Data Quality Report for a sample of CDC data on Covid-19

Introduction

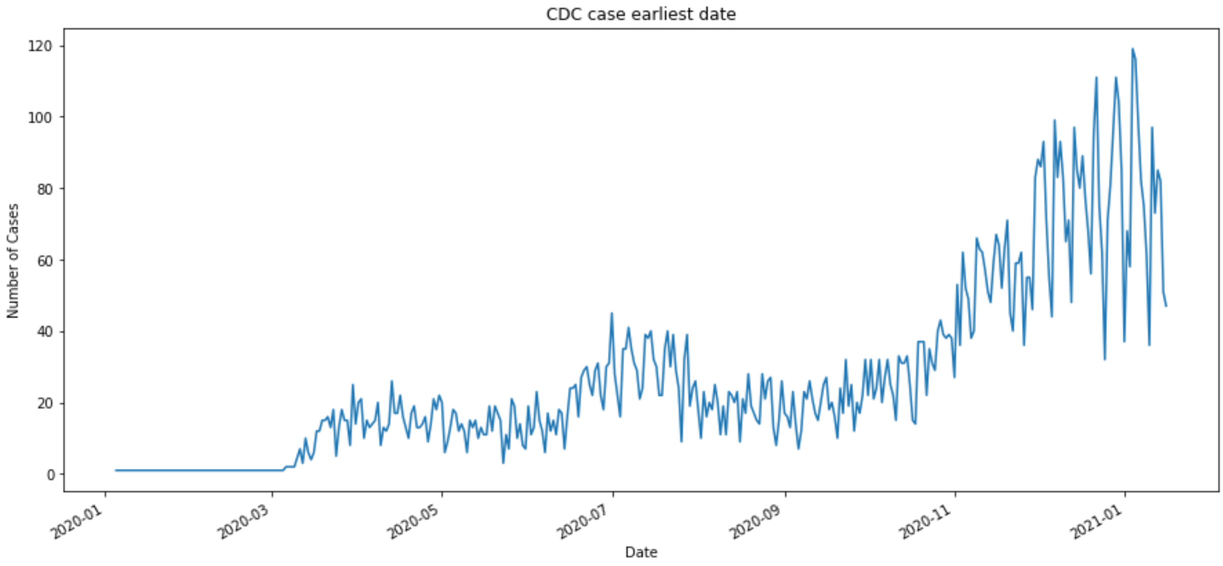
This is a data quality plan for data collected by the CDC is relation to Covid-19. The full dataset has 18,000,000 rows of data. This report will only be based on a subset of this. Therefor it is important to realise that this report may not be representative of the findings were we to examine the full dataset.

The data is presented in a csv file delimited by commas which are the unique values and each line of the document becoming a row in our dataset. The sample of the data that I am analysing has 10,000 rows and 12 columns. On viewing the top five and last five rows of the data I noticed quite a few missing or unknown values, which warranted further investigation. All the data were object types and I concluded that they would be best processed in a categorical manner. There are 7 rows classed as duplicates.

The column headers are as follows:

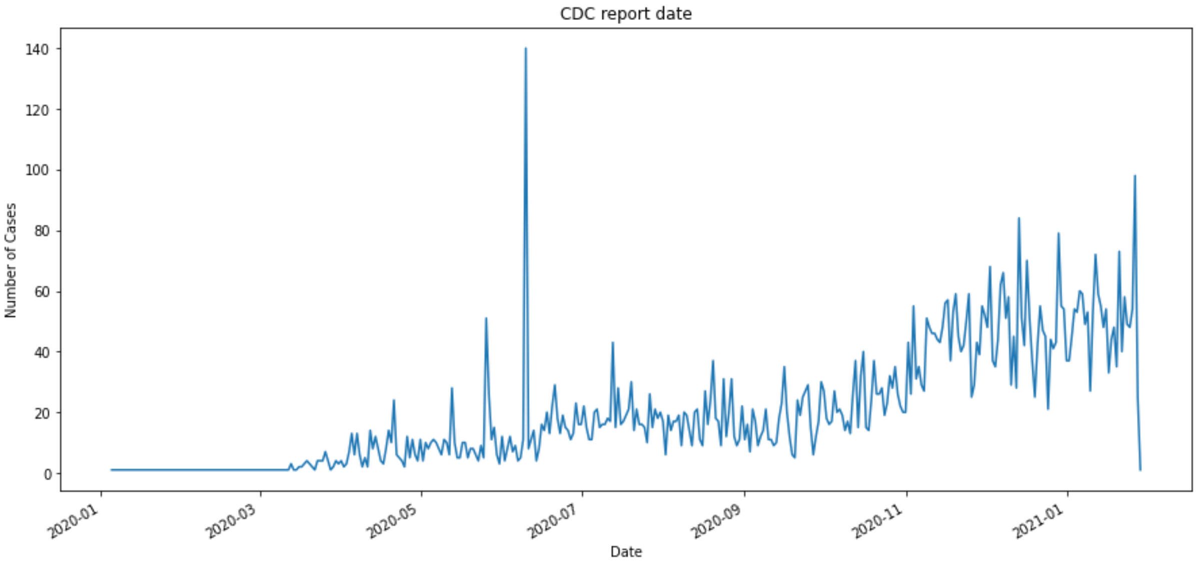
Cdc_case_earliest _dt

There are 322 unique dates for this feature between 2020-01-05 and 2020-01-16. 119 cases were recorded on 2021/01/04 which was the highest number for this feature in this section of the dataset. The dataset is right skewed with numbers very low until July does this indicate a lack of testing. There is only a slight upward trend with the data numbers quite erratic. There seems to be a jump in numbers in December 2020 with the largest jump in the line graph occurring then.



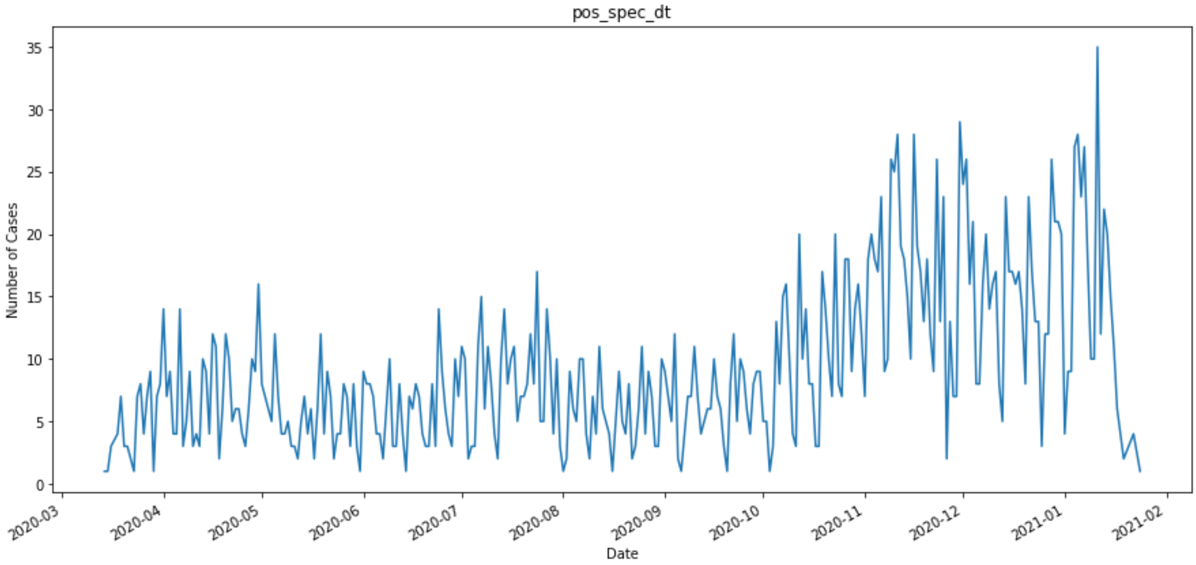
Cdc_report_dt

This feature has 327 unique values between 05/01/2020 and the 29/01/2021. There was a large increase in numbers on the 10/06/2020 with the highest numbers recorded with a very steep drop off on either side. Was there a backlog of cases given that date or was it simply that there were a lot reported that day, this needs further investigation. There is a sharp drop at the end of the graph back to 0 the needs further investigation. There are 2328 missing values for this feature.



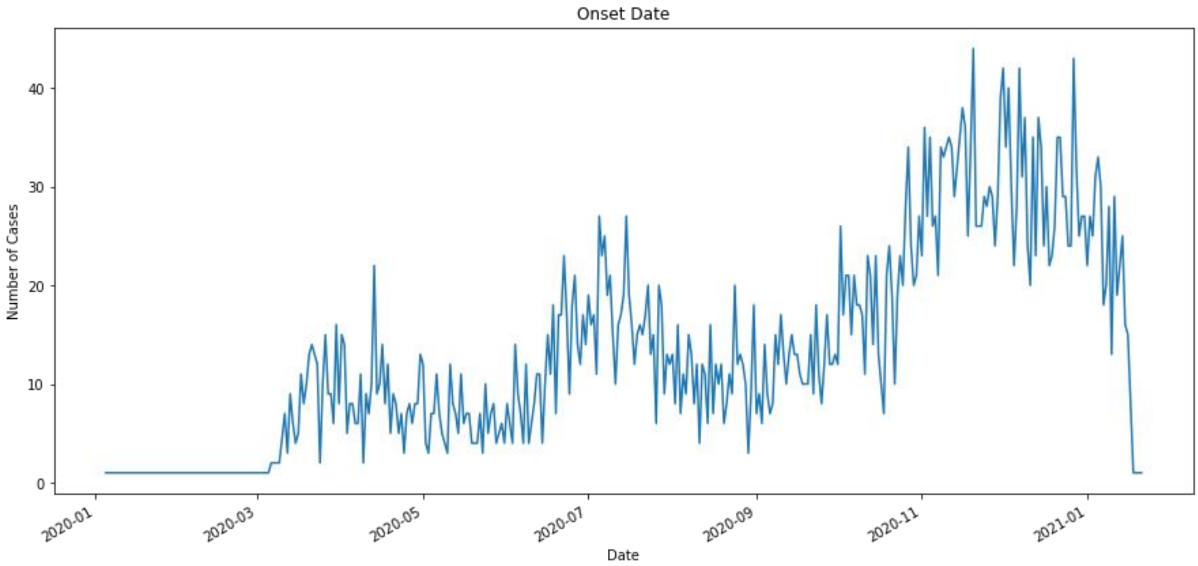
Pos_spec_dt

This feature has a missing value of 71.34%. there are 311 unique dates recorded in this sample between 14/03/2020 and the 24/01/2021.



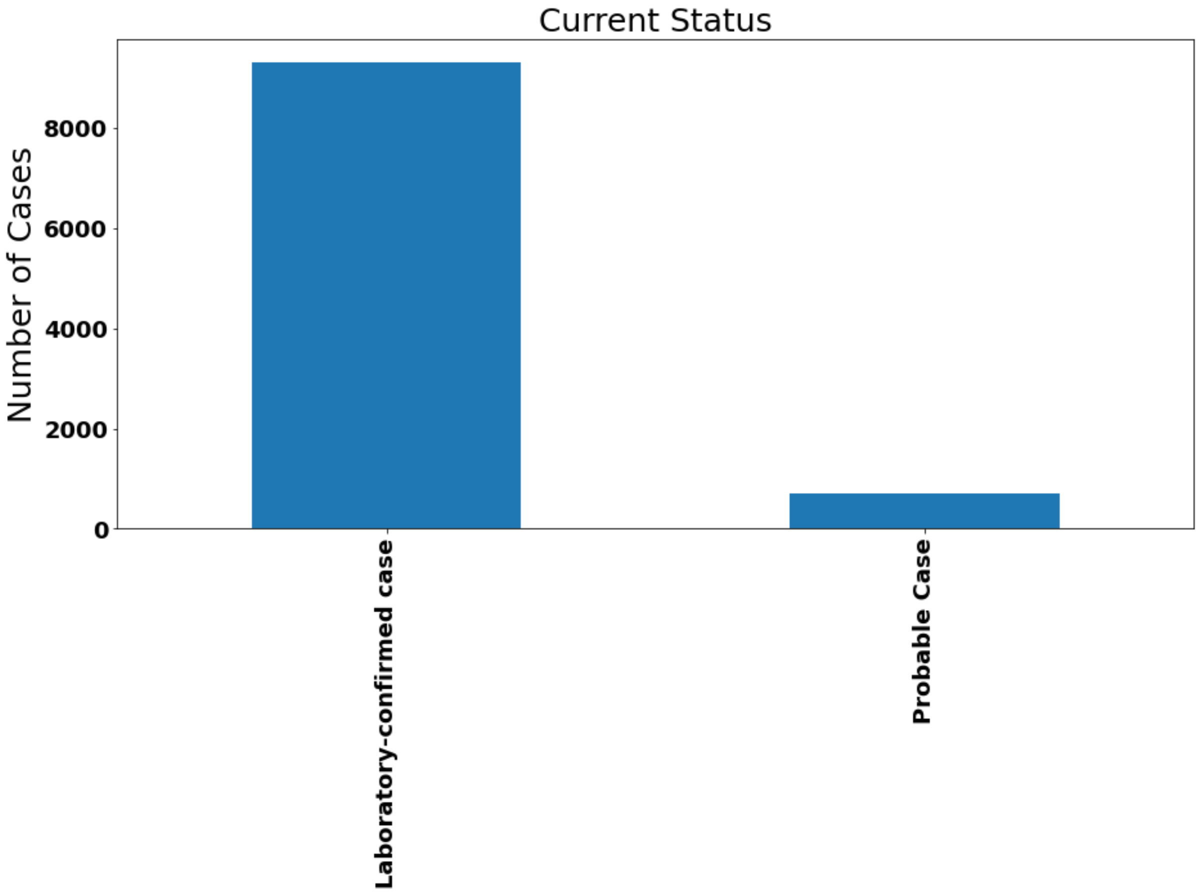
Onset_dt

This feature records the date at which the patient began feeling unwell. It holds records between the dates of 05/01/2020 and the 21/01/2021. There are 324 unique dates in this sample. The date with the highest number of patients noting the onset of symptoms was the 20/11/2020 with 44 specifying that date. Only just over half of the dataset recorded a value for this feature, which means we are missing data for 4958 patients. The end of the line graph ends at 0 why such a dramatic drop is this information still being collected?



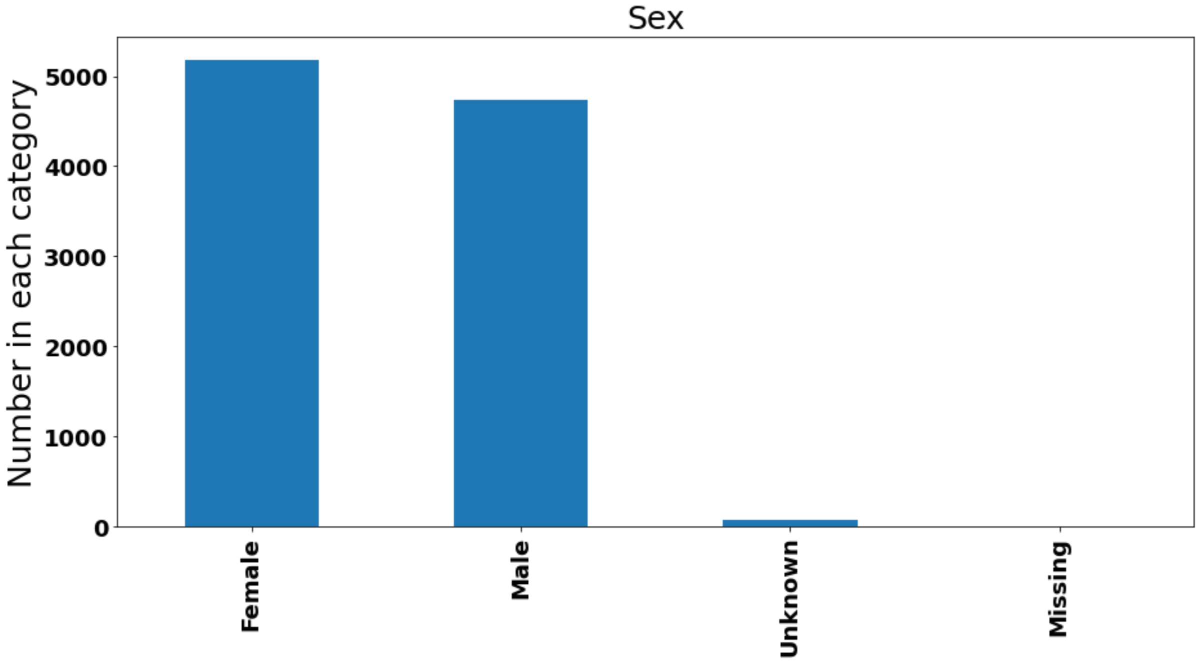
Current_status

This feature records whether the patient has a laboratory confirmed case or probable case. There are only two unique values for this feature. 9307 of the 10,000 patients have a laboratory confirmed case while 693 are designated to be a probable case.



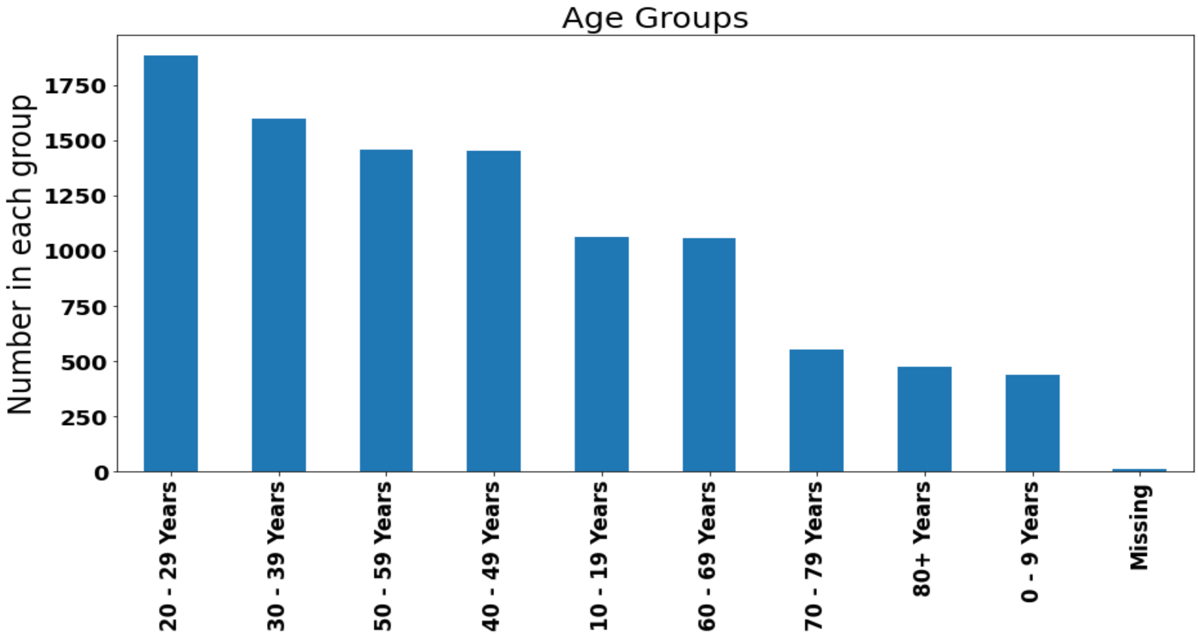
Sex

This feature captures the gender of the patients. It is an important feature as it could show possible trends on effects of the disease by gender. 5183 of the patients are female,4736 are male 74 are unknown and 7 are missing in this data sample.



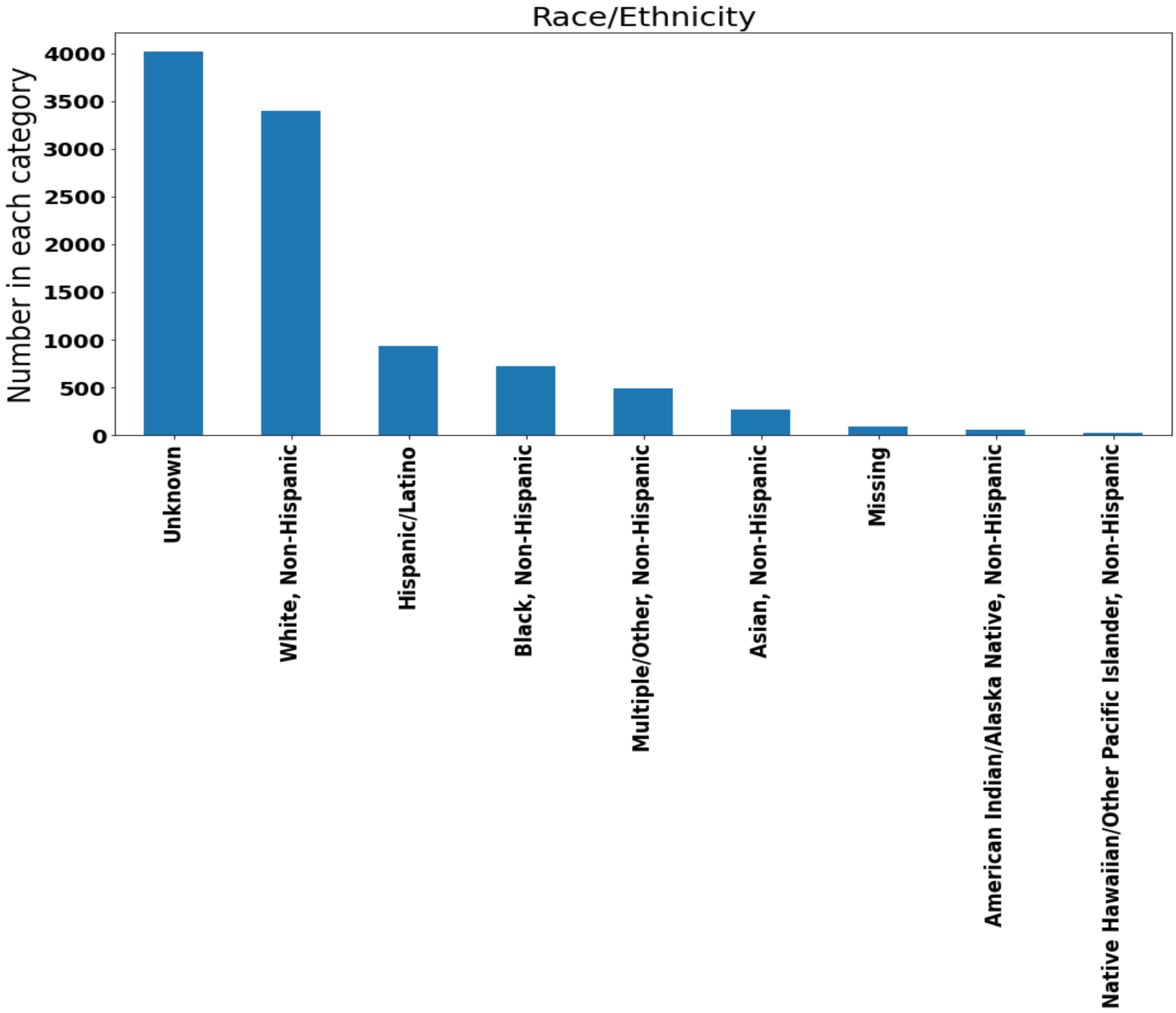
Age_range

Age ranges are divided into 10 distinct groups. The highest number of patients are in the 20–29-year age group. One noticeable thing is that the top four age groups are all working age and perhaps this shows that they are more likely to be exposed. It could also correlate with younger people being more likely to take risks while keeping the more vulnerable members of society children and the elderly safe.



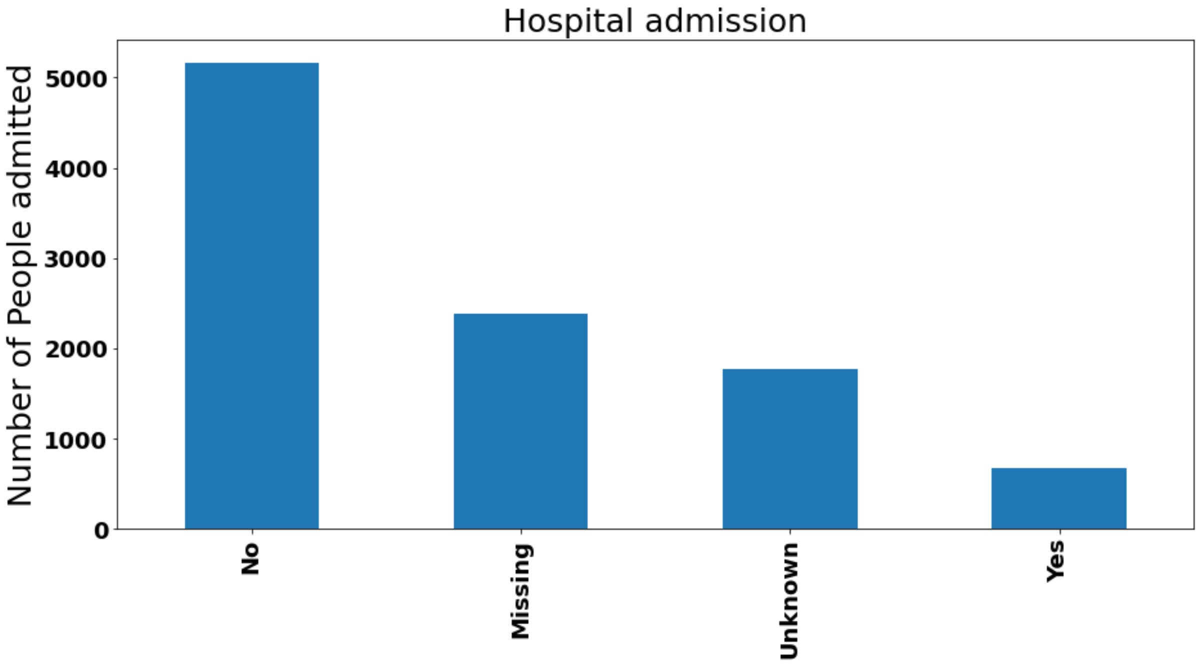
Race_ethnicity_combined

This feature could prove useful in seeing if there are any trends among specific races and ethnicities. There are 9 unique groupings in this feature. 4020 unknown values and a further 91 missing.



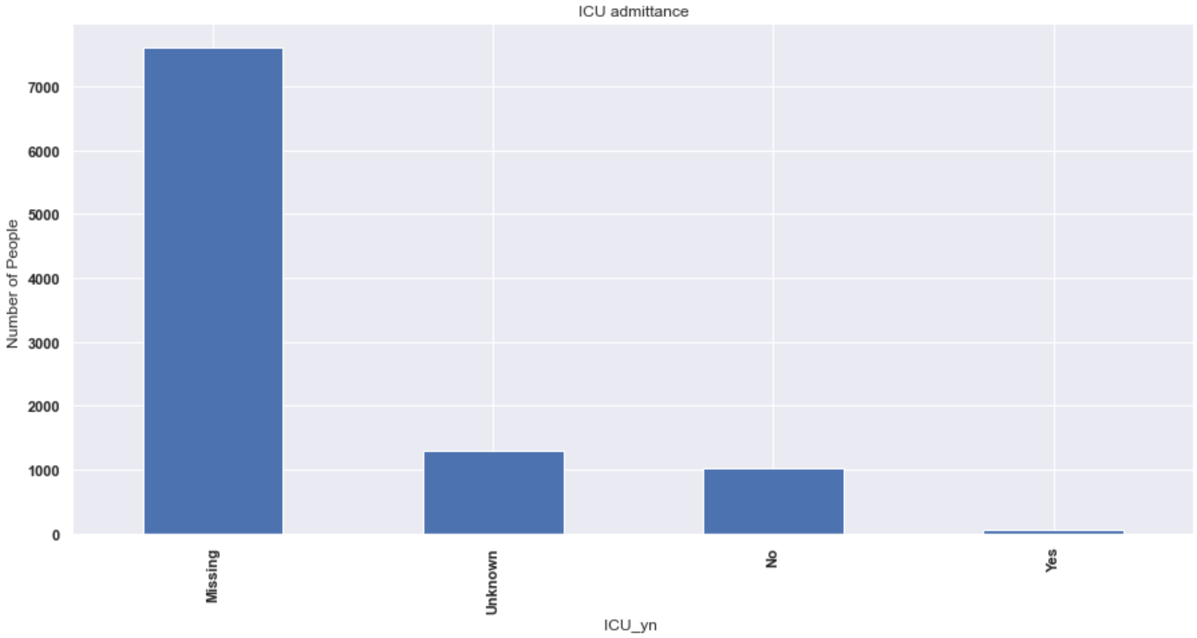
Hosp_yn

5167 of the patients were not admitted to hospital with only 679 of the sample being admitted to hospital. There are 2381 missing values and 1773 unknown values. Tested integrity of data with ICU data making sure that if a person had been in ICU then they should also be yes for this feature. Test passed.



ICU_yn

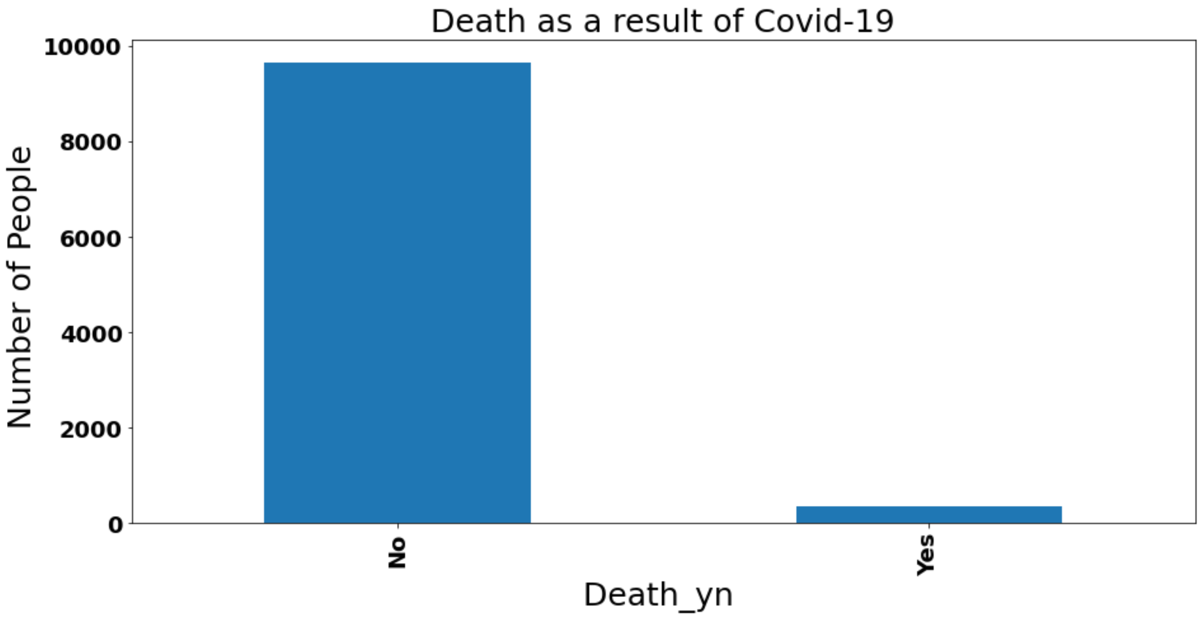
The ICU feature has 64 patients being admitted to ICU with 7603 missing values and a further 1302 values unknown. It could be possible that only people who were admitted to the ICU had this portion filled out, but it seems like a very small number considering the size of the data sample.



Death_yn

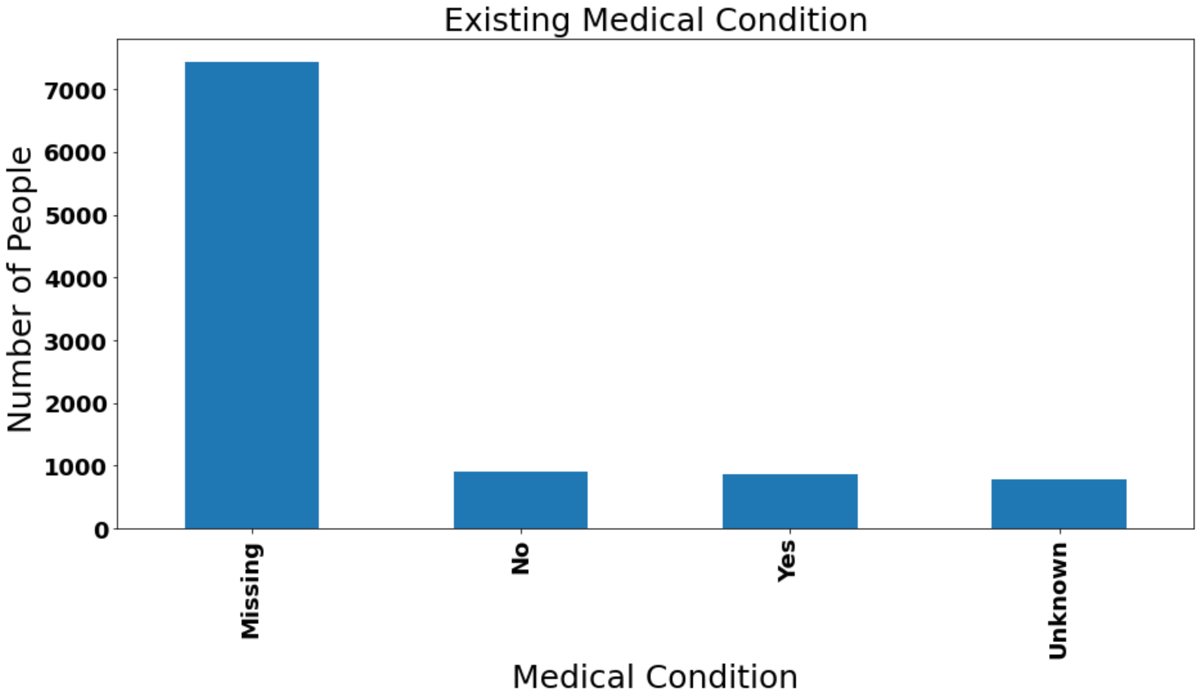
345 patients in this sample have died due to Covid-19. 9655 of patients have survived. According to the WHO Covid-19 website there have been 28700,966 confirmed cases and

521,625 deaths giving a death rate of 1.81%. Our dataset shows a death rate of 3.45% which is significantly higher. (covid19.who.init, accessed 09/03/2021)

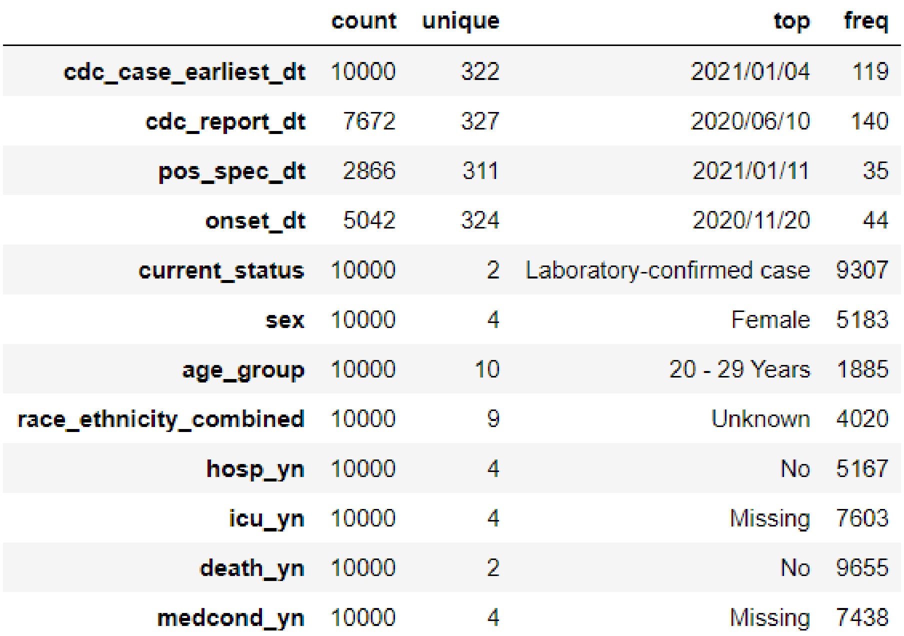


Medcond_yn

This feature is missing 7438 values with a further 788 unknown. 915 of patients do not have an existing medical condition while 859 of patients have an existing medical condition.



Categorical Statistics table



	count	unique	top	freq
cdc_case_earliest_dt	10000	322	2021/01/04	119
cdc_report_dt	7672	327	2020/06/10	140
pos_spec_dt	2866	311	2021/01/11	35
onset_dt	5042	324	2020/11/20	44
current_status	10000	2	Laboratory-confirmed case	9307
sex	10000	4	Female	5183
age_group	10000	10	20 - 29 Years	1885
race_ethnicity_combined	10000	9	Unknown	4020
hosp_yn	10000	4	No	5167
icu_yn	10000	4	Missing	7603
death_yn	10000	2	No	9655
medcond_yn	10000	4	Missing	7438

Bibliography

World Health Organisation. Available at:

https://covid19.who.int/region/amro/country/us (Accessed 09/03/2021)