Market Segmentation Analysis of Electric Vehicles Market in India

Problem Statement:

The Indian electric vehicle (EV) market is rapidly evolving, driven by various factors such as consumer demographics, behaviours, and preferences. This project aims to enhance the existing Market Segmentation Analysis of the Electric Vehicles Market in India by addressing key challenges with additional time and budget. Firstly, the objective is to improve data collection by acquiring additional data points such as charging infrastructure availability, government incentives, regional climate, and EV model-specific features. Additionally, consumer behaviour data, including driving patterns, charging habits, and maintenance behaviour, will be collected. The competitive landscape will also be explored by including data on market shares, competition among EV manufacturers, and customer reviews. Secondly, advanced machine learning models will be employed, including clustering algorithms like K-Means, DBSCAN, or hierarchical clustering, to refine market segmentation based on the expanded dataset. Predictive modelling will help anticipate future market trends and consumer preferences, and recommendation systems will suggest suitable EV models based on consumer profiles. Thirdly, this project will estimate the market size for the overall Indian electric vehicle market using statistical techniques and data extrapolation, considering current and projected trends. Finally, variable identification techniques, such as Recursive Feature Elimination (RFE) or feature importance analysis, will determine the top four variables contributing significantly to market segmentation. This comprehensive approach will provide deeper insights into the electric vehicle market, empowering industry stakeholders to make informed decisions regarding marketing strategies, product development, and policy interventions, and will offer a holistic view of the market's potential. The timeline and budget for this enhanced project will be tailored to its specific requirements, ensuring a thorough and impactful analysis of the Indian electric vehicle market.

Data Collection:

Surveys and interviews will be used to gather data, which will include demographic and behavioural characteristics including age, career, marital status, education, pay, and more as found in the Kaggle-sourced CSV dataset. Market reports will be used to supplement the data. Strategies for sampling will be used to ensure representativeness. Privacy will be protected by ethical considerations, and the Kaggle-sourced CSV dataset will be cleaned and organised for analysis as part of data preprocessing. Detailed documentation will be maintained for transparency and reproducibility.

Data Preprocessing:

To prepare your data for a Market Segmentation Project, you can follow these data preprocessing steps:

- **Data Loading:** Import the dataset into a data analysis tool like Python with pandas.

```
In [4]: df.head()
```

	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

Columns Explanation:

1. Age: This column represents the age of individuals in the dataset. Age is often used as a demographic variable to understand how different age groups interact with products, services, or markets.

2. Profession: This column indicates the profession or occupation of each individual. Profession is a key variable for market segmentation as it can help identify specific occupational groups that may have distinct preferences or needs related to electric vehicles.

3. Marital Status: This column shows whether individuals are single or married. Marital status is a demographic variable that can influence purchasing decisions and family-related factors.

4.Education: This column specifies the highest level of education attained by each individual. Education is another demographic variable that can impact preferences, as more educated individuals may have different attitudes toward technology and environmental concerns.

5.No of Dependents: This column represents the number of dependents (such as children or other family members) that each individual has. The number of dependents can be an essential factor in understanding household dynamics and financial considerations when purchasing an electric vehicle.

6. Personal loan: This column indicates whether individuals have taken a personal loan (yes/no). This variable can provide insights into the financial situation of individuals and their ability to make large purchases like electric vehicles.

7. House Loan: This column shows whether individuals have a house loan (yes/no). Similar to personal loans, having a house loan can impact an individual's financial stability and, consequently, their ability to afford an electric vehicle.

8. Wife Working: This column specifies whether an individual's wife is working (yes/no). This variable relates to household income and can influence purchasing decisions if both partners contribute to the family income.

9. Salary: This column represents the salary of the individuals. Salary is a crucial financial variable as it directly impacts an individual's ability to afford an electric vehicle.

10.Wife Salary: This column indicates the salary of the individual's wife. This variable contributes to the overall household income and can influence affordability and purchasing decisions.

11.Total Salary: This column is the sum of an individual's salary and their wife's salary, providing a measure of the total household income.

12. Make: This column specifies the make or brand of the electric vehicle owned or considered by each individual. This variable is essential for understanding brand preferences and market share.

13. Price: This column indicates the price of the electric vehicle associated with each individual. Price is a critical factor in purchasing decisions, and it helps segment consumers based on their affordability and willingness to invest in an electric vehicle.

These columns represent a mix of demographic, financial, and behavioural variables, which are commonly used in market segmentation analyses to understand consumer behaviour and preferences in the context of the electric vehicle market in India.

- **Data Exploration:** Explore the dataset using functions like `head()`, `info()`, and `describe()` to understand its structure, data types, and summary statistics.

```
In [7]: df.shape

        (99, 13)

In [8]: df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 99 entries, 0 to 98
        Data columns (total 13 columns):
         #   Column            Non-Null Count  Dtype
        ---  ------            --------------  -----
         0   Age               99 non-null     int64
         1   Profession        99 non-null     object
         2   Marrital Status   99 non-null     object
         3   Education         99 non-null     object
         4   No of Dependents  99 non-null     int64
         5   Personal loan     99 non-null     object
         6   House Loan        99 non-null     object
         7   Wife Working      99 non-null     object
         8   Salary            99 non-null     int64
         9   Wife Salary       99 non-null     int64
         10  Total Salary      99 non-null     int64
         11  Make              99 non-null     object
         12  Price             99 non-null     int64
        dtypes: int64(6), object(7)
        memory usage: 10.2+ KB
```

```
In [25]: df.describe()
```

	Age	No of Dependents	Salary	Wife Salary	Total Salary	Price
count	99.000000	99.000000	9.900000e+01	9.900000e+01	9.900000e+01	9.900000e+01
mean	36.313131	2.181818	1.736364e+06	5.343434e+05	2.270707e+06	1.194040e+06
std	6.246054	1.335265	6.736217e+05	6.054450e+05	1.050777e+06	4.376955e+05
min	26.000000	0.000000	2.000000e+05	0.000000e+00	2.000000e+05	1.100000e+05
25%	31.000000	2.000000	1.300000e+06	0.000000e+00	1.550000e+06	8.000000e+05
50%	36.000000	2.000000	1.600000e+06	5.000000e+05	2.100000e+06	1.200000e+06
75%	41.000000	3.000000	2.200000e+06	9.000000e+05	2.700000e+06	1.500000e+06
max	51.000000	4.000000	3.800000e+06	2.100000e+06	5.200000e+06	3.000000e+06

- **Handling Missing Values:** Check for missing values in the dataset. If any are found, decide whether to impute them or remove rows/columns with missing data. Since your dataset is relatively small, removing rows with missing values may not be ideal.

```
In [5]: df.isnull().sum()

Age                 0
Profession          0
Marrital Status     0
Education           0
No of Dependents    0
Personal loan       0
House Loan          0
Wife Working        0
Salary              0
Wife Salary         0
Total Salary        0
Make                0
Price               0
dtype: int64
```

- **Encoding Categorical Variables:** Encode categorical variables like 'Profession,' 'Marital Status,' 'Education,' 'Personal loan,' 'House Loan,' 'Wife Working,' and 'Make' using one-hot encoding or label encoding, depending on the nature of the variable and its relationship with the target variable. For example, 'Profession' and 'Marital Status' can be one-hot encoded, while 'Education' can be label encoded.

```
In [37]: encoded_df.head()
```

	Marrital Status	Education	Personal loan	House Loan	Wife Working	Make	Age	Profession	No of Dependents	Salary	Wife Salary	Total Salary	Price
0	1	1	1	0	0	8	27	1	0	800000	0	800000	800000
1	0	1	1	1	1	1	35	1	2	1400000	600000	2000000	1000000
2	0	0	1	1	0	4	45	0	4	1800000	0	1800000	1200000
3	0	1	0	0	1	2	41	0	3	1600000	600000	2200000	1200000
4	0	1	1	0	1	6	31	1	2	1800000	800000	2600000	1600000

- **Scaling Numerical Features:** Since your dataset includes numerical features like 'Age,' 'No of Dependents,' 'Salary,' 'Wife Salary,' and 'Total Salary,' consider scaling them using standardization (StandardScaler) or Normalization (Normalizer) to ensure they are on the same scale.

- **Feature Selection:** Identify the top features that are most relevant for your market segmentation analysis. You can use techniques like correlation analysis or feature importance analysis to select the most influential variables.
- **Data Visualization:** Visualize your data to gain insights into patterns and relationships. Data visualization tools like Matplotlib or Seaborn can help you create informative plots.

  These steps preprocess the data and set the foundation for effective market segmentation analysis.

- **Exploratory Data Analysis:**

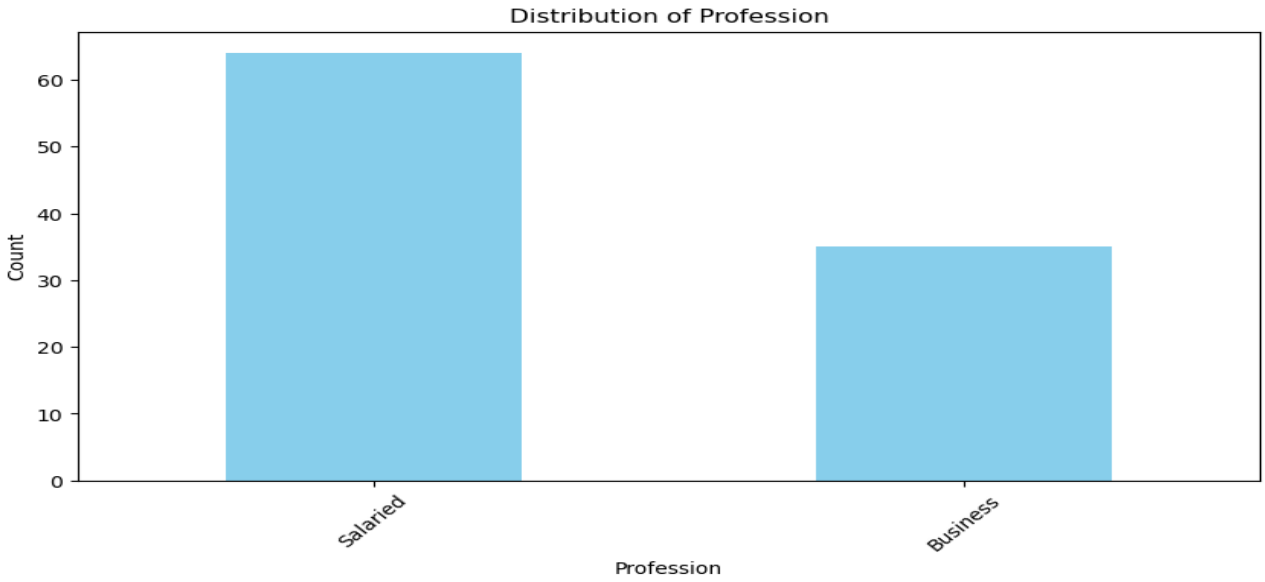
  Data cleaning plays a pivotal role in preparing the dataset for analysis. In this case, we observed that certain variables required attention. Handling missing values and addressing outliers were essential steps to ensure data integrity. Additionally, no duplicate records were identified. These data cleaning measures were taken to enhance the reliability of subsequent analyses.

  Visualizing the data offers valuable insights. Histograms and box plots provide a glimpse into the distribution of numerical variables, helping identify trends or skewness. For instance, these plots reveal variations in salary levels among individuals in the dataset. Categorical variables, like "Profession" or "Education," are explored through bar charts, illustrating the frequency of different categories. Such visualizations facilitate the identification of patterns or disparities within the data.

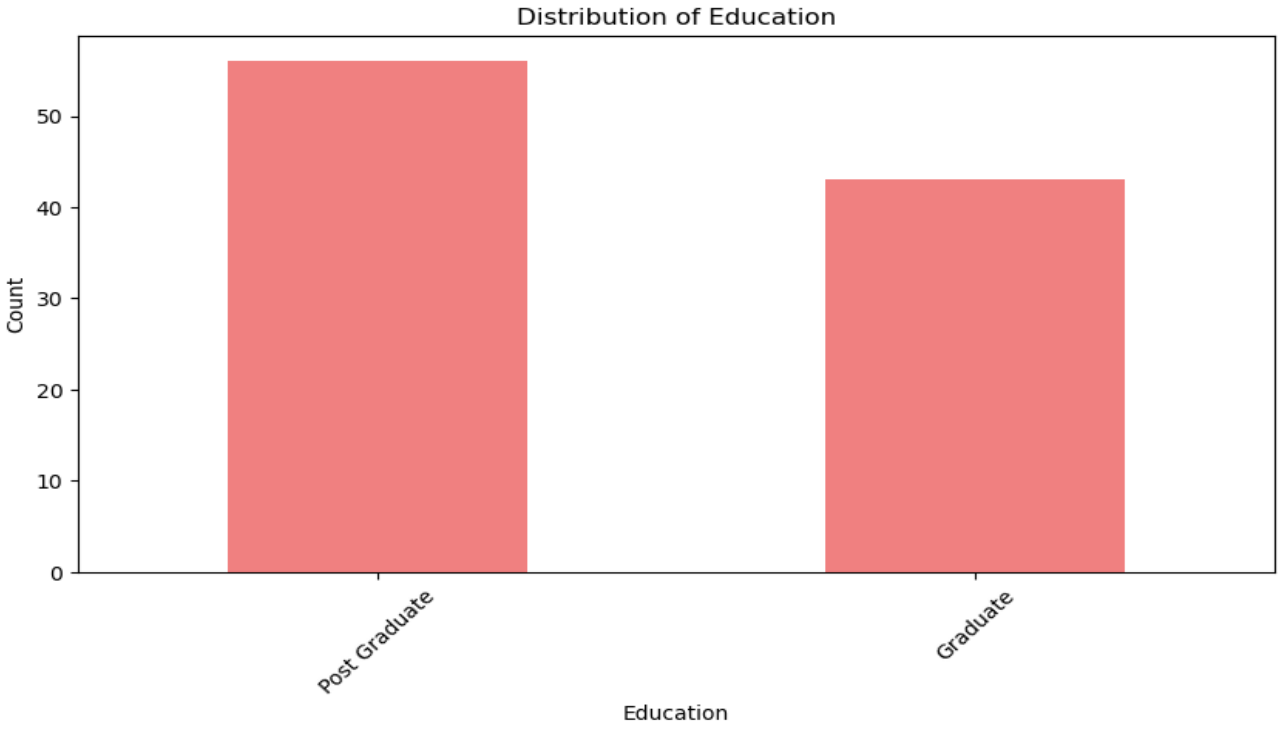
  Moving on to univariate and bivariate analyses, we delve deeper into the relationships between variables. Univariate analysis entails examining variables individually. For numerical attributes like "Age" or "Total Salary," histograms and box plots allow us to understand their distributions. Meanwhile, for categorical attributes like "Education" or "House Loan," bar charts display category proportions. Bivariate analysis explores the interplay between two variables. Scatter plots demonstrate relationships, such as the correlation between "Age" and "Salary."

  Bar plot to visualize the distribution of the "Profession"

  By illustrating how many people in the dataset are employed and how many are compensated, this figure makes it easy to determine whether certain occupations are more or less prevalent in the data.

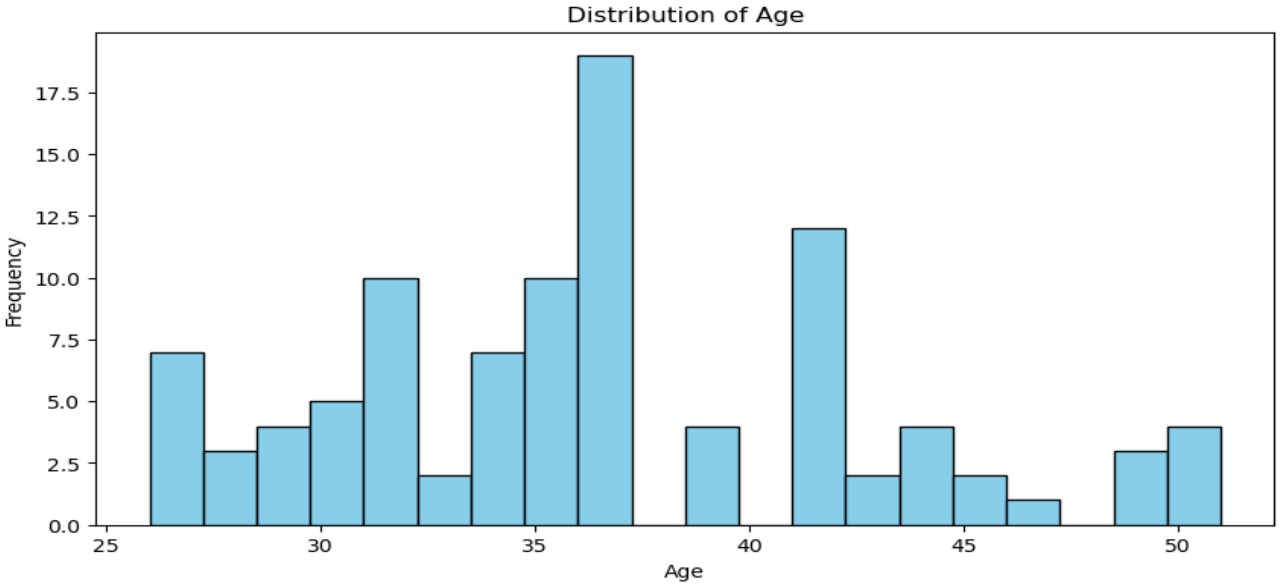


Bar plot to visualize the distribution of the "Education"



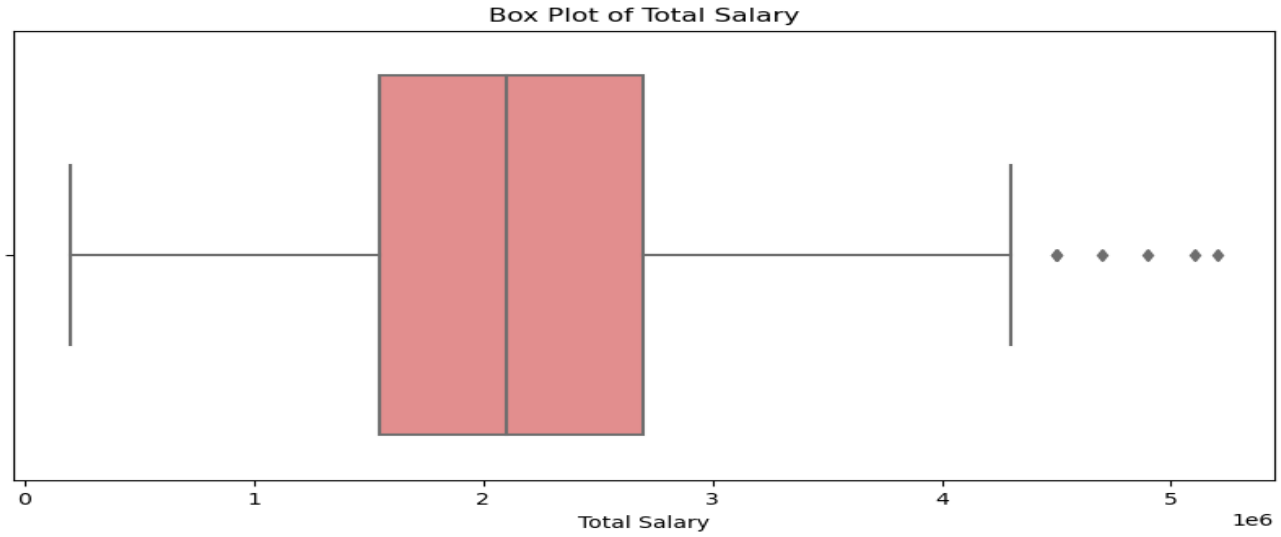
This figure visually represents the percentage of individuals in the dataset who have achieved different educational levels. It provides an easy-to-read overview of how the dataset's subjects are distributed in terms of their education, making it helpful for understanding their educational backgrounds.

Histogram to visualize the distribution of the "Age"



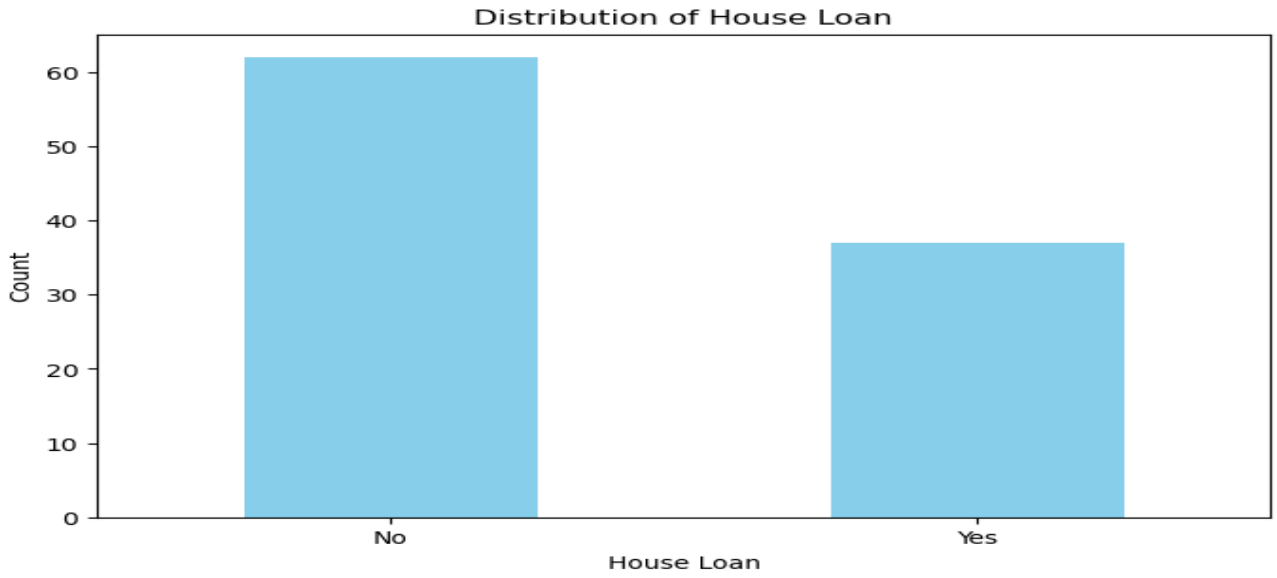
This plot generates a histogram to show how the ages of individuals in the dataset are distributed. Each bar/bin in the histogram represents a range of ages, and the height of each bar indicates how many individuals fall into that age range. This visualization helps you understand the age distribution within the dataset.

Box plot to visualize the distribution of the "Total Salary"



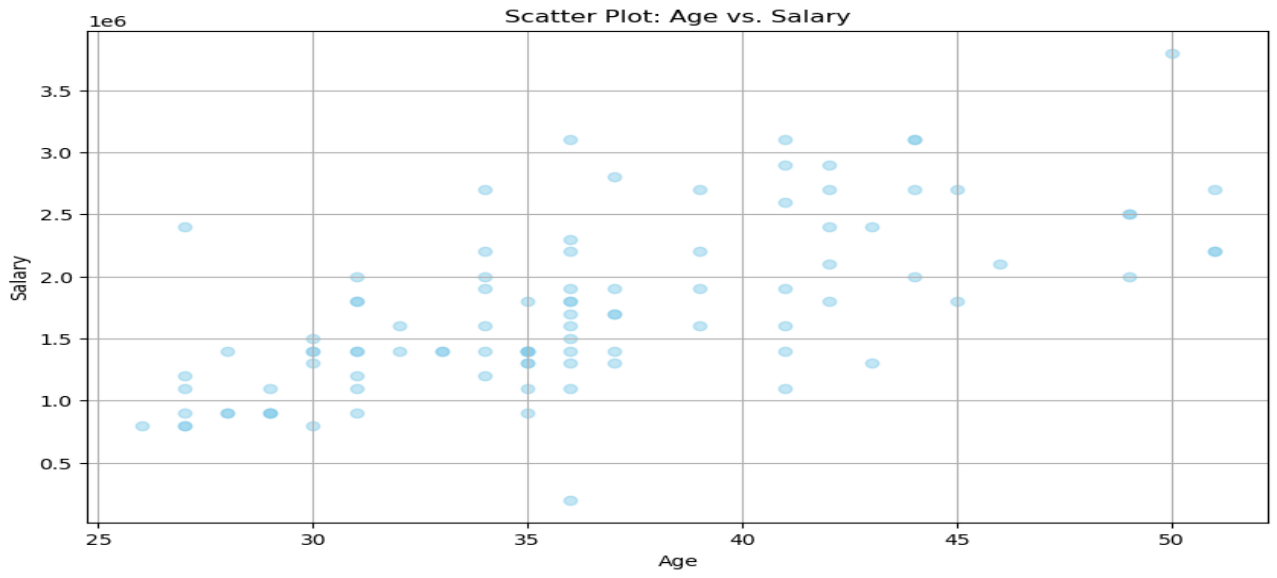
This box plot to provide insights into the distribution of total salaries within the dataset. The box plot includes key statistics such as the median, quartiles, and potential outliers, making it a useful tool for understanding the central tendency and spread of salary data and identifying any unusual observations.

Bar chart to visualize the distribution of the "House Loan"



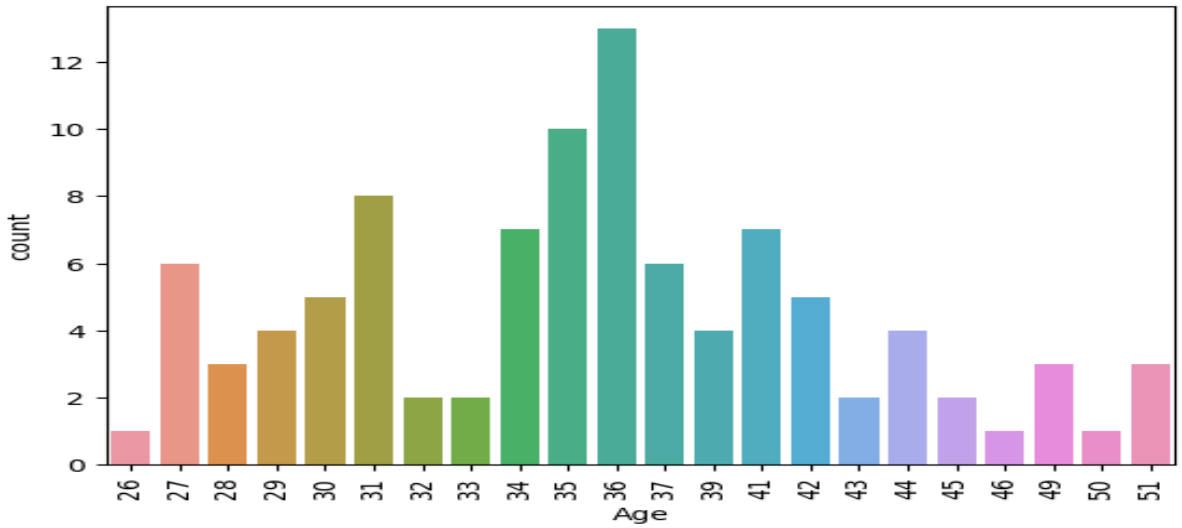
This bar chart that illustrates how house loans are distributed within the dataset. Each bar represents a different category of house loan, and the height of each bar indicates the count of individuals or cases associated with that particular type of house loan. It provides a clear visual representation of the distribution of house loan categories.

Scatter plot to visualize the relationship between the "Age" and "Salary"

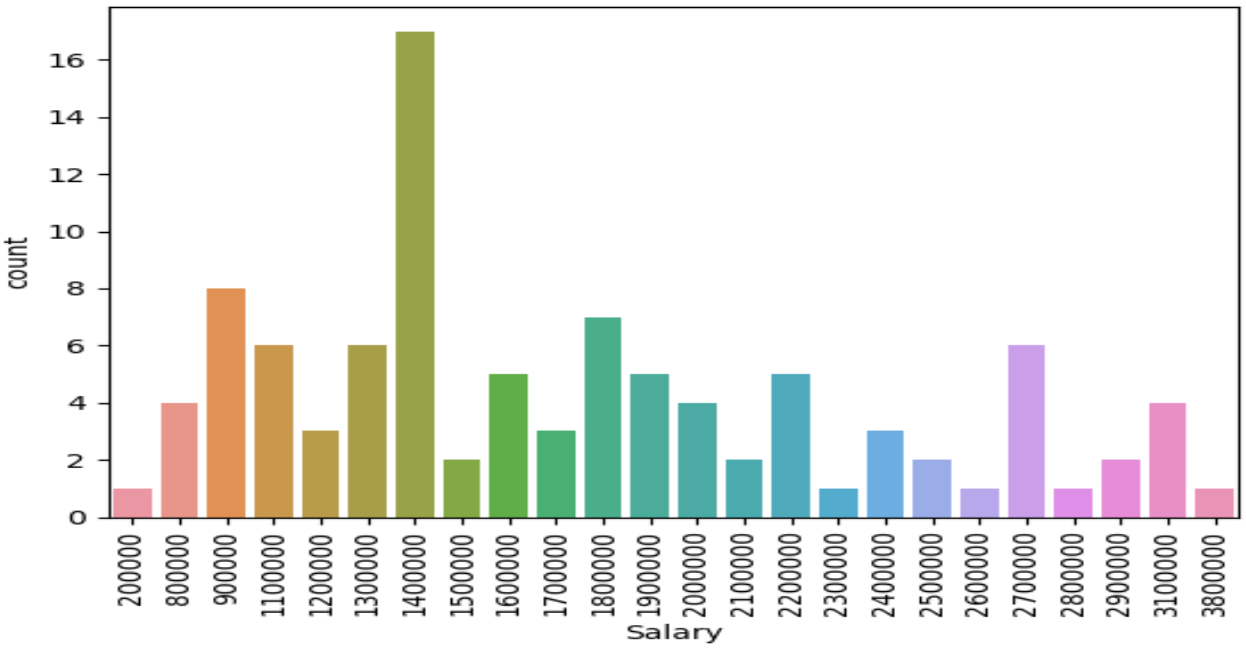


This scatter plot that helps visualize the relationship between age and salary within the dataset. Each data point on the plot represents an individual, with their age on the x-axis and their salary on the y-axis. The scatter plot is useful for identifying patterns or trends in the data, such as whether there is a correlation or relationship between age and salary. This graph demonstrates the highly favourable relationship between salary and age.

Count plot of "Age"

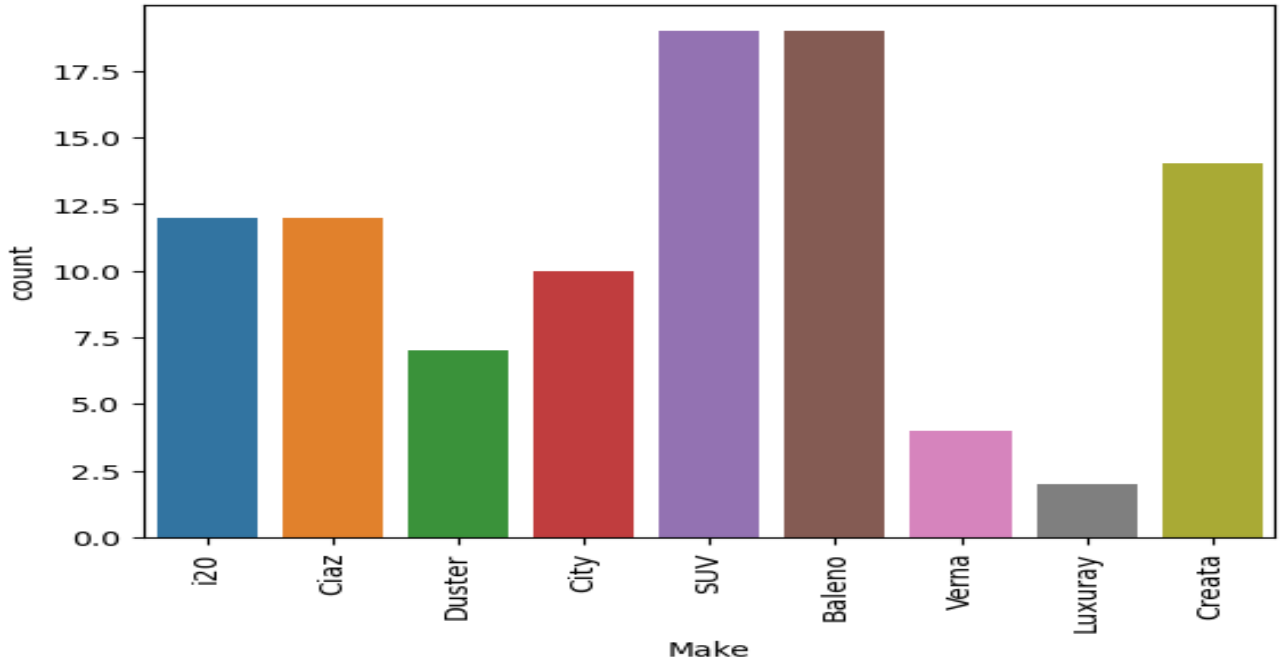


Count plot to visualize the distribution of "Salary"



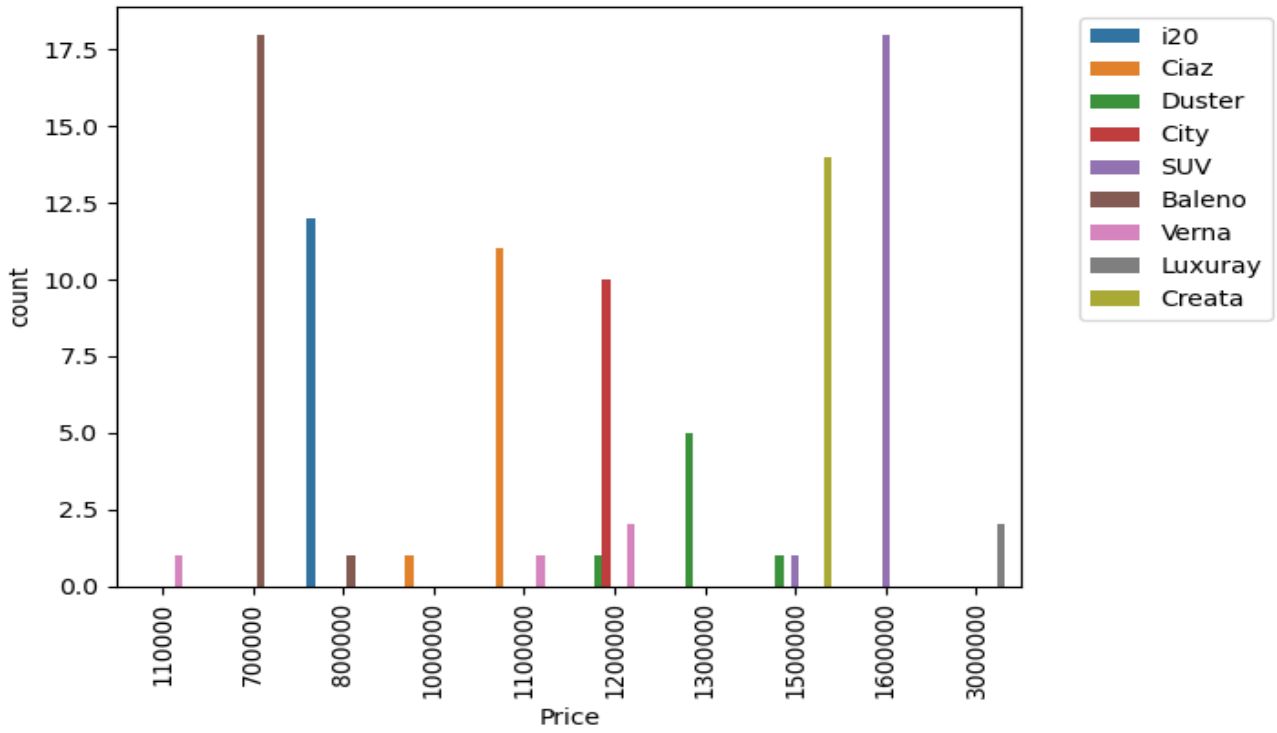
The majority of individuals in the dataset have 14L (lower income), while only a small number of members have the maximum income, according to this plot.

Countplot to visualize the distribution of car “makes”



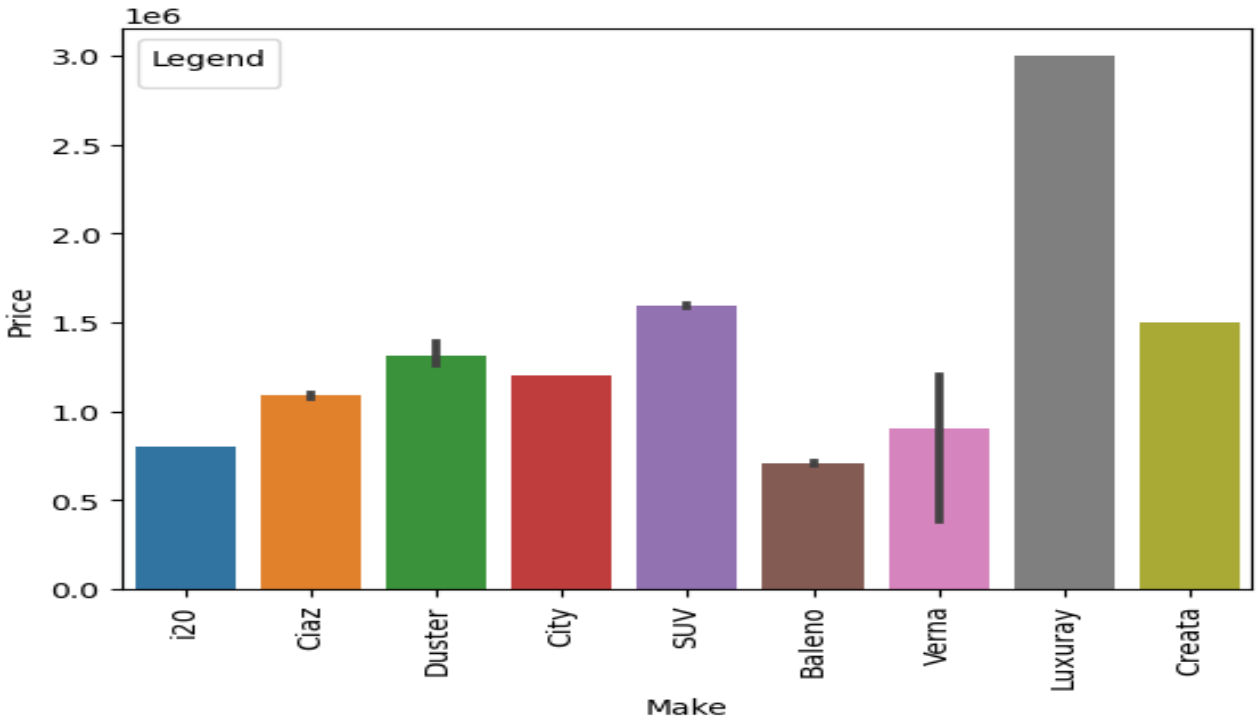
The resulting graph will display the frequency of each distinct car make in the dataset, making it useful for examining the distribution of various car makers in the data. And this graph unmistakably demonstrates that, according to the data, luxury brands are used less frequently than others whereas SUV and Baleno brands are used mostly.

Count plot for the distribution of car makes ('Make') across different price ranges ('Price')

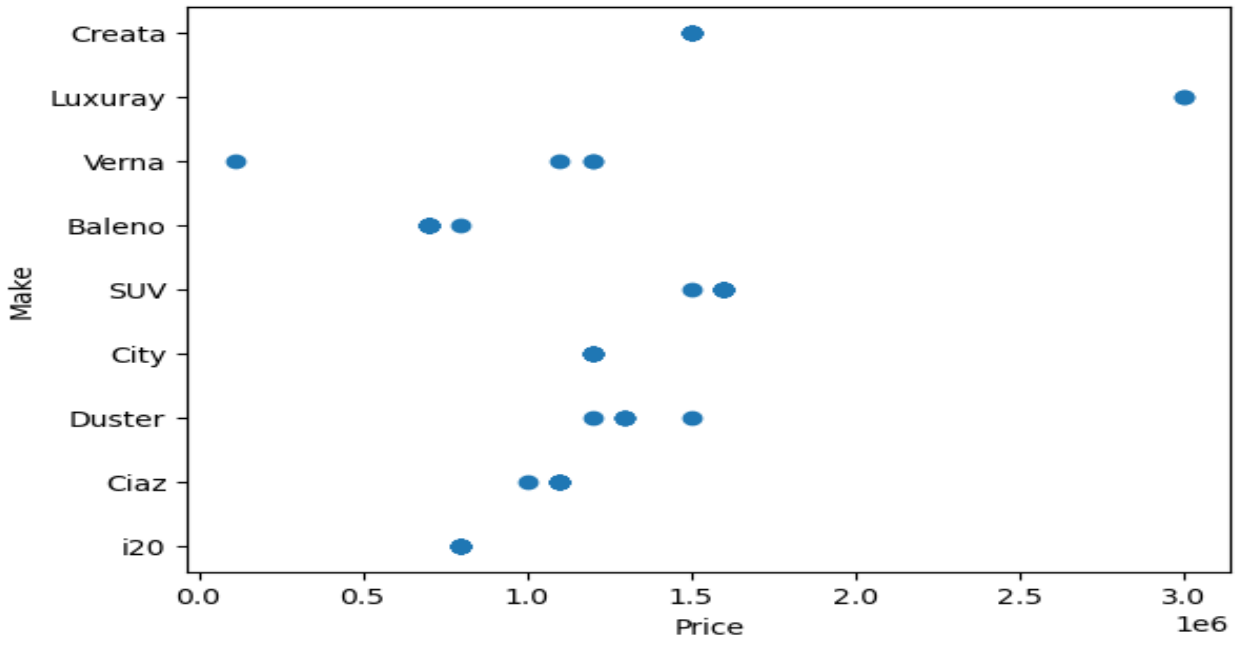


This plot helps viewers understand how car makes are distributed at various price points, making it valuable for assessing the relationship between car make, price, and count within the dataset, which can be useful for market analysis and decision-making.

Barplot that visualizes the relationship between the 'Make' of cars and their 'Price'



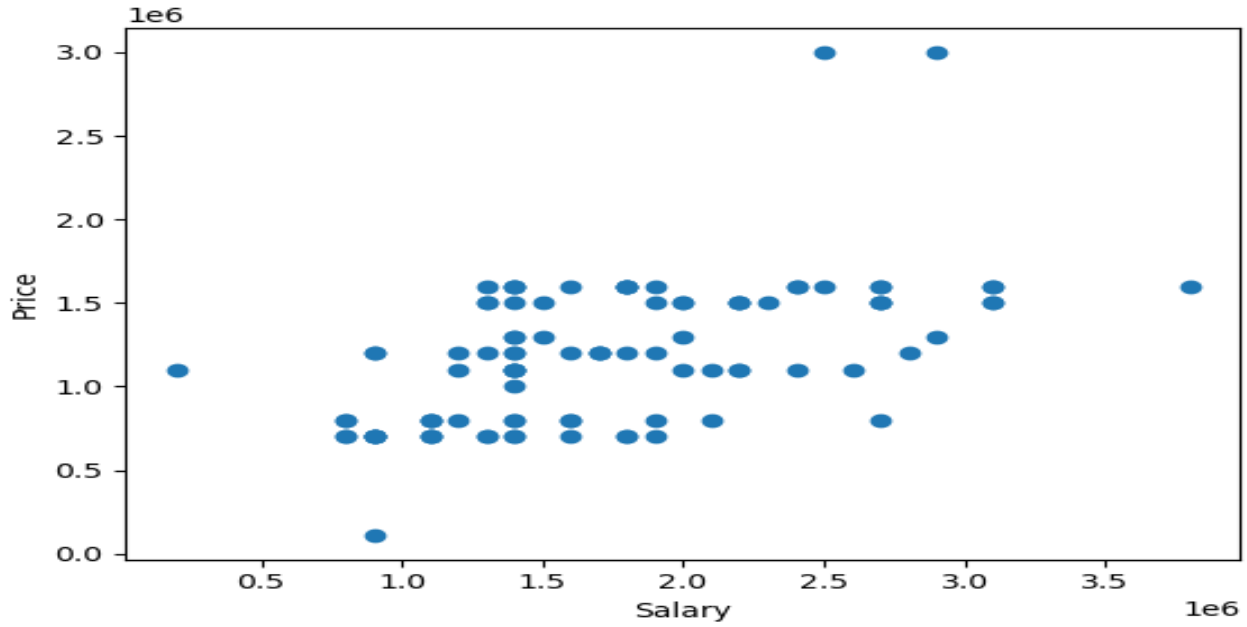
Scatter plot for the visualization of the relationship between 'Make' of cars and 'Price'



From both the plots, we can rapidly determine whether automakers tend to have higher or lower average pricing by visualising this data. It gives you a visual breakdown of how much cars cost across various makes, enabling you to compare costs and get insights about pricing patterns in your dataset.

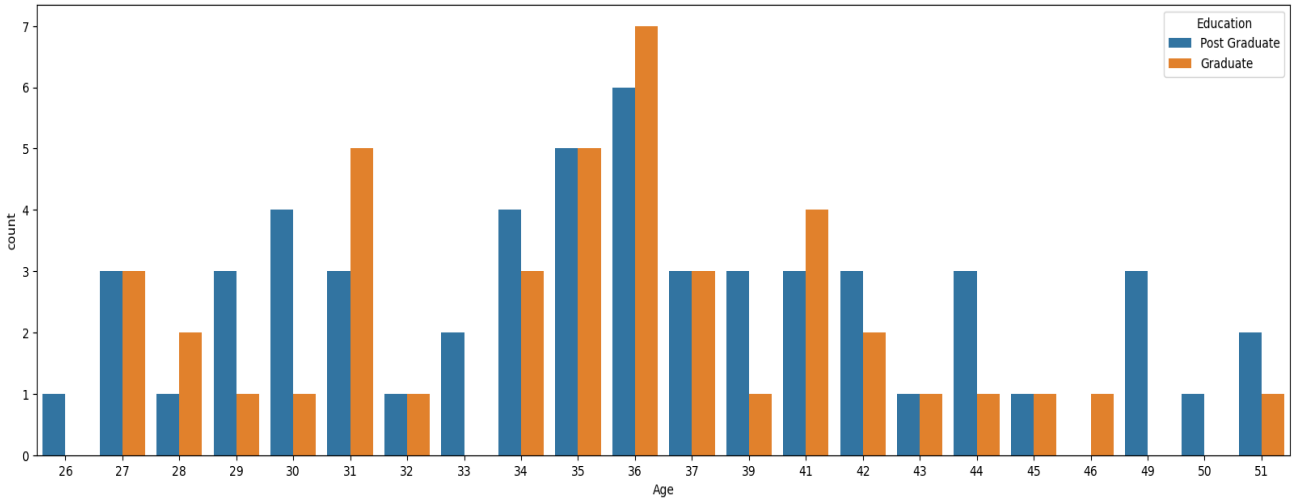
From the above charts, it is evident that luxury brand vehicles are very expensive, with the Baleno, Verna and i20 being the most affordable options.

Scatter plot that to visualize the relationship between ‘Salary' and 'Price'



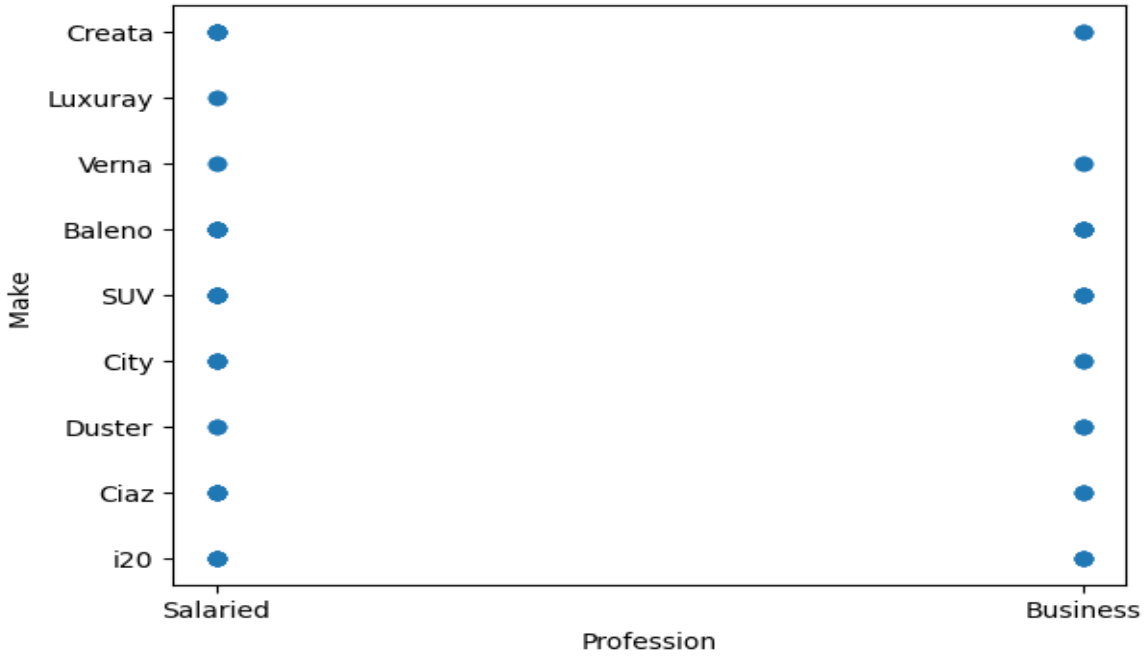
According to the plot, only a small number of people with the highest salaries owned electric vehicles that were the most expensive. The majority of people owned electric vehicles that were less expensive than the average vehicle.

Count plot for relationship between ‘Age’ and ‘Education’



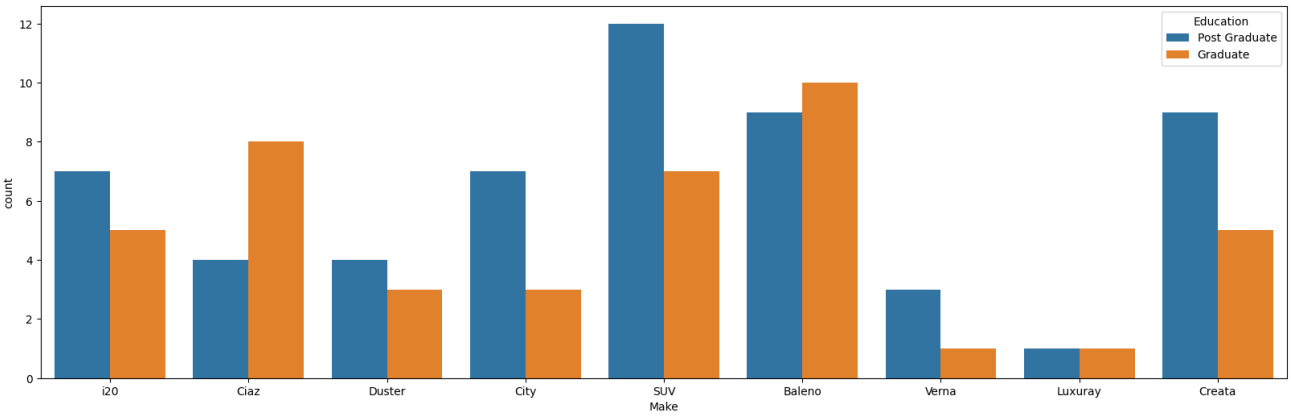
According to the above count plot, the majority of persons in the dataset are graduates who are middle-aged, and the majority of the top post-graduate and graduate students are located exclusively in the middle-aged zone.

Scatter plot between Make and Profession



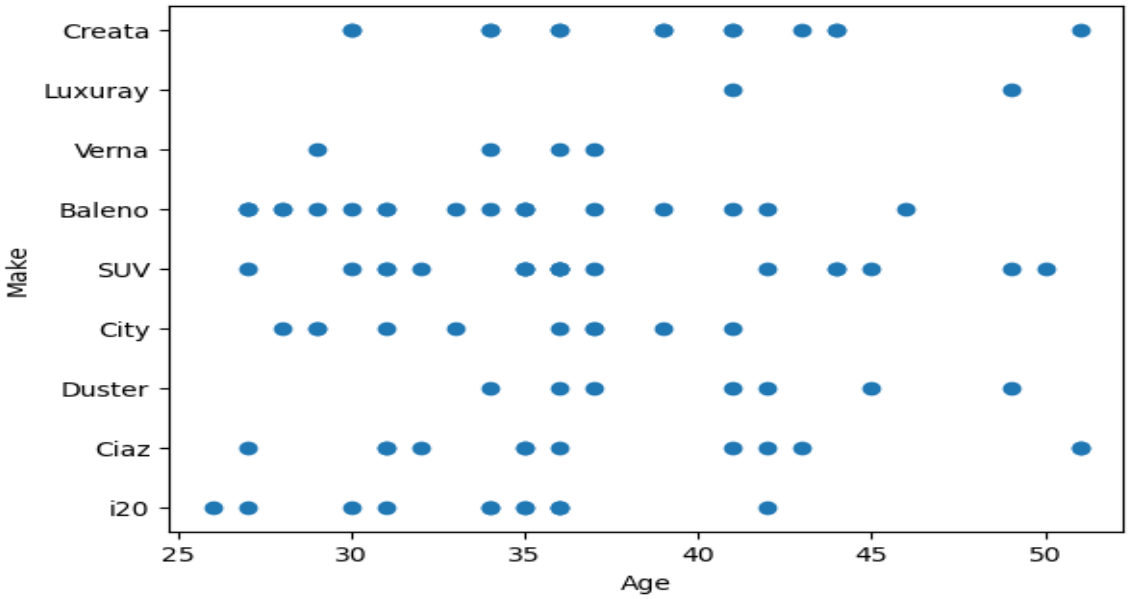
However, it's important to note that a scatter plot may not be the most appropriate choice for visualizing the relationship between 'Profession' and 'Make' unless there is a clear numeric or ordinal relationship between these two variables. Scatter plots are typically used to visualize the relationship between two continuous variables.

Count plot for relationship between 'Make' and 'Education'

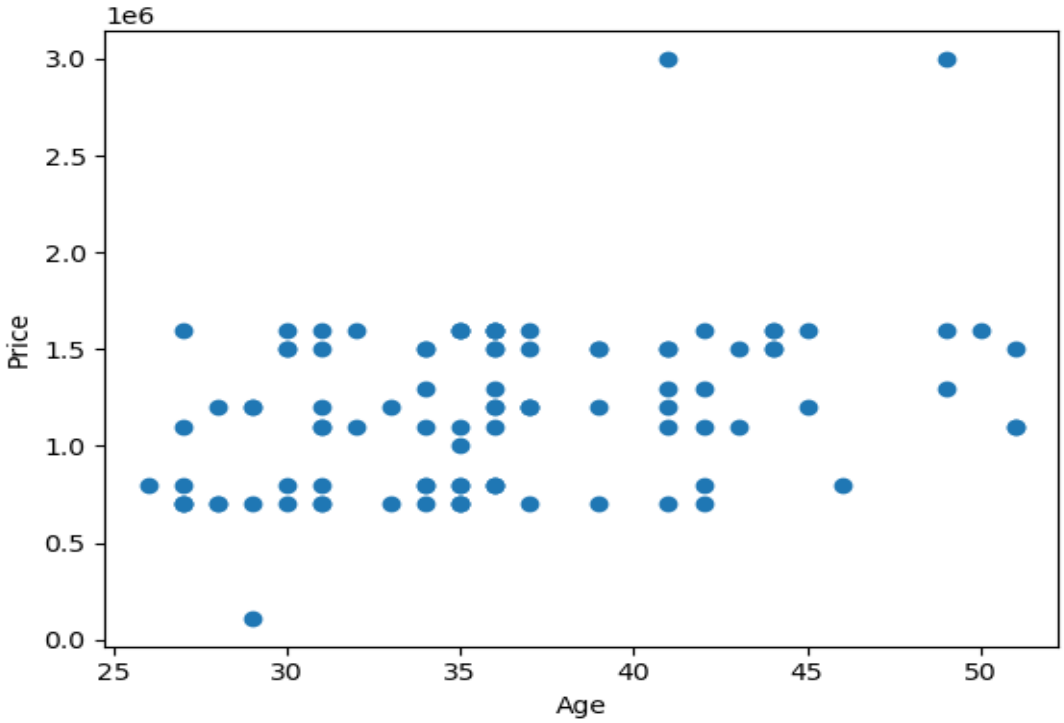


According to the above diagram, the majority of postgraduates own SUVs, and the majority of graduates own Baleno cars, with the least number of postgraduates and graduates owning luxury brand cars.

Scatter plot to visualize the relationship between Age and Make

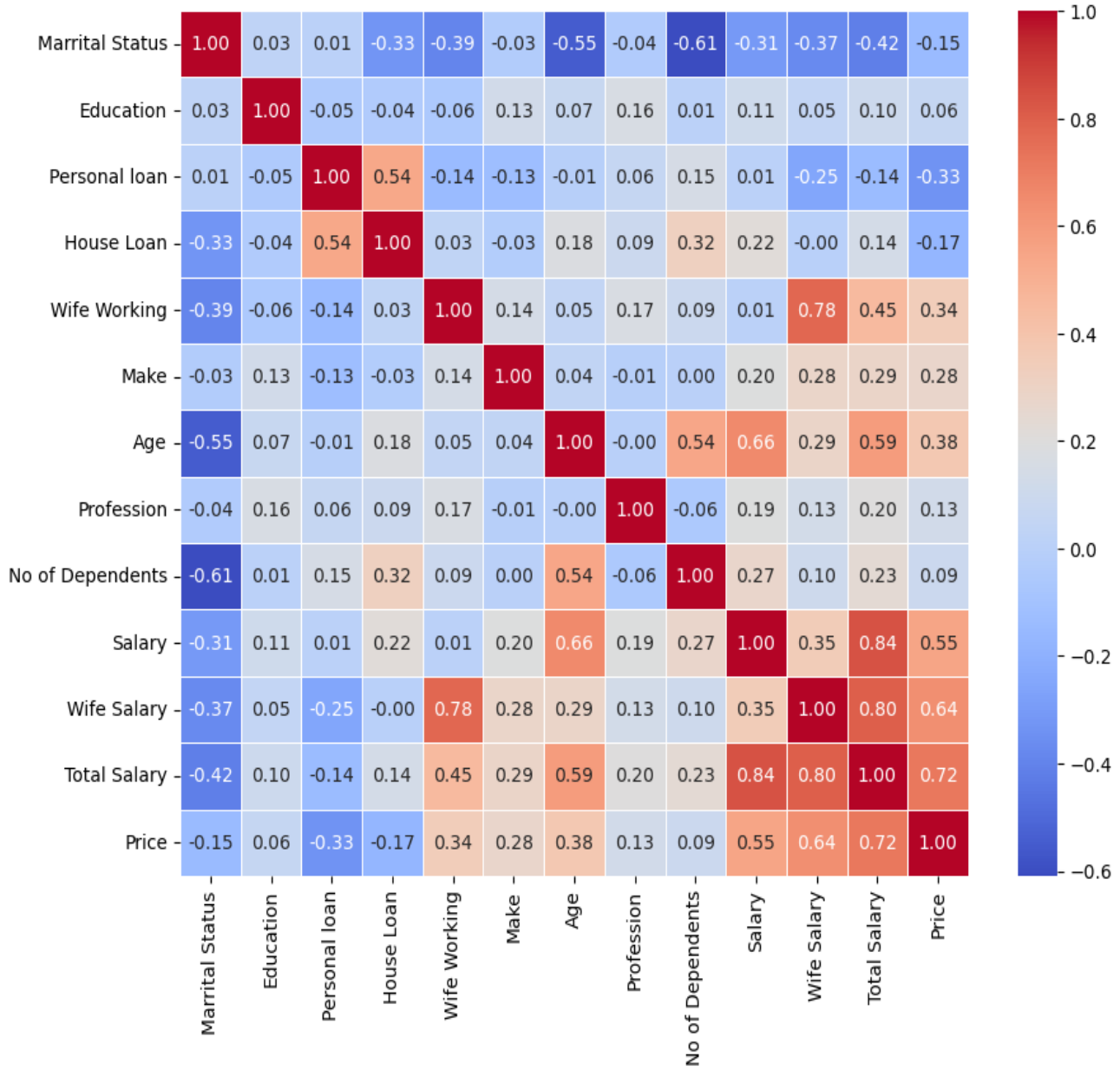


Scatter plot to visualize the relationship between Age and Make



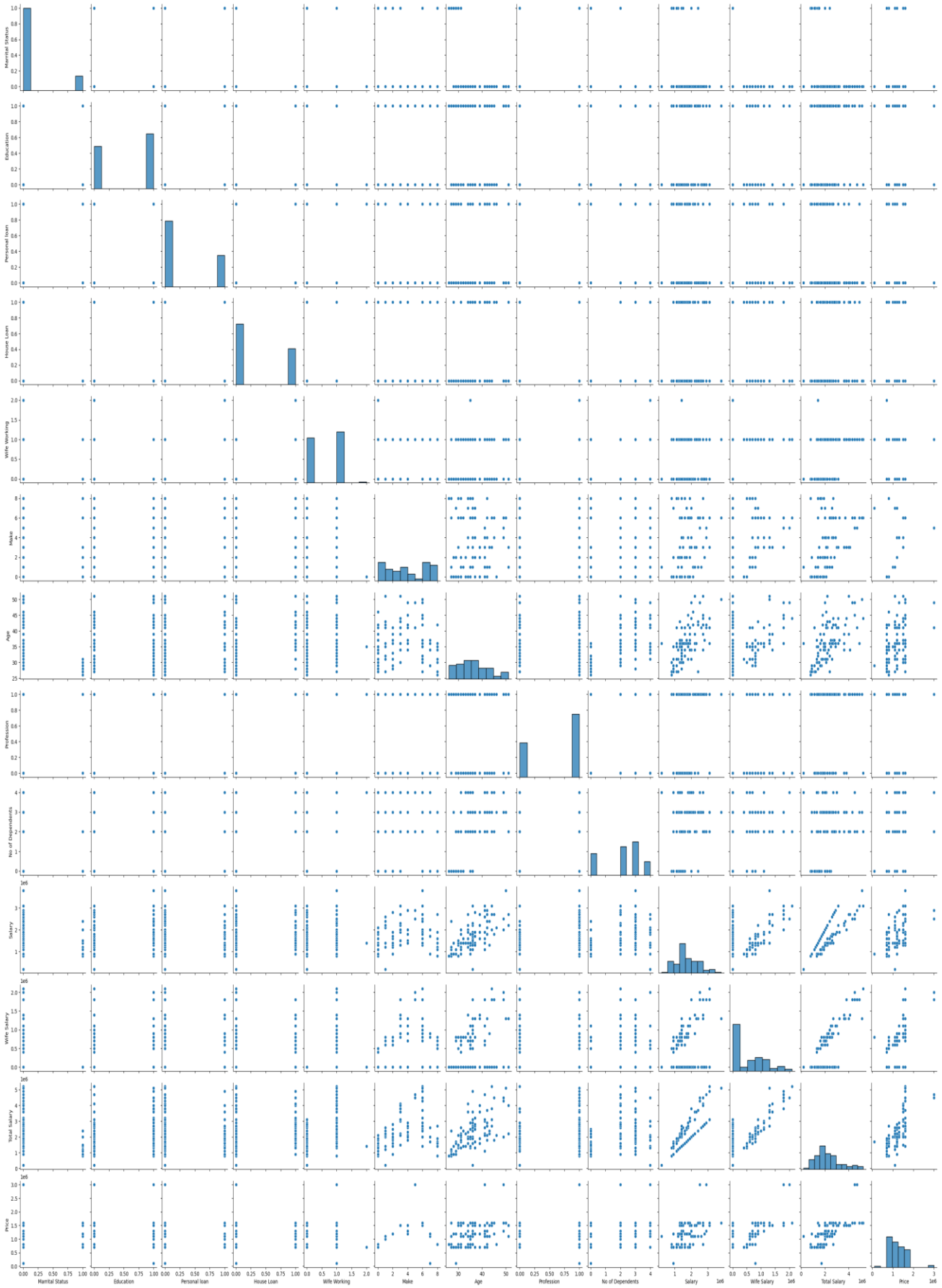
The scatter plot in the image above depicts the association between car age and price. The figure indicates that every member of the all-age groups owned a car that cost about as much as the typical electric car or less.

Heatmap



Overall, this heatmap shows the correlation between all the variables in the data. It facilitates the discovery of patterns and connections in the data, such as the direction and magnitude of any positive or negative correlations between particular variables. This can help you grasp the underlying relationships in the dataset and conduct exploratory data analysis.

Pair plot

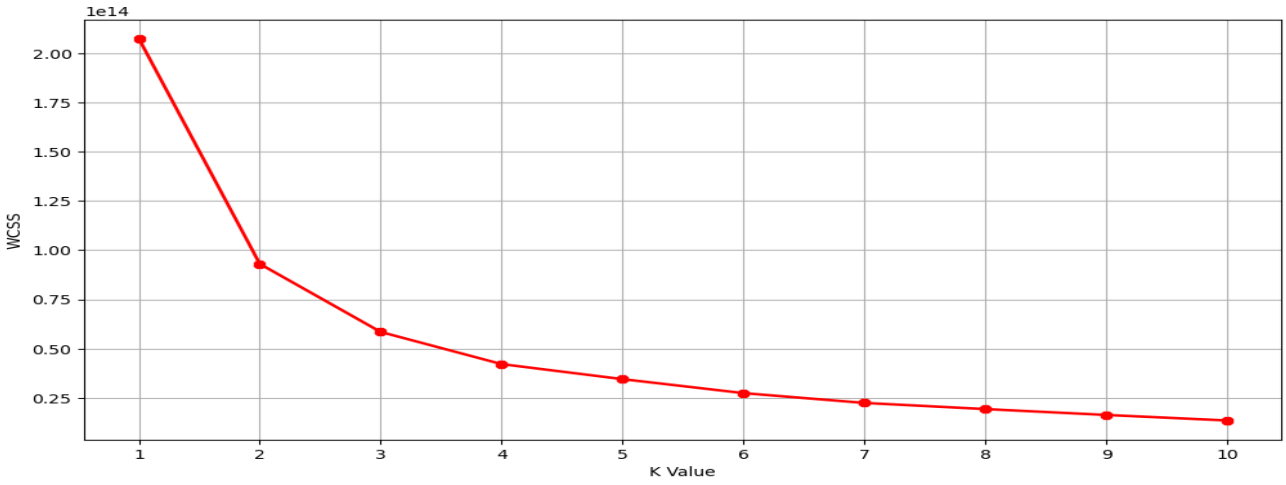

Marital Status
Education
Personal loan
House Loan
Wife Working
Make
Age
Profession
No of Dependents
Salary
Wife Salary
Total Salary
Price
1e6

K Means Clustering for Segment Extraction

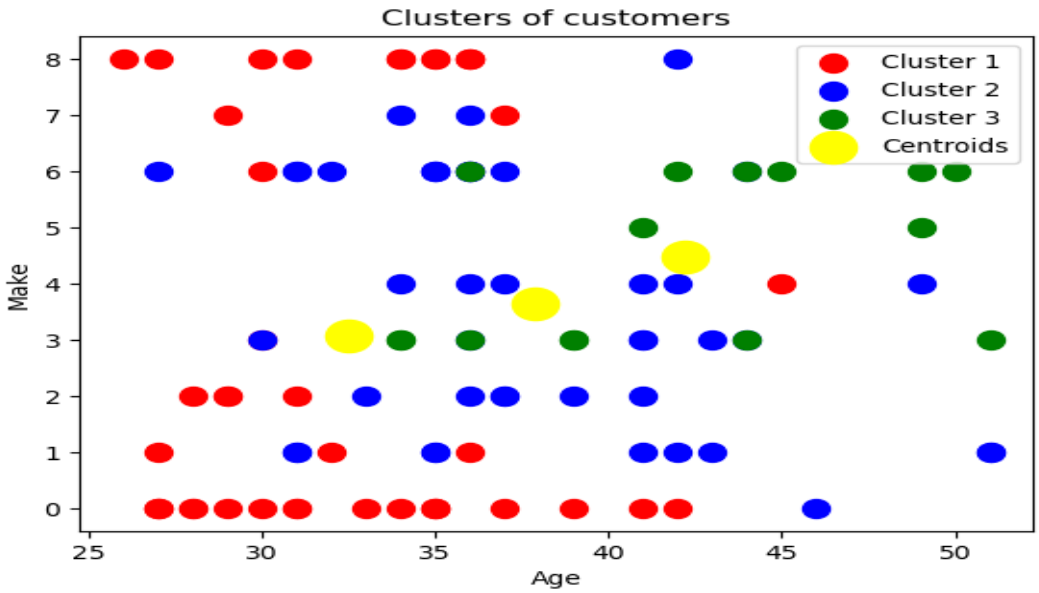
K-means clustering is an unsupervised machine learning technique used to segment data into distinct groups or clusters based on similarity. Inorder to determine the optimal number of clusters (K) using methods like the elbow or silhouette score. Once you apply the K-means algorithm, it assigns each data point to a cluster based on feature similarity. The clusters can then be interpreted, and the characteristics of each segment analyzed. Segment extraction can involve tailored marketing campaigns, customer profiling, anomaly detection, product recommendations, or resource allocation. Evaluation and iteration are crucial aspects to refine the segmentation process. While K-means is a popular choice, selecting the right clustering algorithm depends on your specific data and goals, as there are various clustering techniques available.

Elbow Method

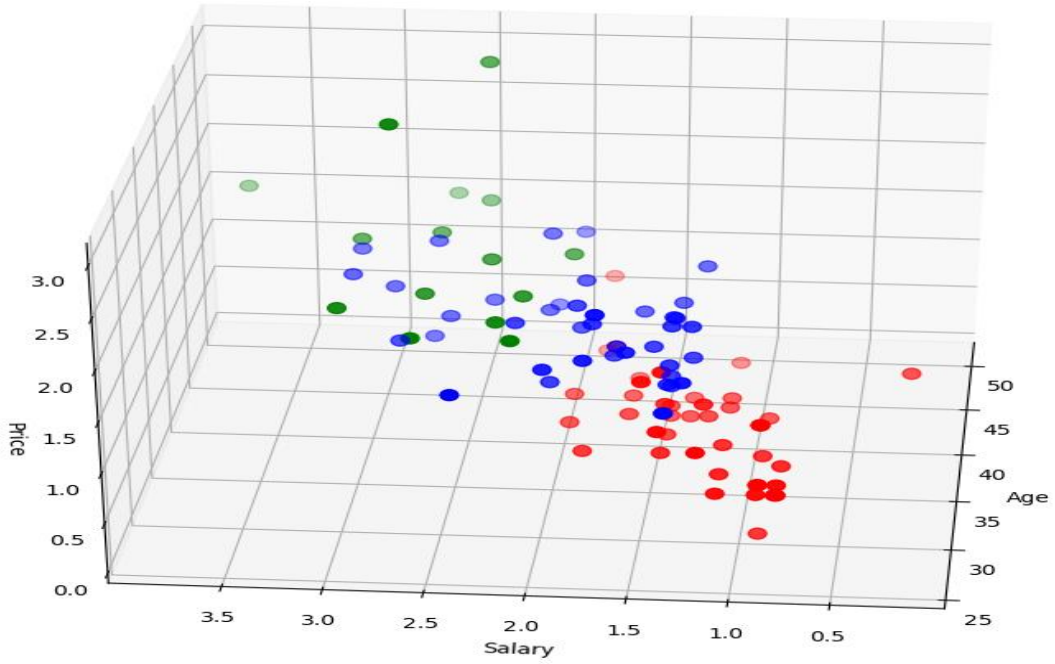
The elbow method is a technique used to determine the optimal number of clusters (K) for a K-means clustering algorithm. It is called the "elbow method" because the resulting plot typically resembles an elbow, and the point where the plot starts to bend is considered the ideal K value. The method works by running the K-means algorithm for a range of K values, usually from 1 to some maximum value, and for each K, it calculates the sum of squared distances (WCSS) between data points and their assigned cluster centroids. The WCSS represents the within-cluster variability, and it tends to decrease as K increases because more clusters allow for a better fit to the data. However, beyond a certain point, adding more clusters doesn't significantly reduce the WCSS, resulting in a diminishing return. The "elbow" in the plot corresponds to this point where the WCSS starts to level off, indicating that adding more clusters doesn't improve the clustering quality significantly. Therefore, the K value at the elbow is often chosen as the optimal number of clusters for your data, as it strikes a balance between minimizing within-cluster variability and avoiding overfitting.



The elbow point is at 3 in the preceding illustration, which corresponds to the number of clusters (ideal k=3).

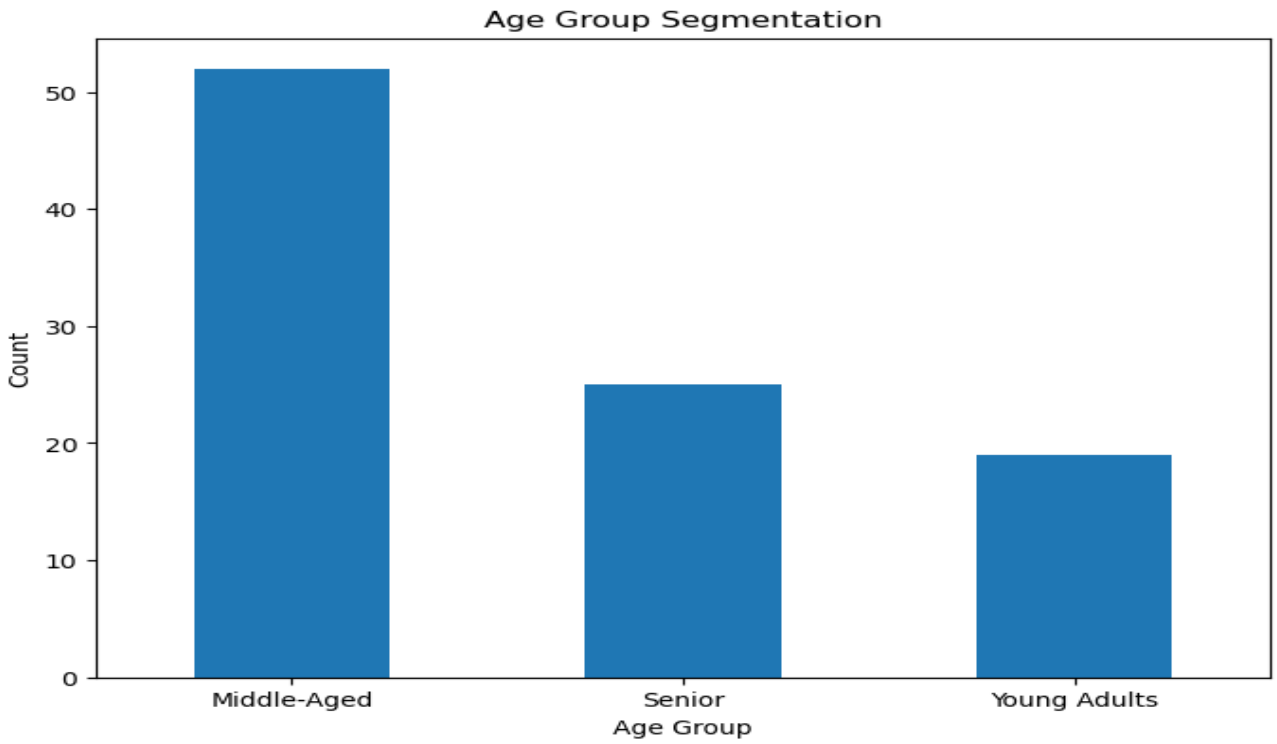


K-means clustering with three clusters (K=3)



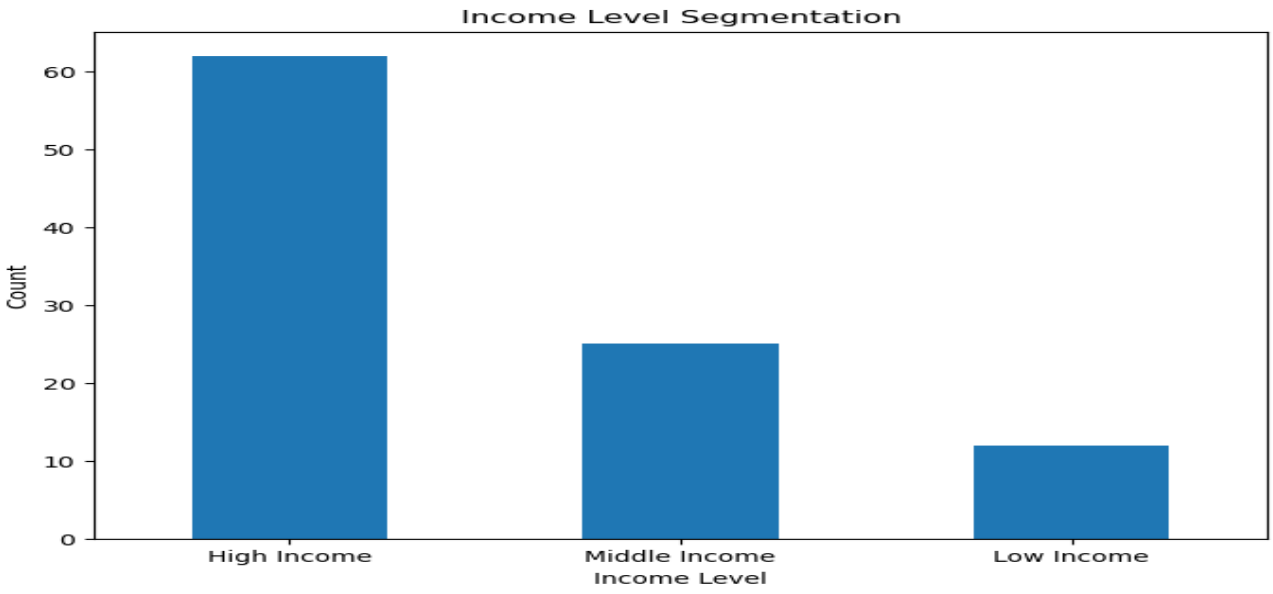
This visualization allows you to observe how data points are distributed in three-dimensional space based on their 'Age,' 'Salary,' and 'Price' attributes, with different colors indicating their respective clusters. It helps you explore the separation and distribution of data points among the three clusters visually.

Bar chart to visualize the segmentation of data into different 'Age Group' categories

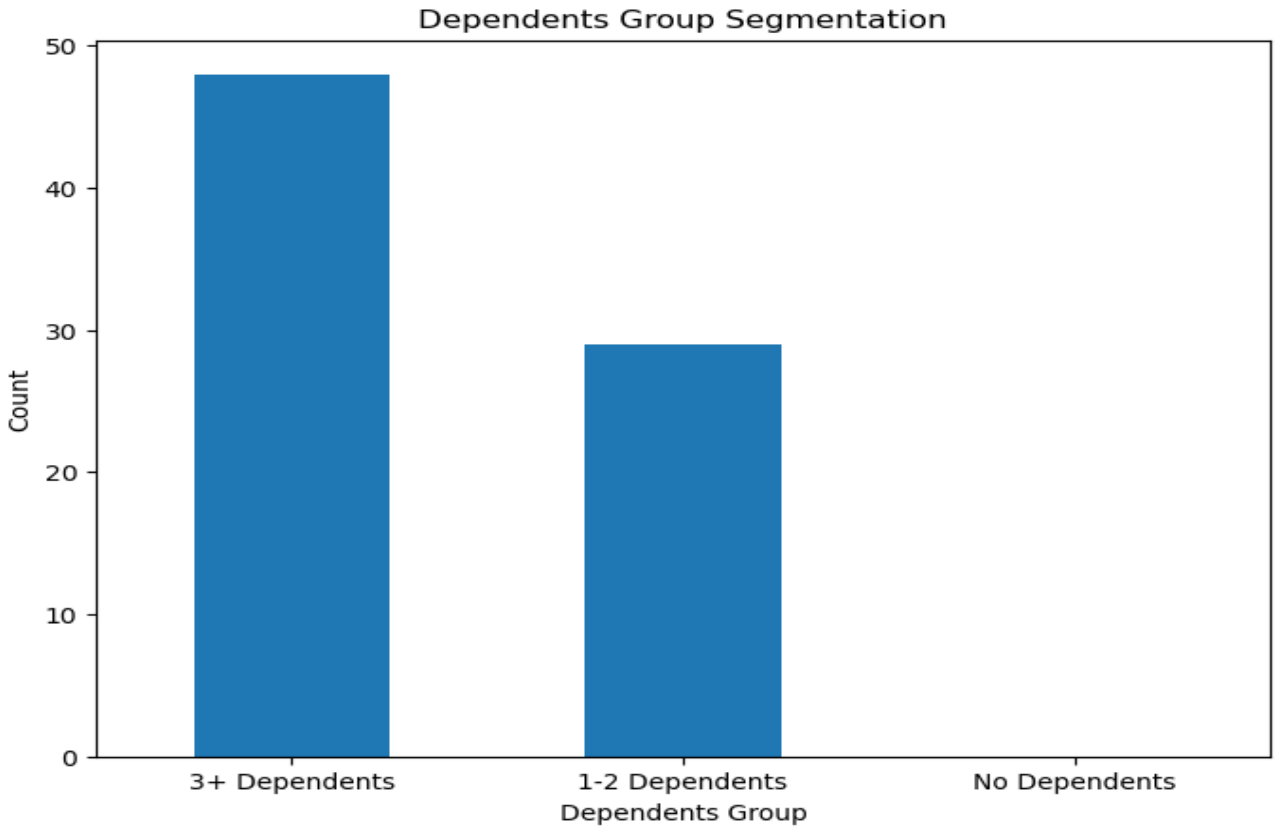


This bar chart provides a visual representation of the distribution of data points across different 'Age Group' categories. Each bar represents the count of data points in a particular age group, allowing you to quickly assess how data is segmented based on age. It's a useful visualization for understanding the demographic composition of your dataset.

Bar chart to visualize the segmentation of data into different 'Income' categories



Bar chart to visualize the segmentation of data into different 'Dependents' categories



Top 4 Variables for Market Segmentation from the Data:

1. Age: Age is a fundamental demographic variable that often correlates with consumer preferences, needs, and purchasing behaviours. It helps in targeting different age groups with tailored marketing strategies and product offerings.

2. Wife Working: The employment status of a spouse or partner can significantly impact household income and expenditure. Knowing whether the wife is working can help segment customers based on their financial capacity and lifestyle.

3.Personal Loan: Understanding whether a customer has a personal loan provides insights into their financial commitments and capacity for additional spending. This variable can be critical for tailoring financial products or services.

4. House Loan: Similar to personal loans, having a house loan can indicate financial stability and long-term financial goals. It helps in segmenting customers based on their housing needs and financial priorities.

By elaborating on the significance of these variables within the market domain, it provide a more compelling answer that demonstrates a deeper understanding of market segmentation strategies in this electric vehicles market segmentation.

Based on the data in the development of electric vehicles (EVs) in India, here are some key points to focus on:

1. Age: Based on Analysing the age distribution of target market middle age (younger) consumers more open to adopting EVs due to environmental concerns and technological appeal. And middle age group people consist of a greater number of post graduates and graduates are more open to adopting EV's. Tailor marketing strategies and product features to attract this demographic.

2. Wife Working: The employment status of spouses can impact household income and purchasing power. If both spouses are working, it may indicate a higher disposable income, making them a potential target market for premium or mid-range EVs.

3. Personal Loan: Understanding the prevalence of personal loans among your target audience can provide insights into their financial capacity and willingness to take on additional financial commitments. Offering EV financing options or incentives could be a strategy to consider.

4. House Loan: Similar to personal loans, the presence of house loans can affect the financial situation of potential EV buyers. Evaluate whether homeownership is associated with greater financial stability and explore financing packages that align with these consumers' needs.

5. Education: Although not included in the top four variables, education levels (e.g., Post Graduate, Graduate) can still be relevant. Higher education levels may correlate with a greater awareness of environmental issues and technological trends, potentially making individuals more inclined to adopt EVs.

6. Marital Status and Dependents: These factors can influence family size and lifestyle choices. Families with more dependents may prioritize practicality and cost-effectiveness, making them potential candidates for EVs with lower operational costs.

7. Salary and Total Salary: Assess the income distribution within your target market. Higher incomes may allow for more expensive EV models, while lower incomes might benefit from incentives or subsidies to make EVs more affordable.

8. Wife Salary: Consider the combined household income and whether both spouses are contributing. Dual-income households may have a stronger purchasing capacity for EVs, particularly if they value sustainability and technological advancement.

9. Make and Price: Research the preferred car makes and price ranges among your target market. Identify which EV models align with these preferences and price points, as this can guide product offerings and marketing strategies.

10.Environmental Awareness: Although not directly provided in the data, assess the level of environmental awareness and concern among your target audience. Develop marketing campaigns that highlight the eco-friendly benefits of EVs to resonate with environmentally conscious consumers.

11. Charging Infrastructure: Beyond the data, consider the availability and accessibility of EV charging infrastructure in India. The expansion of charging networks is crucial for widespread EV adoption.

12. Government Policies and Incentives: Stay updated on government policies, subsidies, and incentives related to EVs in India. These can significantly impact the EV market's growth and consumer adoption rates.

By analysing these factors in conjunction with the data, we can create a more targeted and effective strategy for the development and marketing of electric vehicles in India.