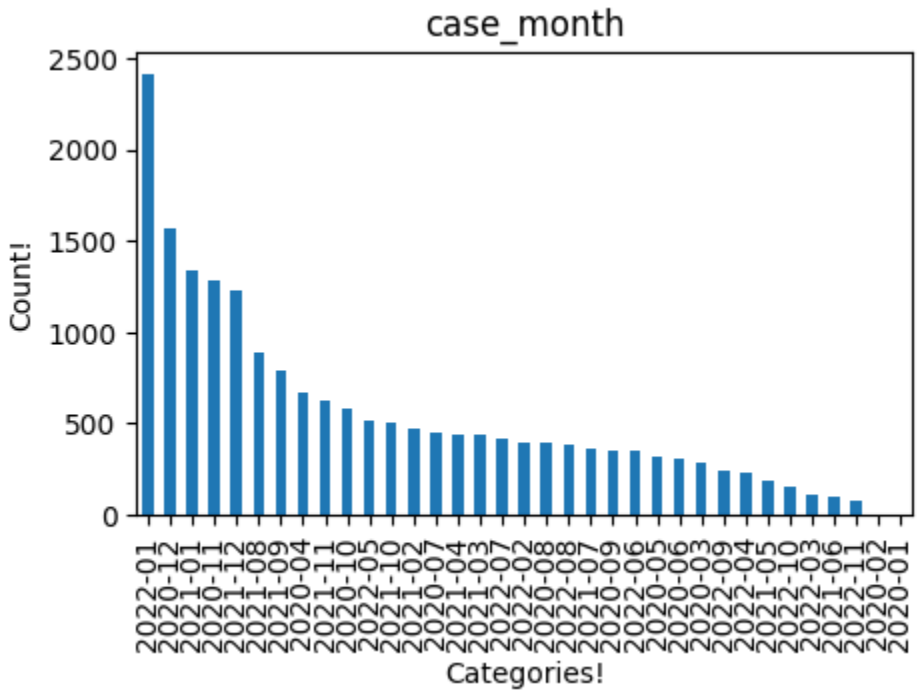
Data Quality report.

Categorical features

	A	B	C	D	E	F	G	H	I	J	K
1		count	missing%	cardinality	mode	modefreq	mode%	secondmc	secondmc	secondmode%	
2	case_mon	18865	0	35	2022-01	2412	12.78558	2020-12	1568	8.311688	
3	state_fips	18865	0	49	36	1919	10.17228	37	1655	8.77286	
4	county_fi	17726	6.037636	1201	12086	336	1.895521	4013	307	1.731919	
5	age_group	18665	1.060164	5	18 to 49 ye	7310	39.16421	65+ years	5769	30.90812	
6	sex	18441	2.247548	4	Female	9616	52.14468	Male	8754	47.47031	
7	race	15969	15.35118	8	White	11691	73.2106	Black	1962	12.2863	
8	ethnicity	15501	17.83196	4	Non-Hispa	11362	73.2985	Unknown	2541	16.39249	
9	process	1699	90.9939	10	Clinical ev	787	46.32137	Laborator	366	21.54208	
10	exposure_	2703	85.67188	3	Yes	1959	72.47503	Unknown	744	27.52497	
11	current_st	18865	0	2	Laborator	16039	85.01988	Probable (	2826	14.98012	
12	symptom_	11119	41.06016	4	Symptoma	8834	79.44959	Unknown	1954	17.57352	
13	hosp_yn	14867	21.19268	4	No	9618	64.69362	Yes	3079	20.7103	
14	icu_yn	4239	77.52982	4	Unknown	2569	60.60392	No	1201	28.33215	
15	death_yn	18865	0	2	No	14287	75.73284	Yes	4578	24.26716	
16	underlyin	1729	90.83488	3	Yes	1713	99.07461	No	16	0.92539	
17											

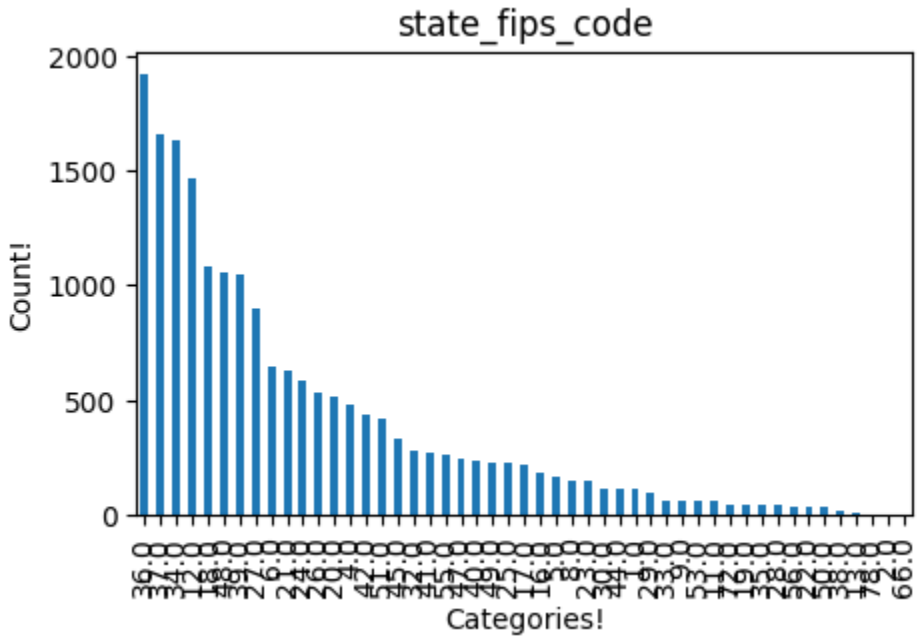
Figure 1 from Categorical22203536.csv file.

Case_month



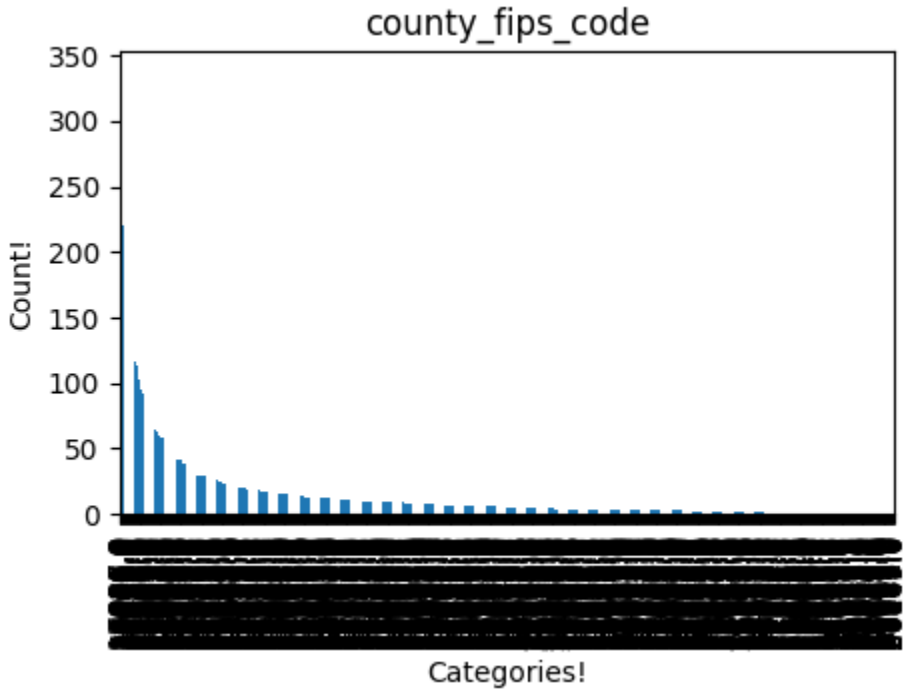
Data relates to cases over 35 different months for covid. Cases are rising exponentially from January 2020 to January 2022.
There is no missing data here so **data quality is good**.

State_fips_code



Feature refers to what state the cases present in. Data is present for 49 of the 50 US states. The modal state is New York .There is no missing data here so **data quality is good.**

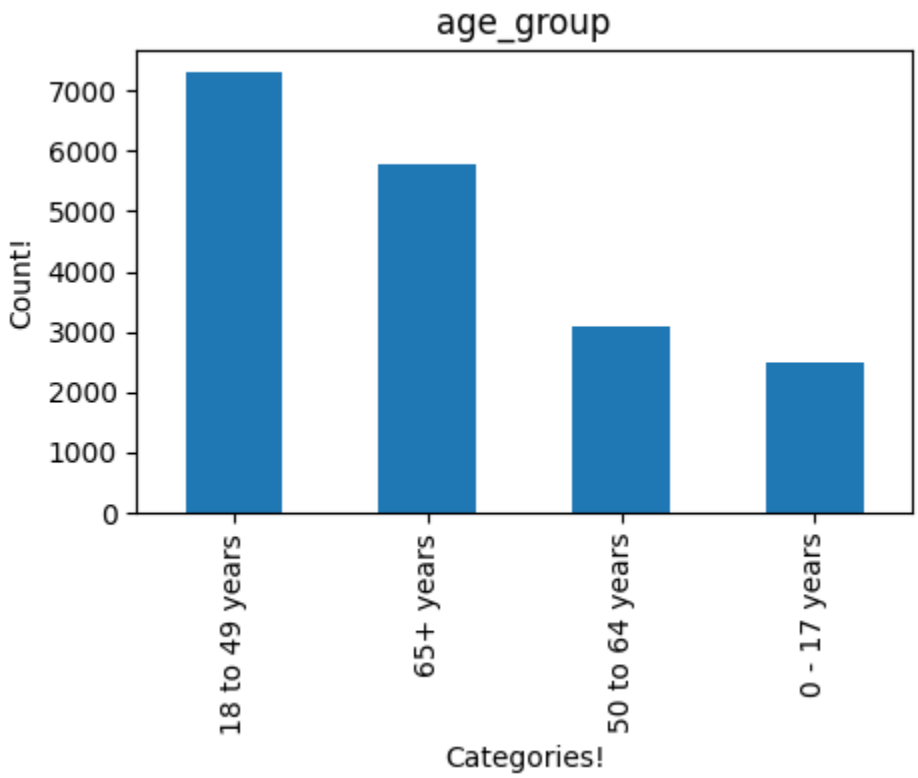
County_fips



County fips codes relate to which US county the cases are in. Some counties have more cases due to higher populations and higher covid spread. The missing data is over 6%. This leads to

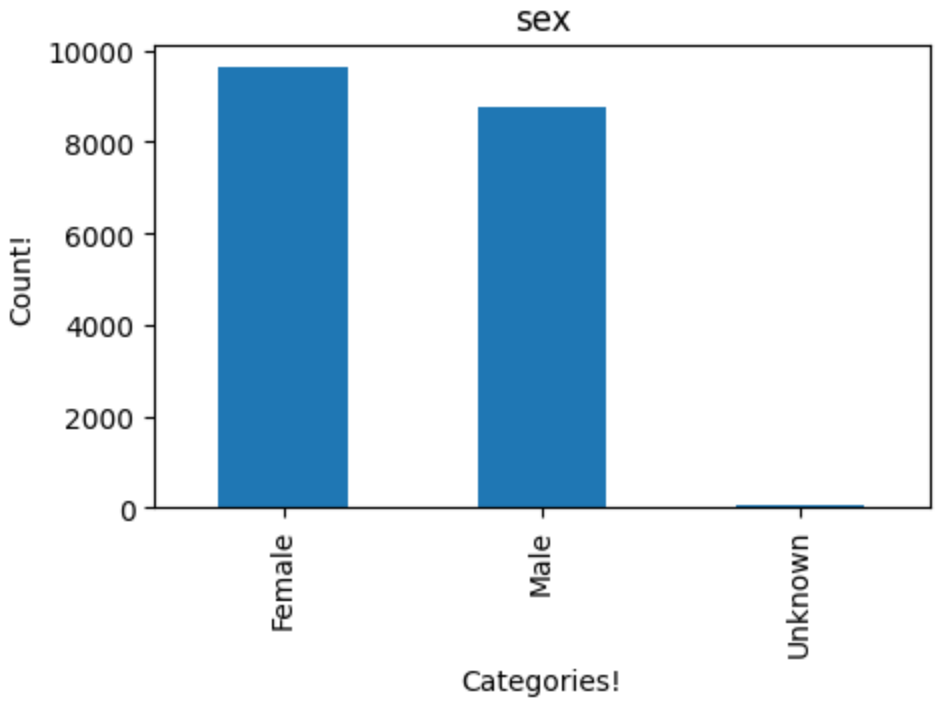
lower accuracy.The cardinality here is very high. This can lead to issues such as overfitting, data sparsity and slow model training times. For this reason **the data quality is poor.**

age _group



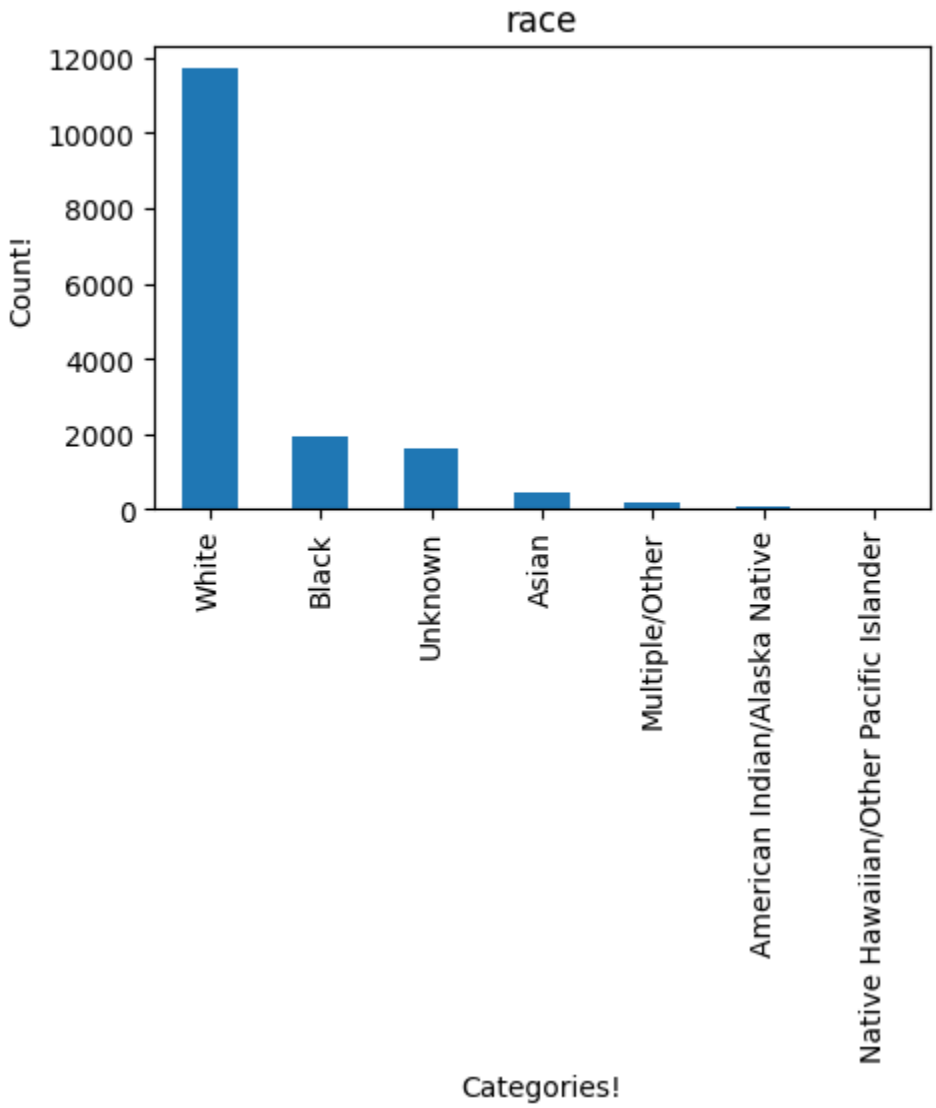
Modal age group is 18-49 years old. Missing data percentage is 1.06%. **Data quality is good**

Sex



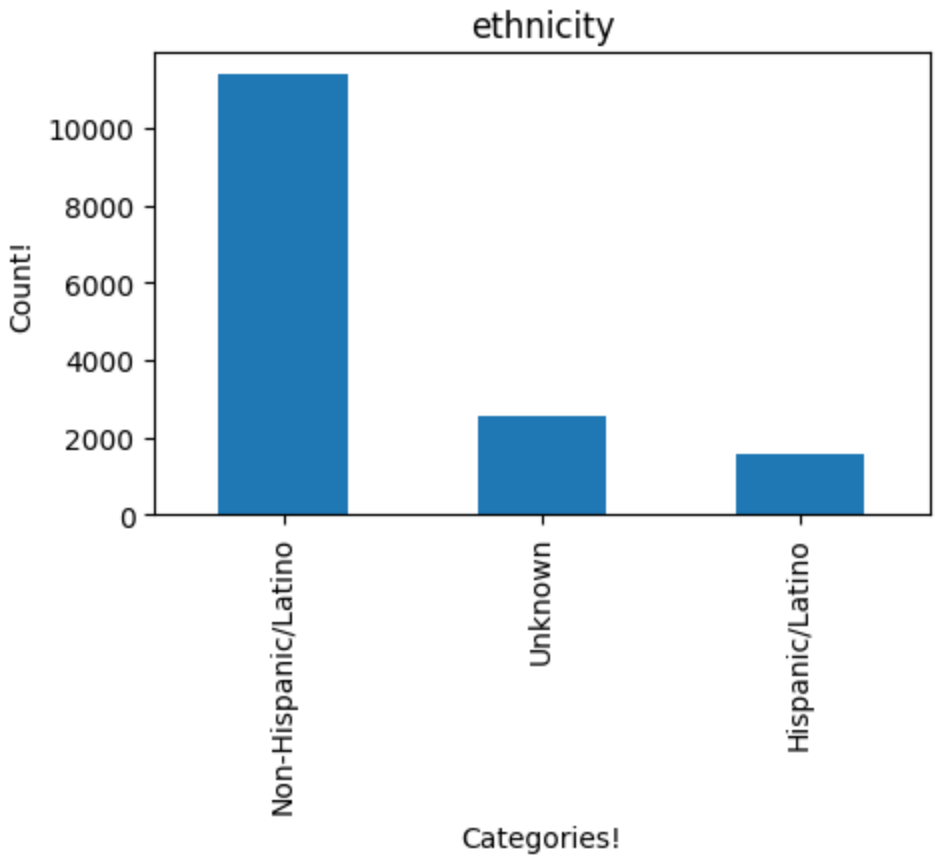
More female cases than male. Small number of unknown sex. Missing data % is 2.25%. This is lower than 5%. **Data quality is good.**

Race



Model race is white. Could make it harder to extrapolate about some minorities due to lower value counts. Missing data also accounts for 15.35 % of the dataset. Missing data causes a number of issues such as lowering the sample size thus lower accuracy, creating bias including imputation bias if the data is imputed. **The data quality is poor.**

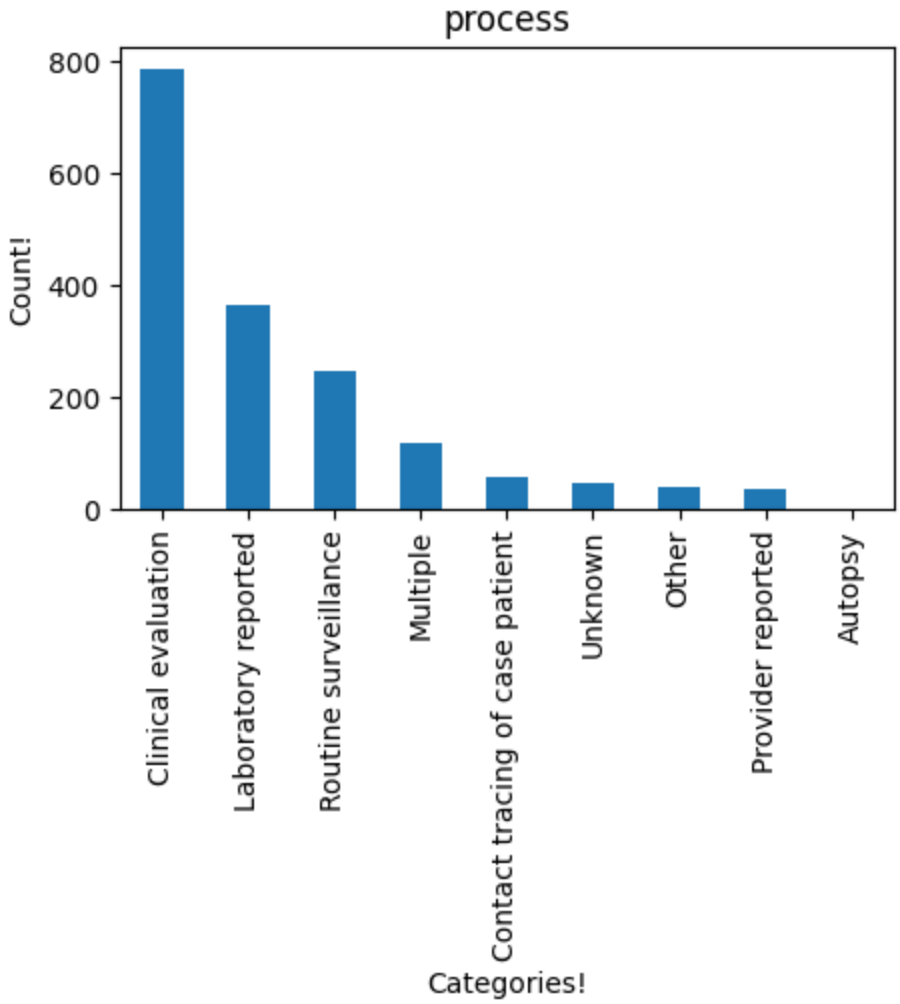
Ethnicity



Not divided into enough categories to draw conclusions about certain groups i.e all whites and blacks would come under non hispanic. The missing data % is 17.83%. Missing data causes a number of issues such as lowering the sample size thus lower accuracy, creating bias including imputation bias if the data is imputed. **The data quality is poor.**

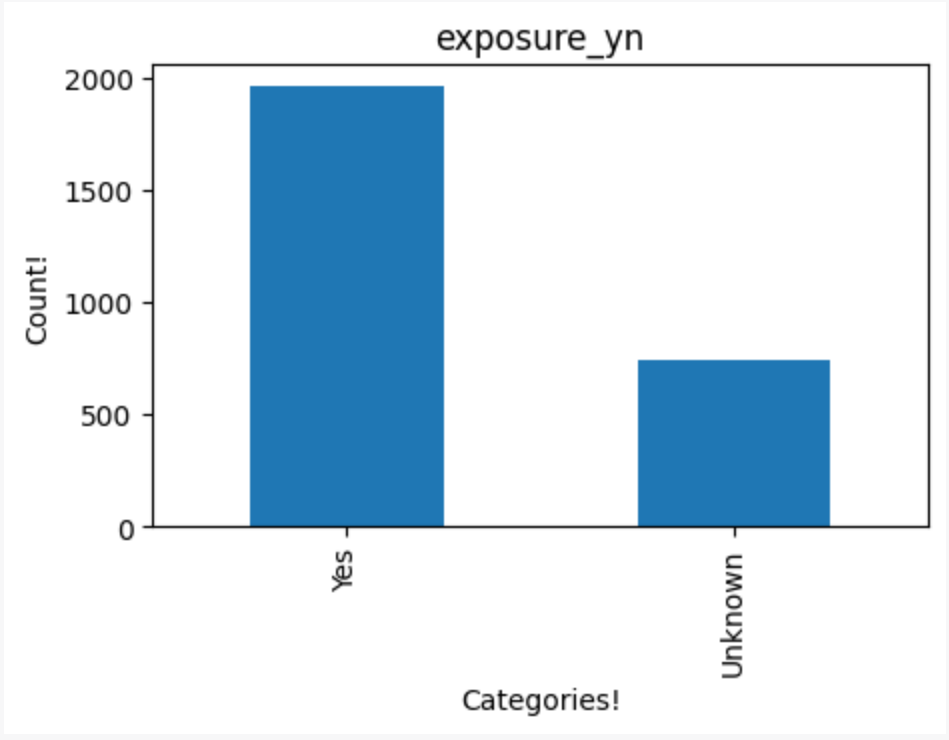
Process

]



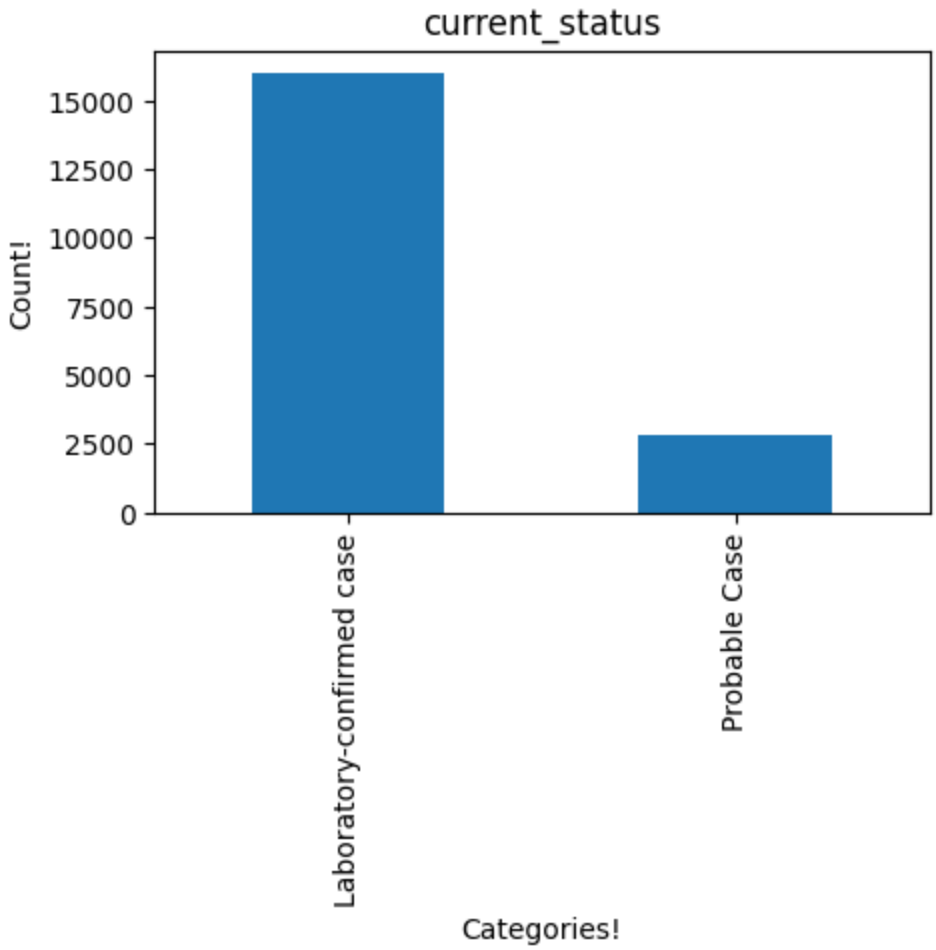
Over 90% of the data is missing.Missing data causes a number of issues such as lowering the sample size thus lower accuracy, creating bias including imputation bias if the data is imputed. We will drop this feature due to very low data accuracy. The **data quality is very poor**

Exposure



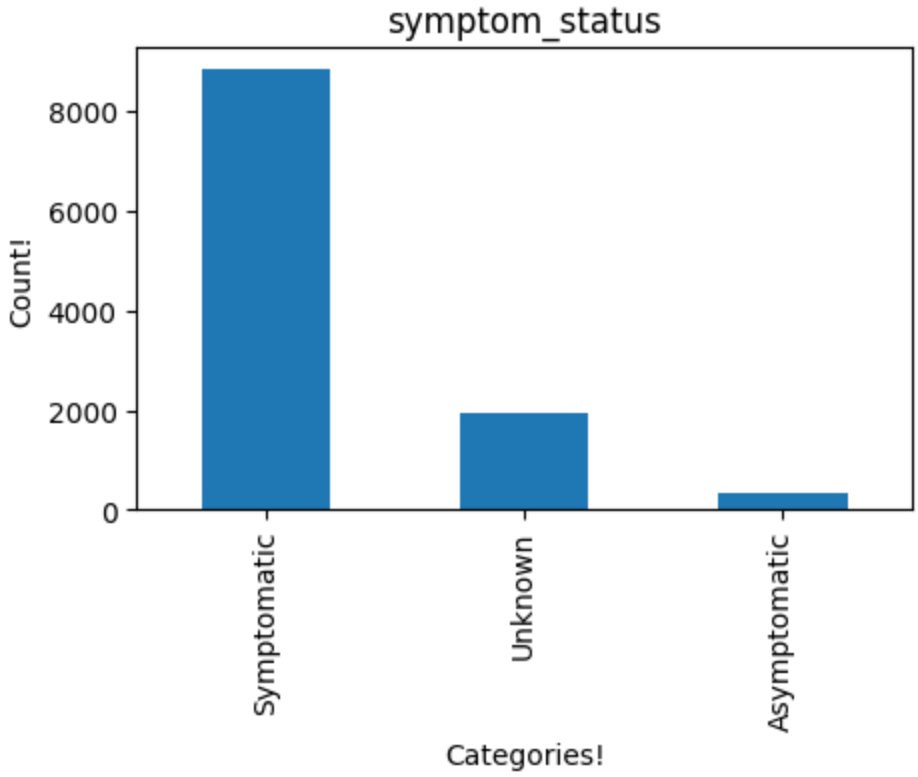
Data relates to whether a known exposure caused a case. Over 85% of the data is missing so the **data quality is poor.** Missing data causes a number of issues such as lowering the sample size thus lower accuracy, creating bias including imputation bias if the data is imputed. We will drop this feature

Current_status



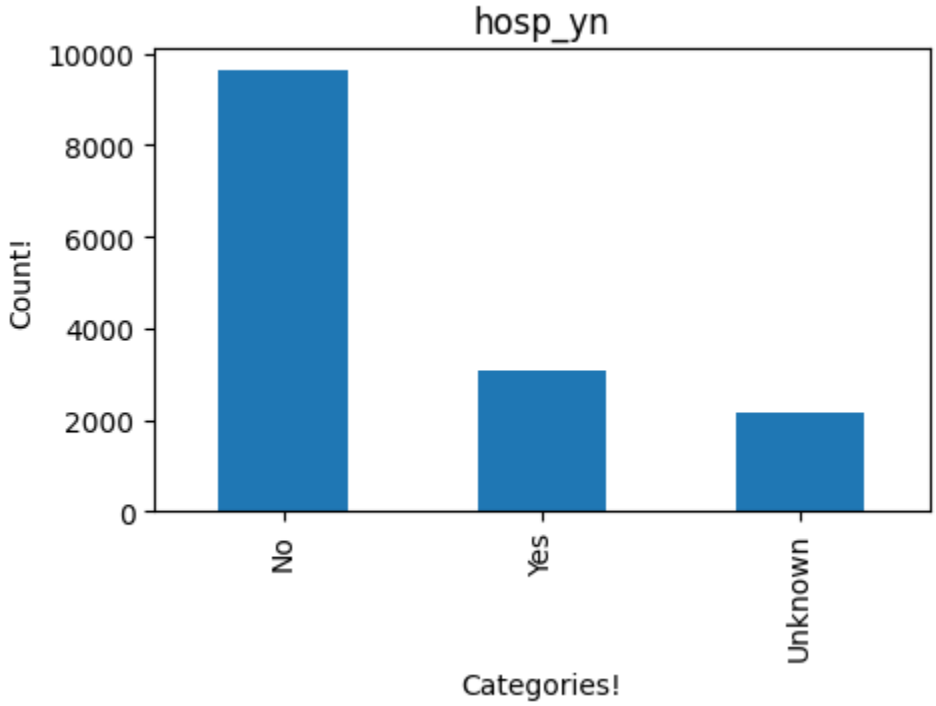
Data relates to whether the current status of a case is Lab confirmed or a probable case. No missing data for this feature. **Data quality is good.**

Symptom_status



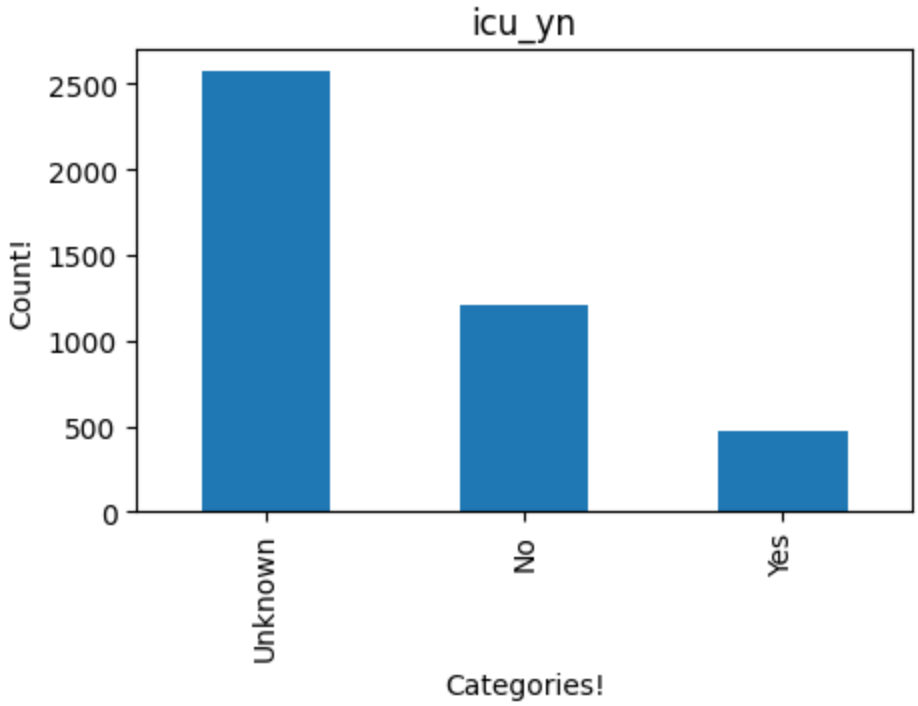
Data relates to whether the current case is symptomatic or asymptomatic. 41% of the data is missing data quality is poor. Missing data causes a number of issues such as lowering the sample size thus lower accuracy, creating bias including imputation bias if the data is imputed. However since the missing data is lower than 50% we will consider imputing the data. **Data quality is poor.**

hos_yn



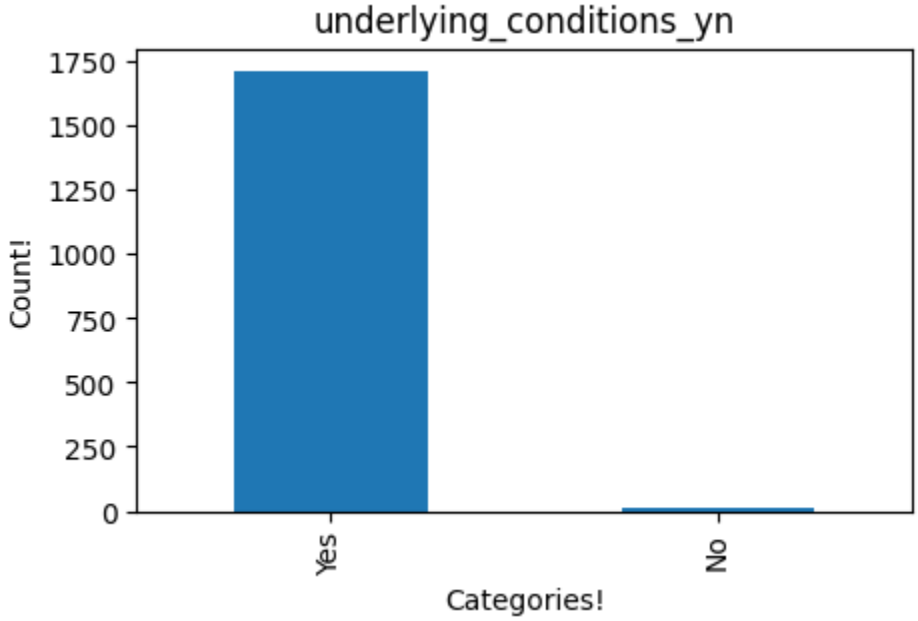
Data relates to whether a case was admitted to the hospital. Missing data accounts for 21.19 % of all data. Missing data causes a number of issues such as lowering the sample size thus lower accuracy, creating bias including imputation bias if the data is imputed. However since the missing data is lower than 50% we will consider imputing the data. **Data quality is poor.**

Icu_yn



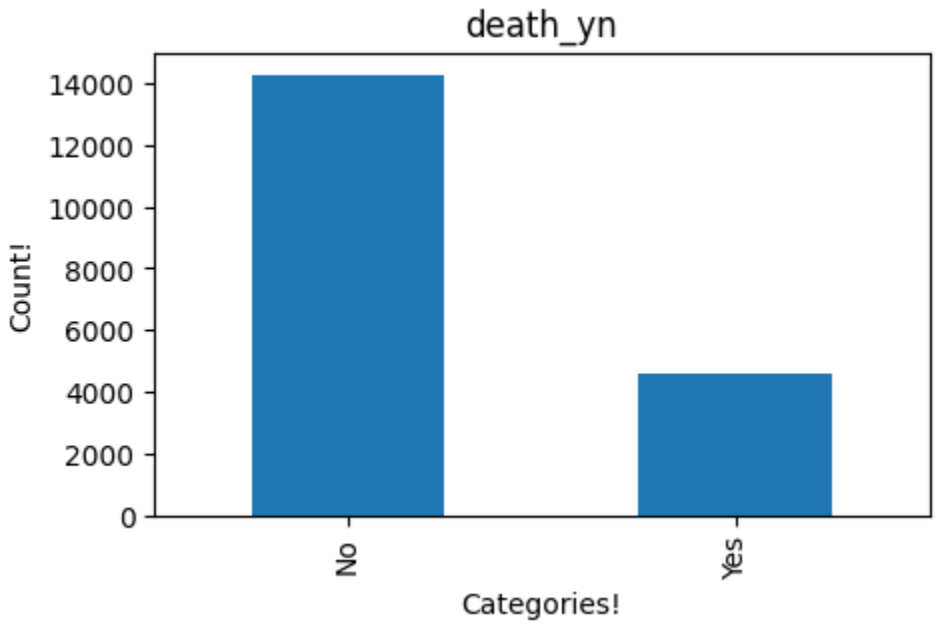
Data relates to whether a patient was admitted to ICU. The **data quality is very poor** since more than 77.5% of the data is missing so even imputed values would not be informative. We will drop this feature.

underlying _conditions



Data relates to whether a case had underlying medical conditions other than covid. More than 90% of the data is missing. Even imputing the values would not be informative. **Data quality is very poor.** We will drop this feature.

death_yn



Feature relates to whether a case died. No missing or unknown data. **Data quality is very good**.

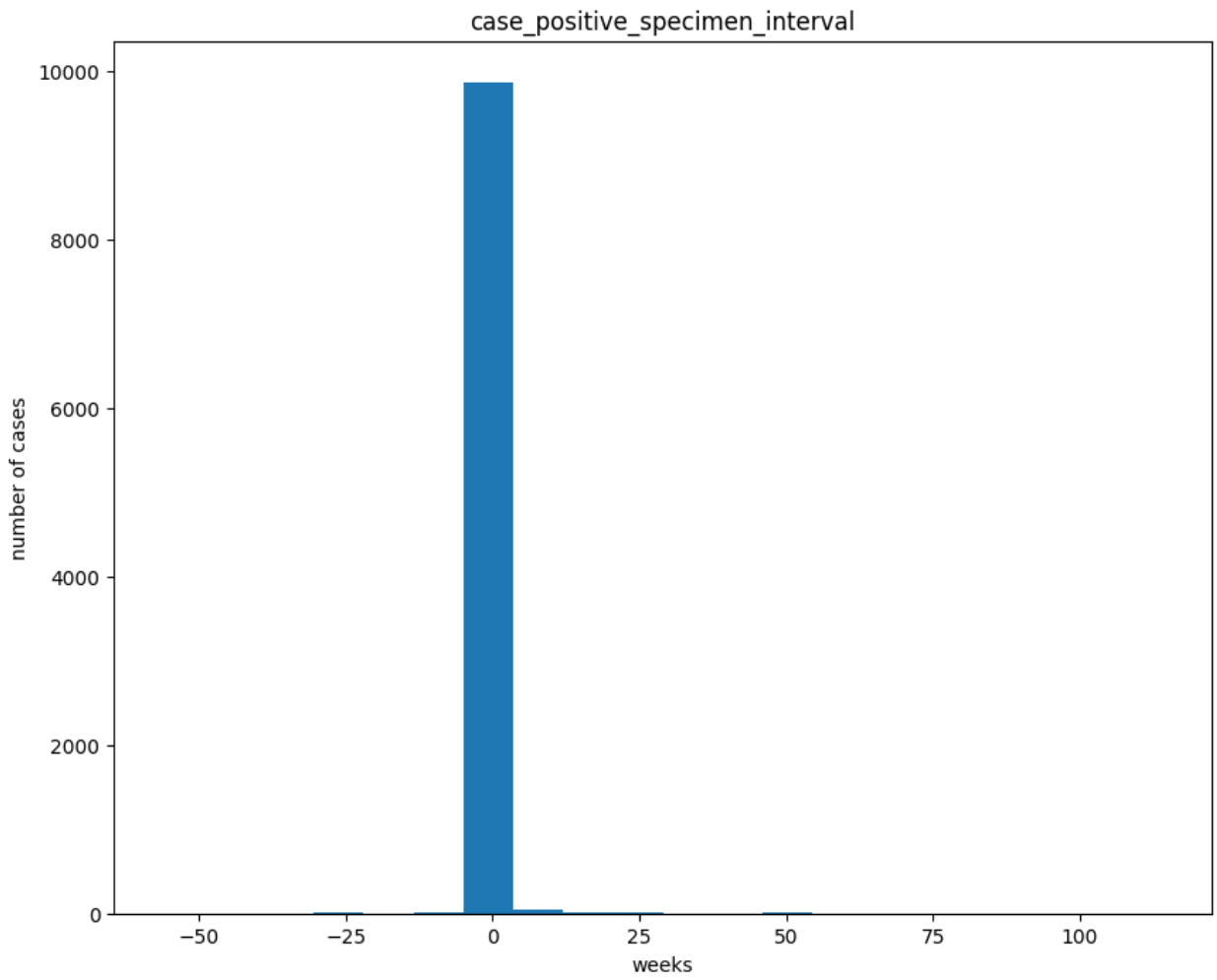
Continuous features

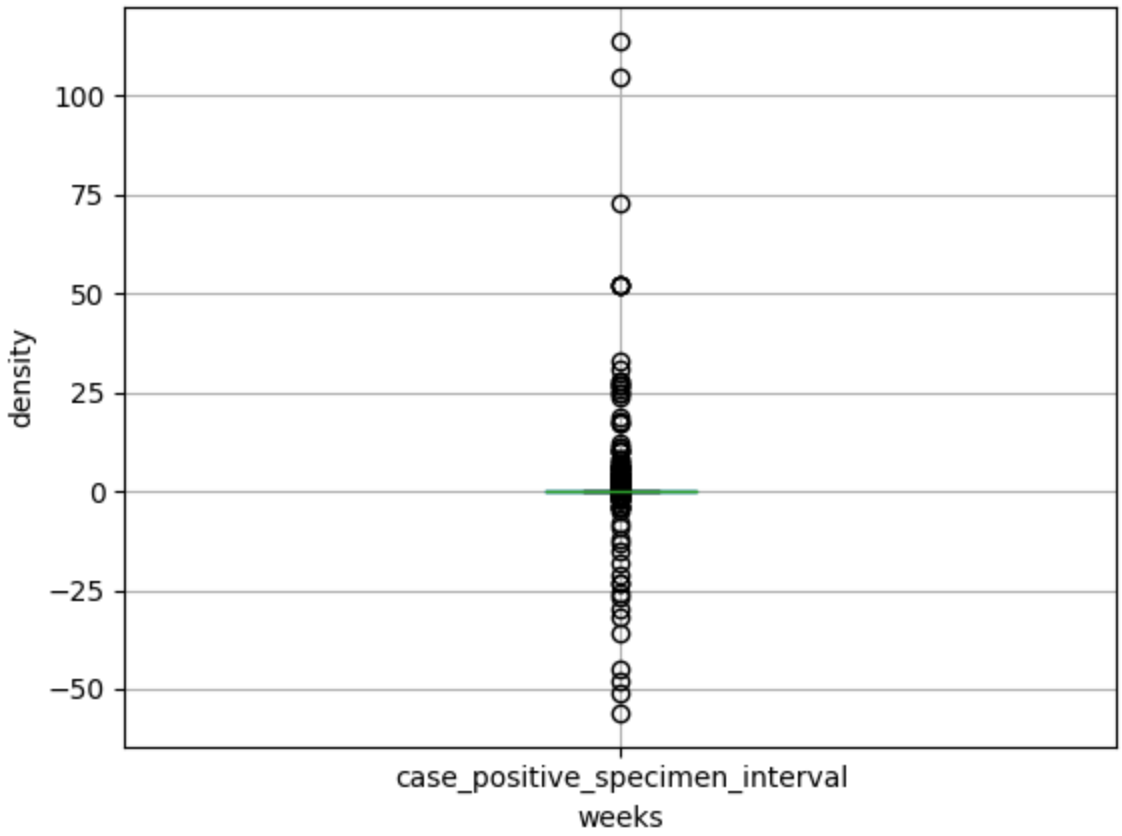
	A	B	C	D	E	F	G	H	I	J	K	L
1	Feature	Count	Missing pe	Cardinalit	Minimum	1st quart	Mean	Median	3rd quart	Max	Standard Deviation	
2	case_posi	9938	47.32043	48	-56	0	0.16613	0	0	114	2.653115	
3	case_onse	8347	55.75404	43	-69	0	-0.06781	0	0	51	1.872193	

Figure taken from continuous22203536.csv. Refer to csv file.

Case_positive_specimen_interval

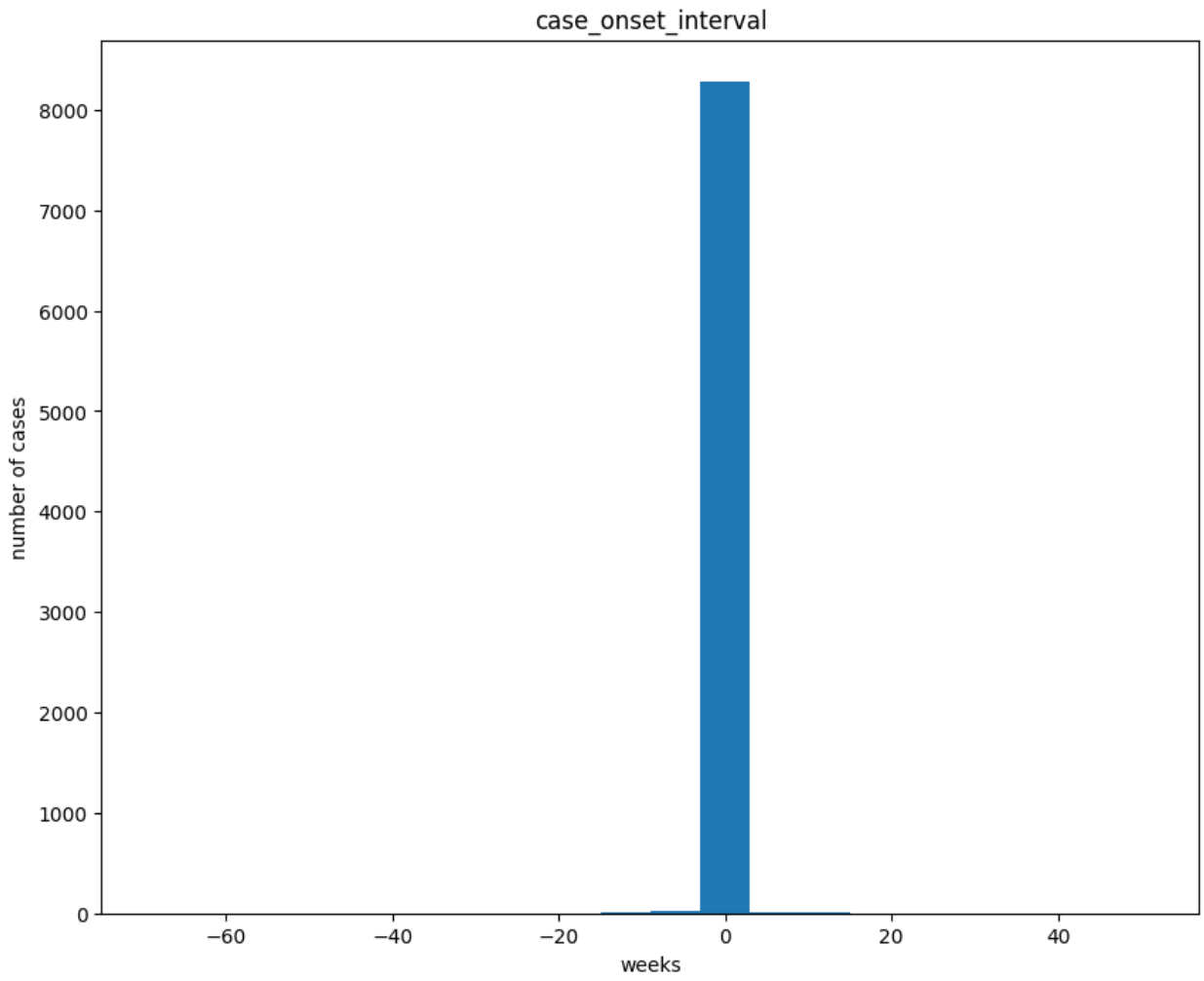
Refers to distance from earliest date to date of positive specimen. As such negative values are logically incoherent. A number of approaches such as excluding values below 0 or getting the absolute value of the negative values could be taken. 47% of data is also missing. **Data quality is very poor**

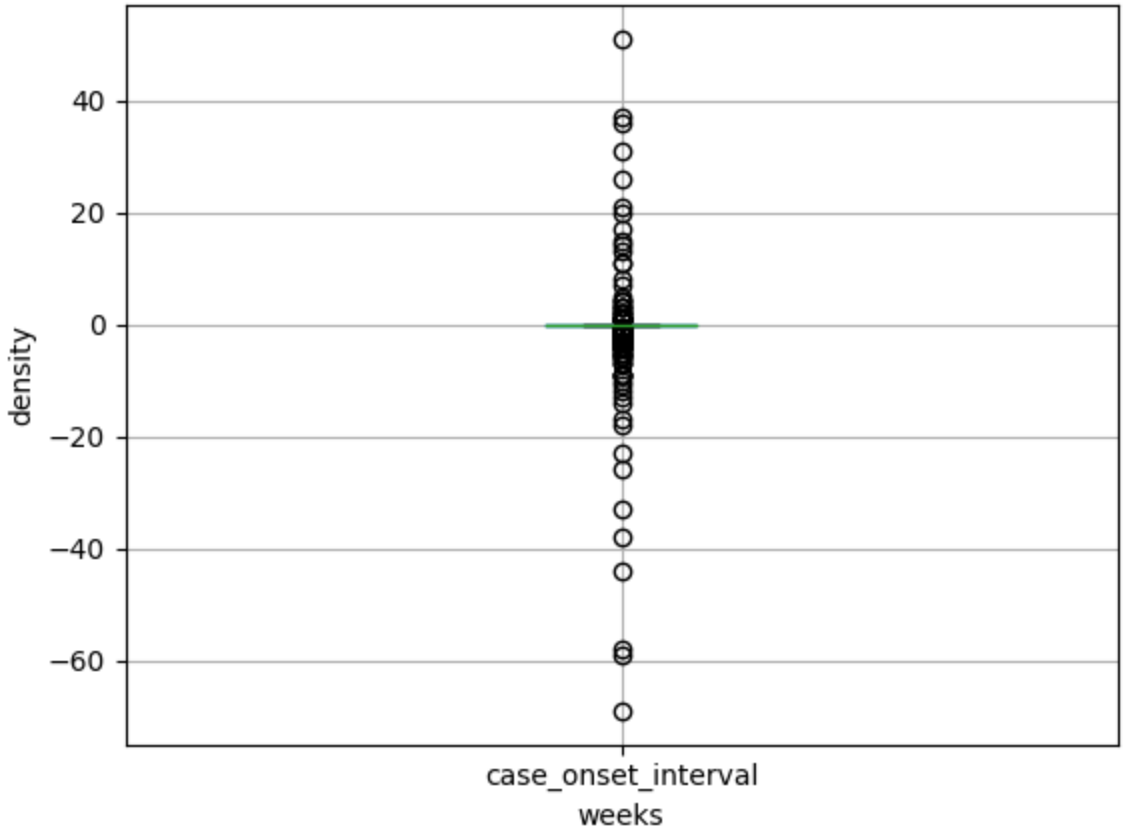

case_positive_specimen_interval
number of cases
10000
8000
6000
4000
2000
0
−50
−25
0
25
50
75
100
weeks



Case_onset_interval

Refers to distance in weeks from earliest date to date of onset of symptoms. Logically incoherent for it to hold negative numbers. A number of approaches such as excluding values below 0 or getting the absolute value of the negative values could have been taken, however since the missing values are over 55% of the dataset we will simply exclude this feature since imputing with less than half the dataset is not informative. **The data quality is very poor.**


case_onset_interval
number of cases
8000
7000
6000
5000
4000
3000
2000
1000
0
−60
−40
−20
0
20
40
weeks



END OF REPORT.