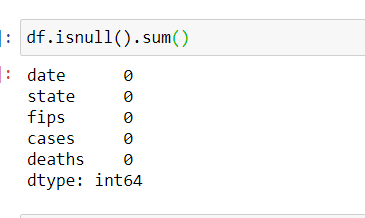
For this project the US states covid data set is considered. In order to analyse the dataset, at first the data set is checked for any kind of inconsistencies such as missing values. When checked using python code, it is found that there are no missing values.

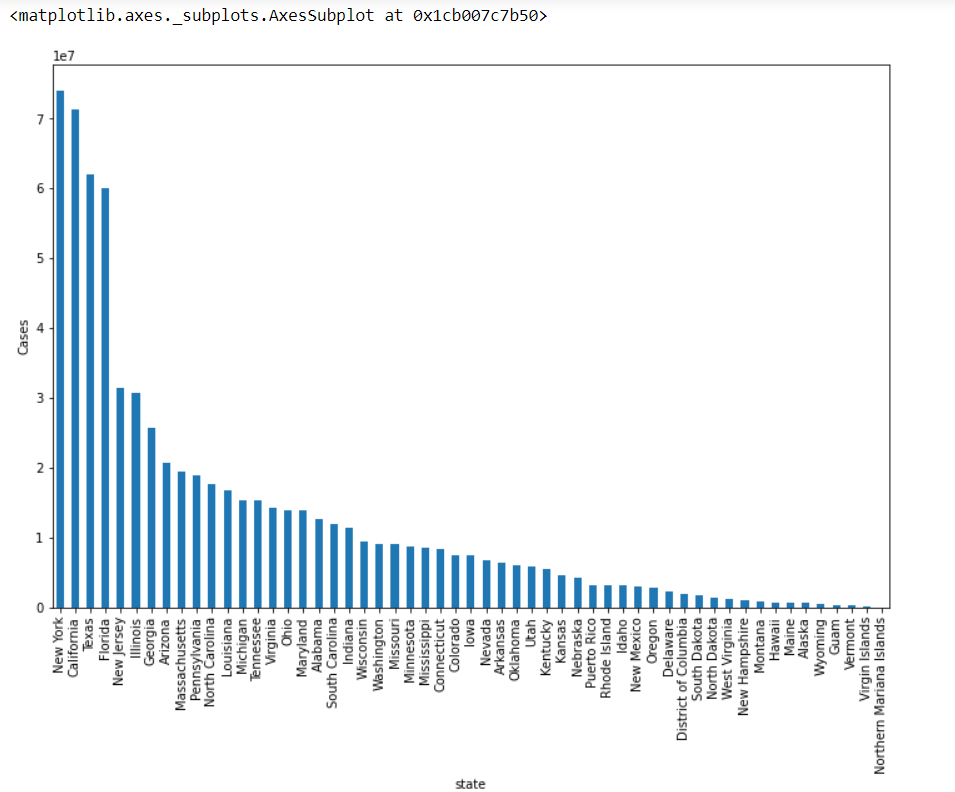


For preparing the data the dropna() function is utilized in the process.

Following table shows the descriptive statistics for the dataset,

|  |  |  |  |
| --- | --- | --- | --- |
|  | fips | cases | deaths |
| count | 12279 | 12279 | 12279 |
| mean | 31.85577 | 56404.21 | 2033.969 |
| std | 18.61078 | 110606 | 4500.918 |
| min | 1 | 1 | 0 |
| 25% | 17 | 1738.5 | 42 |
| 50% | 31 | 14755 | 418 |
| 75% | 46 | 62707 | 1904.5 |
| max | 78 | 857076 | 32879 |

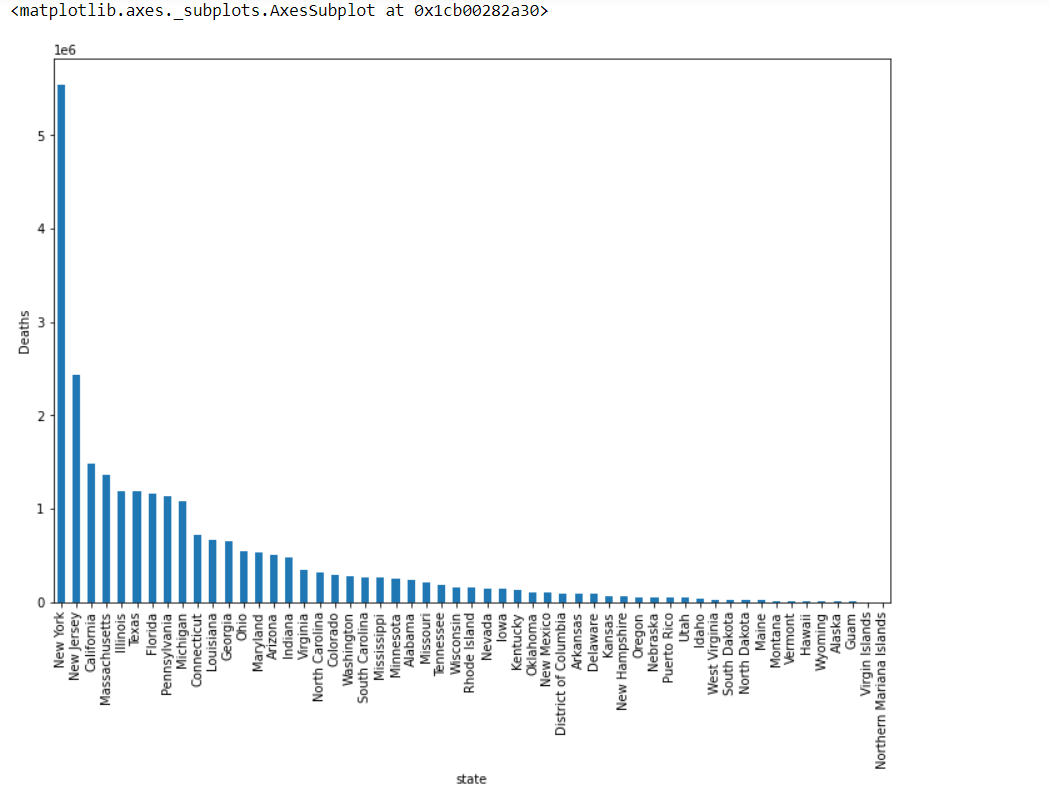
For the recorded cases in the different states, the following plot is created.



Here, as expected New York had the highest number of cases. But surprisingly the cases in California, Texas and Florida is comparatively high too.

Following is the histogram for the death column data.

For the plots of the available data a first the deaths in the different states are plotted.



Highest number of deaths are recorded for New York and lowest is for Northern Mariana Islands.

