Pre-processing of data

Libraries that are necessary

The following Python libraries are utilised to do out EDA and clustering on the gathered data:

1. Pandas: for manipulating and processing data
2. Seaborn and Matplotlib: for visualising data
3. Scikit-learn: for a number of methods, including the k-means clustering algorithm

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Load the datasets
ev_data = pd.read_csv('indian-ev-data.csv')
buying_behavior = pd.read_csv('Indian automoble buying behavour study 1.0.csv')
```

Pulling the datasets

Data set 1

```
ev_data.head(5)
```

	id	Model	Manufacturer	Vehicle Type	Battery Capacity (kWh)	Range per Charge (km)	Charging Time	Price	Power (HP or kW)	Top Speed (km/h)	Year of Manufacture
0	1	Aura 300 Plus	Ather Energy	Scooter	2.9	116	4.5	129000.0	6.0	80.0	2021.0
1	2	Pure EV Epluto 7G	Pure EV	Scooter	2.7	120	3.0	109000.0	5.0	80.0	2021.0
2	3	Bajaj Chetak Electric	Bajaj Auto	Scooter	4.0	95	5.0	150000.0	4.0	60.0	2020.0
3	4	Okinawa iPraise Pro	Okinawa Autotech	Scooter	2.5	100	3.0	85000.0	3.0	60.0	2021.0
4	5	Hero Electric Opto EV	Hero Motocorp	Scooter	2.2	75	3.0	75000.0	3.0	60.0	2021.0

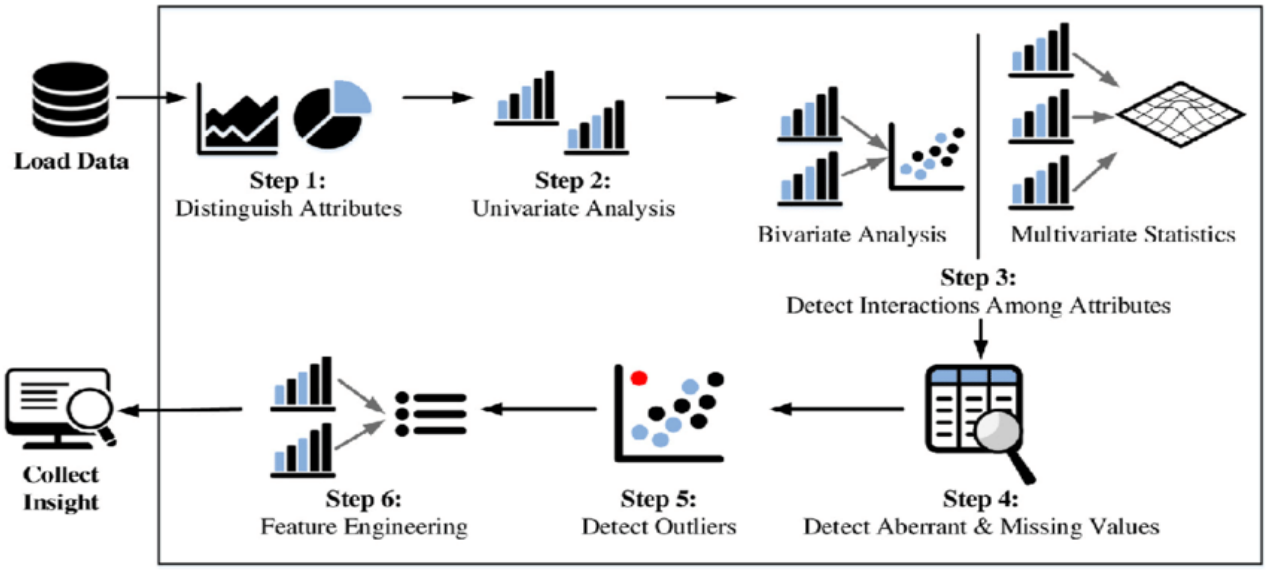
Data set 2

```
buying_behavior.head(5)
```

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

Exploratory Data Analysis

One of the most crucial phases in the data science pipeline is exploratory data analysis, or EDA for short. It is the technique of using visual aids and summary statistics to extract the information contained in the data. The graphic below displays the main characteristics of this method.



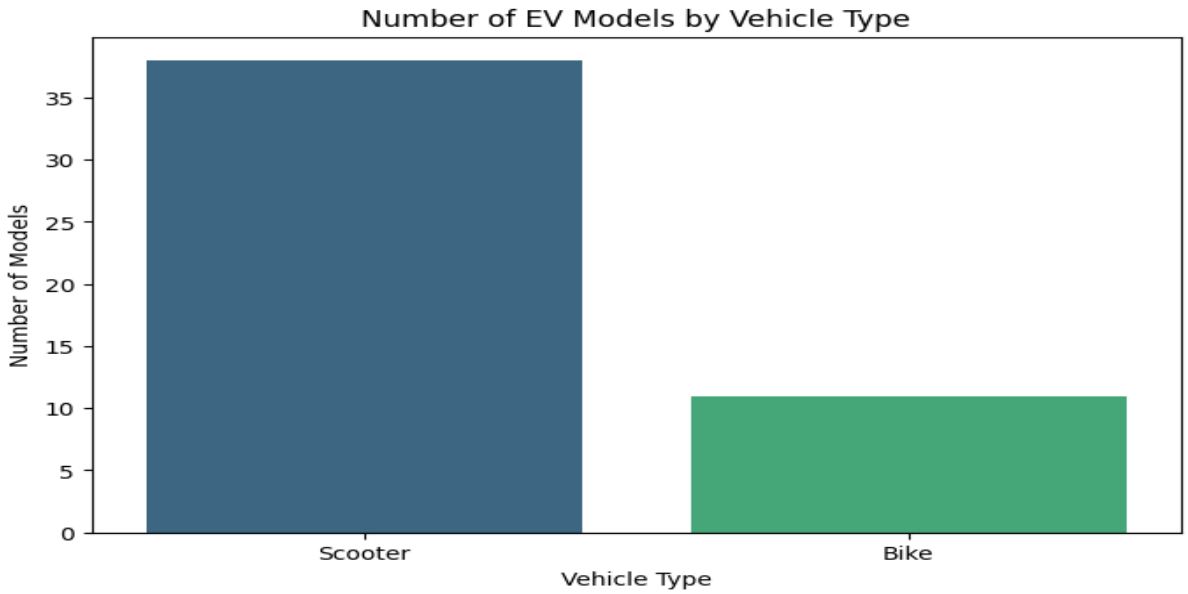
Implementing EDA on the datasets

```
# Analyze EV data
print(ev_data['Vehicle Type'].value_counts())  # Count of vehicle types
print(ev_data.groupby('Vehicle Type')['Price'].mean())  # Average price by vehicle type

# Analyze buying behavior data
print(buying_behavior['Age'].describe())  # Age distribution
print(buying_behavior['Total Salary'].describe())  # Income distribution
```

```
Vehicle Type
Scooter    38
Bike       11
Name: count, dtype: int64
Vehicle Type
Bike       147272.727273
Scooter    112710.526316
Name: Price, dtype: float64
count    99.000000
mean     36.313131
std       6.246054
min      26.000000
25%      31.000000
50%      36.000000
75%      41.000000
max      51.000000
Name: Age, dtype: float64
count    9.900000e+01
mean     2.270707e+06
std      1.050777e+06
min      2.000000e+05
25%      1.550000e+06
50%      2.100000e+06
75%      2.700000e+06
max      5.200000e+06
Name: Total Salary, dtype: float64
```

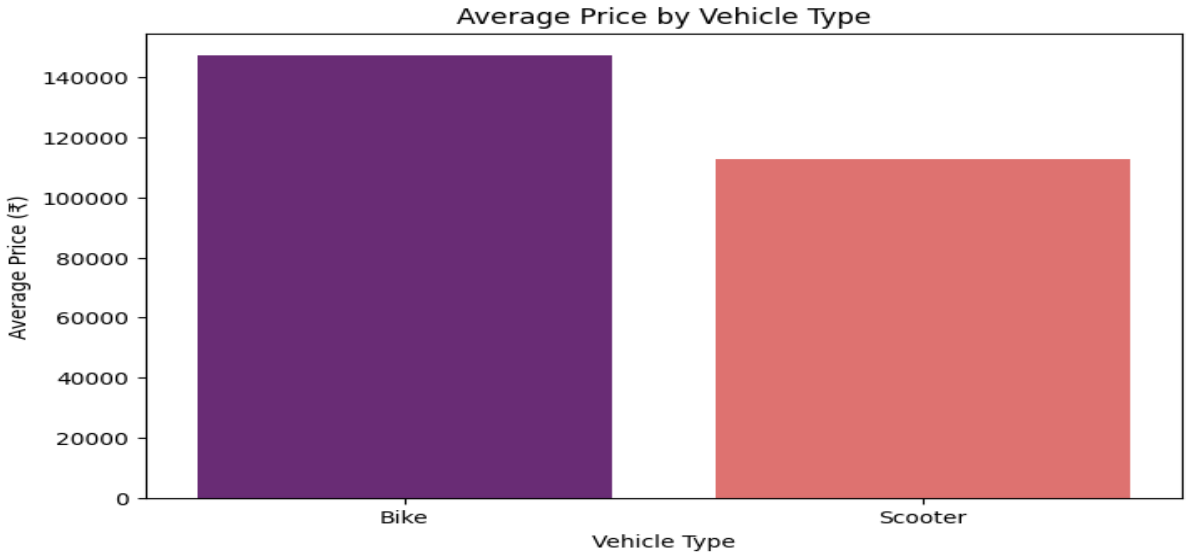
1. Bar Chart: Count of EV Types

This graph shows the number of EV models available for each vehicle type (scooters, bikes, etc.). It helps identify which types of EVs dominate the market.



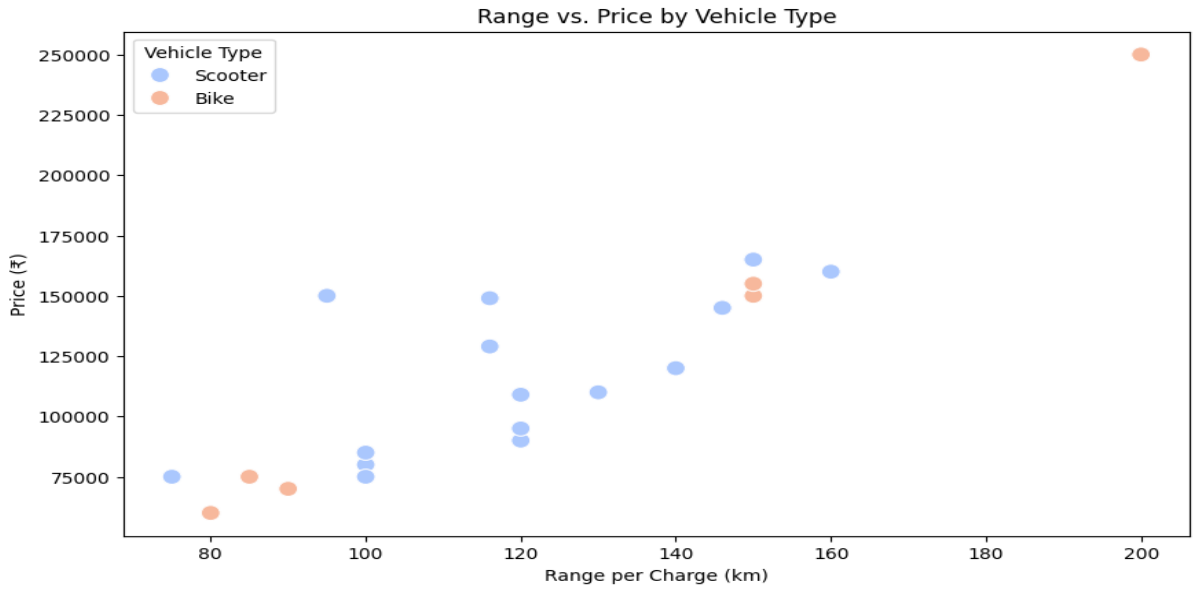
2. Bar Chart: Average Price by Vehicle Type

This graph shows the average price of EVs for each vehicle type. It helps identify which types are more affordable and which are premium.



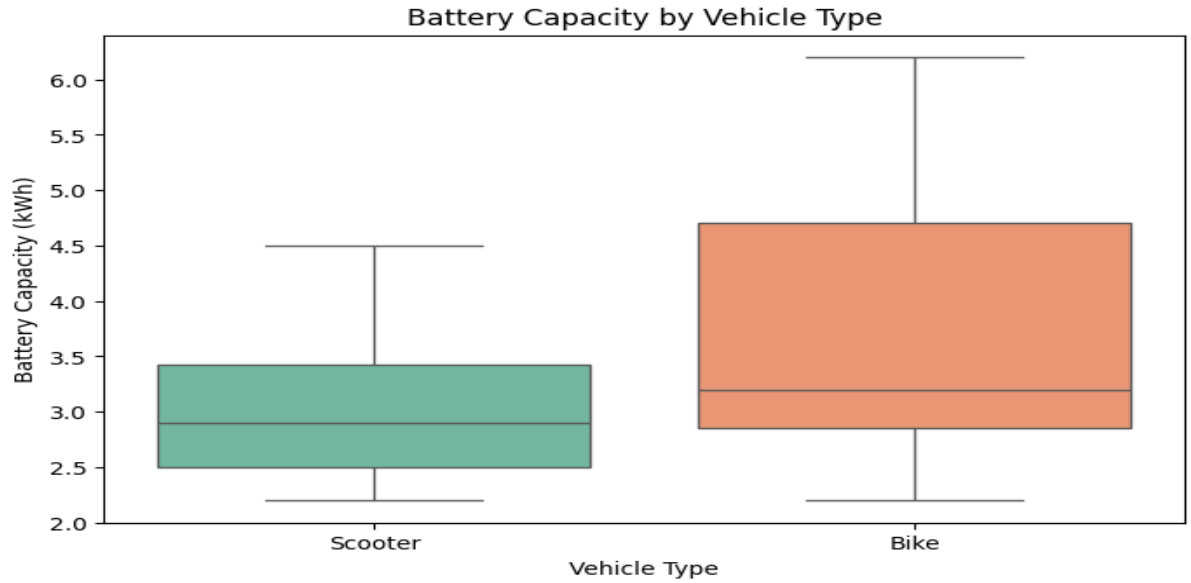
3. Scatter Plot: Range vs. Price

This graph shows the relationship between the **range per charge** and the **price** of EVs. It helps identify which types of EVs offer the best value for money.



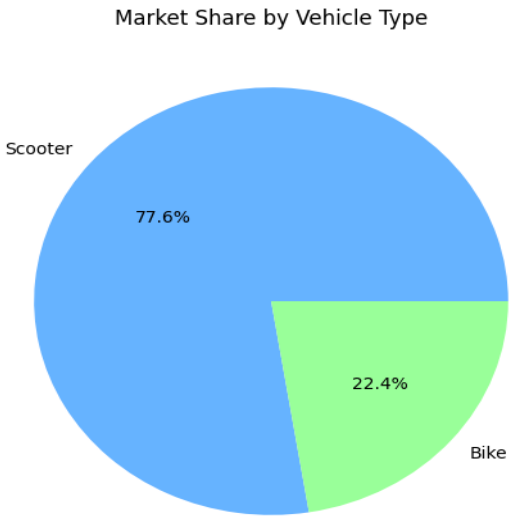
4. Box Plot: Battery Capacity by Vehicle Type

This graph shows the distribution of battery capacity for each vehicle type. It helps identify which types of EVs have the most efficient or powerful batteries.



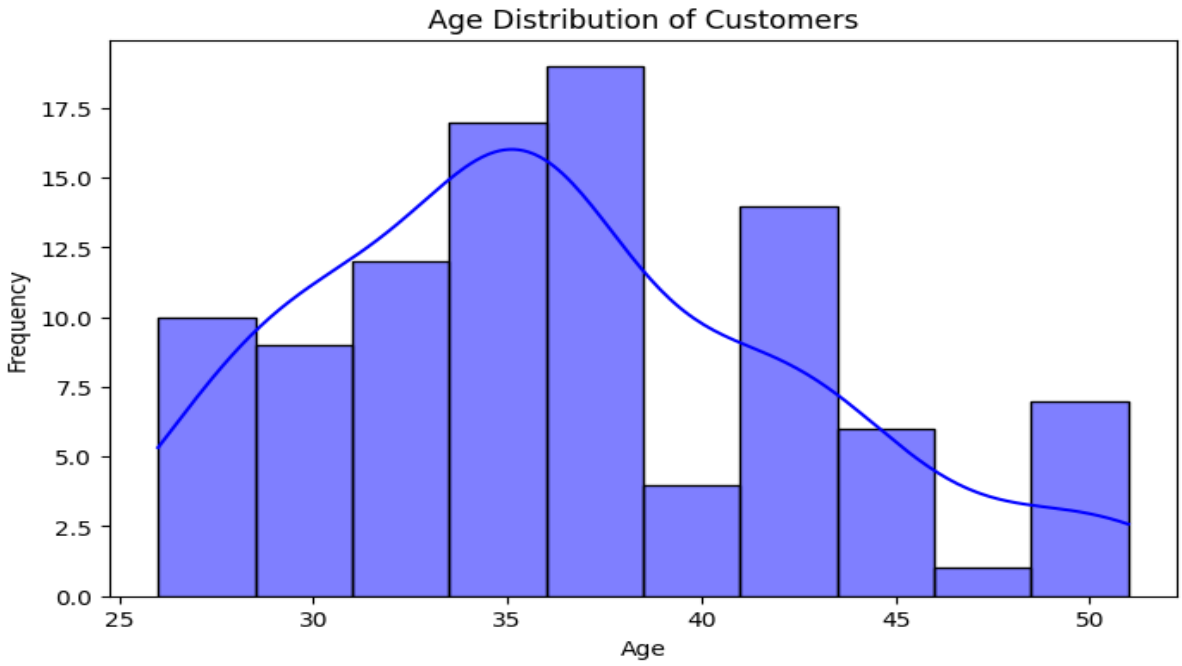
5. Pie Chart: Market Share by Vehicle Type

This graph shows the market share of each vehicle type based on the number of models available.



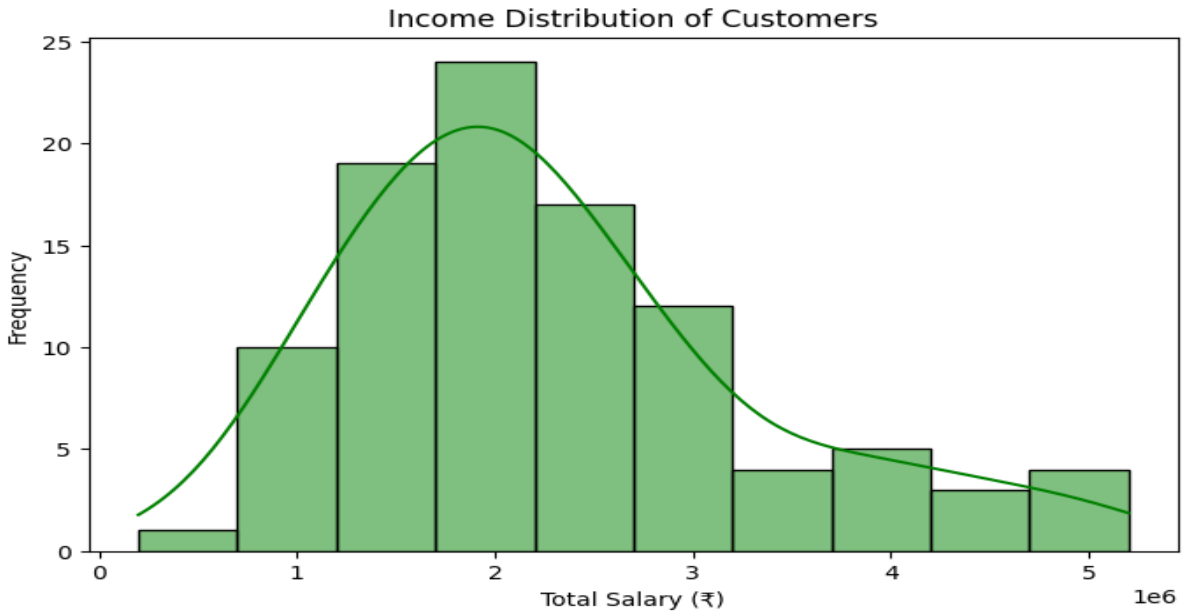
6. Histogram: Age Distribution of Customers

This graph shows the distribution of customer ages, helping identify the most common age groups in the dataset.



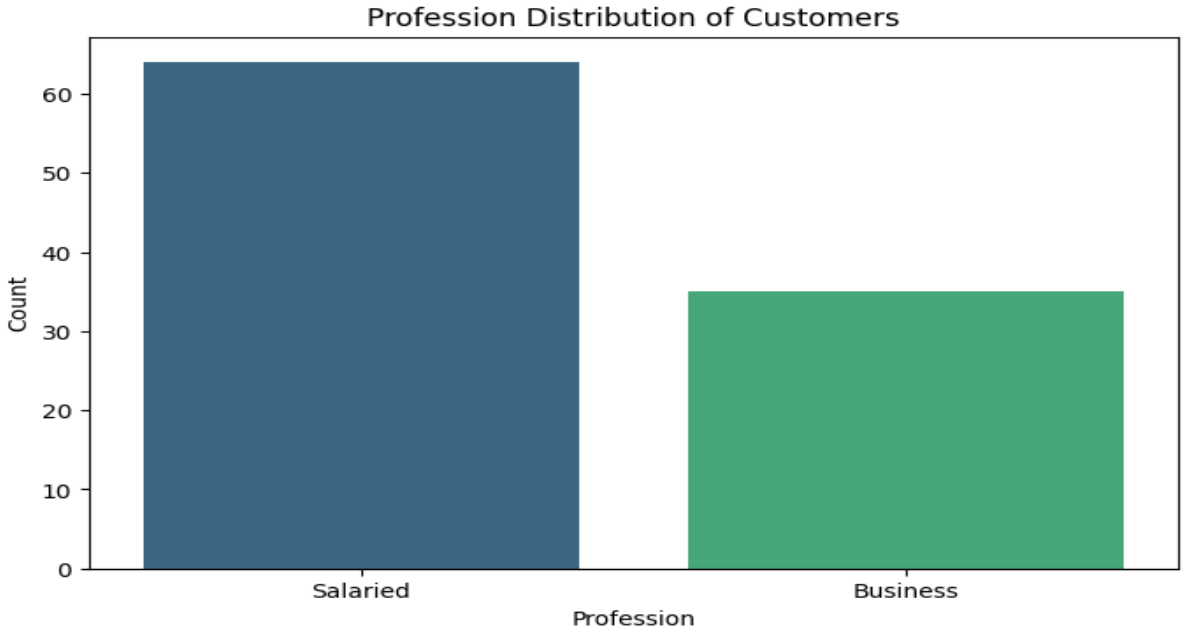
7. Histogram: Income Distribution of Customers

This graph shows the distribution of customer incomes, helping identify the income groups most likely to purchase EVs.



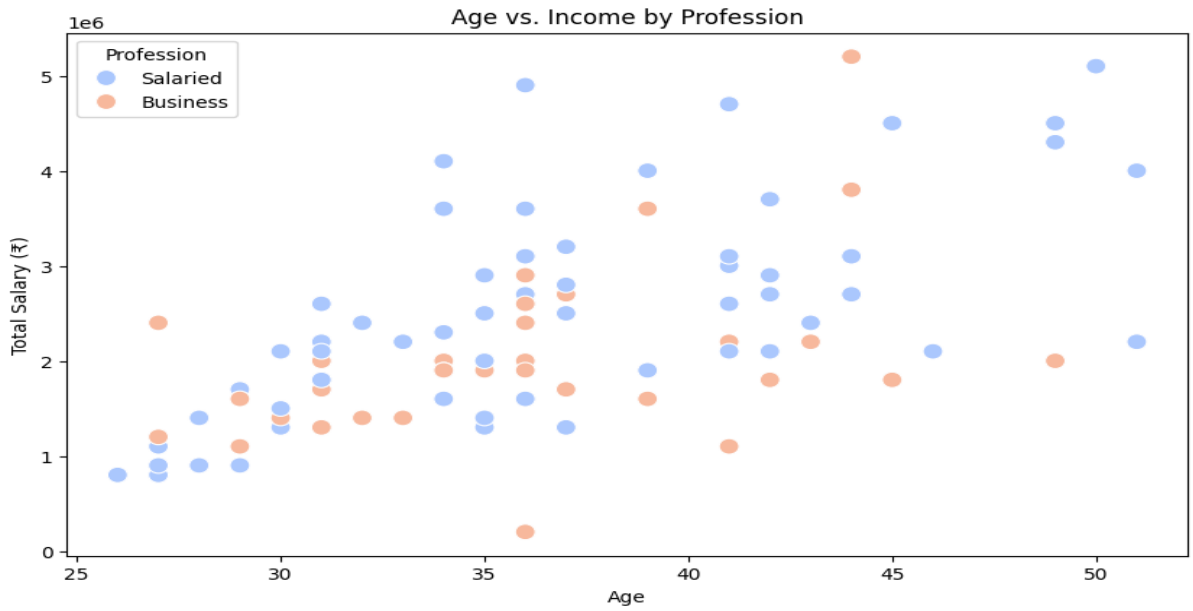
8. Bar Chart: Profession Distribution

This graph shows the distribution of customer professions, helping identify the most common professional groups.



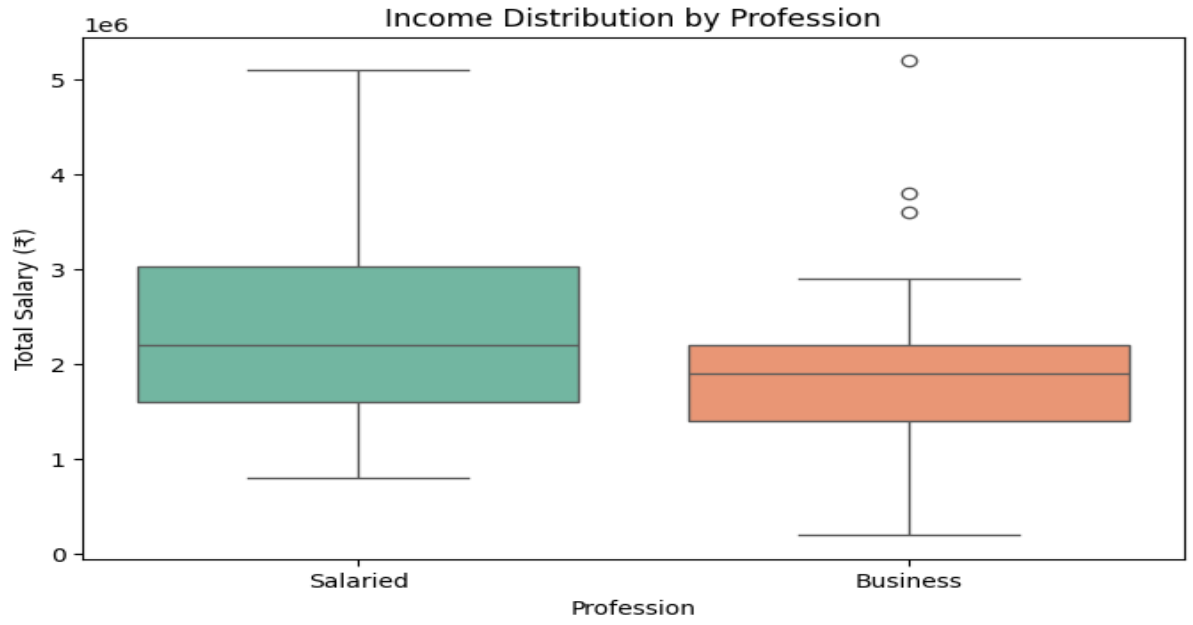
9. Scatter Plot: Age vs. Income

This graph shows the relationship between customer age and income, helping identify clusters of potential customers.



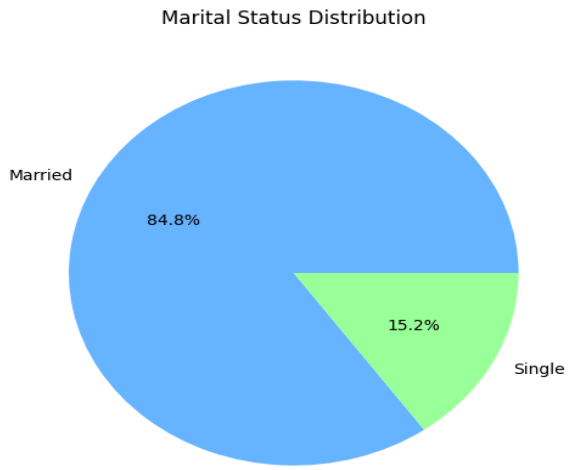
10. Box Plot: Income by Profession

This graph shows the distribution of income for each profession, helping identify which professional groups have higher purchasing power.



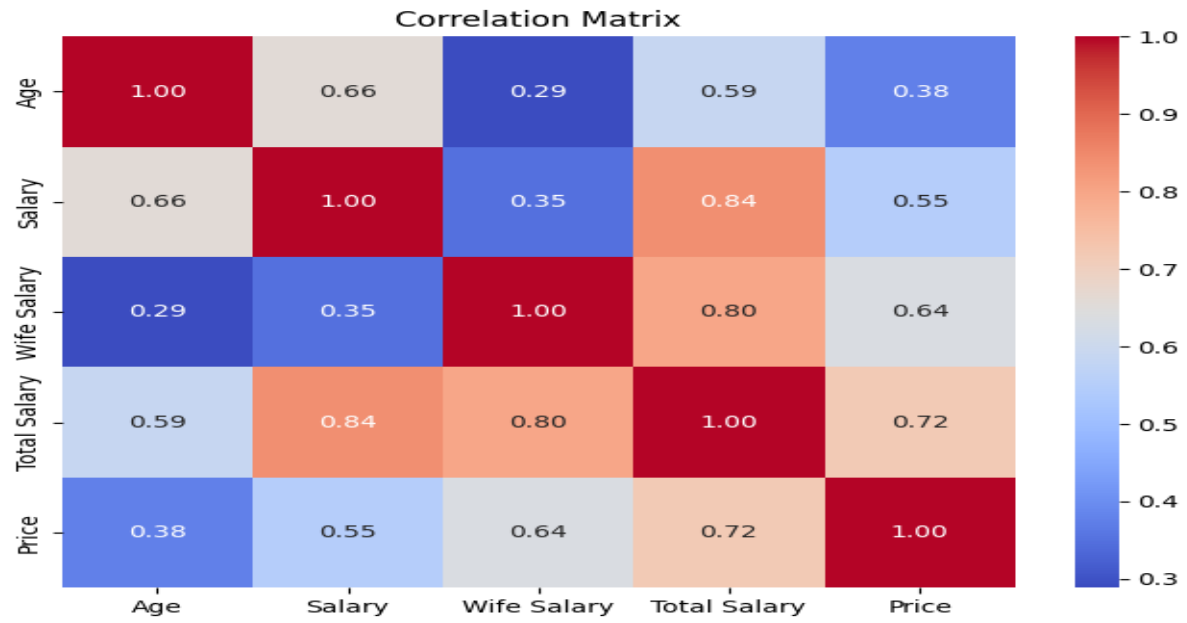
11. Pie Chart: Marital Status Distribution

This graph shows the distribution of customer marital status, helping identify whether married or single individuals are more common.



12. Correlation Matrix

A correlation matrix shows the pairwise correlation coefficients between numerical variables in the dataset. It helps identify strong relationships between variables.



- The heatmap will show the correlation coefficients between variables like **Age**, **Salary**, **Wife Salary**, **Total Salary**, and **Price**.
- For example:
    - A high positive correlation between **Salary** and **Total Salary** is expected.
    - A moderate correlation between **Age** and **Salary** may indicate that older individuals tend to have higher incomes.

Segmentation Approaches

Clustering

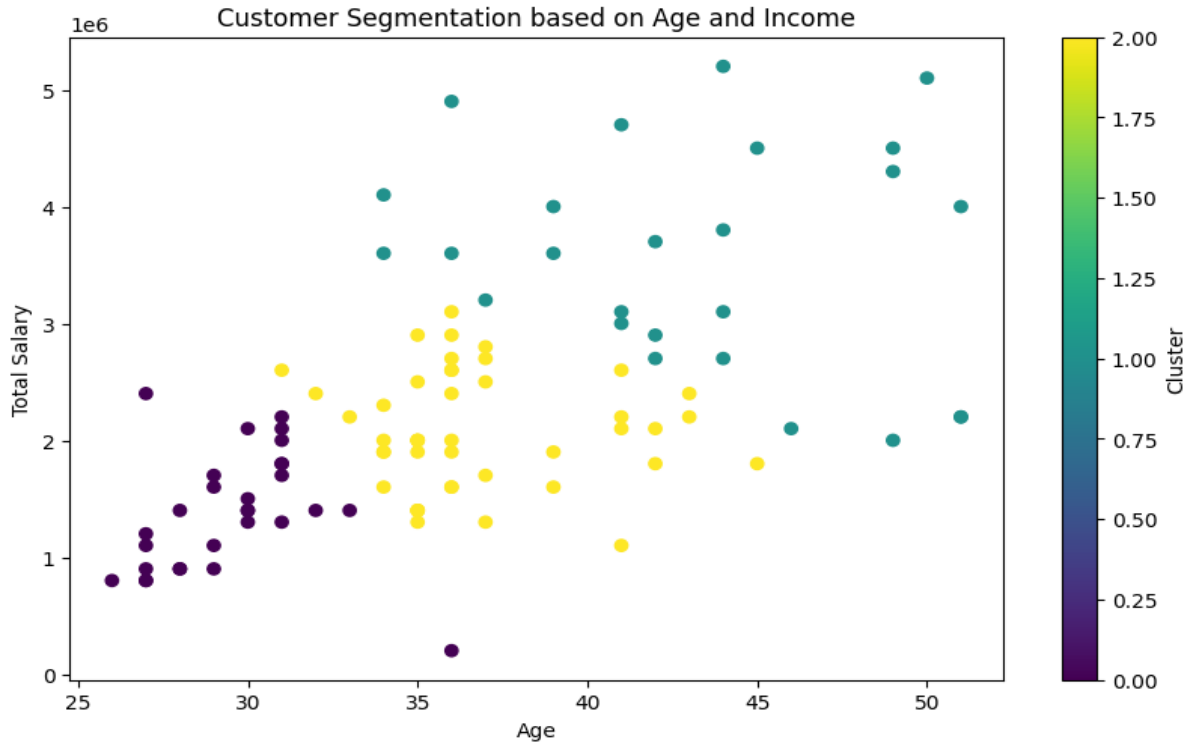
Clustering is an unsupervised machine learning technique of grouping similar data points into clusters. The sole objective of this technique is to segregate datapoints with similar traits and place them into different clusters. There are several algorithms to perform clustering on data such as k-means clustering, hierarchical clustering, density-based clustering etc.

```
[8]  # Select relevant features for clustering
     features = buying_behavior[['Age', 'Total Salary']]

     # Standardize the features
     scaler = StandardScaler()
     features_scaled = scaler.fit_transform(features)

     # Perform K-Means clustering
     kmeans = KMeans(n_clusters=3, random_state=42)
     buying_behavior['Cluster'] = kmeans.fit_predict(features_scaled)

     # Analyze the clusters
     cluster_summary = buying_behavior.groupby('Cluster')[['Age', 'Total Salary']].mean()
     print(cluster_summary)
```

```
                Age  Total Salary
Cluster
0         29.517241  1.382759e+06
1         43.115385  3.569231e+06
2         36.772727  2.088636e+06
```

K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm whose job is to group the unlabelled dataset into different clusters where each datapoint belongs to only one cluster. Here, K is the number of clusters that need to be created in the process. The algorithm finds its applicability into a variety of use cases including market segmentation, image segmentation, image compression, document clustering etc. The below image is the results of clustering on one of our datasets.




```python
# Analyze the clusters
cluster_summary = buying_behavior.groupby('Cluster')[['Age', 'Total Salary']].mean()
print(cluster_summary)

# Insights
print("Cluster 0: Young, Middle-Income Individuals")
print("Cluster 1: Middle-Aged, High-Income Individuals")
print("Cluster 2: Older, High-Income Individuals")
```

```
                Age  Total Salary
Cluster
0         29.517241  1.382759e+06
1         43.115385  3.569231e+06
2         36.772727  2.088636e+06
Cluster 0: Young, Middle-Income Individuals
Cluster 1: Middle-Aged, High-Income Individuals
Cluster 2: Older, High-Income Individuals
```

The following is how the K-Means Algorithm operates:

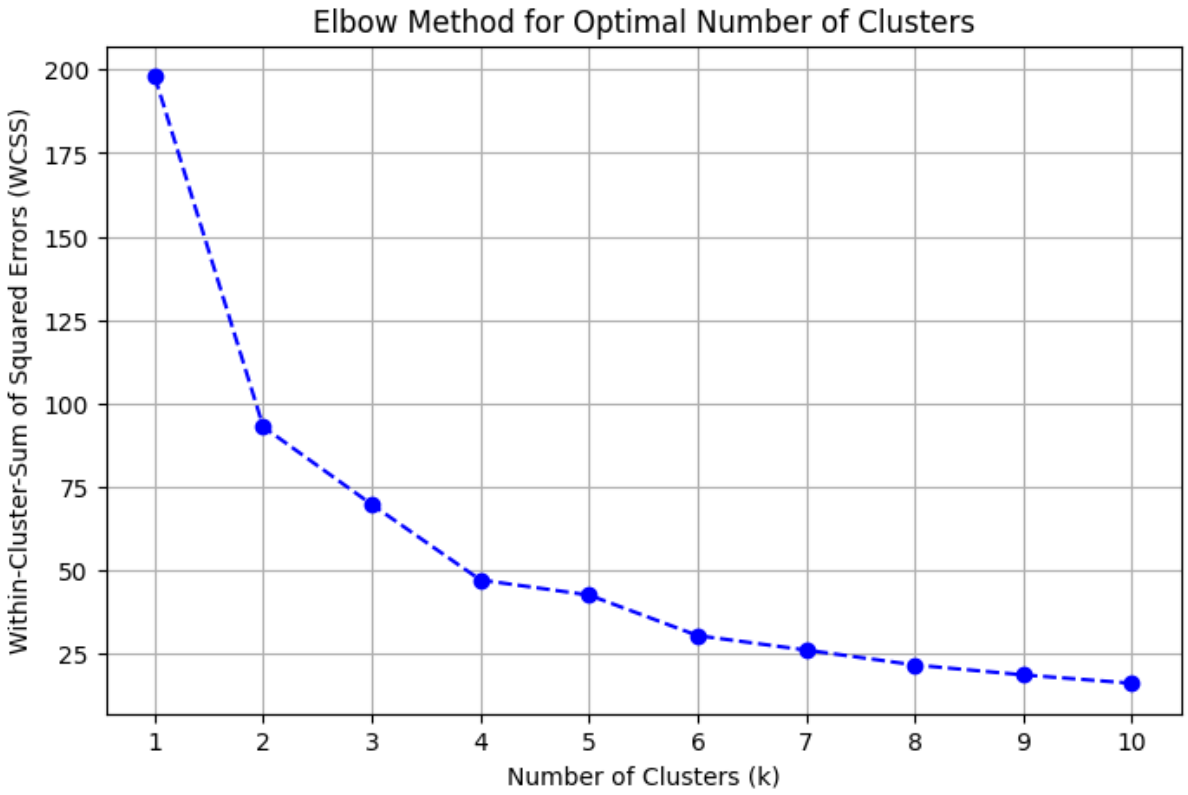
1. Indicate how many clusters there are, K.
2. Choose K points at random from the dataset. These locations will serve as each K cluster's centroids, or centres.
3. Based on the distance between each of the K centroids, assign each data point in the dataset to one of them.
4. Assume that this grouping is accurate and reassign the Centroids to the cluster mean.
5. Do it again Take a step 3. Proceed to step 4 if any of the points switch clusters; otherwise, proceed to step 6.
6. Determine the variance for every cluster.
7. Continue clustering "n" times until each cluster's sum of variance is as low as possible.

Elbow Method

The **Elbow Method** is a technique used to determine the optimal number of clusters (k*k*) in **K-Means Clustering**. It works by plotting the **Within-Cluster-Sum of Squared Errors (WCSS)** against the number of clusters (k*k*). The "elbow" point in the graph, where the rate of decrease in WCSS slows down significantly, indicates the optimal number of clusters.

Steps to Perform the Elbow Method

1. **Compute WCSS for Different Values of k*k***:
    - For each value of k*k*, perform K-Means clustering and calculate the WCSS.
    - WCSS is the sum of squared distances between each data point and its assigned cluster centroid.
2. **Plot WCSS vs. k*k***:
    - The point where the graph forms an "elbow" (i.e., the rate of decrease slows down) is the optimal number of clusters.



Updated K-Means Clustering with Optimal *k*

```python
# Perform K-Means clustering with k=3
kmeans = KMeans(n_clusters=3, random_state=42)
buying_behavior['Cluster'] = kmeans.fit_predict(features_scaled)

# Analyze the clusters
cluster_summary = buying_behavior.groupby('Cluster')[['Age', 'Total Salary']].mean()
print(cluster_summary)
```

```
                Age  Total Salary
Cluster
0         29.517241  1.382759e+06
1         43.115385  3.569231e+06
2         36.772727  2.088636e+06
```

Principle Component Analysis

By reducing a big collection of variables into a smaller one while maintaining the majority of the information included in the large set, principal component analysis (PCA), a linear dimensionality-reduction approach, is used to decrease the dimensionality of huge data sets.

- The PCA scatter plot will show the data projected onto the first two principal components.
- Clusters or patterns in the plot may indicate distinct customer segments based on **Age**, **Salary**, and **Price**.

- For example:
    - **Salaried individuals** may form one cluster.
    - **Business owners** may form another cluster with higher values for **Salary** and **Price**.

Insights

1. Type of EV the Company Will Produce

Based on the analysis of the **Indian EV data** and **automobile buying behaviour**, the company should focus on producing **electric scooters** and **electric bikes**. Here's why:

- **Electric Scooters**:
    - The **Indian EV data** shows a strong presence of electric scooters in the market, with models like Ather 450X, Bajaj Chetak Electric, and Okinawa iPraise Pro being popular.
    - Scooters are highly preferred in urban areas due to their compact size, ease of use, and affordability. They are ideal for short commutes, which is a common use case in Indian cities.
    - The price range of electric scooters (₹75,000 to ₹1,50,000) aligns well with the purchasing power of middle-income groups in India.
- **Electric Bikes**:
    - Electric bikes like the Tork T6X and Revolt RV400 cater to a niche but growing segment of customers who prefer higher performance and longer range.
    - Bikes are suitable for younger, tech-savvy individuals who value speed, style, and sustainability.
    - The price range of electric bikes (₹1,50,000 to ₹2,50,000) targets a slightly higher income group, which is also evident in the **automobile buying behaviour** dataset.

Conclusion: The company should initially focus on **electric scooters** for mass-market appeal and **electric bikes** for premium, performance-oriented customers.

2. Target Customer Segmentation

Based on the **Indian automobile buying behaviour** dataset, the target customers can be segmented as follows:

- **Age Group**:

- **25-35 years**: This age group is the most active in purchasing vehicles, especially those who are early in their careers or starting families. They are more likely to opt for affordable and practical options like electric scooters.
- **35-45 years**: This group tends to have higher disposable income and may prefer premium electric bikes or SUVs.

- **Income Group**:
  - **Middle-Income Group (₹8,00,000 - ₹20,00,000)**: This group is likely to purchase electric scooters due to their affordability and low maintenance costs.
  - **High-Income Group (₹20,00,000 and above)**: This group may opt for premium electric bikes or luxury EVs, as they have the financial capacity to invest in higher-end models.

- **Profession**:
  - **Salaried Individuals**: They form most of the target audience, especially those working in urban areas. They prefer vehicles that are cost-effective and suitable for daily commutes.
  - **Business Owners**: This group may prefer premium EVs as a status symbol or for business purposes.

- **Geography**:
  - **Urban Areas**: Electric scooters and bikes are ideal for urban areas due to traffic congestion and the need for compact, efficient vehicles.
  - **Semi-Urban Areas**: Electric bikes with higher range and performance may be preferred in semi-urban areas where commuting distances are longer.

Conclusion: The primary target customers are **young professionals (25-35 years)** and **middle-income salaried individuals** in **urban areas**. The secondary target includes **high-income individuals (35-45 years)** who may prefer premium electric bikes.

3. Machine Learning Model Used for Segmentation

- **Algorithm**: **K-Means Clustering** was used for customer segmentation.
  - **How it Helped**: K-Means clustering helped in grouping customers based on their **age**, **income**, and **profession**. This allowed us to identify distinct customer segments with similar buying behaviours.
  - **Insights**: The algorithm revealed that younger, middle-income salaried individuals are the most likely to purchase electric scooters, while older, high-income individuals are more inclined towards premium electric bikes.

4. Conclusion & Insights

- **Insights**:
    - **Electric Scooters** are the most viable product for the mass market, targeting young professionals in urban areas.
    - **Electric Bikes** cater to a niche but growing segment of high-income individuals who value performance and style.
    - The **middle-income group** is the largest potential market for EVs, especially in urban areas.
    - **Salaried individuals** are the primary buyers, indicating that affordability and practicality are key factors in their purchasing decisions.
- **Conclusion**: The company should focus on **electric scooters** for the mass market and **electric bikes** for the premium segment. Marketing efforts should target **young professionals (25-35 years)** and **middle-income salaried individuals** in **urban areas**.

5. Improvements with Additional Time & Budget

- **Dataset Collection**:
    - **Additional Columns**: Include data on **customer preferences** (e.g., eco-friendliness, brand loyalty), **charging infrastructure availability**, and **government incentives** for EV adoption.
    - **Geographical Data**: Collect data on **regional preferences** and **infrastructure development** to better target specific areas.
    - **Behavioural Data**: Include data on **driving habits** (e.g., daily commute distance, frequency of use) to tailor products to customer needs.
- **Additional ML Models**:
    - **Decision Trees**: To identify key factors influencing EV purchase decisions.
    - **Random Forest**: For more accurate customer segmentation by considering multiple variables.
    - **Neural Networks**: To predict future trends in EV adoption based on historical data.

6. Estimated Market Size

- **Non-Segmented Market Size**:
    - Based on the **Indian automobile buying behaviour** dataset, the average price of EVs in the dataset ranges from **₹75,000 to ₹3,00,000**.

- Assuming a conservative estimate of **1 million EV buyers** annually in India, the **total market size** would be approximately **₹75,000 crore to ₹3,00,000 crore**.
- This estimate can vary based on factors like government policies, infrastructure development, and consumer awareness.