**Report on Advanced Python Project**

* First, I imported all the necessary libraries and then read the data set.
* The data set contains 640 records and 61 columns which includes two categorical columns State and Area Name.
* From the summary of the given data set, the Dist.Code contains the values are the serial numbers from 1 to 640. So, I feel that the variable Dist.Code is of no use and hence dropped the column Dist.Code.
* After that I performed Univariate Analysis for all the numerical columns and the categorical columns.
* For the numerical columns, I constructed the histograms to study the shape of the distribution and observed that all the numerical columns except the State Code are positively skewed.
* Also, I constructed the box plots for all the numerical columns to identify the outliers and observed that all the numerical columns except the State Code have outliers.
* For the categorical columns, I constructed the bar plots and observed that Uttar Pradesh has maximum frequency, Lakshadweep, Chandigarh, and Dadra & Nagar Haveli have the minimum frequency. The areas Raigarh, Bijapur, Aurangabad, Hamirpur and Bilaspur have the maximum frequency and the rest were having the minimum frequency.
* After that I performed a bi-variate analysis and studied the correlation between all the numerical columns.
* From the heat map, I observed that there are consider number of features that are highly correlated.
* Then I did the outlier treatment.
* Then before proceeding to the PCA, I normalized all the numerical features by applying z scores.
* After that I performed some statistical tests to know whether PCA is recommendable or not for the given data.
* From the Bartletts Test of Sphericity, I got the p-value as 0.0 which means that at least one pair of the variables are correlated and hence PCA is recommended.
* From the KMO Test, the measure of sampling adequacy (MSA) = 0.93 > 0.7 which means that there is a considerable reduction in the dimension and extraction of meaningful components.
* Finally, I performed the PCA.
* From the cumulative variance explained in percentages, we observe that more than 84% of the variance is explained by 3 Principal Components, around 93% of the variance is explained by 6 Principal Components, around 97% of the variance is explained by 12 Principal Components.