**Data Quality Report Initial Findings**

**Problem: Analyse the dataset to build a solution for death risk prediction.**

Table of Contents

[Overview 2](#_Toc132715258)

[1. Numeric Columns 3](#_Toc132715259)

[a. case\_positive\_specimen\_interval 3](#_Toc132715260)

[b. case\_onset\_interval 4](#_Toc132715261)

[1.2 Timestamp Column 5](#_Toc132715262)

[a. case\_month 5](#_Toc132715263)

[2. Categorical Columns 6](#_Toc132715264)

[a. res\_state 7](#_Toc132715265)

[b. state\_fips\_code 7](#_Toc132715266)

[c. res\_county 7](#_Toc132715267)

[d. country\_fips\_code 7](#_Toc132715268)

[e. age\_group 8](#_Toc132715269)

[f. sex 8](#_Toc132715270)

[g. race 9](#_Toc132715271)

[h. ethnicity 9](#_Toc132715272)

[i. process 9](#_Toc132715273)

[j. exposure\_yn 10](#_Toc132715274)

[k. current\_status 10](#_Toc132715275)

[l. symptom\_status 11](#_Toc132715276)

[m. hosp\_yn 12](#_Toc132715277)

[n. icu\_yn 12](#_Toc132715278)

[o. death\_yn 13](#_Toc132715279)

[p. underlying\_conditions\_yn 13](#_Toc132715280)

[3. Summary 14](#_Toc132715281)

**Overview**

This report outlines the initial findings based on the cleaned dataset (UPDATED-covid19-cdc-22200359.csv). It will summarise data and address the data quality issues and provide recommendation for the data quality plan. The dataset has numeric features with a high significant of missing data and the report will analyse each feature.

**1. Numeric Columns**

**Graphical user interface, application

Description automatically generated**

**Data Visualisation for Numeric Features**

Chart

Description automatically generated with low confidenceChart

Description automatically generated

## a. case\_positive\_specimen\_interval

This feature shows the weeks between the earliest date and the date of the first positive specimen collected. The first, and third quartiles and median are 0.0, which means most of the date the first positive specimen was collected, and the earliest date is the same. There are negative numbers, which means that the first specimen is collected earlier than the case is reported, which doesn’t make much sense. We can deduce that most of the data is 0, which is constant after removing the outliers.

There is 46% of data missing from this dataset, which represents almost half of the data. This is a high amount of data missing that might lead to inaccurate data analysis. Furthermore, this feature is also of less importance and has little effect on target outcome. Thus, we could ignore the fact that there are abnormal negative numbers in this feature as the feature will be dropped.

Recommendation: Drop feature

## b. case\_onset\_interval

This feature shows the weeks between the earliest date and the date of symptoms first offset. The first, and third quartiles and median are 0.0, which means that the majority of the date symptoms onset, and the earliest date are the same. There are negative numbers, which might mean that the first positive specimens are collected before the earliest date case was reported to the CDC. The negative numbers are explainable in this case. We can deduce that the majority of the data is 0, which is constant after removing the outliers.

There is 55.27% missing from this data. This is a high amount of data missing that might lead to inaccurate data analysis. Furthermore, this feature is also of less importance and has little effect on target outcome.

Recommendation: Drop feature

# **1.2 Timestamp Column**

The ‘case\_month’ feature is a special datatype that is a time series. It could be categorical or continuous. In this case, we treat it as continuous and see how the trend is for the date range on ‘case\_month’. However, this time series only applies to the context of the pandemic (cannot be used to predict infectious disease rate by itself, cannot make inferences eg. xxx increases as the date increases)

There is no missing data in this feature. We analyse it with a boxplot to see each month’s frequency, but it doesn’t make much sense as the frequency is relative to months, which is categorical. Thus, a line plot to see the pattern of the number of cases recorded is more reasonable for data visualisation.

|  |
| --- |
| a. case\_month A picture containing graphical user interface  Description automatically generated  Chart, box and whisker chart  Description automatically generated  Below is an arranged bar plot based on time:  Chart, line chart, histogram  Description automatically generated  Recommendation: na |

**2. Categorical Columns**

Graphical user interface, text

Description automatically generated

|  |
| --- |
| a. res\_state The “*res\_state”* data has 0 missing thus has no problem by itself but is a duplicated column for: “*state\_fips\_code”.*  Recommendation: Drop feature |
| b. state\_fips\_code state fips code is actually contains in county fips code, and can be removed in this case as the first numbers in county fips code include state fips code, thus considered as duplicate information available (reference https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt)  Recommendation: drop feature |
| c. res\_county The “*res\_county”* data has 0 missing thus has no problem by itself but is a duplicated column for: “*county\_fips\_code”.*  Recommendation: Drop feature |
| d. country\_fips\_code 6% missing value, replace with the mode of the column.  Recommendation: Replace null with mode value |

|  |
| --- |
| e. age\_group   There is about 0.6% missing data for this feature, which is a relatively low missing value in this dataset. No significant problem observed.  Recommendation: na |
| f. sex   There are 4 unique data records in this feature, including “unknown” and “missing”. There is about 2% missing data from this feature. “Unknown” meaning is vague, and the data also represented a small amount of data according to the bar plot. This should be considered and interpreted based on the context when conducting analysis related to this feature. In this problem, we could combine null, “unknown” and “missing” together to “missing” for further analysis categorisation.  Recommendation: Combine null, “unknown” and “missing” column |

|  |
| --- |
| g. race   There are 11.77% missing data in this feature. There is also 1669 “unknown” data represented. “Unknown” meaning is vague and should be considered and interpreted based on the context when conducting analysis related to this feature. In our problem, we could combine null, “unknown” and “missing” together to “missing” for further analysis categorisation.  Recommendation: Combine null, “unknown” and “missing” column |
| h. ethnicity   There are about 13% missing data in this feature. There is also 2541 “unknown” data represented. “Unknown” meaning is vague and should be considered and interpreted based on the context when conducting analysis related to this feature. In our problem, we could combine null, “unknown” and “missing” together to “missing” for further analysis categorisation.  Recommendation: Combine null, “unknown” and “missing” column |

|  |
| --- |
| i. process   Process has “missing” as its top frequent category according to its bar chart. There is a high amount of data missing that might lead to inaccurate data analysis. This feature is also in less importance for tackling the problem. Drop column as this feature has little effect on target outcome and has most values missing.  Recommendation: Drop feature |
| j. exposure\_yn   Exposure has “missing” as its top frequent category according to its bar chart. The missing category is significant. This is a high amount of data missing that might lead to inaccurate data analysis. There is also a vague “unknown” data records. This feature might be important in tacking problems that is related to the spread of disease, but not death, thus is less importance for tackling the problem. Drop column as this feature has little effect on target outcome and has most values missing.  Recommendation: Drop feature |

|  |
| --- |
| k. current\_status   No problem identified with the data in this feature.  Recommendation: na |
| l. symptom\_status   Based on the barplot, there is a significant missing data of the symptom. There is also a vague “unknown” column present in the data records. This should be considered and interpreted based on the context when conducting analysis related to this feature. In our problem, we could combine null, “unknown” and “missing” together to “missing” for further analysis categorisation.    Recommendation: Combine null, “unknown” and “missing” column |

|  |
| --- |
| m. hosp\_yn   The missing data should be considered and interpreted based on the context when conducting analysis related to this feature. In our problem, we could combine null, “unknown” and “missing” together to “missing” for further analysis categorisation.  Recommendation: Combine null, “unknown” and “missing” column |
| n. icu\_yn   *icu\_yn* has “missing” as its top frequent category according to its bar chart. This feature could have effect on targeted outcome, thus cannot be dropped yet. We could combine null, “unknown” and “missing” together to “missing” for further analysis categorisation.  Recommendation: Combine null, “unknown” and “missing” column |

|  |
| --- |
| o. death\_yn   There are no problems observed in this feature data.  Recommendation: na |
| p. underlying\_conditions\_yn   Has high missing data values. However, this feature could have effect on targeted outcome, thus cannot be dropped yet. Convert null to missing, and take missing values into account when analysing data  Recommendation: Combine null, “unknown” and “missing” column |

# **3. Summary**

Below is a list of actions required summarised that should be included in the data quality plan:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Quality Issue** | **Data Quality Plan** |
| case\_month | **-** | **-** |
| case\_onset\_interval | large amount null values | drop feature |
| case\_positive\_specimen\_interval | large amount null values | drop feature |
| state\_fips\_code | duplicate feature | drop feature |
| res\_state | duplicate feature | drop feature |
| country\_fips\_code | null values | replace null with mode |
| res\_county | duplicate feature | drop feature |
| age\_group | null/missing/unknown values | combine “unknown”and null to “missing” column |
| sex | null/missing/unknown values | combine “unknown”and null to“missing” column |
| race | null/missing/unknown values | combine “unknown”and null to“missing” column |
| ethnicity | null/missing/unknown values | combine “unknown”and null to“missing” column |
| process | large amount null values | drop feature |
| exposure\_yn | large amount null values | drop feature |
| current\_status | - | - |
| symptom\_status | large amount null values | drop feature |
| hosp\_yn | null/missing/unknown values | combine “unknown”and null to“missing” column |
| icu\_yn | null/missing/unknown values | combine “unknown”and null to“missing” column |
| underlying\_conditions\_yn | null/missing/unknown values | combine “unknown”and null to“missing” column |
| death\_yn | - | - |