**Student Name: Muhammad Shoaib Khan  
Roll Number: DS-007/2021**

**1 Data Pre-processing and Preparation Report a short summary of the pre-processing steps you applied to the data to make it usable for analysis. This should at a minimum include discussion of outliers and normalization (which features, what kind of normalization, if any, is needed and appropriate?). What do you do with "Day", "State" and "State ID"?**

Ans 1) For **Pre-processing**, we clean columns [‘Resident Population 2020 Census’, ‘Population Density 2020 Census’] by removing ‘,’ from the column values using the command:

data[‘Col\_Name'] = data[‘Col\_Name'].replace({'\$': '', ',': ''}, regex=True)

Then we convert their types form object to int or float.

For Normalization we use two different methods:

1. from sklearn.preprocessing import MinMaxScaler
2. from sklearn.preprocessing import StandardScaler
3. from sklearn.preprocessing import scale

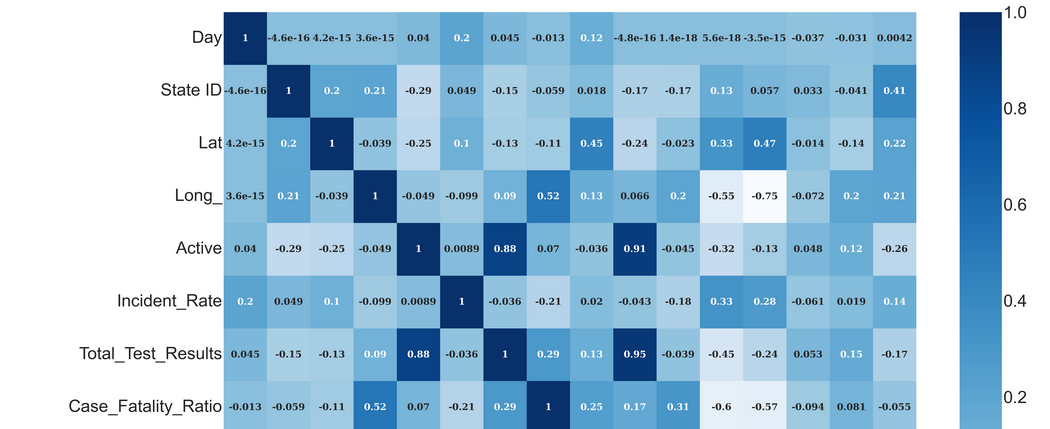
**sklearn.preprocessing.MinMaxScaler**This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

**sklearn.preprocessing.StandardScale**Standardize features by removing the mean and scaling to unit variance.

**sklearn.preprocessing.scale**Standardize a dataset along any axis.

Center to the mean and component wise scale to unit variance.

**Which Feature:**



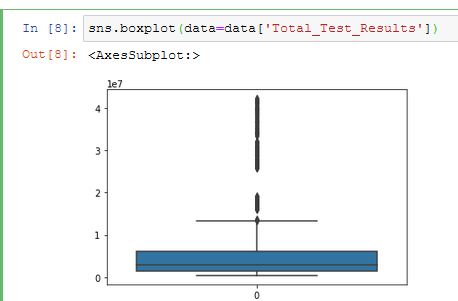
Features which are strongly correlated with other features like ‘Day’, ‘Density Rank 2020 Census’, ‘Resident Population 2020 Census’, ‘Total\_Test\_Results’ etc

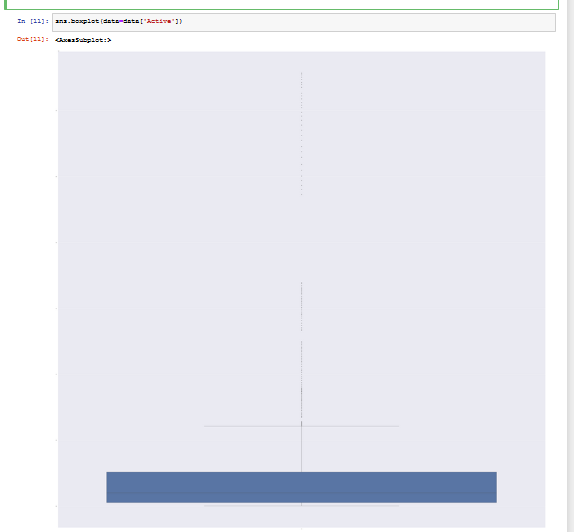
**Reason of Apply Normalization and Standardization:**

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

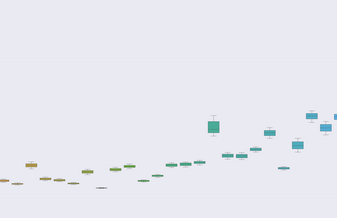
We ues normalization so that values come between the range of 0 and 1 which makes applying PCA easier.so that data values can be scale around the mean. It values remain at the different scale then because of ranges different we cannot compare values easily and thoroughly. Like some value is in thousand and some in million. So because of we can do different values comparison easily we scale and standardize the values

Ans 2) Then we check if there are any outliers present in the data file by using boxplot. Firstly, on displaying boxplot on all values it seems that there are huge number of outliers present in the two columns[‘Active’ and ‘Total\_Test\_Results’]





but when we draw these two columns boxplots with respect to the population of the State they seems to be normal.



Ans 3) We dropped the ‘State’ Column for applying PCA and use ‘State ID’ only. Day column is used for grouping of data.

First 2 components were only used to show the explained variation among the most significant variables.

The 4 components in this analysis gave a cumulative 64.5% of the total explained variation which means that these 4 components are important upto 64.5% significance. The distance of each principal component are found out and then the explained variation is obtained which are the eigenvalues. The most significant principal component has explained variation of 24.8%.

These explained and cumulative variances were then plotted.