Data Preprocessing:

in data preprocessing we use different machine learning libraries, which help to understand the dataset and analyze are follow:
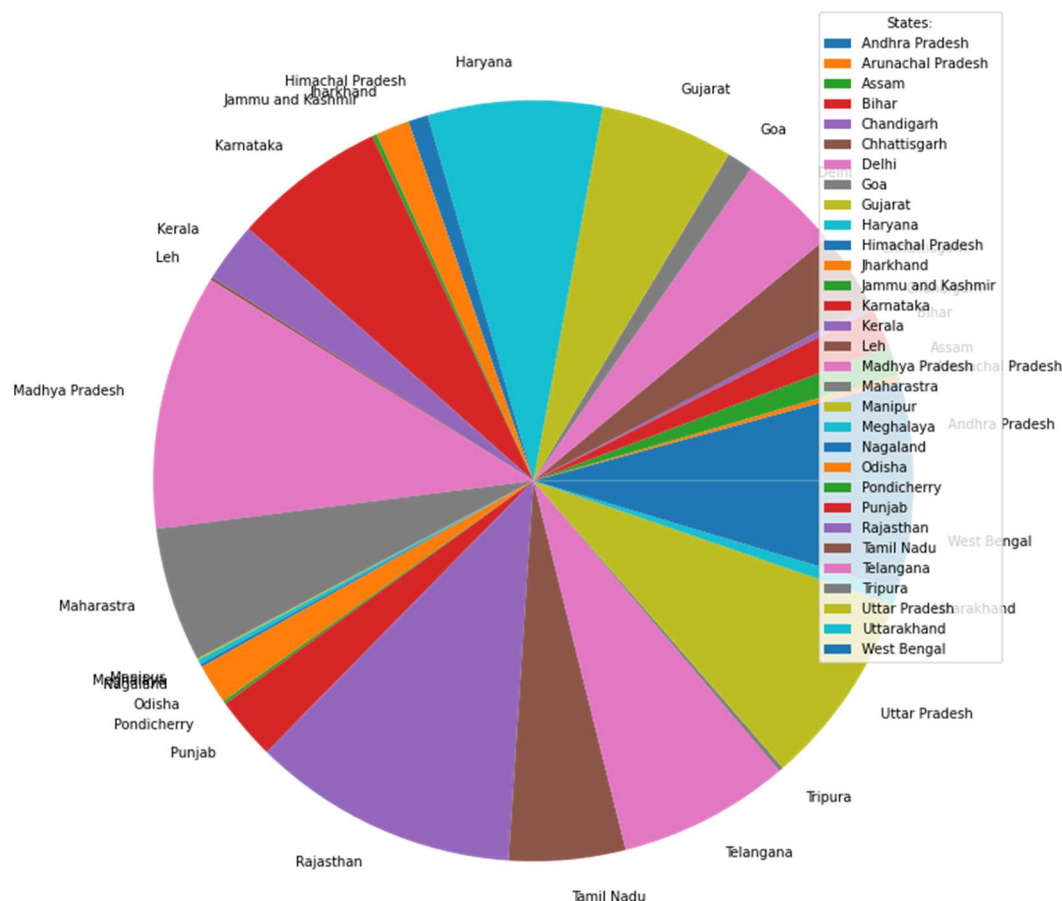
pandas

NumPy

Seaborn

Matplotlib

First of all, import all libraries, Once all these libraries were imported, we started understanding the datasets.

The pie chart visualization of the dataset helped us to understand that No of ROs were EV Charging Facilities of different states. The pie charts are given below-

a dataset that tells us the percentages of EVs distributed across each state as of December 2021.

Total Number of Active Electric Vehicles
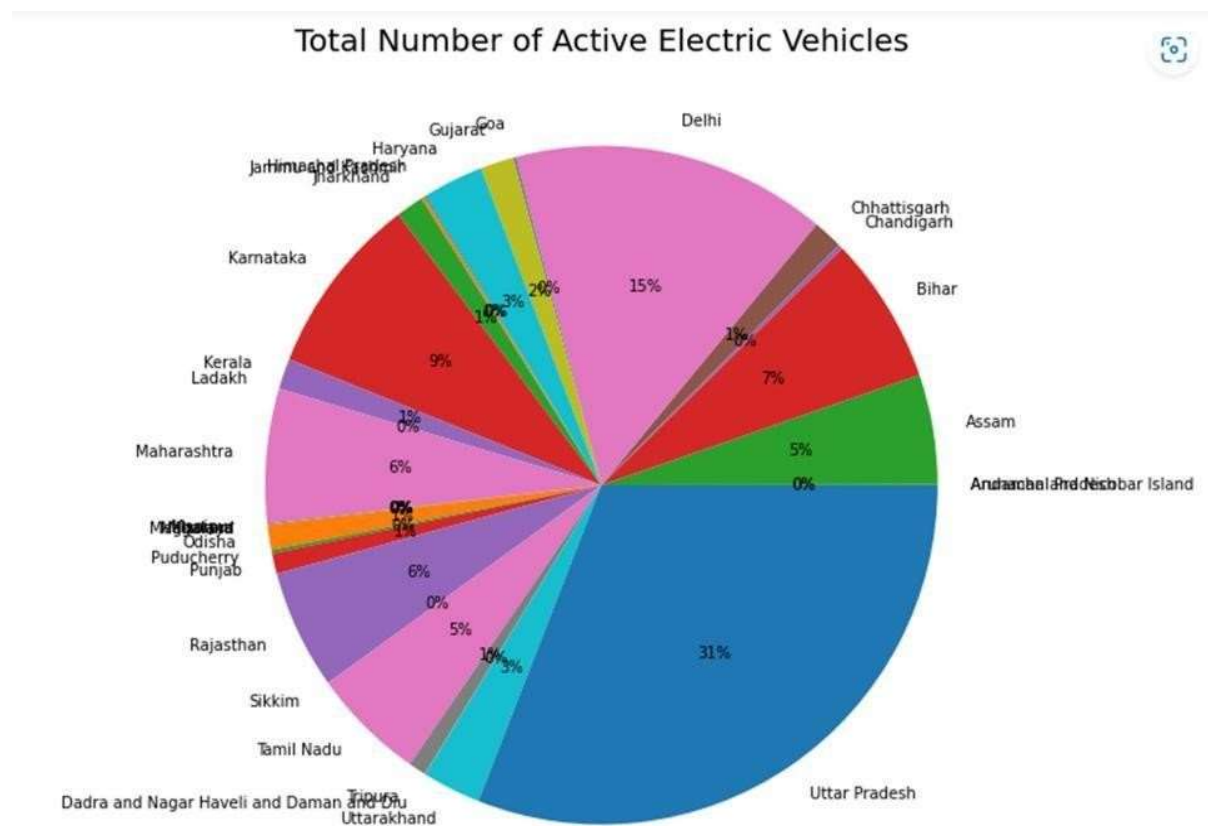
From the above pie chart, it can be clearly seen the most number of EVs present in the country is in Uttar Pradesh followed by Delhi, Karnataka, Maharashtra, etc.

a dataset df_total_sales that tells us the total number of vehicle sales across each state we visualize using a bar chart and see state Karnataka has more sales than other states. The bar charts are given below. And sort in ascending order based on the maximum number of sales see the table given below.
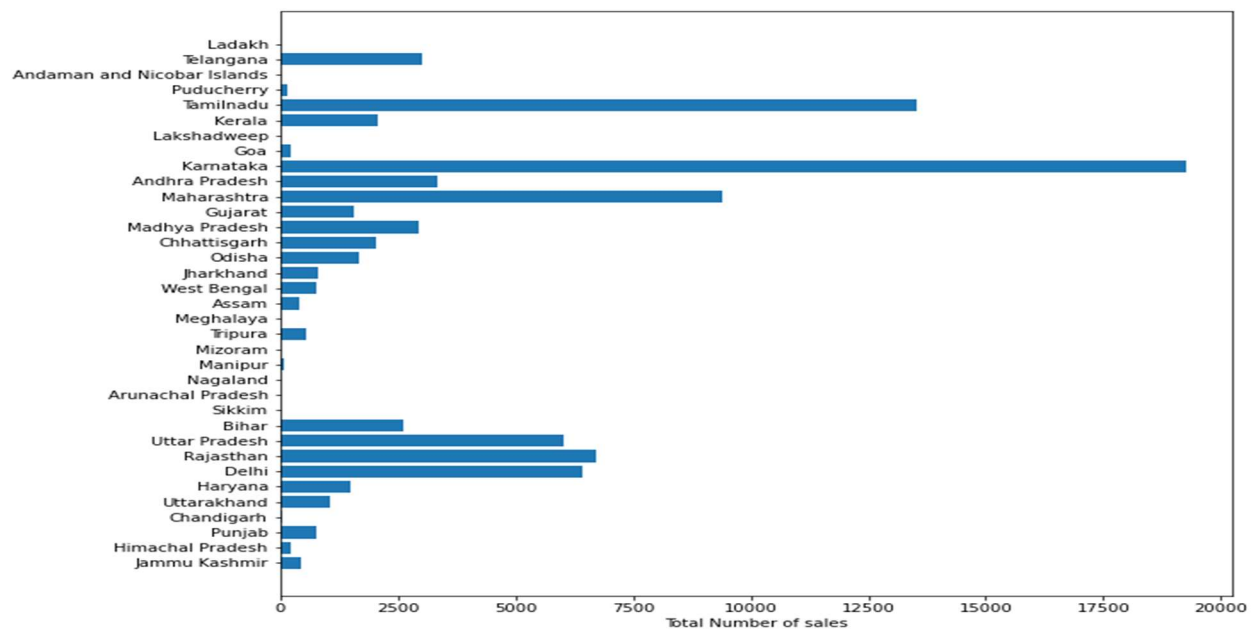
Unnamed: 0	Sl. No.	State/UT	Total Number of Invoice/Sales
26	26	27	Karnataka	19270
30	30	31	Tamilnadu	13515
24	24	25	Maharashtra	9393
7	7	8	Rajasthan	6721
6	6	7	Delhi	6413
8	8	9	Uttar Pradesh	6022
25	25	26	Andhra Pradesh	3325
33	33	34	Telangana	3031
22	22	23	Madhya Pradesh	2953
9	9	10	Bihar	2615

Top 10 States sale wise

After analyzing both the above pie charts it is felt these states will be a more suitable place for an EV start-up to start its operations because it has a significant percentage of Electric Vehicles and also its EV sales have been very high in recent times. Hence it is felt that starting the start-up operations from Karnataka will help the start-up company build a strong base in the Indian EV market.

Once the geographical location for the opening up of the start-up has been decided it is important to understand the Car Buying Behaviour of Indian Consumers. It will help the start-up to target the correct segment and thereby

start manufacturing the correct type of vehicle in order to attract their target consumers. The first look at the dataset df_income is as follows –

|	Unnamed: 0	Age	Profession	Marrital Status	Education	No of Dependents	Personal loan	House Loan	Wife Working	Salary	Wife Salary	Total Salary	Make	Price
0	0	27	Salaried	Single	Post Graduate	0	Yes	No	No	800000	0	800000	i20	800000
1	1	35	Salaried	Married	Post Graduate	2	Yes	Yes	Yes	1400000	600000	2000000	Ciaz	1000000
2	2	45	Business	Married	Graduate	4	Yes	Yes	No	1800000	0	1800000	Duster	1200000
3	3	41	Business	Married	Post Graduate	3	No	No	Yes	1600000	600000	2200000	City	1200000
4	4	31	Salaried	Married	Post Graduate	2	Yes	No	Yes	1800000	800000	2600000	SUV	1600000

The visualization of the dataset df_income will help to provide a lot of useful information related to the dataset df_income. It can allow us to analyze the Car Buying Behaviour of Indian Consumers and hence make lots of inferences from it. Hence the visualization of the dataset df_income containing

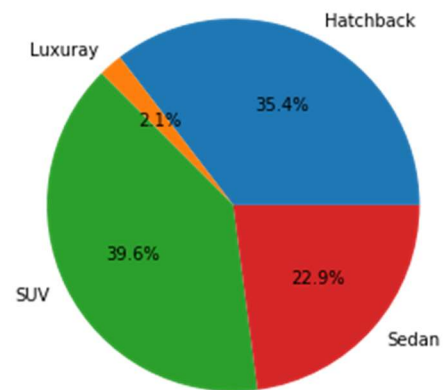Indian Consumer`s Car Buying Behaviour is as follows –
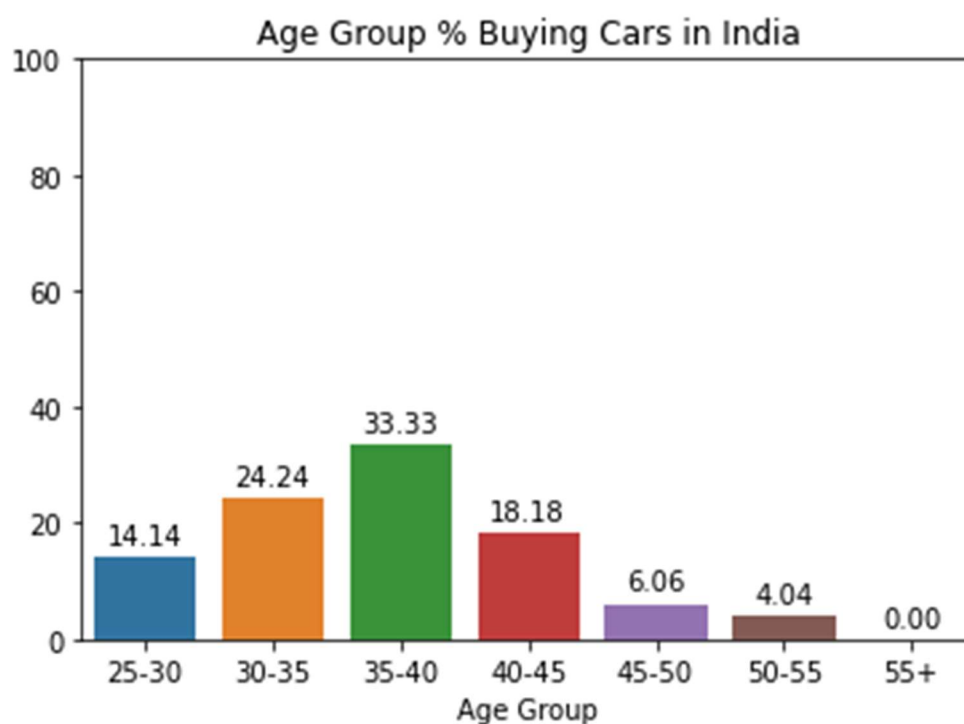
The above bar graph shows that the maximum number of people who are buying cars in India lies in the Age-Groups of 30 – 45 years. Hence targeting this age group seems to be a good option for the Company.

Married and No.of.Dependents>=3 'Car Make' chart
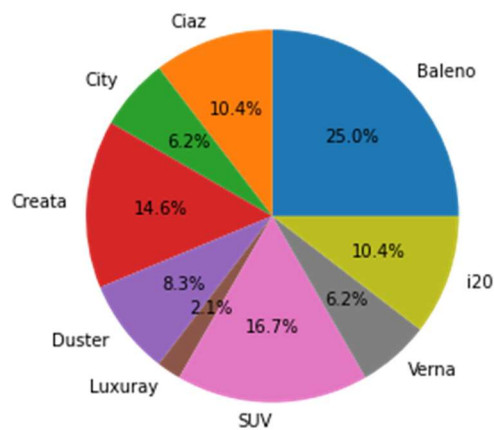
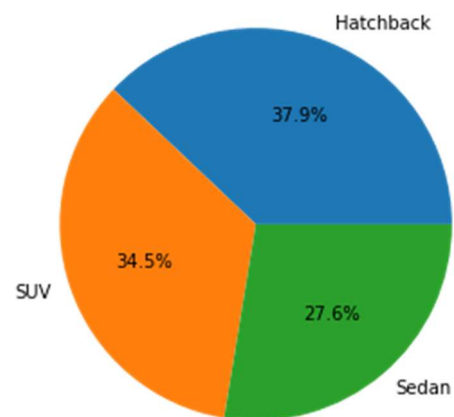
Married and No.of.Dependents>=3 'Car Type' chart

The above pie chart shows that Married People who have a number of dependents more than 3 people tend to buy SUVs more than other Car types. This will give an idea to the Company that it will be a good idea to start manufacturing an Electric SUV in order to target these people.

Salaried and Married
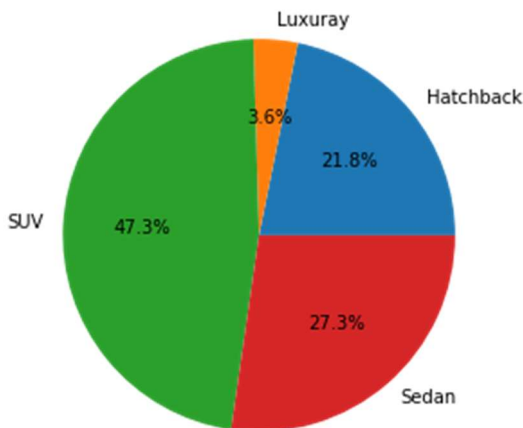
Business and Married

The above pie chart shows that of Consumers who are salaried and married about half of them tend to buy an SUV whereas about 35% of the consumers who are married and have their own Businesses also tend to buy SUVs as well.

The above pie charts compare the Financial Status of Indian Consumers. It shows that all those people whose financial status is low end up buying a Sedan car type as it is affordable for them. Whereas consumers whose financial status is Middle Class tend to buy Hatchback car type more as compared to an SUV or a

Sedan. On the other hand, consumers whose Financial Status is high tend to buy SUVs more often than Sedans or Hatchbacks. Hence the Financial Status of a consumer greatly affects their choice of Car. The company can use this information to target consumers of a particular target segment and then can make an Electric Car of the particular type which can woo its target consumers.

Financial Status='Low' Car Buying Behaviour | Financial Status='Middle' Car Buying Behaviour | Financial Status='High' Car Buying Behaviour
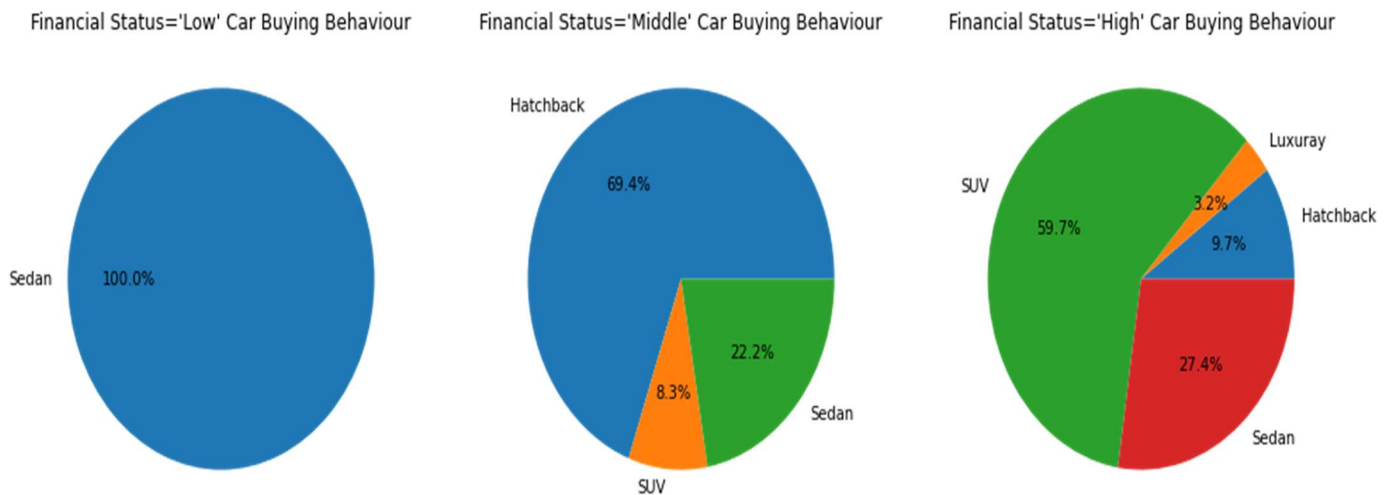
Then the histograms of the data were drawn below –
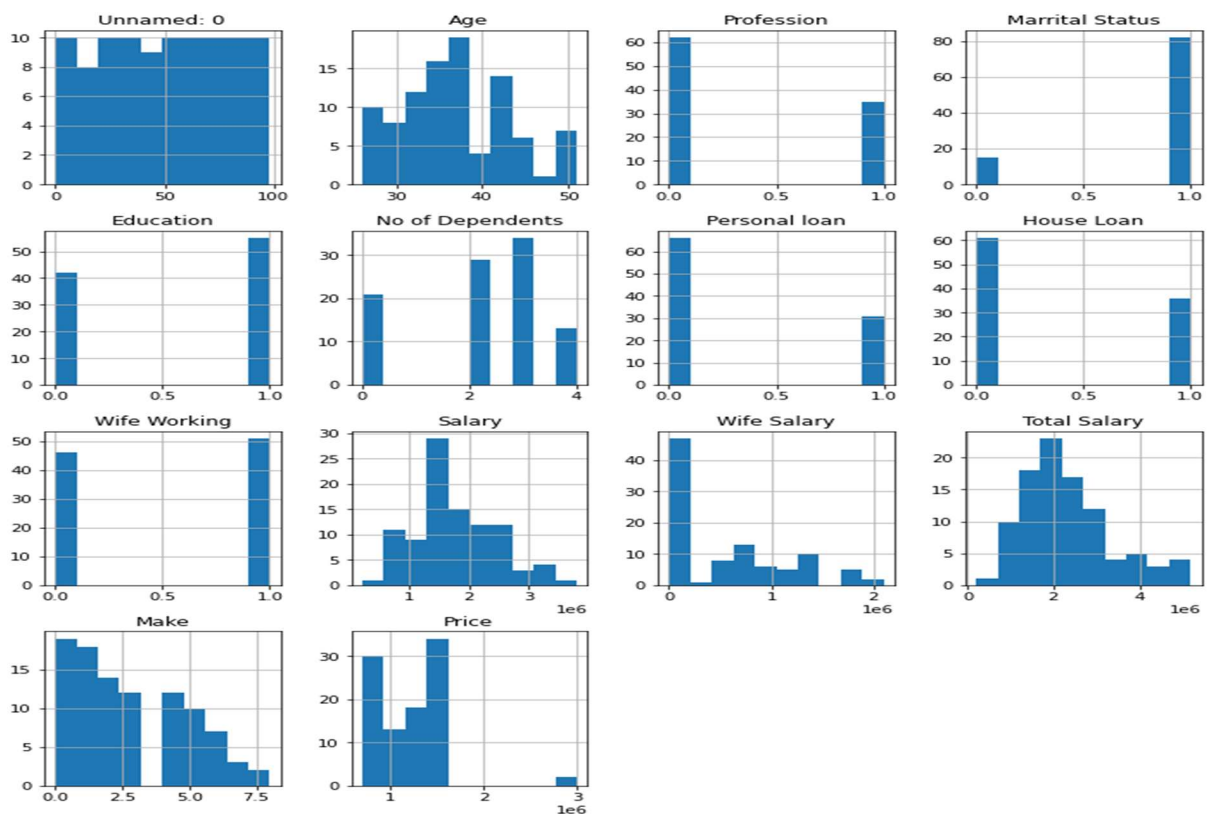
Once the histogram was plotted then the Principal Component Analysis(PCA) of the dataset was done. The PCA is a statistical process that converts a set of Correlated features into a set of linearly uncorrelated features with the help of Orthogonal Transformation. These new Transformed Features are called Principal Components.

This process reduces the dimensions of the dataset and makes it cost-effective to carry out clustering operations.

The Principal Components of the Indian Consumer Car Buying Behavior dataset is as follows –
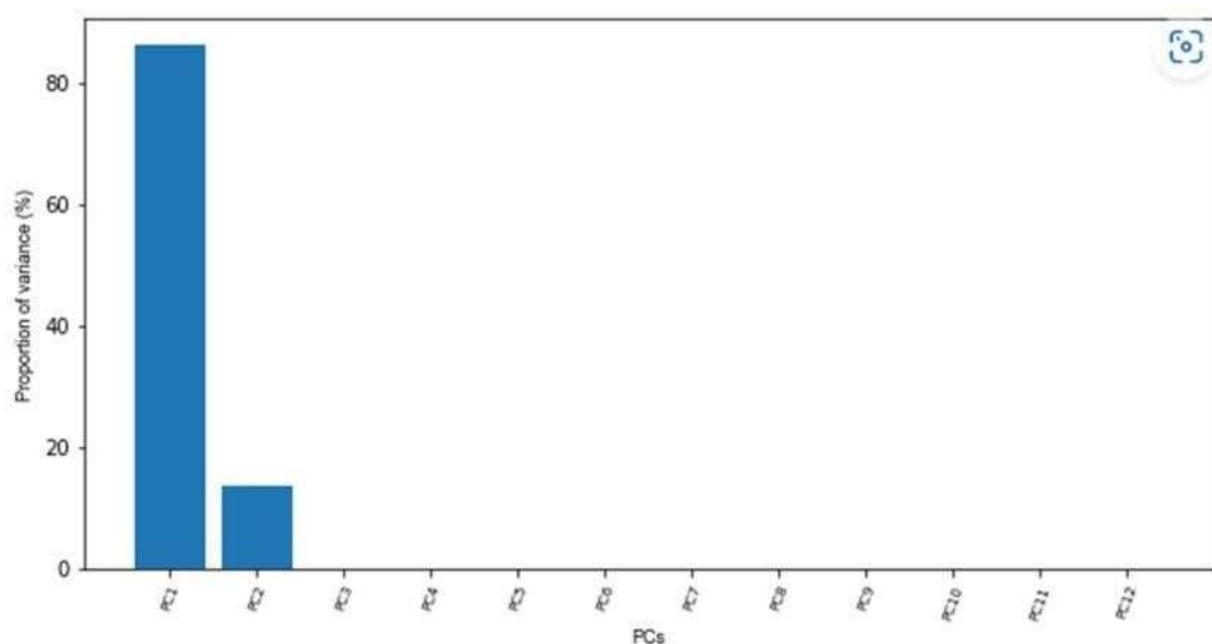
From the above bar chart, it can be clearly seen that the first 2 Principal Components themselves account for 100% of the variance in the data.

Also, the heatmap between the original variables and the principal components can be shown as follows –

variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Age	2.9e-06	3.4e-06	-0.98	-0.12	0.14	-0.021	-0.0079	-0.009	0.0047	0.0099	0.013	2.8e-11
Profession	-8e-08	-5.2e-08	-0.018	-0.0039	-0.011	0.22	0.53	0.8	-0.054	-0.13	0.056	7.7e-11
Marital Status	1.2e-07	-5.1e-08	-0.034	0.01	-0.13	-0.069	0.037	-0.092	-0.4	-0.63	-0.64	-4.4e-11
Education	3.7e-08	6.3e-08	-0.001	-0.0037	-0.0013	0.1	-0.84	0.51	-0.073	-0.095	0.0057	-2.1e-11
No of Dependents	2.5e-07	2.9e-07	-0.14	0.027	-0.96	0.18	-0.016	-0.039	0.083	0.08	0.057	-1.2e-11
Personal loan	-4.7e-08	2.1e-07	-2.3e-05	-0.02	-0.097	-0.66	0.0028	0.19	0.67	-0.25	-0.11	-2.6e-11
House Loan	5.6e-08	1.6e-07	-0.0068	0.0054	-0.15	-0.69	0.033	0.17	-0.6	0.33	0.13	5.1e-11
Wife Working	1.9e-07	-7.2e-07	0.0015	0.0041	-0.028	-0.072	-0.0061	-0.14	-0.14	-0.63	0.74	9.7e-11
Salary	0.44	0.69	3.6e-06	4.1e-07	-1.8e-07	1.8e-07	1.1e-07	-3.1e-08	-9.6e-08	-3.3e-07	2.5e-07	0.58
Wife Salary	0.37	-0.73	-1.4e-06	-3.9e-08	8.7e-08	-1.4e-07	-4.1e-08	1.3e-07	1.9e-07	4.4e-07	-3.7e-07	0.58
Total Salary	0.82	-0.042	2.2e-06	3.7e-07	-9.6e-08	3.7e-08	7.3e-08	1e-07	9.5e-08	1.1e-07	-1.2e-07	-0.58
Make	-1.3e-07	1.3e-07	-0.12	0.99	0.044	-0.014	-0.0021	0.0093	0.019	0.00076	0.0011	-2.9e-12

From the above heatmap it can be inferred that for the first principal component Salary, the Wife's Salary, and Total Salary are the original variables that are highly correlated with it, and similarly for other principal components as well.