Data Quality Report – Initial Findings

Peijin JIANG 22202041

1.Overview

This report aims to present the preliminary findings obtained from the CDC dataset sample provided. The report intends to provide an overview of the data, including a comprehensive summary of the various data quality issues that were identified during the analysis. Furthermore, the report will discuss the steps taken to address these issues, ensuring the data's accuracy and reliability.

In addition to the data summary and quality assessment, the report will also include a detailed description of the features observed in the dataset. This will be presented in the form of feature summaries and boxplots that are used to effectively visualize the data.

It is important to note that the analyses presented in this report are based on the thoroughly cleaned data. This ensures that any erroneous or incomplete data that could potentially skew the results has been removed, and the conclusions drawn from the analysis are reliable and accurate.

2.Summary

① To ensure the logical integrity of the data, various tests were carried out, which uncovered a total of 11 cases of illogical data. For instance, continuous features should not have negative values, and their presence is deemed illogical. Any such illogical data must be addressed and evaluated by domain experts. For further details, refer to the Logical Integrity section of this report.

② The dataset contains a considerable number of null, "unknown," and "missing" values. Following a thorough data analysis, it was discovered that the "res_county," "county_fips_code," "case_positive_specimen_interval," "case_onset_interval," and "underlying_conditions_yn" features all have null values. The "age_group" feature has both null and "unknown" values, whereas "sex," "race," and "ethnicity" features contain null, "unknown," and "missing" values. Additionally, the "process," "exposure_yn," "symptom_status," "hosp_yn," and "icu_yn" features have "unknown" and "missing" values, while the "case_month," "res_state," "state_fips_code," "current_status," and "death_yn" features have no null, "unknown," or "missing" values.
Null, "unknown," and "missing" values for all features signify that the data is unavailable and hold no other significance. The nuances of their meanings do not impact the data analysis process and must be treated as null values in the future to indicate that the data has not been mastered. Not treating them as null values would negatively impact the intuitive understanding of "unavailable data."

③ Carrying out statistics with "unavailable data" uniformly can result in a low proportion of effectively mastered values for certain characteristics.

④ The dataset does not contain any constant columns, and all columns have a cardinality greater than one. Therefore, no further processing is necessary. The cleaned dataset is stored in a single CSV file named covid19-cdc-22202041-2.csv .

⑤ Duplicate rows are present in the dataset, with 2000 rows of duplicate data accounting for accuracy 10% of the total dataset. Deleting these rows is necessary since the possibility exists that the duplicate data originates from independent patients.

⑥ The dataset lacks a primary key, which reduces the quality of the original data and the uniqueness between different data. This is also one of the primary reasons for the appearance of duplicate rows.

⑦ For categorical values, I recommend making some modifications.

⑧ The feature set contains numerous outliers, which require further investigation. Although these values appear plausible initially, it is crucial to examine them thoroughly.

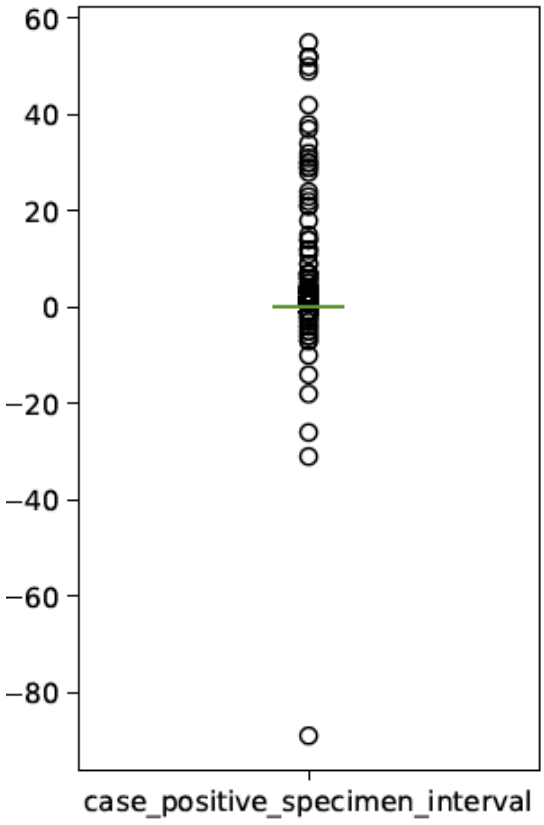
3.Review Logical Integrity

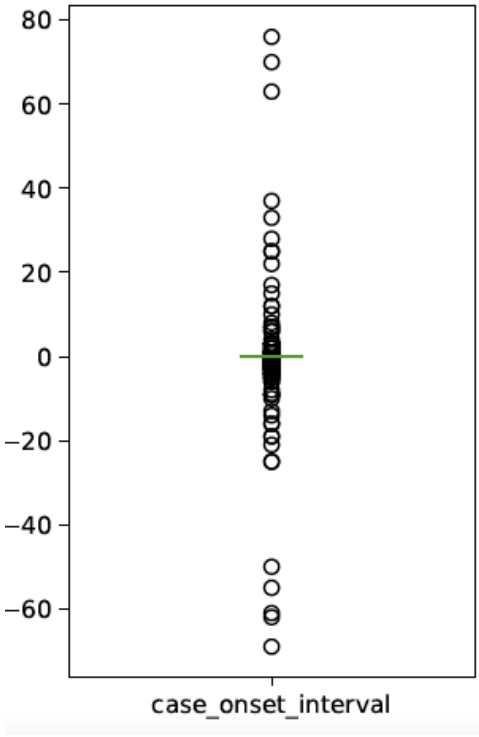
If the value of case_positive_specimen_interval is negative, it indicates that the positive result occurred prior to the patient being tested. Similarly, if the value of case_onset_interval is negative, it means that the case was confirmed prior to the patient exhibiting symptoms. The presence of negative values in these cases is logically unreasonable and requires further investigation.

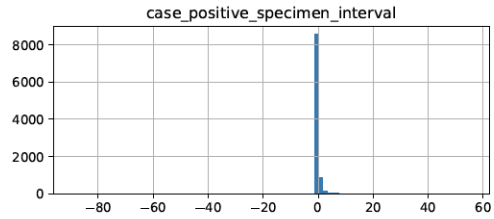
There are several possible reasons for the appearance of negative values. Firstly, it could be due to an error in calculation, where the subtrahend and minuend are mistakenly reversed. Alternatively, it could be a result of incorrect data entry during the initial stage. Additionally, discrepancies in date formats can also lead to the occurrence of negative values. Lastly, inaccurate reporting or processing delays can also result in negative values.

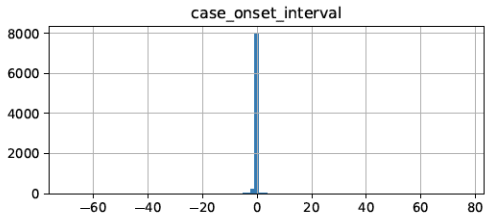
While it is possible that some patients do not exhibit any symptoms but still test positive for the disease, this does not explain the presence of negative values in case_onset_interval. Upon closer examination of the data, it is clear that this explanation does not hold, as asymptomatic cases should not result in negative values for case_onset_interval.

4.Review Continuous Features









There are only the above two continuous features, which represent the time from the time when the sample was sent to the positive test and the time from the onset of symptoms to the positive time. From these two data, we can conclude:

1. In most cases, the detection time and onset time are both shorter than 5 days.

2. In the case_positive_specimen_interval data, part of the data is concentrated in 20-30 days, guessing that the case was tested positive again.

3. The negative value may be due to the reverse writing of the date of detection and date of onset, resulting in a calculation error. It may also have no practical significance in clinical practice and can be discarded.

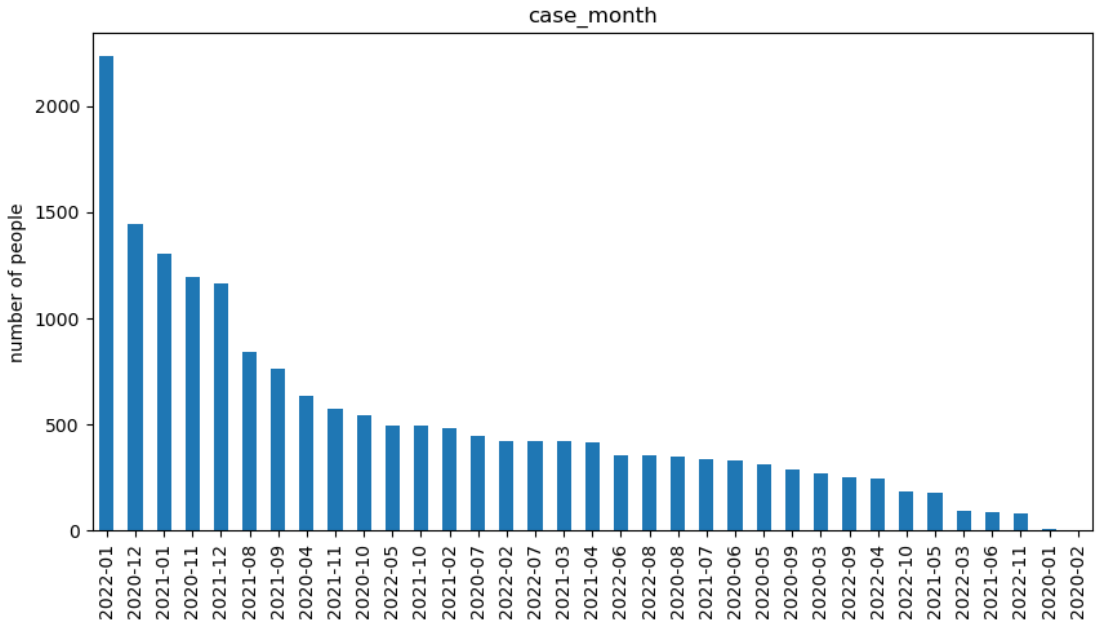
4.When processing, you can consider rounding off negative values or taking absolute values of negative values.

5.Some discrete values that are too extreme can be discarded.

5.Review Categorical Features

There are 17 categorical features in the dataset, all of which are describing a single object.

Let us conduct a brief analysis of some of the more interesting and clearly directional content.



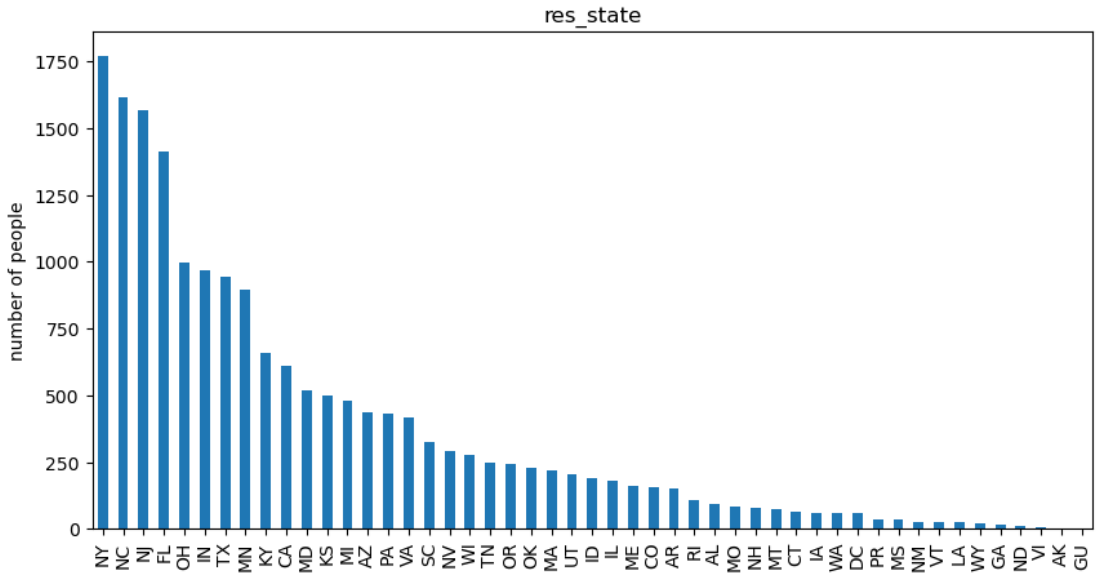
1. The development of the peak has experienced many peaks and valleys since its initial formation.

2. The origin time is January 2020, and then gradually grows to form the first peak: the end of 2020.

3. After June 2021, a second peak will gradually form, peak in January 2022, and then decline.

4. During each peak and valley period, the data fluctuates and does not increase or decrease linearly.

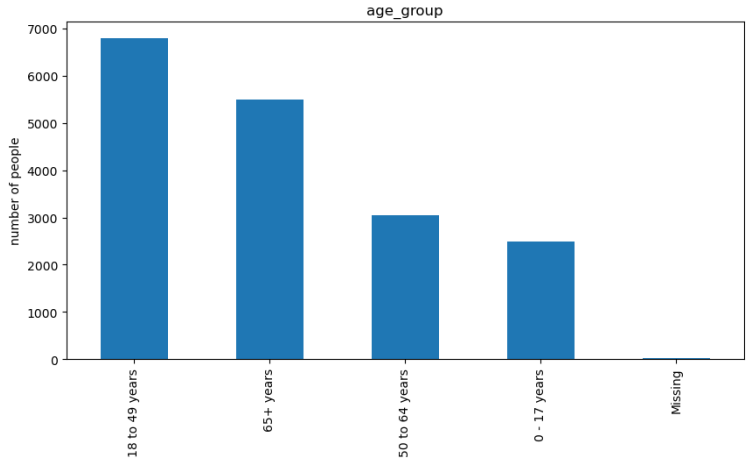
5. Therefore, it is not the best option to use a bar chart to represent this table. It may be more reasonable and easier to observe by constructing a bar chart or line chart in the time order of year-month.

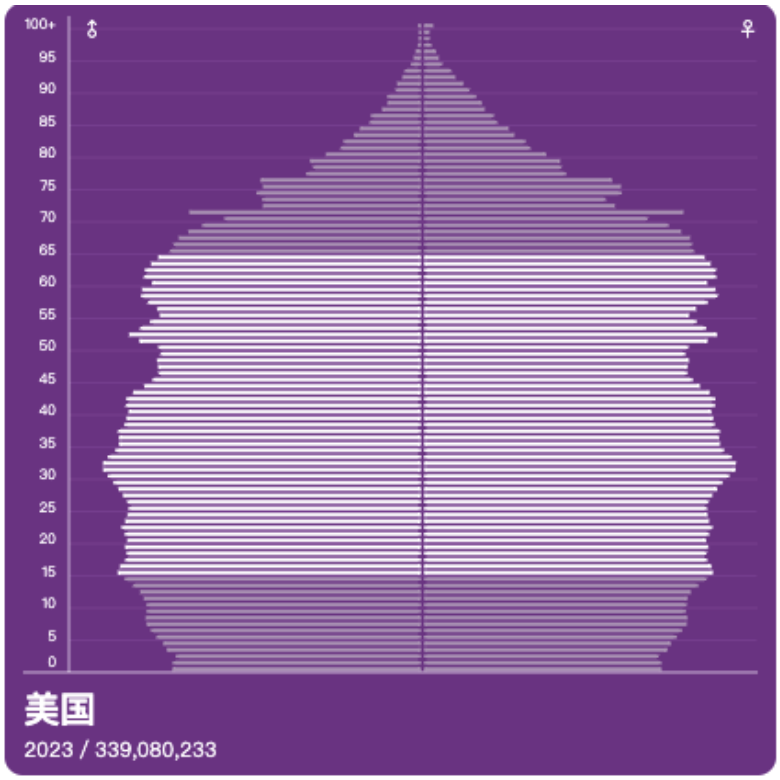


1. The total number of cases in the eastern and western coastal areas is obviously more.

2. The number of cases is positively correlated with the population base.

3. The number of cases is positively correlated with population density in a certain sense. It shows that an excessively high population density may bring about more case transmission, resulting in a higher infection rate and total infected population.

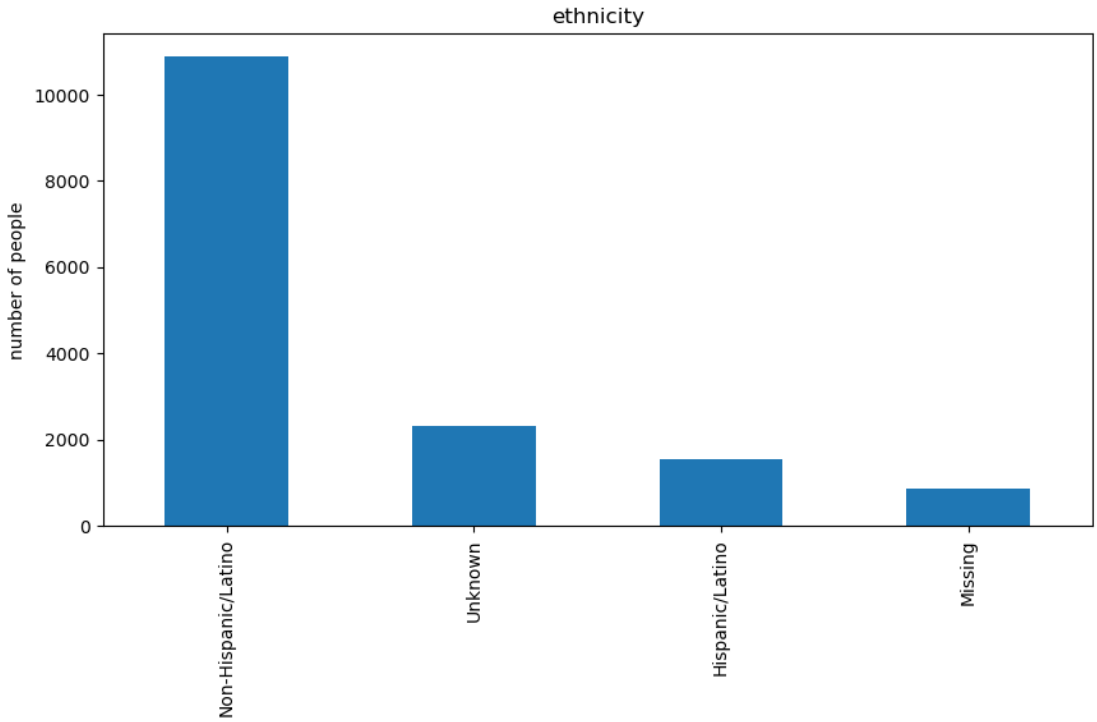




1. The figure above shows the age distribution of the case population in this document. For the convenience of comparison, the figure below introduces the latest US population pyramid (the picture comes from the US government website).

2. It can be seen from the table that the number of young and middle-aged people who are infected is equivalent to the proportion and absolute number of their population in the total population.

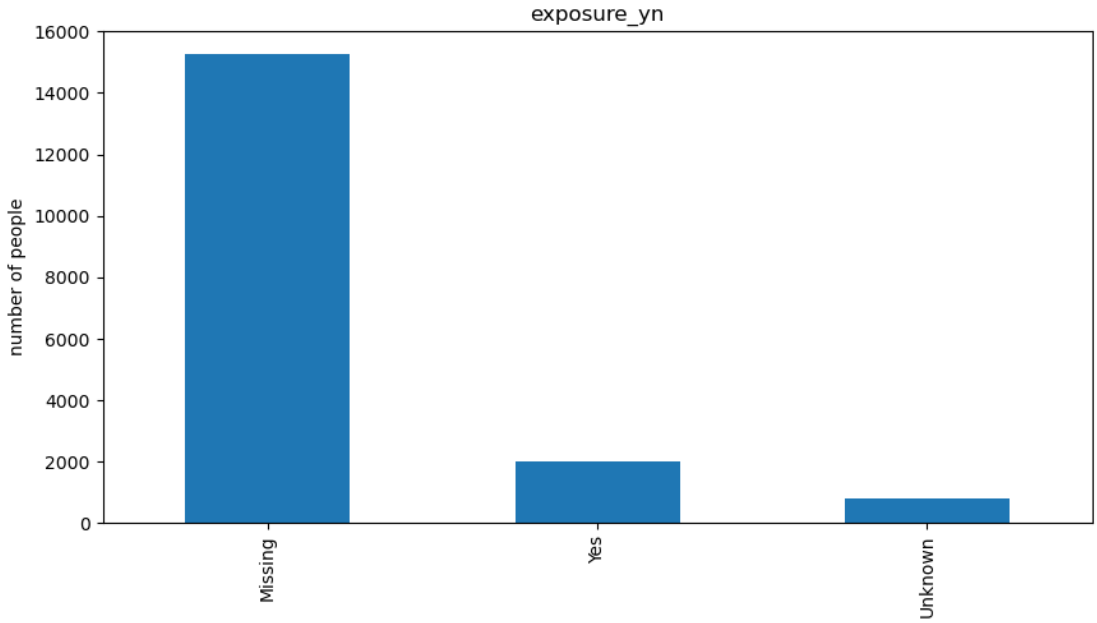
3. The total number of infected people over the age of 65 far exceeds their absolute number and population proportion, indicating that they are more susceptible to infection by Covid-19.



1. This data is very interesting. It counts the proportion of Hispanic-Latino Americans in the total number of cases.

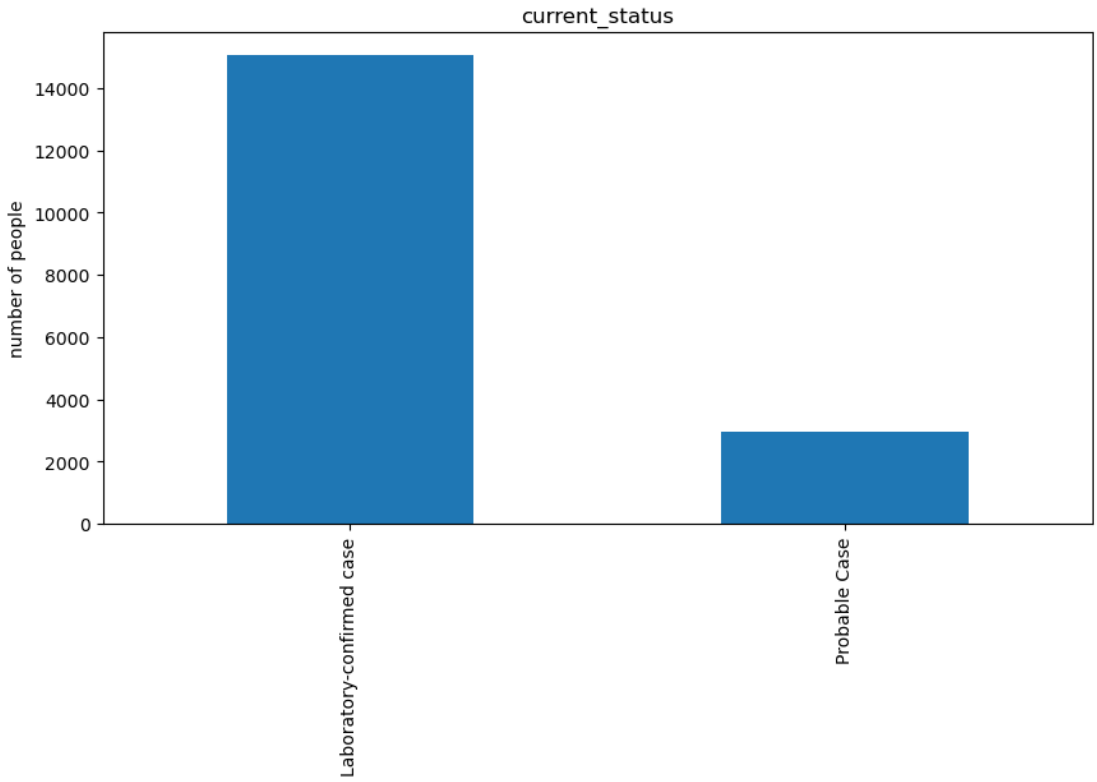
2. After data query, we can know that there are about 50 million Latinos and Hispanics in the US population, accounting for about 15% of the US population.

3. In this statistic, the proportion of Hispanic-Latinos among the positives is about 13%.



1. This data may not be meaningful, because the proportion of Missing is too high.

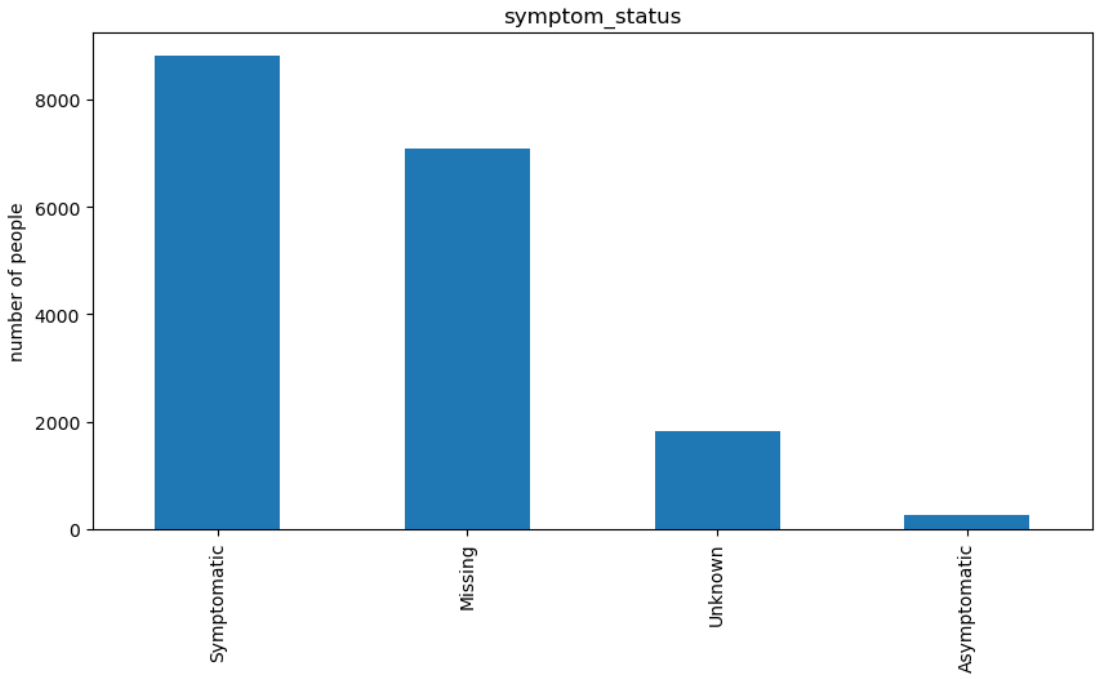
2. Speculation: Missing here may also be expressed as no clear exposure history.



1. If necessary, according to the data in this table, all the people can be split into two tables: confirmed cases and suspected cases.

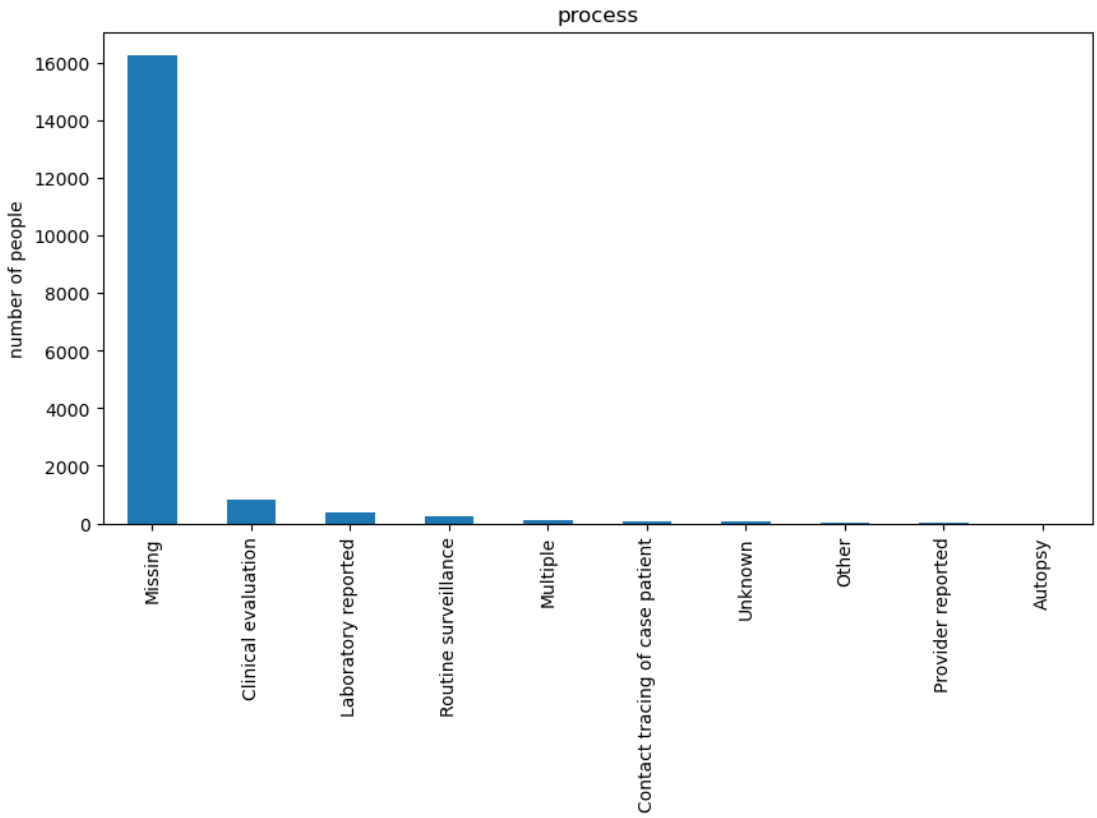
2. When analyzing certain characteristics, only confirmed cases or suspected cases can be analyzed.

3. In addition, the confirmed rate of suspected cases can also be analyzed again.

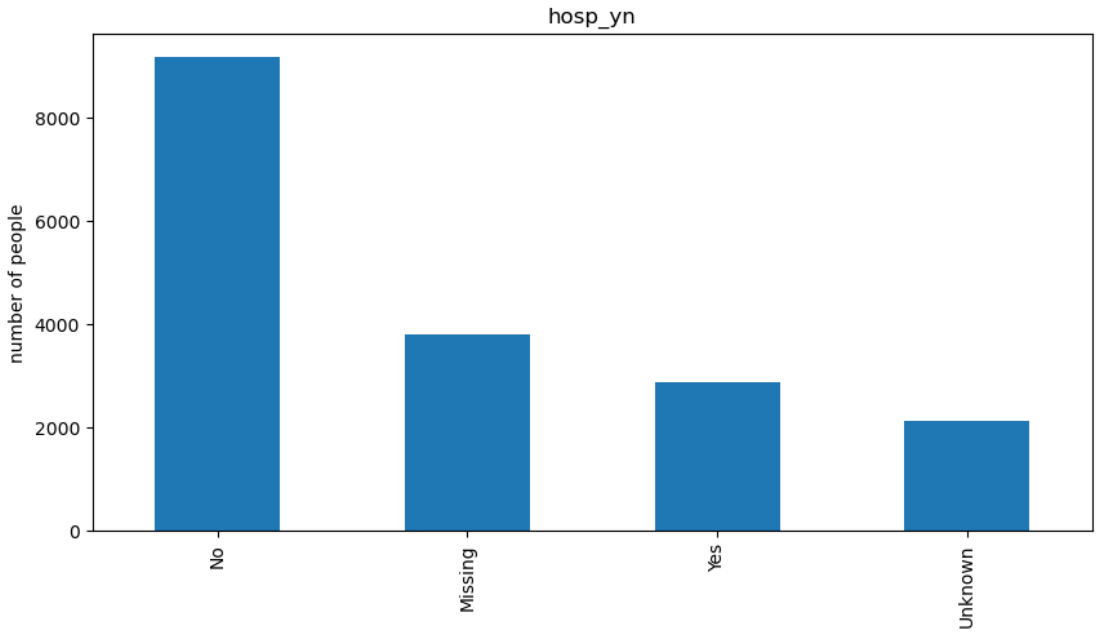


1. This data is easy to find, and symptomatic cases account for the vast majority.

2. The values of Missing and Unknown can be combined.

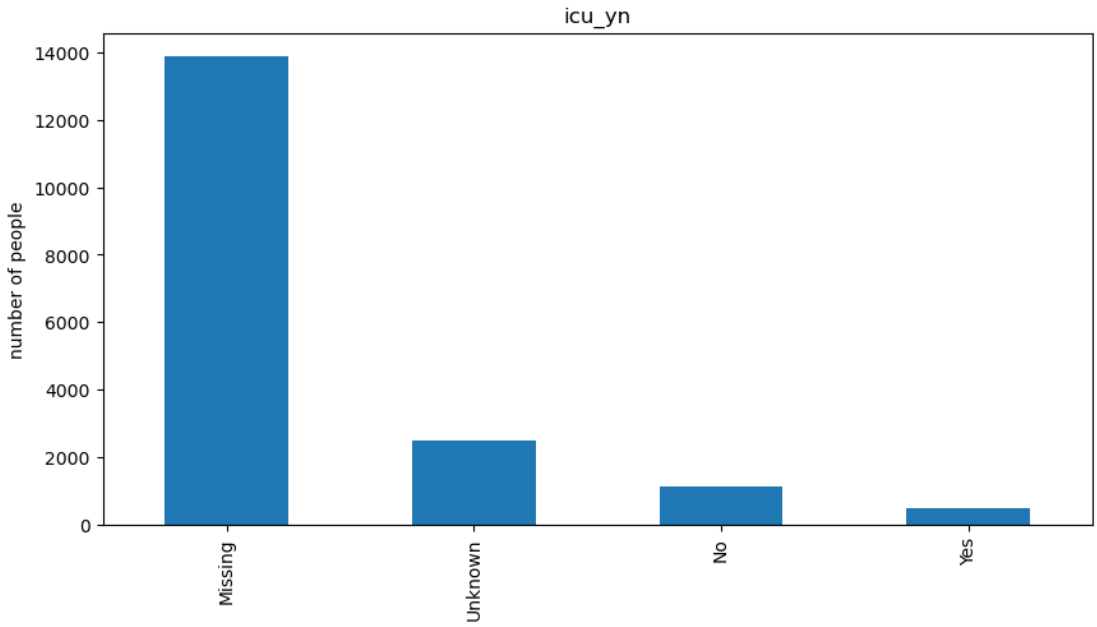


The proportion of Missing values in this table is too high, and the correlation with other values is too low, so this table can be considered discarded.



1. Unknown and Missing in this table can be combined

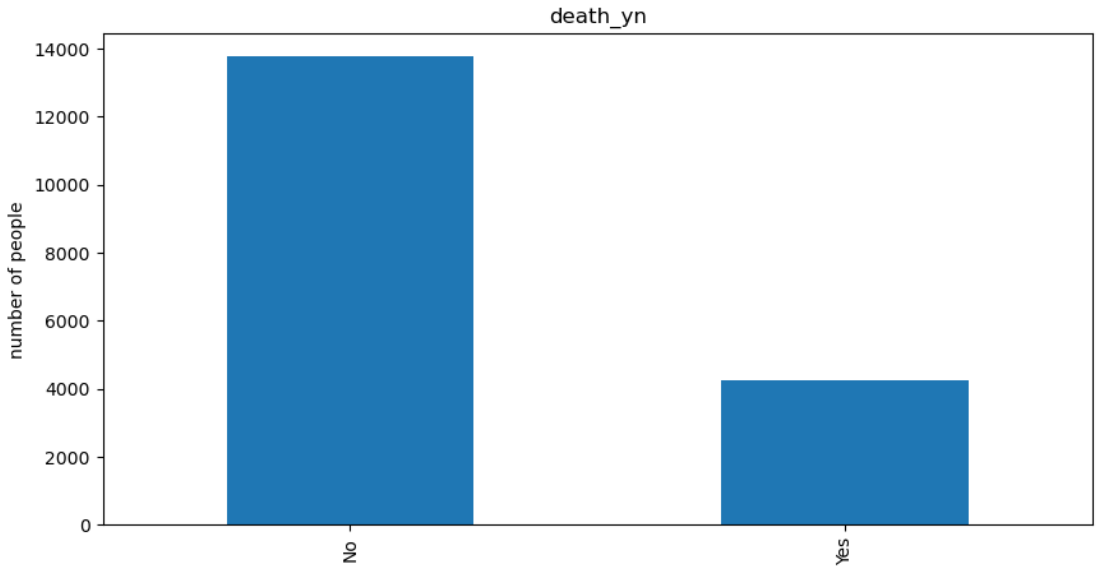
2. The hospitalization rate is about 21.4%



1. Unknown and Missing occupy most of the values in this table, and these two values can be combined.

2. The proportion of Yes is about 15.7% of No

3. If you need to use this table, you should further count unknown values.



The death data in this table account for about 24.1% of the total number of people. If this data needs to be used, more samples should be further evaluated, otherwise serious bias may occur.

6.Appendix

Please see containing folder for files generated visualizing feature relation.