Data Quality Report – Initial Findings

1. Overview

This report will outline the initial findings based on the cleaned dataset (Homework_dataset_cleaned.csv). It will summarize the data, describe the various data quality issues observed and how they will be addressed. Please see appendix for background information to this dataset. Appendix includes explanations, histograms, plots used to visualize the data.

On the first indication, there were some issues with the dataset. We had many missing values and null values that needed to be dealt with. In total there was 2327 null values for cdc_report_dt, 7163 null values for pos_spec_dt, 4936 null values for onset_dt. Furthermore, there was 15 missing values for column "sex", 11 missing values for "age_group", 249 missing values for "race_ethnicity_combined", 2333 missing values for "hosp_yn", 7637 missing values for "icu_yn", and lastly 7431 missing values for "medcond_yn". In general, the biggest issue with this dataset was a lot of missing values.

Furthermore, I ran a duplication test, and I found 790 rows that were categorized as duplicates. However, after studying the dataset further, I decided to keep working with the duplicates. The reason for this was, there was no primary key, or unique identifier for the rows. Having this in mind, I was reluctant to remove the duplicates, as I assume these duplicates are actually unique people, therefor valuable data.

2. Summary

Analyzing the data, I observed a lot of missing data. Firstly, for the first four columns which consisted of dates, I used analysis to determine if there were null values, and there were a lot of missing values. For the rest of the features, the null values were written as "Missing", therefor I needed to run a test to check how many "Missing" values there were in the dataset.

Furthermore, I made a percentage data frame to visualize the percentage of missing data in each column for the categorical and continuous features by running a test.
Thereafter, I made a cardinality check for all features, and they were all > 1, and those were included for both categorical and the continuous features.

3. Reviewing Continuous Features

- cdc_report_dt
  - This represent the calculated date representing intitial date case was reported to CDC. CDC themselves, recommend using cdc_case_earliest_dt as this column is depreciated.

- cdc_case_earliest_dt
  - This represent the calculated date which is the earliest available date for the record. Column has no missing values and reference was made to this column for analysis.

- pos_spec_dt
  - This column represent the date of the first positive specimen collection

- onset_dt
  - This column represent the first date the person had symptoms (if symptomatic)

3.1 Histograms

The histogram can be found in the accompanying pdf and in the appendix.

4. Reviewing Categorical Features

- current_status
  - Represents the case status, if the person is a confirmed case or a probable case

- sex

    - Represents the sex of the person

- age_group
    - Represents age groups that the person fits within based on age

- race_ethnicity_combined
    - Represents race and ethnicity combined

- hosp_yn
    - Represents weather the person was hospitalized or not
- icu_yn
    - Represents weather the person was admitted or the ICU or not

- death_yn
    - Represents the death status of the person

- medcond_yn
    - Represents the presence of underlying comorbidity or disease

Look at appendix for more detailed information on the features.

4.2 Bar Plots

The bar plots can be found in the accompanying pdf and in the appendix. As there are 7 features an in-depth review will not be carried out here. In general, we could identify many missing values for three of the features that affected the distribution. Distribution between sexes differed with a couple percentages. For race_ethnicity_combined the biggest portion of the data was white/non-Hispanic and the second biggest was Hispanic. Furthermore, the biggest age group were 20-29 years while the second biggest was 30-39 years.
More detailed information about all features can be found in the descriptive statistics

5. Action to take

9 main actions will be taken, summarized below.

- cdc_report_dt will be dropped as its depreciated
- pos_spect_dt will be dropped as it has 71.63% missing values
- onset_dt will be dropped as it has 49.36% missing values, and the cdc_report_dt will be used as reference to date as suggested by CDC.
- sex has 0.15% missing values, those missing values will be replaced with the most frequent value for sex.
- age_group has 0.11% missing values, those missing values will be replaced with the most frequent value for age_group
- race_ethnicity_combined has 0.11% missing values, those missing values will be replaced with the most frequent value for race_ethnicity_combined
- hosp_yn has 23.33% missing values, those missing values will be replaced with the most frequent value for hosp_yn
- icu_yn has 76.37% missing values for this reason the feature will be dropped
- med_cond has 74.31% missing values for this reason this feature will be dropped as well

6. References

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

7. Appendix

7.1 Continuous Features

Descriptive Statistics

	count	mean	min	25%	50%	75%	max	%missing	card
cdc_case_earliest_dt	10000	2020-10-04 16:25:58.080000000	2020-01-01	2020-07-25	2020-11-07	2020-12-15	2021-01-16	0.00	319
cdc_report_dt	7673	2020-10-16 05:30:29.323602176	2020-03-04	2020-08-14	2020-11-11	2020-12-21	2021-01-29	23.27	322
pos_spec_dt	2837	2020-09-17 07:27:10.595699712	2020-03-13	2020-07-03	2020-10-17	2020-12-04	2021-01-24	71.63	313
onset_dt	5064	2020-09-22 05:01:08.246445568	2020-01-01	2020-07-15	2020-10-21	2020-12-03	2021-01-27	49.36	322

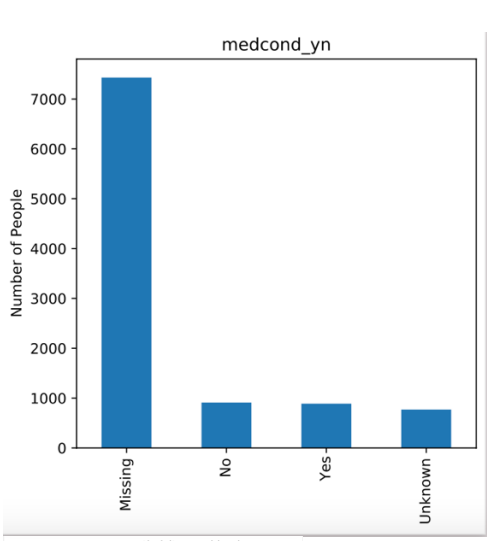
7.2 Categorical Features

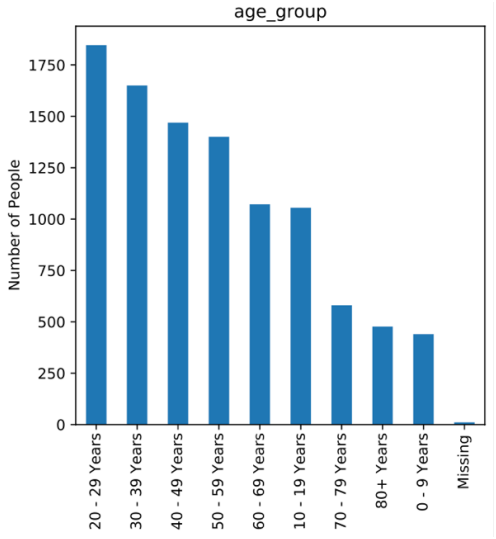
Descriptive Statistics

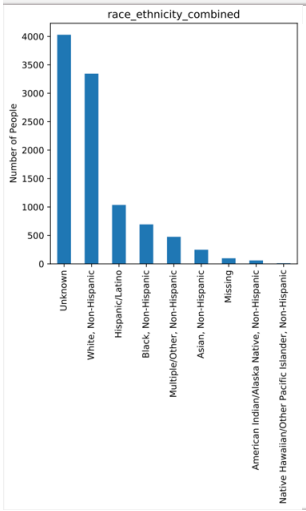
	count	unique	top	freq	mode	%mode	2ndmode	%2ndmode	%missing	%unknown	card
current_status	10000	2	Laboratory-confirmed case	9307	Laboratory-confirmed case	0.9307	Probable Case	0.0693	0.00	0.00	2
sex	10000	4	Female	5213	Female	0.5213	Male	0.4706	0.15	0.66	4
age_group	10000	10	20 - 29 Years	1846	20 - 29 Years	0.1846	30 - 39 Years	0.1650	0.11	0.00	10
race_ethnicity_combined	10000	9	Unknown	4030	Unknown	0.4030	White, Non-Hispanic	0.3344	0.99	40.30	9
hosp_yn	10000	4	No	5221	No	0.5221	Missing	0.2333	23.33	17.92	4
icu_yn	10000	4	Missing	7637	Missing	0.7637	Unknown	0.1300	76.37	13.00	4
death_yn	10000	2	No	9684	No	0.9684	Yes	0.0316	0.00	0.00	2
medcond_yn	10000	4	Missing	7431	Missing	0.7431	No	0.0911	74.31	7.70	4

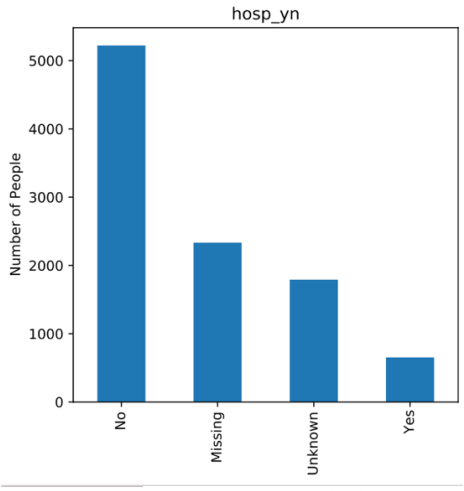
7.3 Box Plots & Histograms

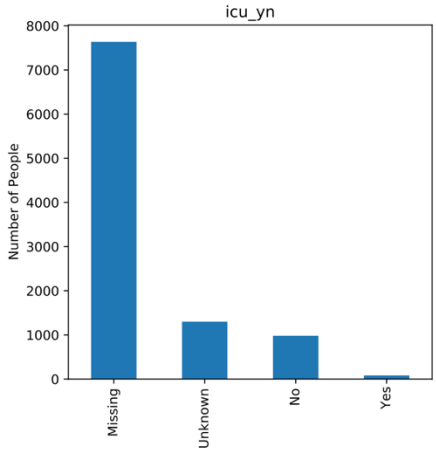
See below summary of box plots and histograms. Accompanying pdfs will show larger plots.

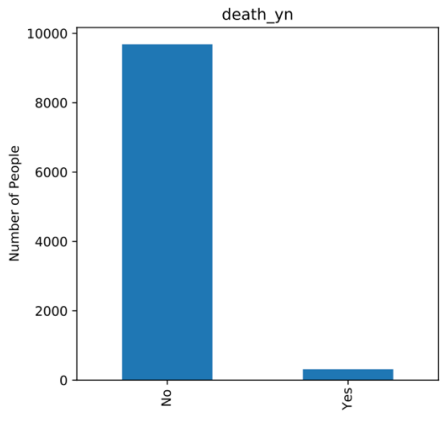


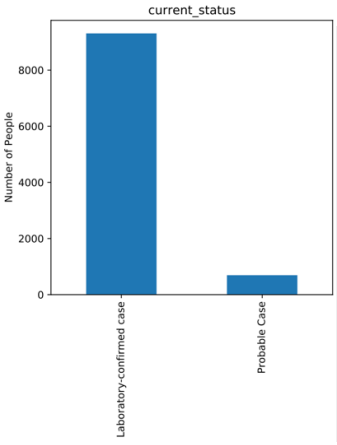


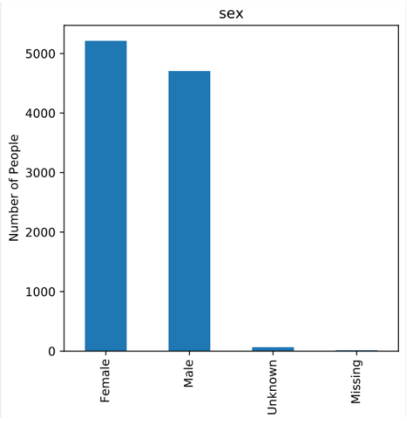


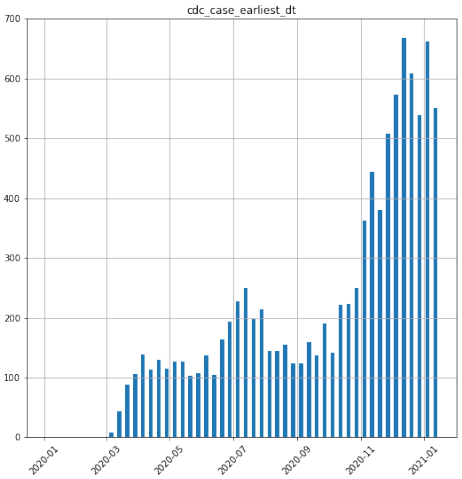


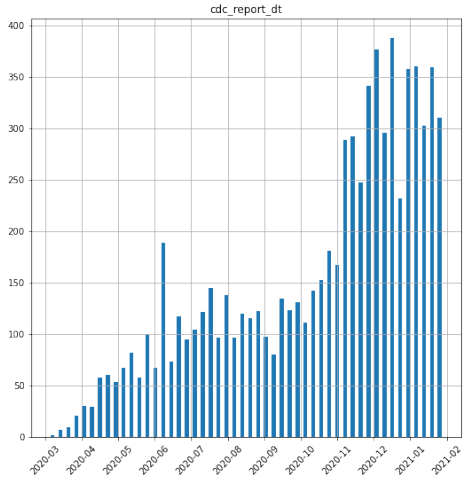


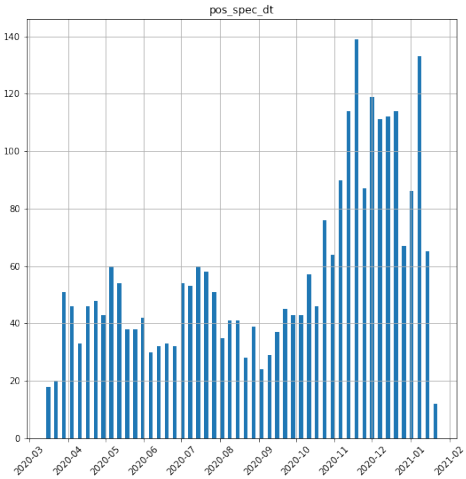


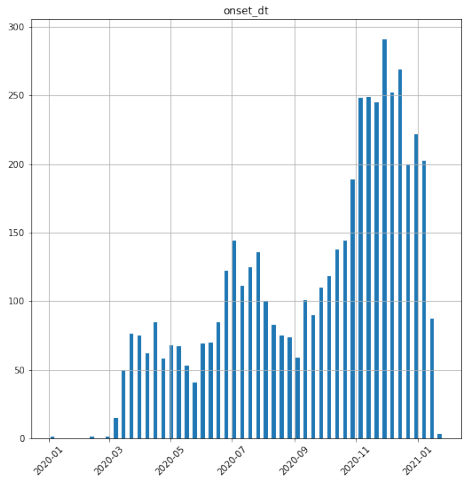





cdc_case_earliest_dt
700
600
500
400
300
200
100
0
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01


cdc_report_dt
400
350
300
250
200
150
100
50
0
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02


pos_spec_dt
140
120
100
80
60
40
20
0
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02


onset_dt
300
250
200
150
100
50
0
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01