        - In terms of identifying trends, the age group of the person in the entry is useful information.
        - 99.9% of the data for this column is available, so not a lot of data will be lost in dropping these rows.

7. References

[1] Source for the data – the Center for Disease Control and Prevention

https://covid.cdc.gov/covid-data-tracker/ - datatracker-home

[2] Source for the dataset worked on

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

8. Appendix

8.1. Terminology and Assumptions

- It is assumed that Actions to take section will be carried out in order, where values are replaced before any columns are dropped.

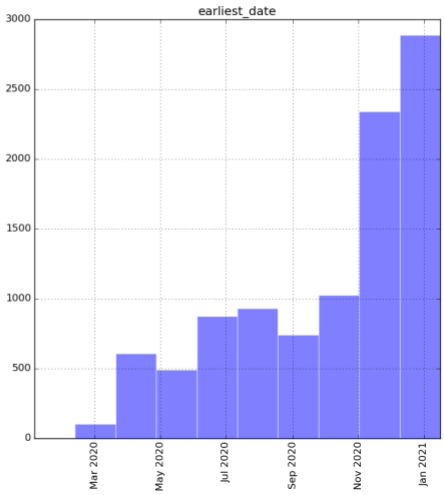
8.2. Continuous Features

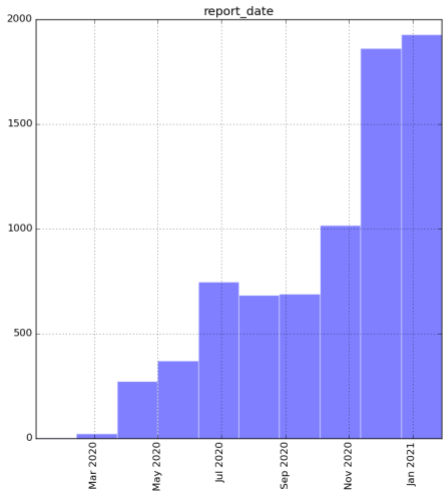
Feature	count	unique	top	freq	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode	%Missing/Unknown
status	10000	2	Laboratory-confirmed	9340	Laboratory-confirmed	9340	93.4	Probable Case	660	6.6	0
sex	10000	4	Female	5238	Female	5238	52.38	Male	4680	46.8	0.82
age_group	10000	10	20 - 29 Years	1860	20 - 29 Years	1860	18.6	30 - 39 Years	1605	16.05	0.1
race_ethnicity	10000	9	Unknown	4143	Unknown	4143	41.43	White, Non-Hispanic	3284	32.84	42.54
hosp	10000	4	No	5223	No	5223	52.23	Missing	2352	23.52	41.25
icu	10000	4	Missing	7695	Missing	7695	76.95	Unknown	1240	12.4	89.35
death	10000	2	No	9680	No	9680	96.8	Yes	320	3.2	0
medcond	10000	4	Missing	7532	Missing	7532	75.32	No	885	8.85	82.83

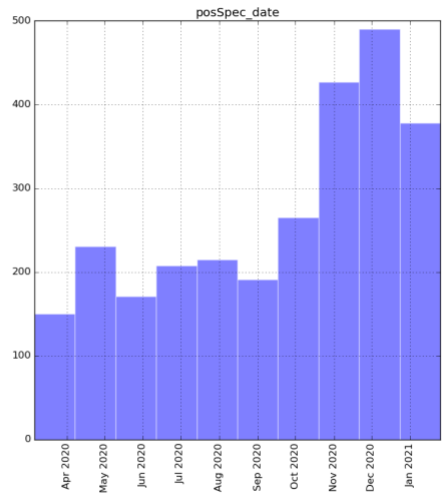
8.3. Categorical Features

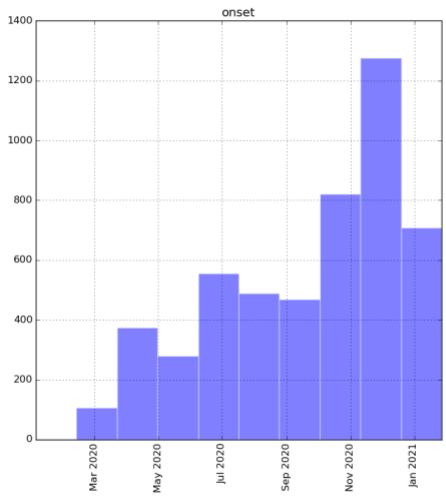
Feature	count	mean	min	25%	50%	75%	max	%missing	cardinality
earliest_date	10000	52:45.1	00:00.0	00:00.0	00:00.0	00:00.0	00:00.0	0	325
report_date	7592	49:15.1	00:00.0	00:00.0	00:00.0	00:00.0	00:00.0	24.08	326
posSpec_date	2726	41:06.9	00:00.0	00:00.0	00:00.0	00:00.0	00:00.0	72.74	312
onset	5084	36:09.6	00:00.0	00:00.0	00:00.0	00:00.0	00:00.0	49.16	326

8.4. Histograms

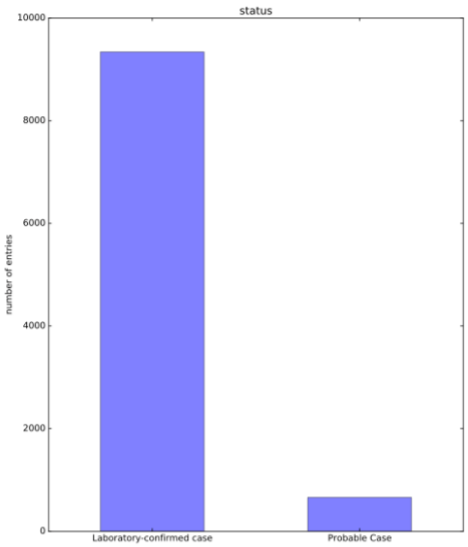


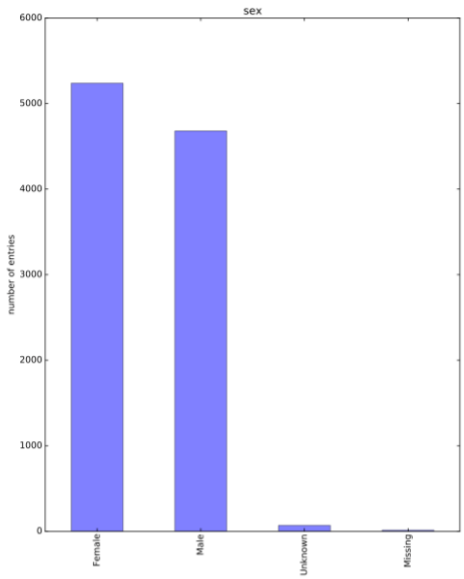


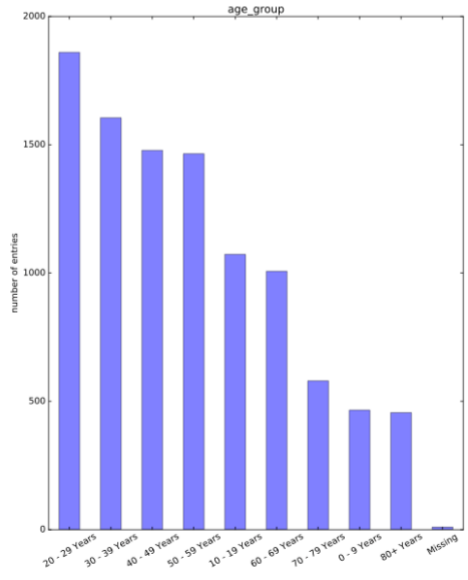


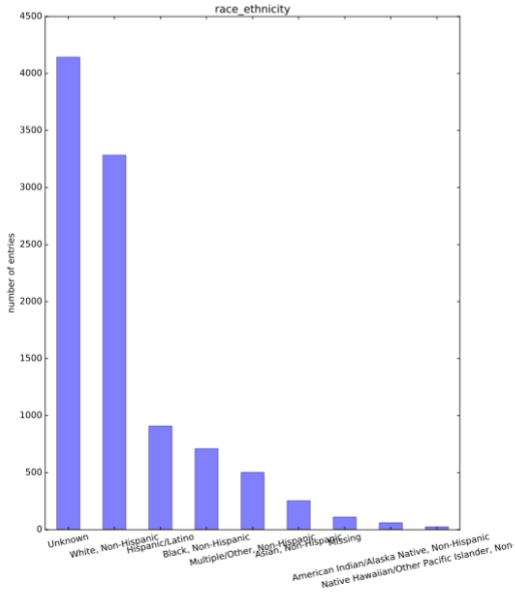


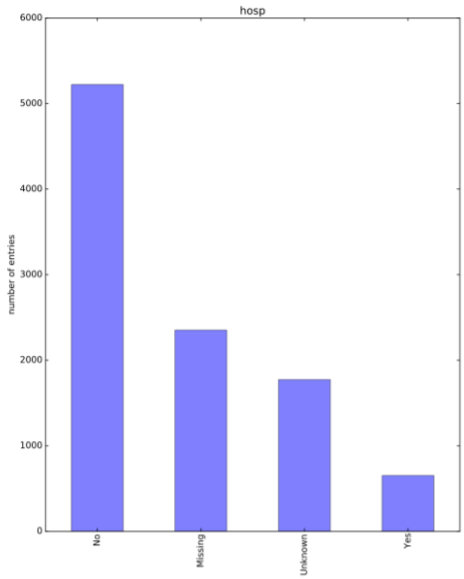
8.5. Bar charts

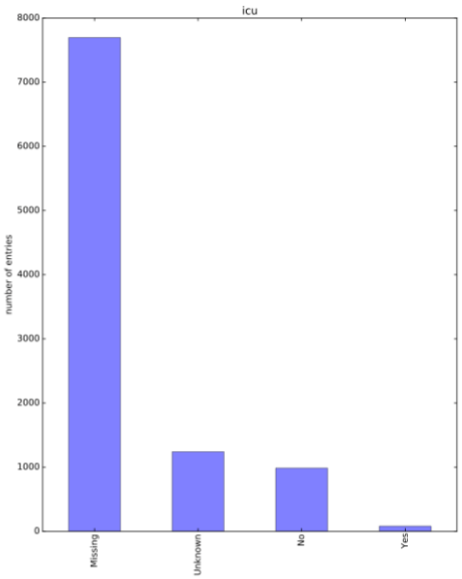


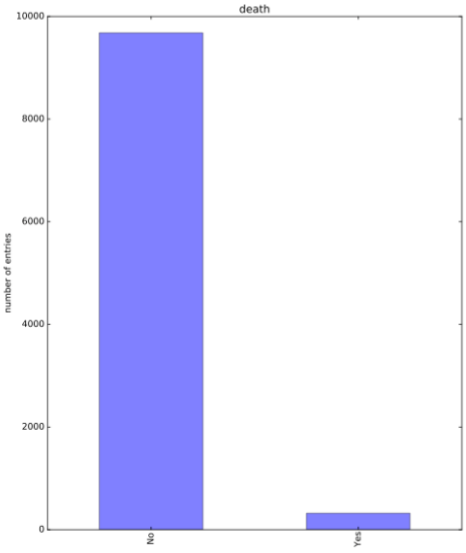


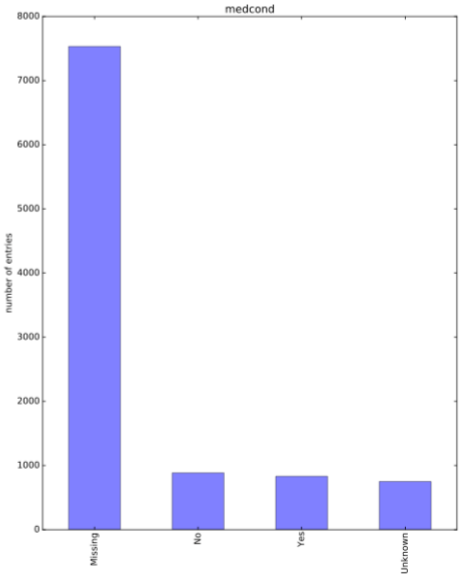





hosp
number of entries
6000
5000
4000
3000
2000
1000
0
No
Missing
Unknown
Yes


icu
number of entries
8000
7000
6000
5000
4000
3000
2000
1000
0
Missing
Unknown
No
Yes


death
number of entries
10000
8000
6000
4000
2000
0
No
Yes


medcond
number of entries
8000
7000
6000
5000
4000
3000
2000
1000
0
Missing
No
Yes
Unknown