COMP47350 Data Analytics Homework I

Name **Anil Vercruysse**
Student ID **22202474**

DATA QUALITY REPORT

1. Introduction

The COVID-19 pandemic has had a significant impact on the global population, with the United States being one of the most affected countries. In this report, we analyse the quality of a sample of the public data released by the Centers for Disease Control and Prevention (CDC) on COVID-19 cases in the United States. This dataset includes information such as demographic characteristics, exposure history, disease severity indicators, and survival outcomes.

The CDC is a health protection agency in the United States that provides deidentified individual-case data daily, submitted using standardised case reporting forms. Our sample of the dataset was obtained from the CDC COVID-19 data tracker, which is updated daily with new information. The dataset includes both categorical and continuous features that were classified according to their type.

The underlying objective of this analysis is to clean the data set with a view to develop a data analytics solution for death risk prediction using the data. The Data Quality Report presented here will first provide a summary of the initial dataset (section 2) and describe the initial data cleaning measures (section 3) implemented. The Data Analysis and Visualisation section will present the results of the data quality analysis (section 4). This will involve a detailed analysis of each feature, including any patterns and data quality issues that were identified, as well as potential strategies for handling them. Finally, the Recommendations section will summarise the findings (section 5) of the report.

2. Data Set Summary

The dataset used in this analysis comprises 20,000 rows and 19 columns, containing both categorical and continuous variables. The dataset provides a sample of COVID-19 cases in the United States that occurred between January 2020 and November 2022, and includes the survival outcome for every case. Most of the data was collected by the CDC using standardised case reporting forms, apart from the two continuous features that are calculated based on other data points.

The dataset presents the following limitations: Firstly, not all features are complete, which affects the completeness of the data. Additionally, there are no unique identifiers for individual cases, which makes it difficult to identify true duplicates. Furthermore, missing values are inconsistent, such as a 'Missing' value sometimes being referred to as 'Unknown' or left blank.

The following features were classified as categorical: case month, residence state and county, fips code, age group, sex, race, ethnicity, current and symptom status, hospitalisation and ICU admission status, death status, and underlying medical conditions. On the other hand, the following features were categorised as continuous: case_positive_specimen_interval (“CPSI”) and case_onset_interval (“COI”).

3. Initial Data Cleaning

Prior to conducting a quality analysis of the dataset, an initial data cleaning process was implemented. This process was guided by the assignment instructions and involved the elimination of duplicate rows and the removal of redundant columns. The initial dataset consisted of 20,000 rows without unique identifiers to differentiate individual case records. Utilising the Pandas Python Library, 1,112 duplicate rows were identified and subsequently removed based on the following argumentation:

> *(i) Data entry errors*: The use of standardised case reporting forms for recording cases suggests a potential source of data entry errors that may.compromise data quality by impacting its accuracy.
>
> *(ii) Low probability of true duplicates*: Considering that each case has 17 distinct features (excluding both FIPS codes, which were deemed redundant as explained later), the likelihood of encountering two cases with identical features is unlikely from a combinatorial perspective.
>
> *(iii) Low proportion of duplicates*: Approximately 5.5% of rows were duplicates (1,112 / 20,000), meaning that their removal would have a relatively low impact on sample size.
>
> *(iv) Absence of unique identifiers*: The lack of unique identifiers for each case in the dataset makes it challenging to ascertain whether these rows are true duplicates.

The option to retain the duplicate rows was also considered based on the following reasons:

> *(i) Loss of potential trends*: If the removed rows were in fact true duplicates, their elimination might obscure trends within the data (e.g., correlations between cases with identical features).
>
> *(ii) Reduced sample size*: Removing duplicates diminishes the overall sample size, potentially leading to a loss of statistical power.

Ultimately, the reasoning presented above led to the decision that eliminating duplicate rows would enhance the accuracy and relevance of the dataset for the quality analysis. Along with the removal of duplicate rows, two columns related to FIPS codes were discarded since they are redundant with the state and county columns.

The initial cleaning process resulted in a refined dataset containing 18,888 rows and 17 columns.

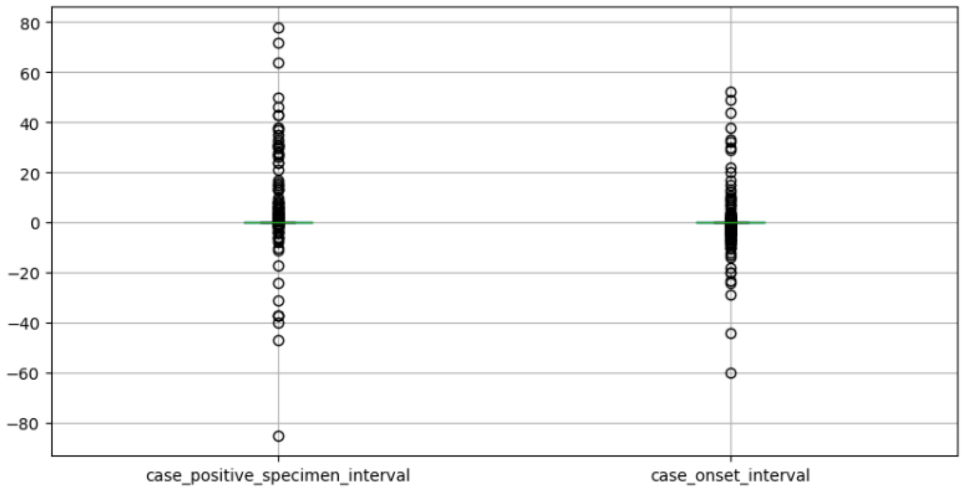
4. Data Quality Analysis and Visualisation

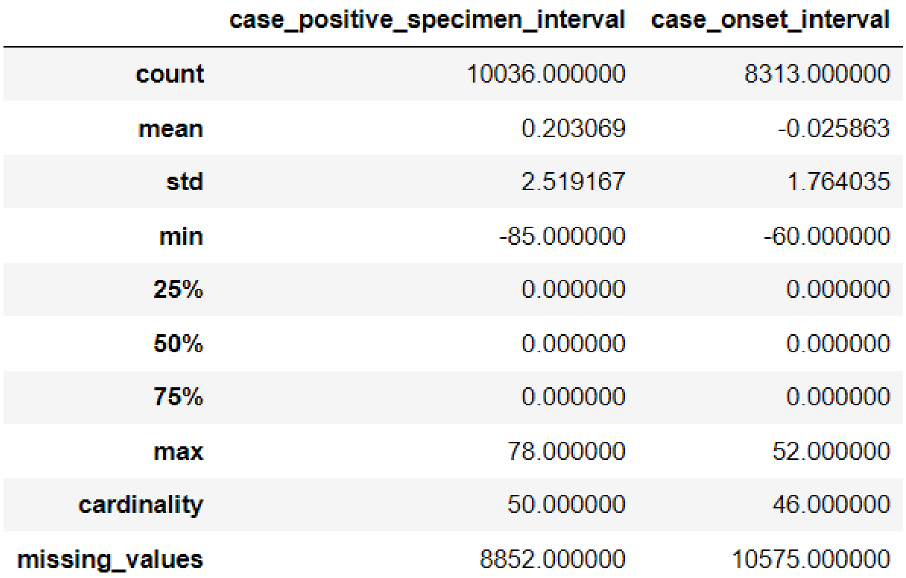
This section highlights the findings of our data quality analysis and presents data visualisations. Our goal is to reveal insights into the data, identify data quality issues, and suggest solutions.

A general observation is the dataset's inconsistent representation of missing values, appearing as 'Missing', 'Unknown', or a blank entry. We propose standardising missing values to 'NaN' for a more consistent data set and the avoidance of mismatches between string and integer data types later on.

a. Continuous Features

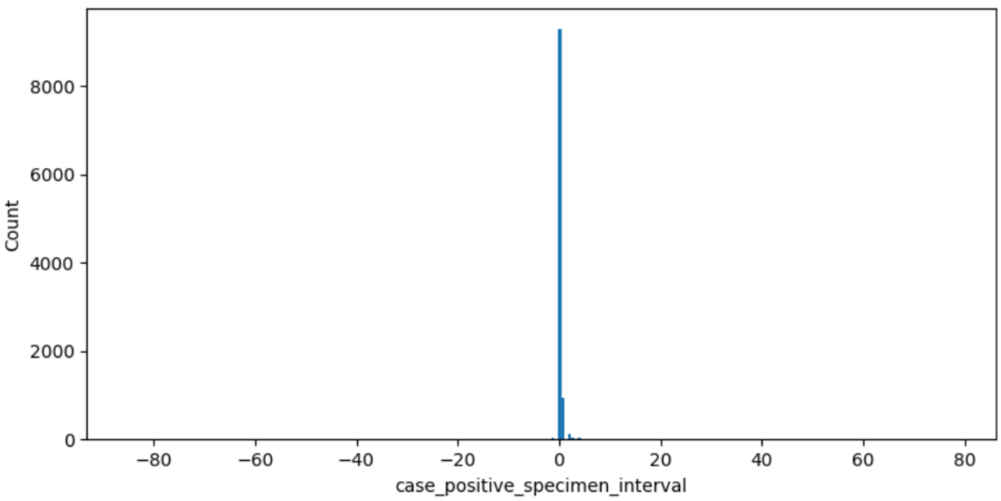
The box plot and descriptive statistics below show information for two continuous variables: CPSI and COI, both expressed in weeks. CPSI measures the weeks between a positive diagnostic test specimen collection and the date of reporting, while COI measures weeks between symptom onset and the date of reporting.



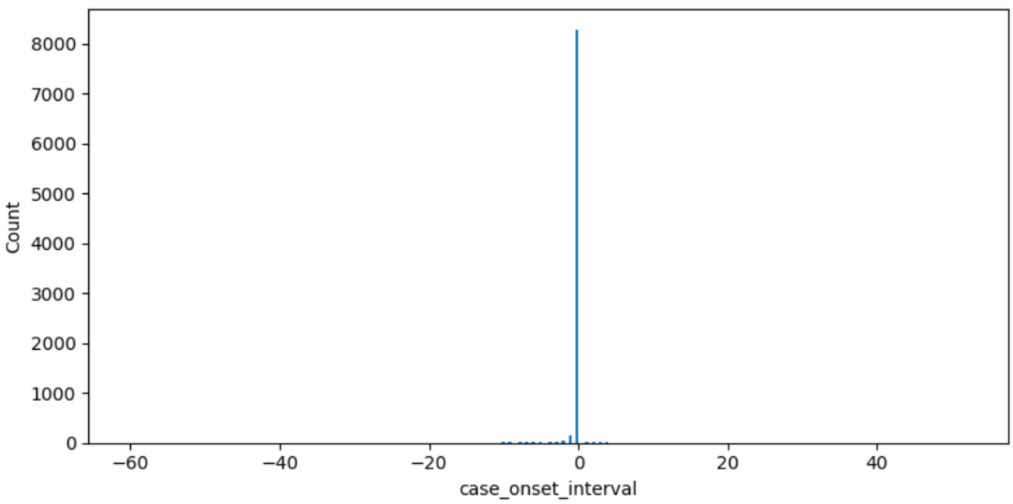


	case_positive_specimen_interval	case_onset_interval
count	10036.000000	8313.000000
mean	0.203069	-0.025863
std	2.519167	1.764035
min	-85.000000	-60.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	78.000000	52.000000
cardinality	50.000000	46.000000
missing_values	8852.000000	10575.000000

Both variables often have mean values close to 0, suggesting efficient COVID-19 testing and reporting processes. The cardinality reveals 50 unique values for CPSI and 46 for COI, indicating limited variability given the data set’s size. In addition, we observe that both features have a significant amount of missing values.



CPSI has a count of 10,036 and a mean of 0.203 weeks (~1.4 days), with a standard deviation of 2.519 weeks. The minimum and maximum values are -85 and 78 weeks, respectively, indicating potential data quality issues such as illogical negative values and possible outliers.



COI has a count of 8,313, a mean of -0.026 weeks, a standard deviation of 1.764 weeks, and minimum and maximum values of -60 and 52 weeks. Similar data quality issues arise, including illogical negative values, high missing values, and possible outliers.

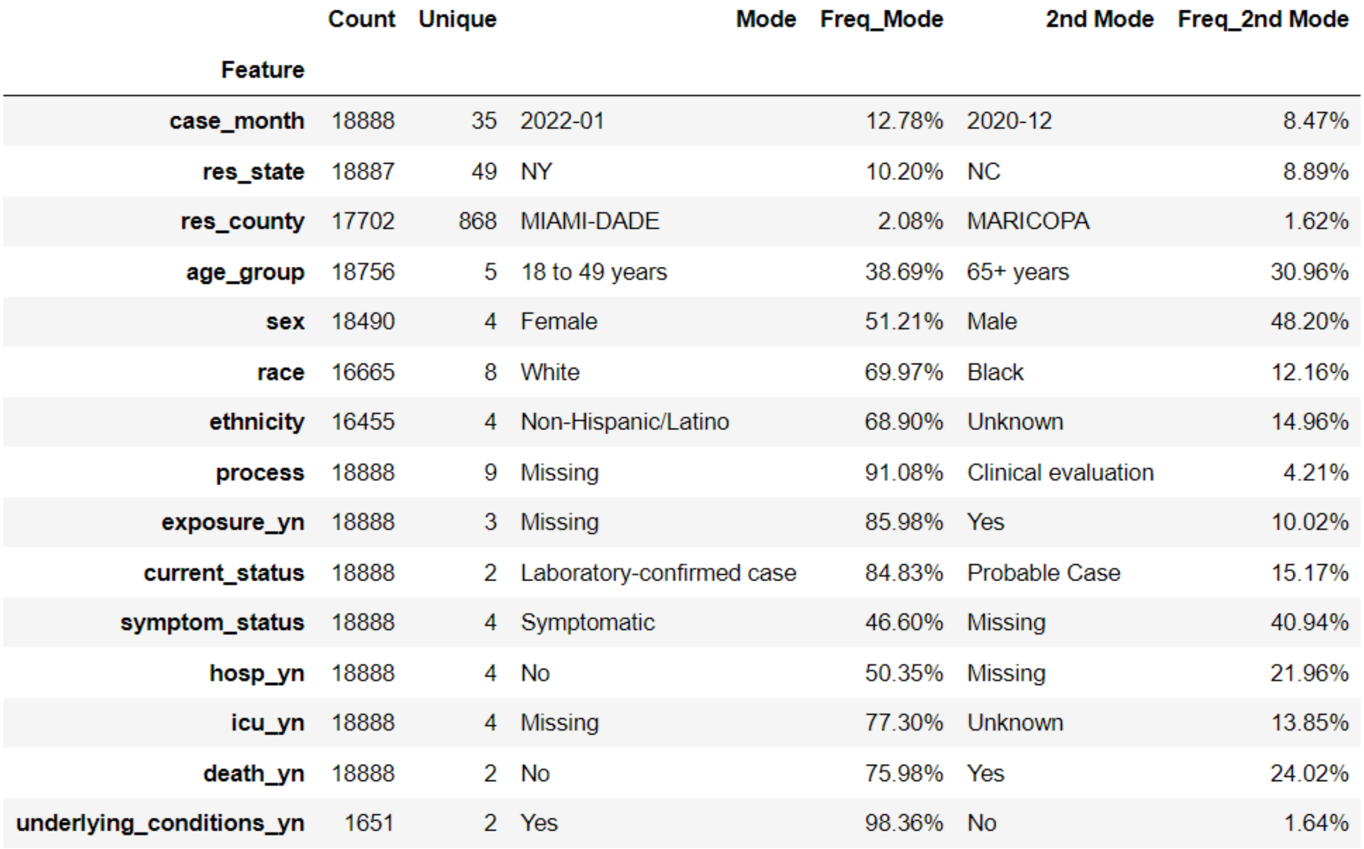
We observe that negative values for CPSI and COI are illogical since they suggest events occurring in reverse order. Negative values in CPSI would mean reporting occurred before a diagnostic test specimen was collected, which is implausible. Similarly, negative values in COI would indicate the illness was reported before symptoms appeared, which is also implausible.

Negative values might stem from data entry errors or inconsistencies in data collection. We recommend replacing negative values with 'NaN', as this avoids introducing illogical data into the set. The alternative of changing them to positive values was considered but deemed inappropriate as we cannot verify if these values are due to simple sign errors. Imputation and row dropping was considered to remedy the high missing values, but deemed inappropriate due to sample size concerns.

The relatively wide distribution between maximum and minimum values for both continuous features suggest that they have outliers, as is also reflected on the box plot above. However, no measures will be taken in respect of outliers as they may provide important information on extreme features for CPSI and COI. The option to clamp outlier data was evaluated, with the intent of substituting extreme values, but the informative significance of these outliers was deemed more important.

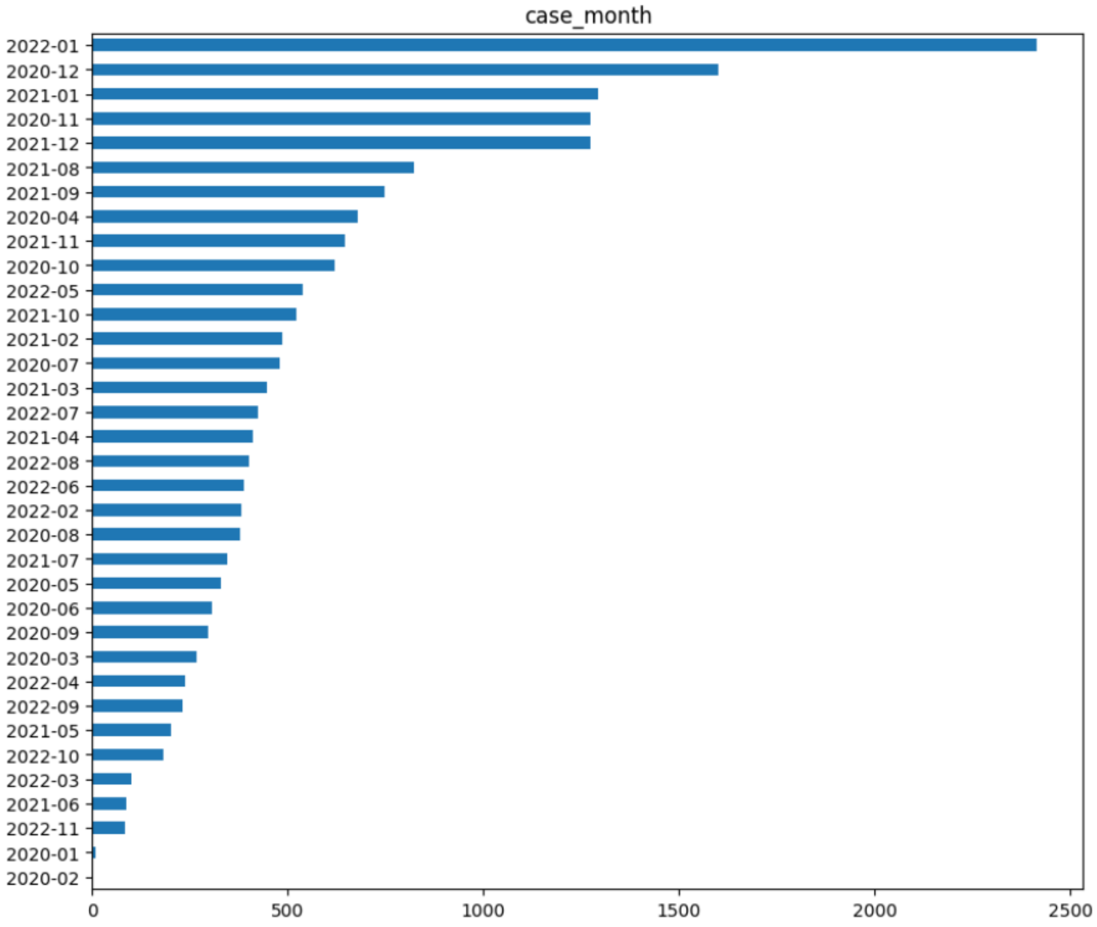
b. Categorical Features

In this section, we analyse each categorical feature using statistics and visualisations, identify data quality issues, and propose solutions to address them.



	Count	Unique	Mode	Freq_Mode	2nd Mode	Freq_2nd Mode
Feature						
case_month	18888	35	2022-01	12.78%	2020-12	8.47%
res_state	18887	49	NY	10.20%	NC	8.89%
res_county	17702	868	MIAMI-DADE	2.08%	MARICOPA	1.62%
age_group	18756	5	18 to 49 years	38.69%	65+ years	30.96%
sex	18490	4	Female	51.21%	Male	48.20%
race	16665	8	White	69.97%	Black	12.16%
ethnicity	16455	4	Non-Hispanic/Latino	68.90%	Unknown	14.96%
process	18888	9	Missing	91.08%	Clinical evaluation	4.21%
exposure_yn	18888	3	Missing	85.98%	Yes	10.02%
current_status	18888	2	Laboratory-confirmed case	84.83%	Probable Case	15.17%
symptom_status	18888	4	Symptomatic	46.60%	Missing	40.94%
hosp_yn	18888	4	No	50.35%	Missing	21.96%
icu_yn	18888	4	Missing	77.30%	Unknown	13.85%
death_yn	18888	2	No	75.98%	Yes	24.02%
underlying_conditions_yn	1651	2	Yes	98.36%	No	1.64%

(i) Case month

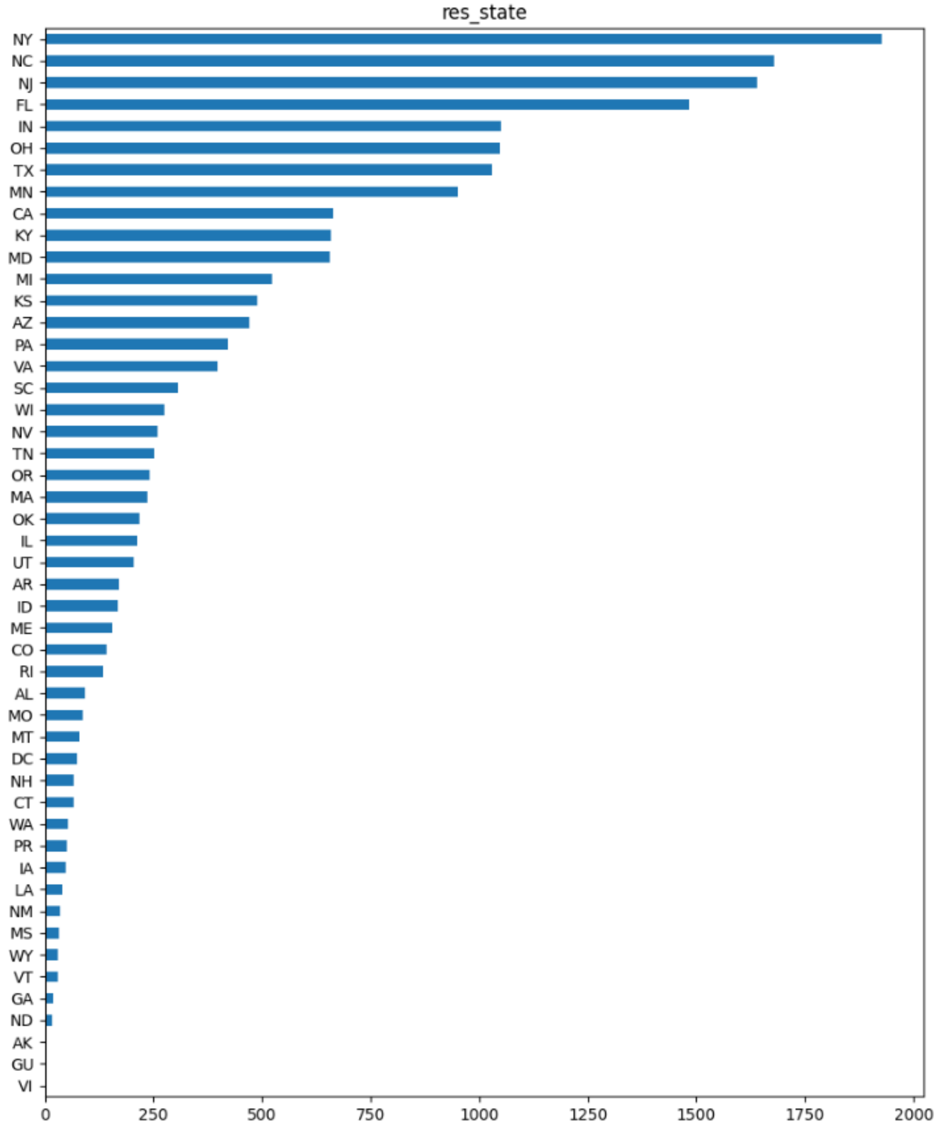


The dataset's case month feature contains 35 unique values, representing a 35-month period. The most frequent value, '2022-01,' accounts for 12.78% of total cases. Examining case distribution reveals higher case counts during winter months (November to February) and lower counts in summer months (May to August), aligning with typical respiratory infection seasonality.

No missing values exist for case month indicating dataset completeness for all 35 months. December 2020, the month with the second-highest case count, likely reflects heightened social activity during the holidays. Similarly, increased case counts in January 2021 and 2022 can be linked to indoor gatherings during winter.

We note that case month could be an important feature for creating an analytics solution for death prediction, as it may help capture seasonal patterns in case occurrences and fatalities.

(ii) State of residence

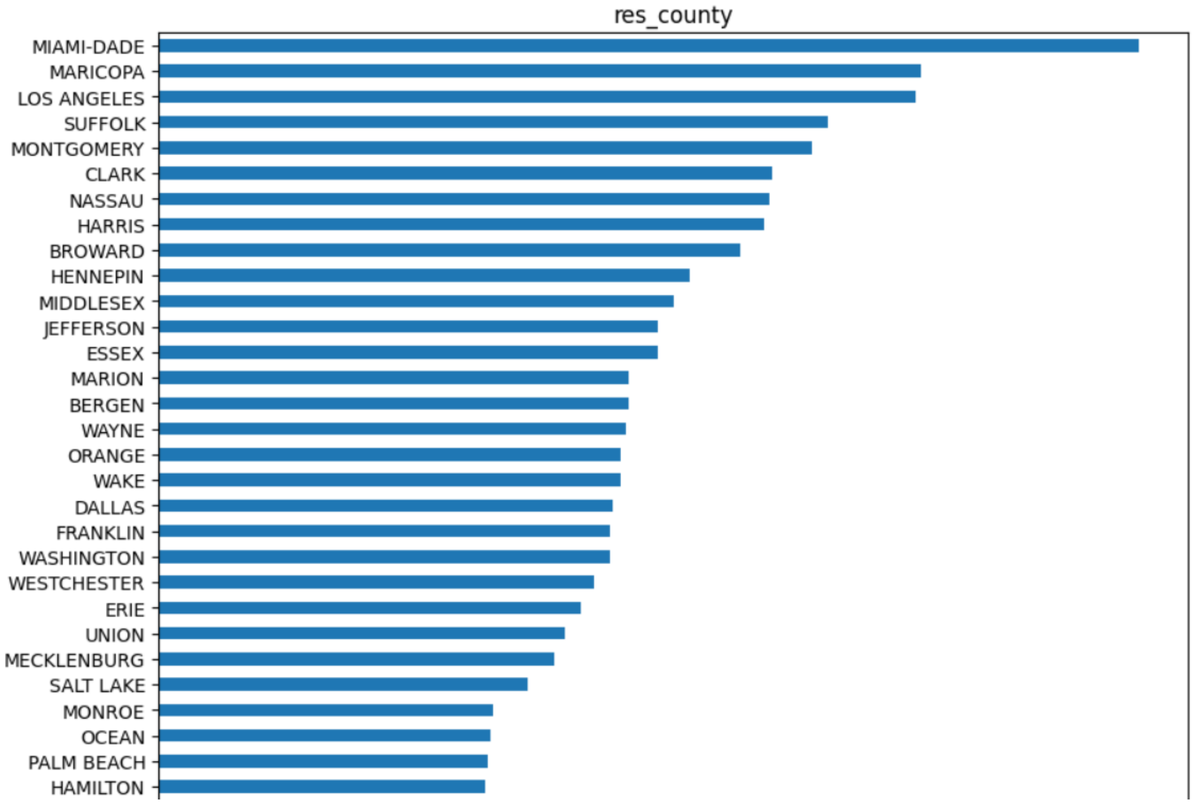


The state of residence feature in the dataset represents the state where COVID-19 cases were reported, with 49 unique values. It provides information on the geographic distribution of COVID-19 cases in the United States. New York has the highest number of reported cases, followed by North Carolina and New Jersey, while the Virgin Islands report the lowest case count.

The data's distribution suggests that factors such as population size and density, testing availability, and international travel may influence the varying case counts across different states. Larger and denser states, like New York and California, report higher case numbers, while smaller states with lower population density, such as Vermont and Wyoming, register fewer cases.

Regarding data quality issues, there is one missing value for the feature, representing a case with an unknown state location. This single missing value will not significantly impact the dataset's completeness and accuracy, as it represents just one out of 18,888 cases, so no data quality recommendations are proposed to address this issue. The state of residence is an important factor for death prediction, as it can reflect regional variations in healthcare infrastructure, public health policies, and socioeconomic factors that may influence the likelihood of severe outcomes from COVID-19.

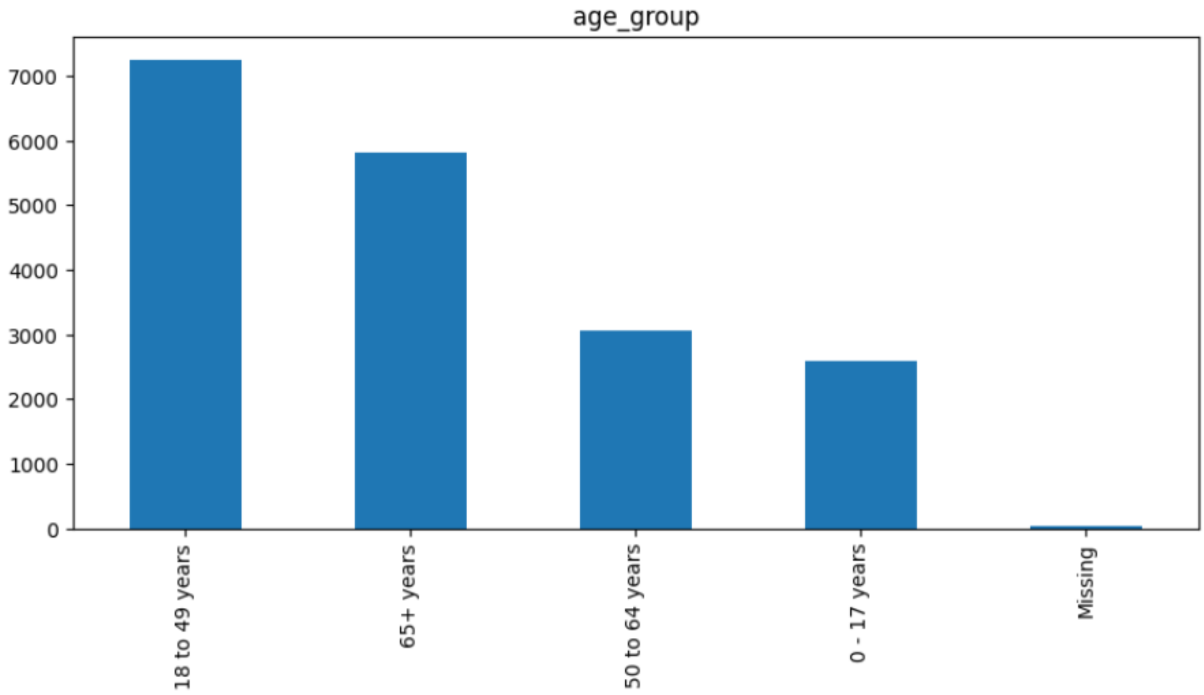
(iii) County of residence



The county of residence feature has 868 unique values, indicating that the dataset covers cases from a large number of counties in the US and a high cardinality. Note that only the top 30 counties are shown on the plot graph above, due to space constraints. Miami-Dade County has the highest concentration of cases (369 cases, 1.95%), followed by Maricopa (287 cases) and Los Angeles (285 cases).

A data quality issue is the absence of county data for about 6.3% of cases, which is relatively high, although we note that the location of these cases can be approximated by the state feature. Their removal would impact the sample size, so we recommend imputing the missing values with an equal distribution county case distribution (i.e. replace them based on the statistics for the county feature but ensuring that the state features match the newly attributed county). Furthermore, the high cardinality of this feature (868) was considered, but it was found that it is not due to inconsistent formatting (all counties have the same format in the file), so no action is recommended in this respect. County of residence is informative for death prediction as it may reveal local disparities in healthcare access, public health measures, and social determinants of health. Although it overlaps with the state of residence feature, it could account for potential interactions between regional and local factors.

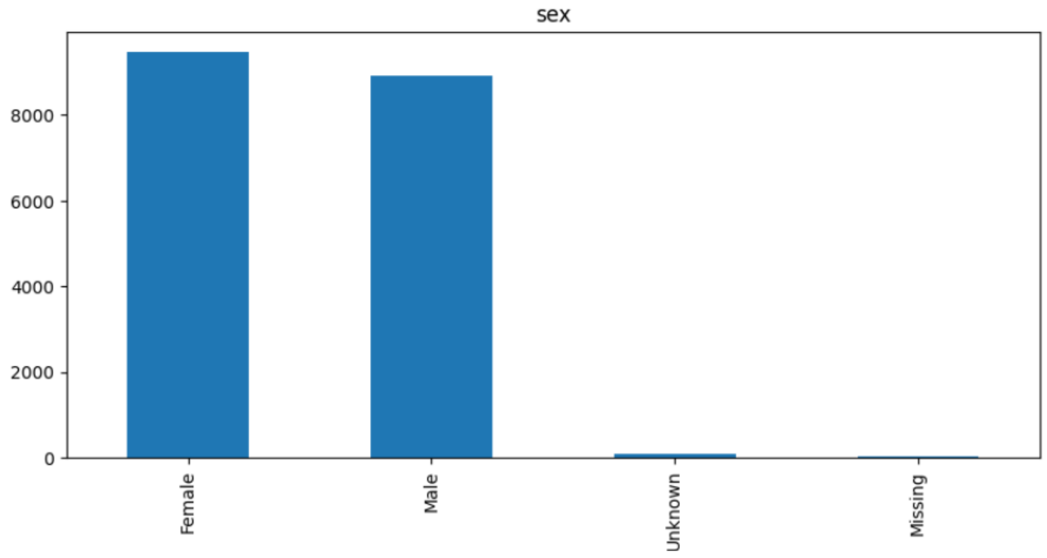
(iv) Age group



The age group feature contains five distinct categories, with '18 to 49 years' being the most prevalent (38.42% of cases). This suggests a significant number of COVID-19 infections occurred among individuals aged 18-49. Examining the other age categories, the '65+ years' group constitutes 31.36% of cases, indicating a considerable proportion of infections among older individuals. The '50 to 64 years' group accounts for 15.99% of cases, while the '0 - 17 years' category has the lowest frequency at 13.54%.

The feature has 29 missing values, therefore it is advised to remove the rows with these missing values. While we could keep these rows or impute their value based on the age group statistics, their removal is relatively immaterial. The feature is crucial for death prediction as old age is a well-established risk factor for COVID-19 severity and mortality.

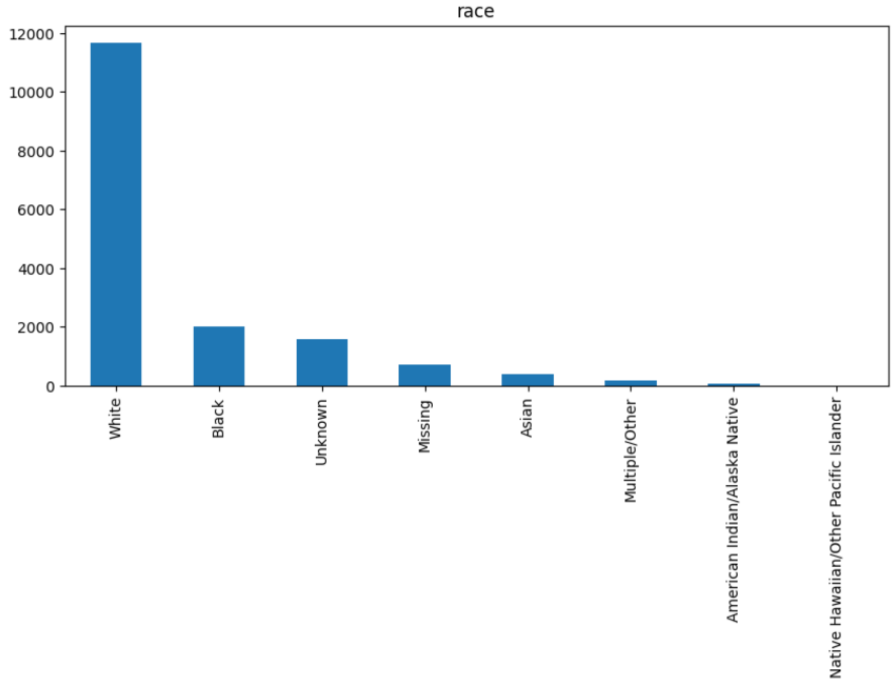
(v) Sex



The sex feature contains four distinct categories: Female, Male, Unknown, and Missing. Females make up 50.13% of cases, while Males represent 47.19%, indicating a fairly balanced distribution of COVID-19 cases between males and females in the dataset. We note that the sex distribution may not reflect the true population due to factors such as testing biases (e.g., women might be more likely to get tested for COVID-19 than men).

As the Unknown and Missing cases are of low prevalence, it is advised to impute these values based on the existing statistics for the sex feature. We could consider dropping the rows or retaining them, but this would respectively reduce the sample size or reduce the quality of the data set. Moreover, examining the connection between sex and COVID-19 death prediction could provide valuable insights for the analysis.

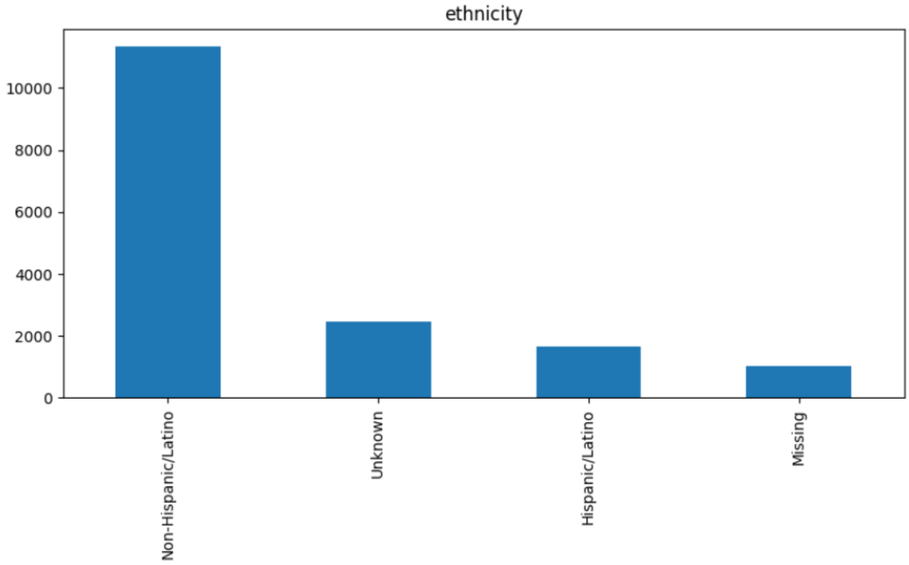
(vi) Race



The race feature comprises 8 unique values, with 'White' being the predominant category, making up 61.73% of cases. The 'Black' category ranks as the second most frequent value at 10.68%, signifying a considerable portion of cases among individuals identifying as Black. The 'Unknown' category, with a frequency of 8.24%, points to a lack of self-identification or incomplete data reporting. The remaining categories have relatively low frequencies.

The primary data quality concern for this feature revolves around the numerous Unknown or Missing values (totaling at 24% with all the different ‘missing’ formats). The removal of these rows would significantly impact the sample size, so it is instead advised to impute these values based on the existing race statistics. This feature can be a significant factor in death prediction, for instance highlighting disparities in healthcare access or socioeconomic status among different racial groups.

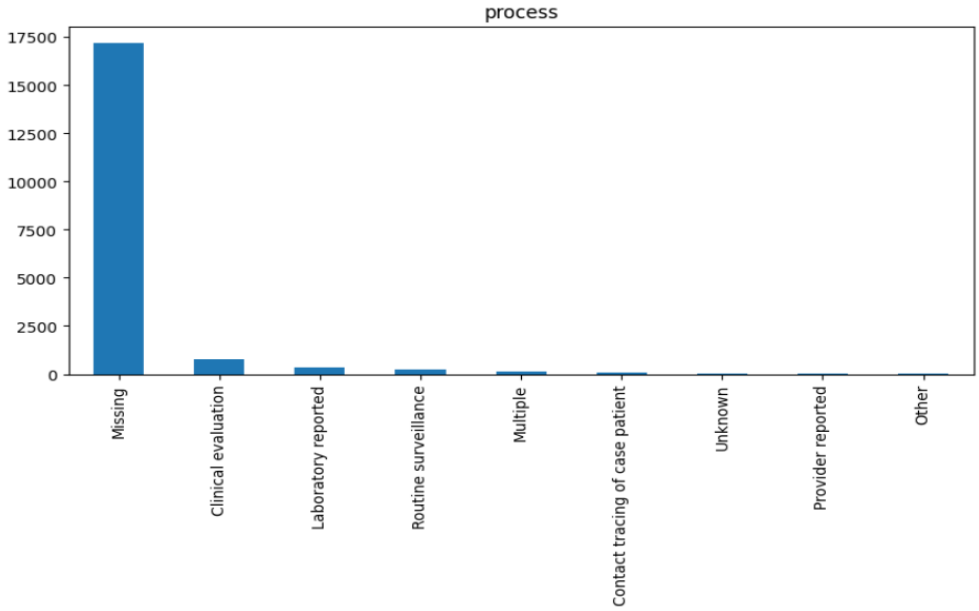
(vii) Ethnicity



The 'ethnicity' feature in the dataset contains 4 unique values, with 'Non-Hispanic/Latino' being the most prevalent at 60.02%. The 'Unknown' category has a substantial frequency of 23.26%, and the 'Missing' category accounts for 9.59%, potentially reflecting incomplete reporting or self-identification issues. The 'Hispanic/Latino' category represents 15.52% of cases.

Missing data may impact the distribution among different ethnic groups. Seeing the high number of Unknown or Missing values, removing them would significantly reduce the sample size. Therefore, it is recommended to retain the relevant rows and instead impute the missing and unknown values based on the non-missing distribution that we have for this feature.

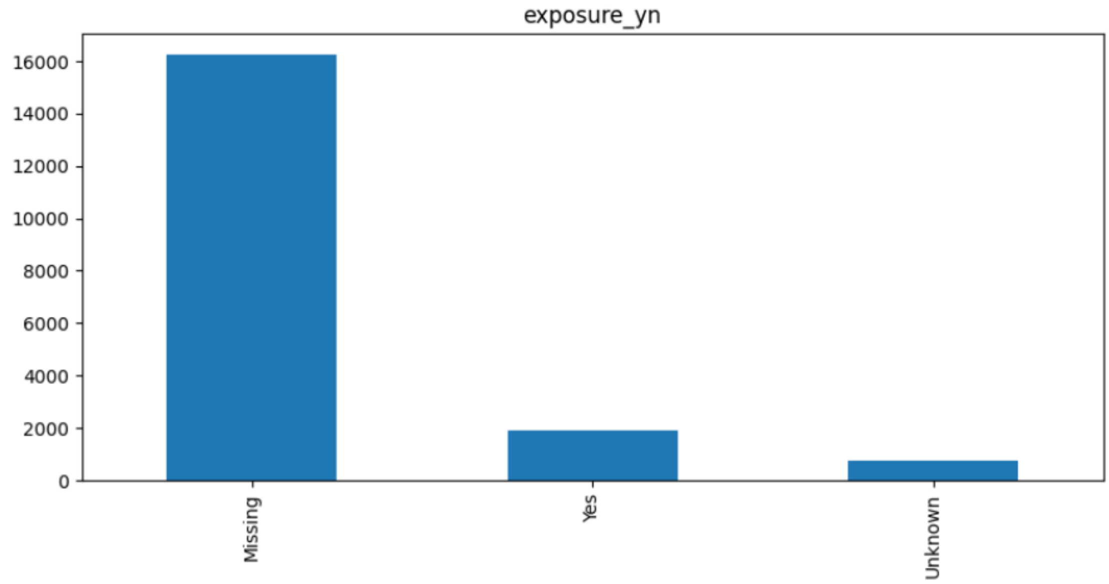
(viii) Case identification process



This feature shows the process by which cases were identified and encompasses nine unique values, with 'Missing' as the most frequent value, accounting for 91.08% of entries. Its high proportion of missing values may substantially impact the dataset's quality.

Given the majority of values are missing, drawing meaningful conclusions from this feature is challenging. Seeing the goal of the data cleaning exercise, it seems recommendable to drop this feature altogether as the process by which a case was detected has a limited value in terms of its ability to provide valuable death prediction insights.

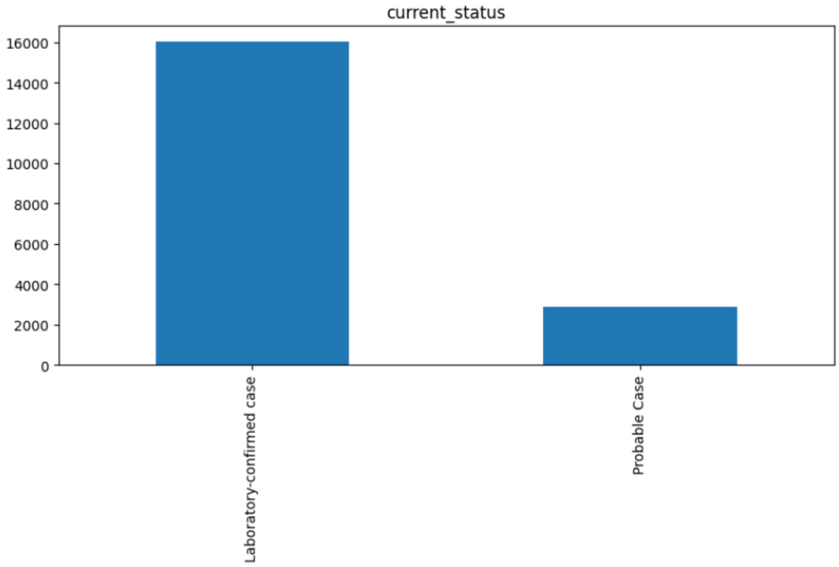
(ix) Exposure status



This feature shows whether the patient has had particular exposures with a known COVID-19 case in the 14 days prior to their illness onset, and consists of three unique values, with 'Missing' as the most prominent, accounting for 85.98% of entries.

The remaining values in the 'exposure_yn' feature, 'Yes' and 'Unknown,' have frequencies of 10.02% and 4.00%, respectively. Given the aim of this data cleaning exercise, it is advisable to remove this feature. Indeed, the exposure information doesn't offer valuable insights for death prediction, as contracting COVID-19 inherently implies exposure to the virus. The removal or imputation of these rows is also inappropriate given the 16996 missing or unknown values.

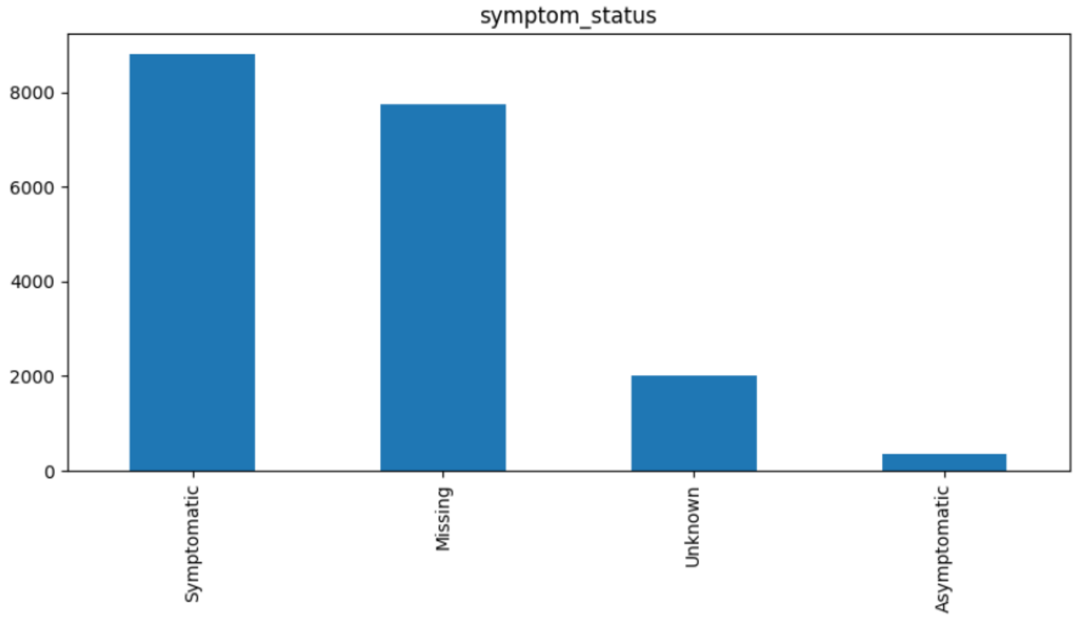
(x) Current status



This feature has two unique values (0 Missing), with 'Laboratory-confirmed case' being the most frequent at 84.83%. This implies that the majority of cases were confirmed through laboratory testing, which generally yields accurate and reliable data.

The other value, 'Probable Case,' has a frequency of 15.17%, suggesting that some cases were not confirmed via laboratory testing but met the criteria for a probable case. No further data cleaning measures are recommended under this feature, since it is complete and might provide death prediction insights. The confirmation status of a COVID-19 case should indeed be retained as a feature for death prediction, as it provides context on the reliability of case data.

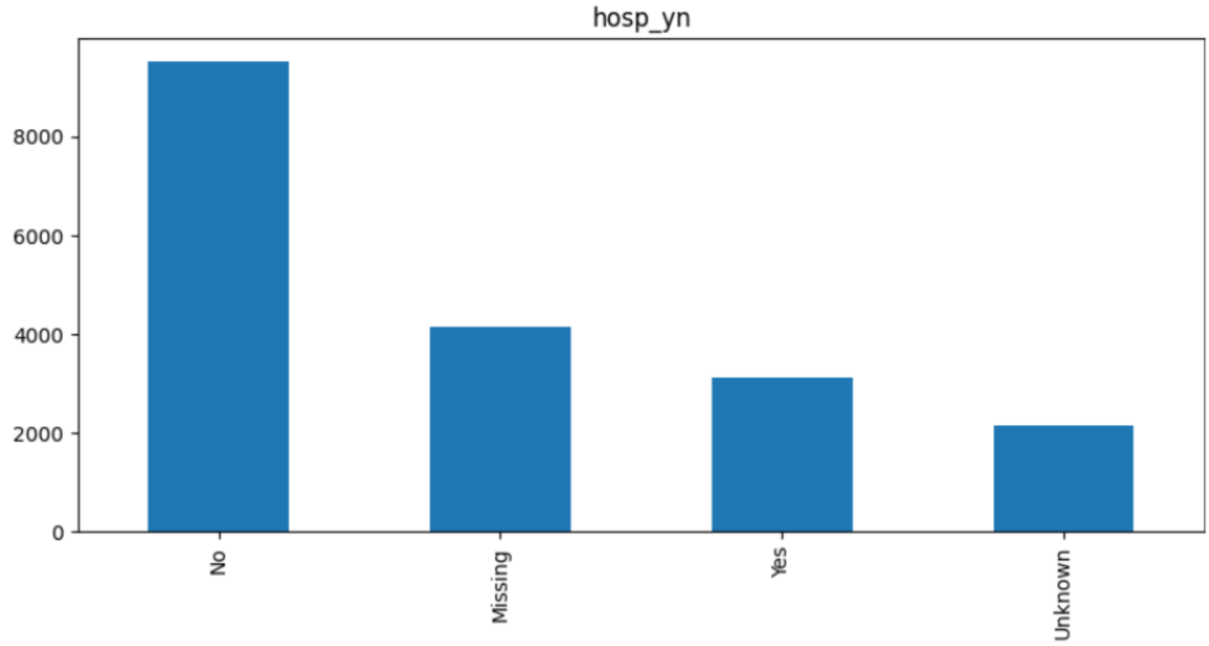
(xi) Symptom status



The symptom status feature in the dataset has 4 unique values, with 'Symptomatic' being the most frequent at 46.60%. This indicates that a significant proportion of cases are symptomatic. However, the accuracy of analyses relying on symptom status might be affected by a high number of missing and unknown values in this feature (49.76%). It should also be noted that symptom status may change over time, which means that cases that were asymptomatic at the time of reporting could have become symptomatic at a later point in time.

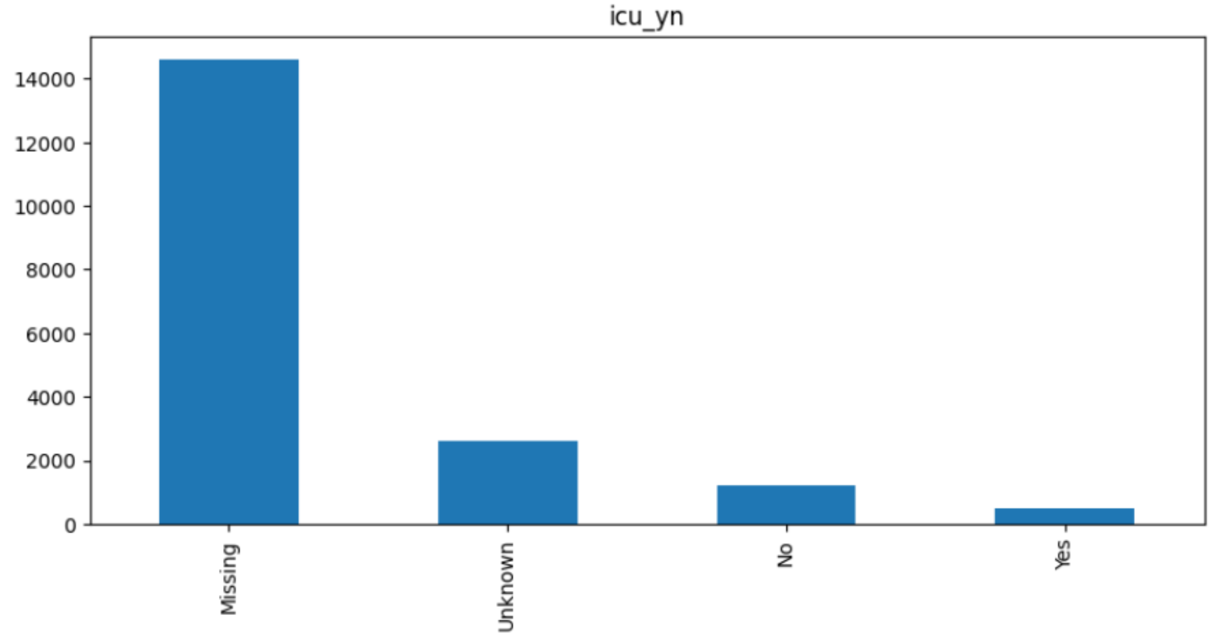
Symptom status can be important, as it provides information on the severity of the disease, which can be related to the risk of death. However, we recommend that this feature is dropped for two reasons: 50% of the data missing makes this feature difficult to apply in a future correlation analysis, and the symptom status feature is changeable and therefore cannot be effectively captured reliably.

(xii) Hospitalisation outcome



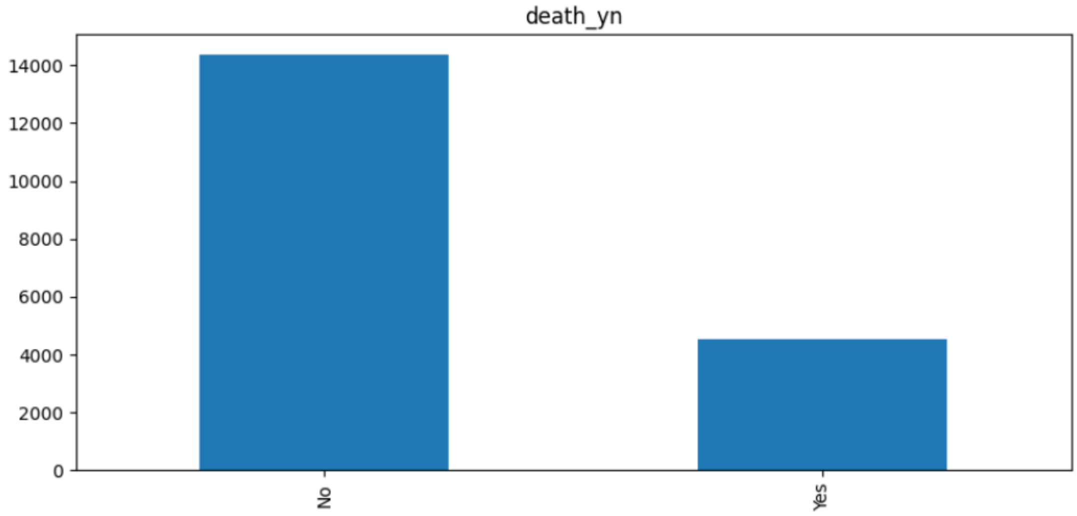
The hospitalisation feature in the dataset reveals that 50.35% of cases did not necessitate hospitalisation, and 24.04% of cases did require it. This information is essential for comprehending the severity of cases within the dataset, which is particularly relevant for death prediction. We also note that a considerable proportion of cases, 33.25%, have an unknown or missing hospitalisation status. Seeing the relevance of this feature for death prediction and the relatively high amount of non-missing values, it is recommended to keep this feature but to complete it by imputation. While it is generally not recommended to apply imputation where +30% of data is missing, we are relatively close to the threshold here and this feature has a potentially high informative value.

(xiii) ICU outcome



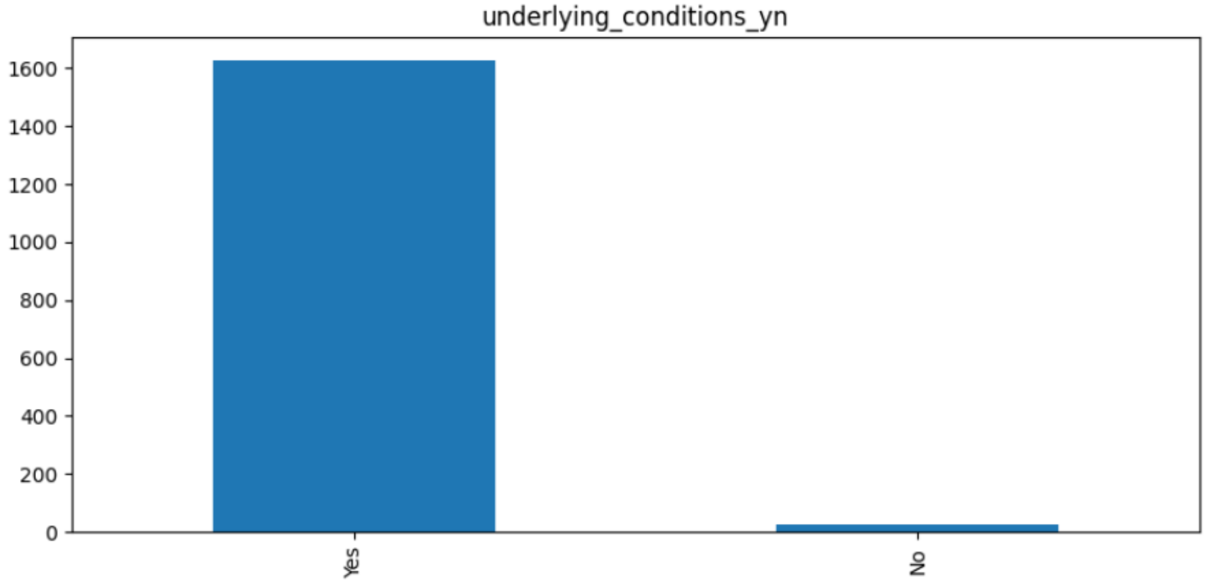
The ICU feature has a substantial proportion of missing and unknown values, constituting 91.15% of the data. Among the non-missing values, the majority of cases (71.15%) did not necessitate ICU admission. Given that the relation between the risk of death and an ICU outcome could bring valuable insights, it is difficult to make a recommendation for this feature. However, since 91.15% of values are missing, it is inappropriate to drop the relevant rows and impute their value. It is therefore recommended that this feature be dropped, seeing that no other measures can be applied to improve its quality and that it is of limited value given its current state.

(xiv) Death outcome



The death feature indicates that 75.98% of cases in the dataset did not result in death, while 24.02% of cases did, and does not have any missing values. This feature is essential to determine the severity and outcomes of COVID-19 cases. No data cleaning measures are recommended for this feature, as it is a key comparator for death prediction analysis.

(xv) Underlying conditions



The underlying conditions feature is only reported for a small subset of cases (1651), which may limit its usefulness in understanding the impact of underlying health conditions on COVID-19 cases. Among the non-missing values, the vast majority (98.38%) of cases had underlying health conditions.

The relation between the risk of death and underlying conditions could bring valuable insights. However, since 91.26% of values are missing, it is inappropriate to drop the relevant rows and impute their value. It is therefore recommended that this feature be dropped, seeing that no other measures can be applied to improve its quality.

5. Recommendations

In the context of the data quality report for the CDC's COVID-19 dataset, our objective was to assess the dataset's quality and identify potential data quality issues, with the aim of preparing the data for a data analytics solution for death risk prediction. Our analysis resulted in the following key observations and recommendations, which are restated in the Data Quality Plan:

1. **Inconsistent representation of missing values** was identified across the dataset. We recommended standardising missing values to 'NaN' for consistency.
2. **Outliers in the data of both continuous features** were identified, but no measures are recommended as these can provide insights into extreme cases.
3. A significant amount of **missing values was found for continuous features**, but the removal of the relevant rows would reduce the sample size too significantly, and imputation was inappropriate.
4. **Illogical negative values were detected for both continuous features**. We recommended replacing these negative values with 'NaN'.
5. **1186 rows with missing values were identified for the County of residence** feature. Imputation based on the non-missing distribution was recommended to remedy this, keeping in mind that the state feature should match the new value.
6. **A relatively high cardinality of 868 was identified for the County of residence** feature, but no recommendation was made in this respect since the high cardinality is real and not due to a formatting issue.
7. **29 rows with missing values for the age group feature** were identified, and their deletion was recommended seeing the importance of this factor for death prediction and the low impact on sample size.
8. **398 rows with missing values for the sex feature** were identified, which we will impute based on the non-missing sex feature distribution.
9. **4534 rows with missing values were identified for the race feature**, which we will impute based on the non-missing race feature distribution.
10. **5911 rows with missing values were identified for the ethnicity feature**, which we will impute based on the non-missing ethnicity feature distribution.
11. It is recommended that the **case identification process feature is dropped** as it is highly incomplete and brings no value to a death prediction solution.
12. It is recommended that the **exposure status feature is dropped** as it is highly incomplete and brings no value to a death prediction solution (contraction of COVID inherently implies being exposed to it).
13. It is recommended that the **symptom status feature is dropped** as it is highly incomplete and changeable over time.
14. **6218 rows with missing values were identified for hospitalisation** outcome, which we will impute based on the non-missing hospitalisation feature distribution.
15. The **features underlying conditions and ICU outcome were dropped**, due to a significant number of rows with missing values (17216 and 17237 respectively).

The findings highlighted above and throughout the report are reiterated in the Data Quality Plan. The data quality plan outlines handling strategies for each feature's data quality issues. By following the above recommendations, the dataset’s quality will be improved and provide a more robust foundation for the development of data analytics solutions for death risk prediction.