Data Quality Report
Name: Xuhui An
Student Number: 20211294

Data Quality Report–Initial Fundings

1. Overview
   The initial findings based on the cleaned dataset (new file.csv) will be presented in this report. The data will be summarised, along with a description of the different data quality issues found and how they will be resolved.Moreover, feature summaries, histograms, and box plots that were utilised to visualise the data are included.

2. Summary
   The dataset contains 17 categorical characteristics and 2 continuous features. In the initial stages of the clean data process, I made a few changes to this database: I deleted duplicate records (1910 rows), and I deleted 2 constant feature called "underlying conditions_yn" and ”exposure_yn” that had over 90% null records, with the remaining records all being "Yes". There are several features that actually have 50% or even 70% null or "Missing" data; nonetheless, I choose to leave them in the dataset without making any adjustments in order to maintain the dataset's truth. For instance, there are separately 55.6% and 46.3% of null data in the "case_onset_interval" and "case_positive_specimen_interval" characteristics, I’ve searched there’re basically 2 kinds of solution to fill them: first is to fill them with mean value, I didn’t use that because majority of this feature is concentrate in 1-5, so obviously that will make the dataset lose reality; second is fill them with the majority data, but I passed on that option as well because there are too many null values, and filling them with just one particular value would also cause the dataset to lose reality. Also, there sia very last option–delete them, I didn’t do that because delete means I’ll lost thousands rows of data, which is obviously not a right action.
   Moreover, I additionally include the categorical feature options "Missing" and "Unknown" in the dataset for the following reasons. I leave them for "Missing" value since that was the report from default option. In addition to the above mentioned reasons, there are additional multiple values for "Unknown" values in the dataset. For instance, in the process feature, even if there are 96.4% "Missing" values, there are still 8 other values.

3. Review of categorical data:
   I sort all category data into 3 groups, here’s details of every group and the corresponding reasons:
   - First group:
     I grouped these five characteristics together because the number of some of their rows fell below a certain threshold, which is barely visible on the chart. Thus, when creating the plot chart, I took them out.
     - ➔ Case_month:
       The dataset's data for this column is the most complete, with no null, Missing, or Unknown values. The highest count in this feature is "2022-01," followed by "2020-12," "2021-01," "2021-12," and "2020-11" at the second level (over 1000), the state fips code and the res_state The two features will be combined because they essentially

Data Quality Report
Name: Xuhui An
Student Number: 20211294

state the same thing—that there is only one null data and no missing or unknown data. I eliminated months with counts below 30.

- ➔ Res_state and state_fips_code:
  These two attributes were combined for the sole reason that they essentially describe the same thing. There is only one null row between them both and no "Missing" or "Unknown" rows. I eliminated states with counts below 40 for the same reason as earlier.
- ➔ Res_county and county_fips_code:
  same reason as above, there are 1099 null rows and no Missing or Unknown data. Again, I eliminated counties with counts below 30.

- Second group:
  Reason I put these 9 columns together is they all have a lot of "Missing" and "Unknown" rows, although the two options appeared on the original report, I still deleted them from the chart because it is pointless to analyse data that cannot provide you with any information. Besides, they may make the final chart look unsightly. Hence, while creating the plots, I erased them from the data rather than reserving them.
  - ➔ Age_group:
    133 null data with 22 “Missing” and “Unknown” data. Among all age groups, people between 18 and 49 years old have the highest number(7233).
  - ➔ race:
    2287 null data with 2225 “Missing” and “Unknown” data. White people have the highest number(11674).
  - ➔ ethnicity:
    2460 null data with 3519 “Missing” and “Unknown” data. Non-Hispanic/Latino people have the highest number(11324).
  - ➔ process:
    0 null data with 17267 “Missing” and “Unknown” data. Clinical evaluation process has the highest number(over 100).
  - ➔ symptom_status:
    0 null data with 9813 “Missing” and “Unknown” data. Most people(8819) are Symptomatic while filling the report.
  - ➔ current_status:
    One of the two complete data in this dataset, with no null, “Missing” or “Unknown” data. Most people(15987) are confirmed by the Laboratory.
  - ➔ hosp_yn:
    0 null data with 6415 “Missing” and “Unknown” data. ”No” is the majority answer of this data(9383).
  - ➔ icu_yn:
    0 null data with 17328 “Missing” and “Unknown” data. ”No” is the majority answer of this data(over 1000). Around 500 people are in icu while filling the report.
  - ➔ death_yn:
    The other complete data in this dataset. Most people saved their life while over 5000 people lost their life during Covid.
- Third group:

Data Quality Report
Name: Xuhui An
Student Number: 20211294

Last group data is the two columns that I personally think it's useless to do analysis to them:

- ➔ exposure_yn:
  This feature has no null data, but has 16993 "Missing" and the rest are all "Yes", which can be seen as a constant feature.
- ➔ underlying_conditions_yn:
  This feature has 17157 null data, over 90 % of the dataset.

4. Review of continuous data:
   There are just two continuous features in this dataset, case_positive_specimen_interval and case_onset_interval, both of which should not have negative values because they measure the time between the earliest date and the first time a positive specimen was collected. Yet, there are 62 and 291 negative values in them independently. This is likely because two dates were presumably written backwards, so I merely took their absolute values.
   - ➔ case_positive_specimen_interval:
     10581 valid data, max value is 102 weeks and mean value is 0.25 week.
   - ➔ case_onset_interval:
     10581 valid data, max value is 70 weeks and mean value is 0.07 week.

5. Appendix:
   5.1 Terminology:
   - Case_positive_specimen_interval: Weeks between earliest date and date of first positive specimen collection
   - Case_onset_interval: Weeks between earliest date and date of symptom onset.

   5.2 continuous features:
   Descriptive statics

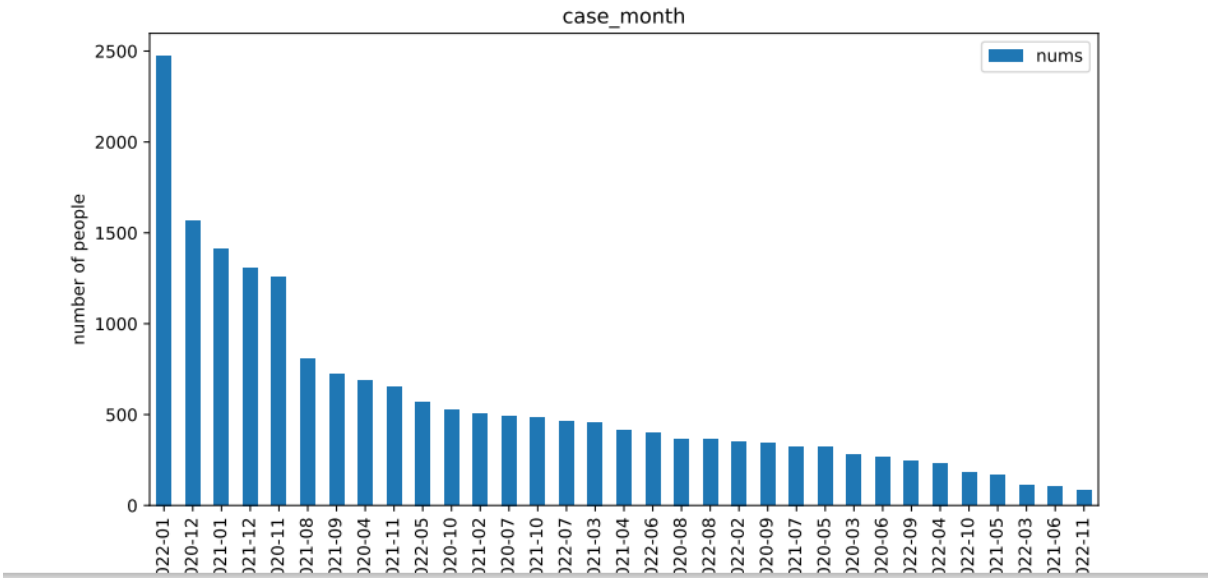
	count	mean	std	min	25%	50%	75%	max
case_positive_specimen_interval	10581.0	0.25111	2.417403	0.0	0.0	0.0	0.0	102.0
case_onset_interval	8666.0	0.160512	2.065118	0.0	0.0	0.0	0.0	105.0

   5.2 categorial features:
   Descriptive statics

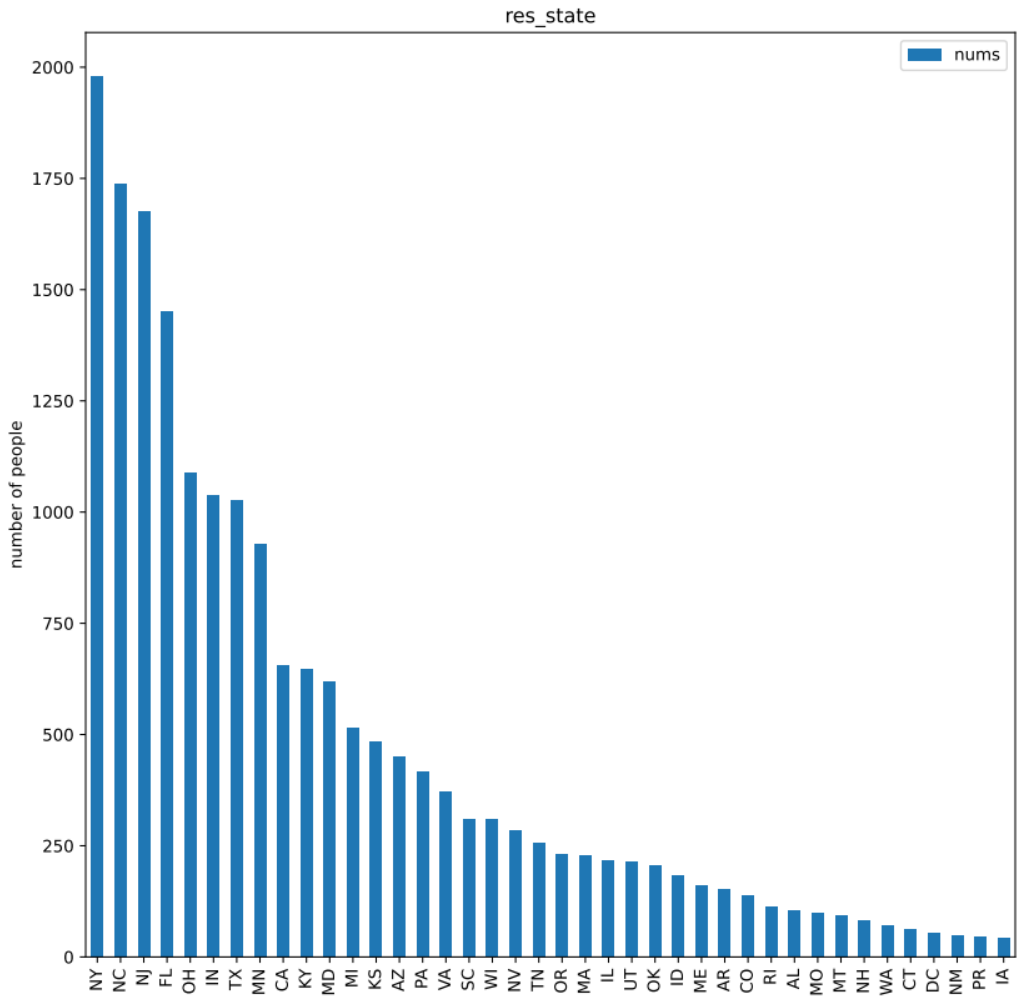
Data Quality Report
Name: Xuhui An
Student Number: 20211294

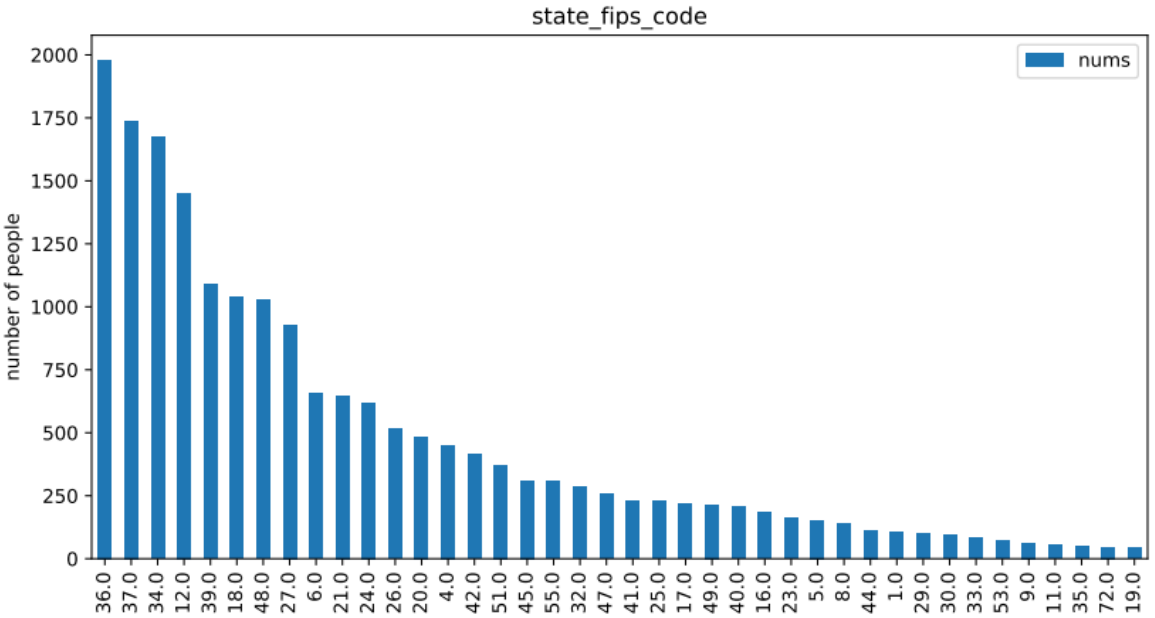
	count	unique	top	freq
case_month	20000	35	2022-01	2735
res_state	19999	49	NY	2174
state_fips_code	19999.0	49.0	36.0	2174.0
res_county	18860	859	MIAMI-DADE	408
county_fips_code	18860.0	1218.0	12086.0	408.0
age_group	19867	5	18 to 49 years	7709
sex	19580	4	Female	10092
race	17658	8	White	12350
ethnicity	17483	4	Non-Hispanic/Latino	11935
process	20000	10	Missing	18261
current_status	20000	2	Laboratory-confirmed case	17005
symptom_status	20000	4	Symptomatic	9064
hosp_yn	20000	4	No	9721
icu_yn	20000	4	Missing	15492
death_yn	20000	2	No	15000

5.3 categorial features:
See below summary of box plots and histograms. Accompanying pdfs will show larger plots.

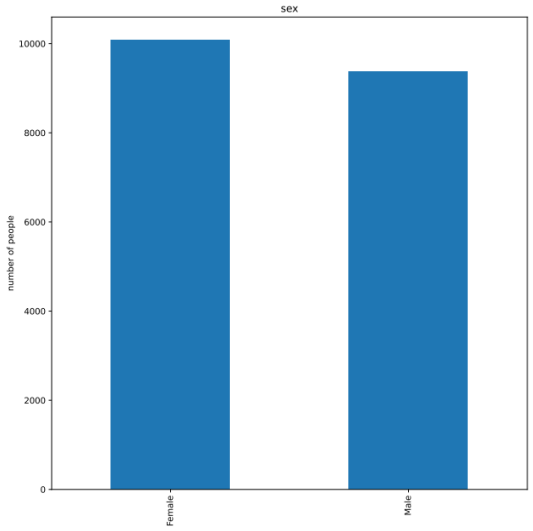


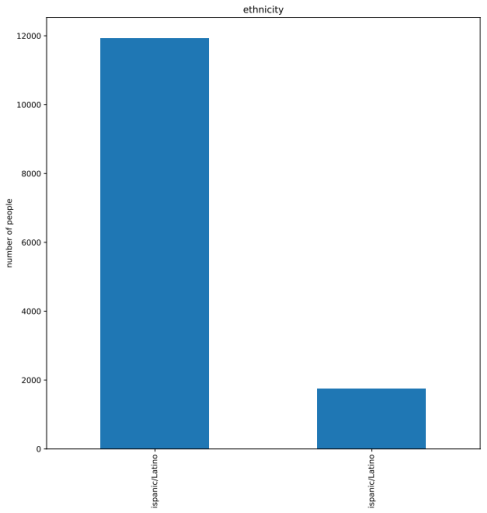
Data Quality Report
Name: Xuhui An
Student Number: 20211294

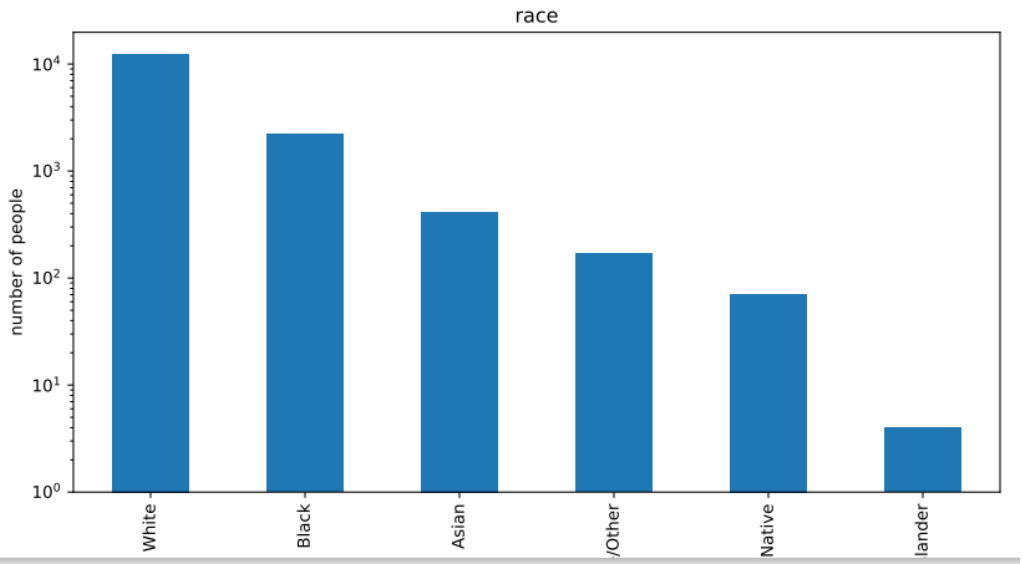


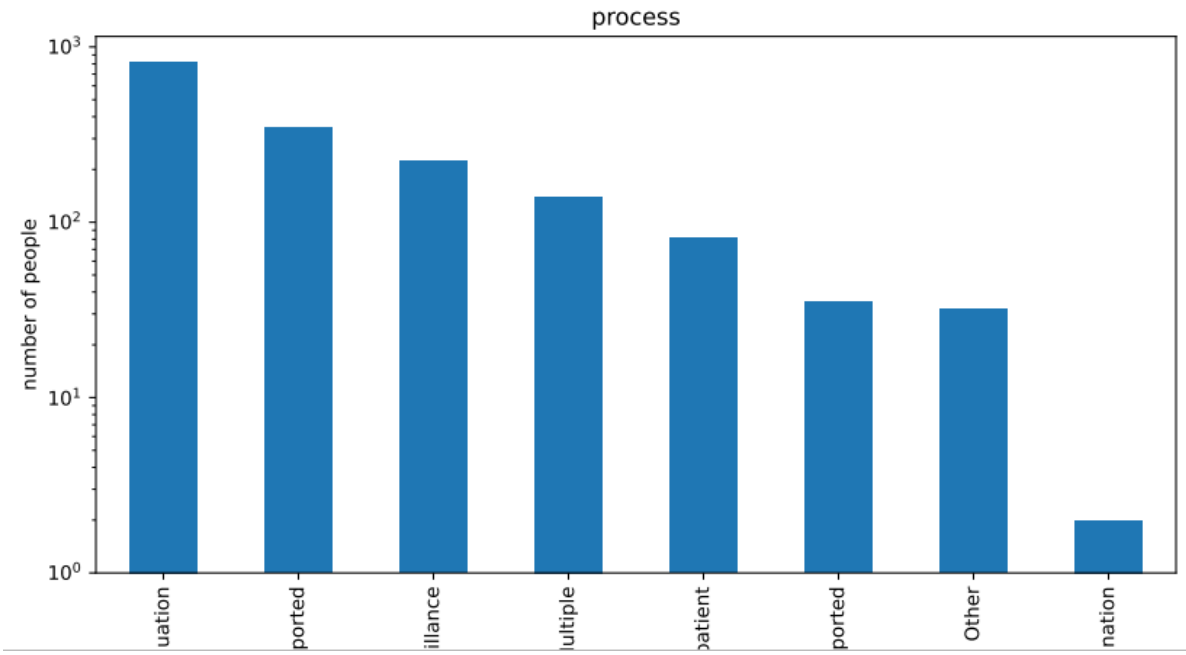


Data Quality Report
Name: Xuhui An
Student Number: 20211294

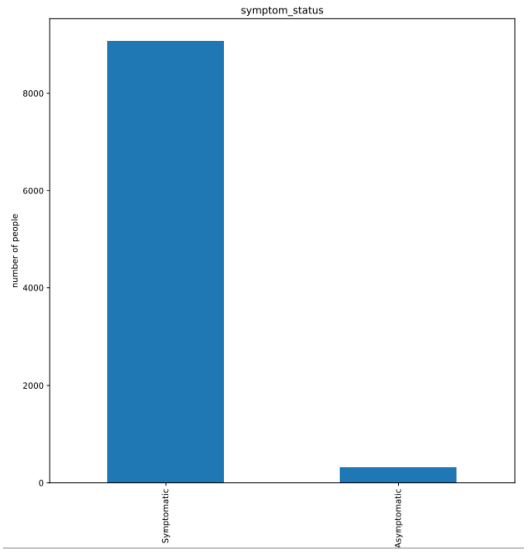


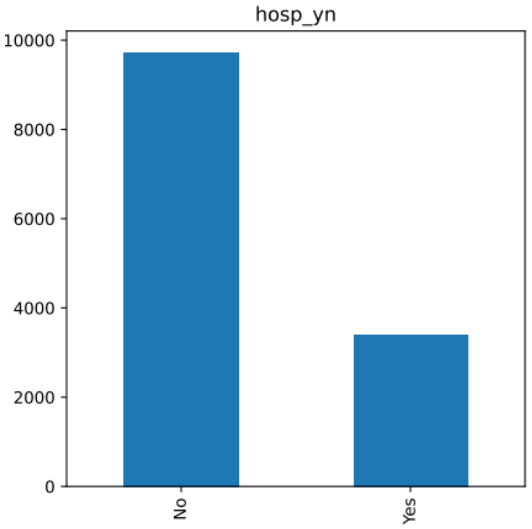


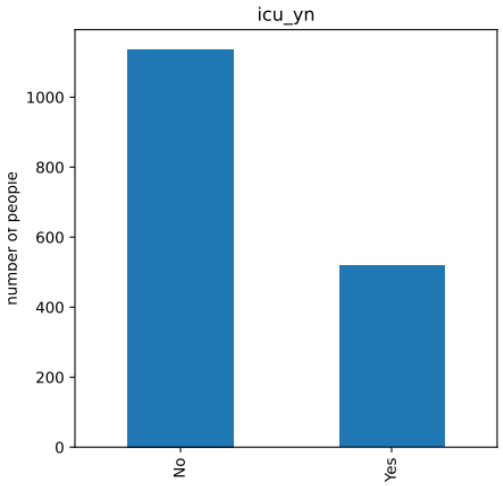


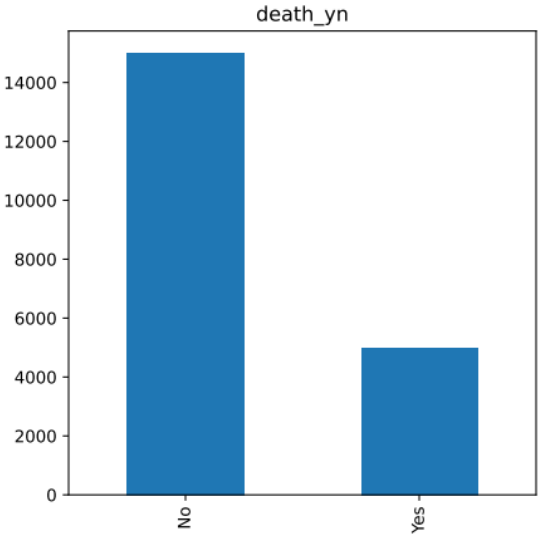


Data Quality Report
Name: Xuhui An
Student Number: 20211294

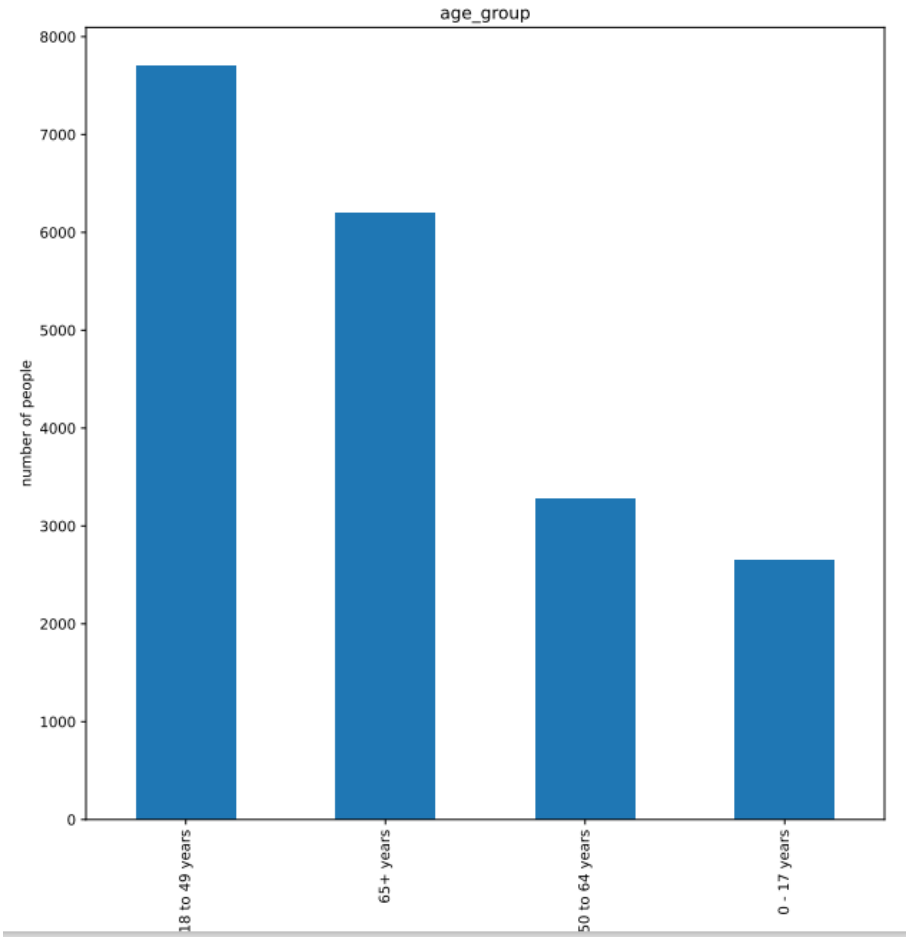


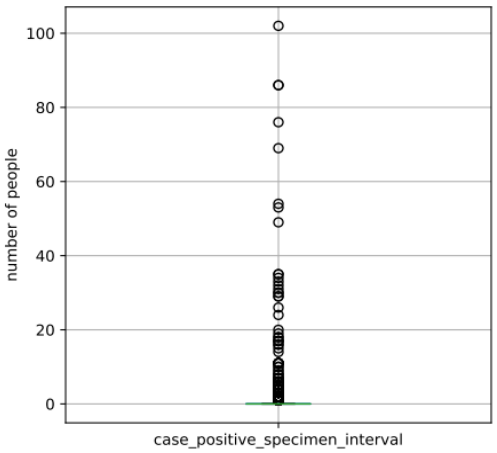


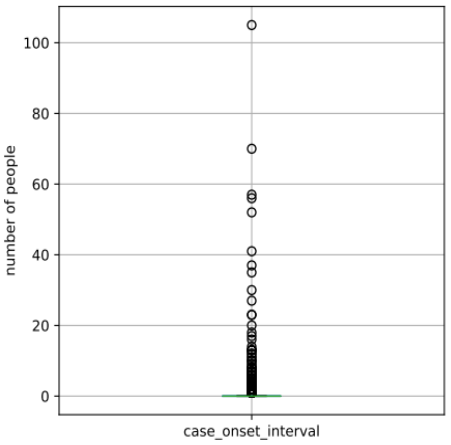




Data Quality Report
Name: Xuhui An
Student Number: 20211294







Data Quality Report
Name: Xuhui An
Student Number: 20211294