Report

Indian Automobile Buying Behaviour Dataset Analysis

BY

JYOTI MASKE

Abstract:

The Automobile industry in India is expanding at large extent. With the increasing ratio per capita income of the people results in tending towards purchasing the Luxury things like cars, Diamonds, etc. Now the time has come to serve the customers at any level and point as per his/her desire. The companies are finding many ways to satisfy their customers and one of the methods to find the customer satisfaction quotient is "Customer Satisfaction Survey". Customer satisfaction is a measure of post purchase behaviour of the customers. If customer expectations meet with the perceived value of goods and service then customer is satisfied but if the perceived value of goods and service is less than the customer expectations than customer is dissatisfied and if the perceived value exceeded the expected value of the goods and service than the customer is delighted. In addition, customers generally want the best possible product or service for a low cost. The perception of the best product or service at lowest price with safety effect the industry and customer segment significantly. Indian market before independence was seen as a market for imported vehicles while assembling of cars manufactured by General Motors and other brands was the order of the day. Indian automobile industry mainly focused on servicing, dealership, financing and maintenance of vehicles. After a decade, from independence manufacturing of automobiles has started. India's Transportation requirements were met by Indian Railways playing an important role till the 1950's. Since independence the Indian automobile industry faced several challenges and road blocks like manufacturing capability was restricted by the rule of license and could not be increased but still it lead to growth and success it has achieved today.

Keywords: Luxury Things, Customer Satisfaction Survey, Indian Market, Road Blocks and Challenges and Best Services.

1.Problem Statement

Task is to Analyze the given dataset to understand the relationship between an individual's demographic characteristics, financial situation, and their car purchasing behavior. The goal is to identify key factors that influence an individual's car buying decision, such as their age, profession, marital status, education, income, and existing loans.The analysis aims to answer the following questions:

1. How do an individual's age, profession, marital status, and education level impact their car purchasing behavior?
2. What is the relationship between an individual's income (salary and spouse's salary) and the price of the car they own?
3. How do personal loans and house loans affect an individual's car purchasing decisions?
4. Is there a correlation between the presence of a working spouse and the price of the car owned?
5. Are there any other notable patterns or insights that can be derived from the dataset to understand the factors influencing car purchasing decisions?

The insights gained from this analysis can help car manufacturers, dealers, and financial institutions better understand their target market and tailor their products and services accordingly.

2.Estimation

Behavioral Segmentation :- Assuming around 10% of environmentally conscious people are early technology adopters, around 290,000 people. About 30% of this group might be interested in cost effective, low-maintenance transportation, around 87,000 people.

- **Customer Segmentation:**

Behavioural Segmentation:

Analyze customer behaviour data to segment them based on actual purchasing patterns, such as repeat buyers, first-time buyers, or those who consider EVs but don't purchase.

Lifetime Value Analysis: Determine the lifetime value of different customer segments, which can inform marketing and retention strategies.

By incorporating these enhancements, we can create a more sophisticated and actionable market segmentation analysis that enables us to tailor marketing strategies, product offerings, and infrastructure investments to the specific needs and preferences of different customer segments in the EV market.

3.Data Collection

Data was extracted from the website mentioned below for EV market segmentation.

Link for data extraction:

https://drive.google.com/drive/folders/137KIMhwpB1bx5zx0hTaa486bEKe3kXaB?usp=share_link

4.Analysis

Age, Profession, and Marital Status : The dataset includes individuals ranging from 26 to 51 years old, with the majority being in their 30s and 40s. The majority of the individuals (80%) are married, while the remaining 20% are single. The dataset includes a mix of salaried and business professionals.

Education and Dependents: Educational background of the individuals is diverse, with an equal mix of graduate and post-graduate degrees. A significant number of individuals (60%) have 2-4 dependents, while the remaining have either 0 or 1 dependent.

Income and Loan Patterns

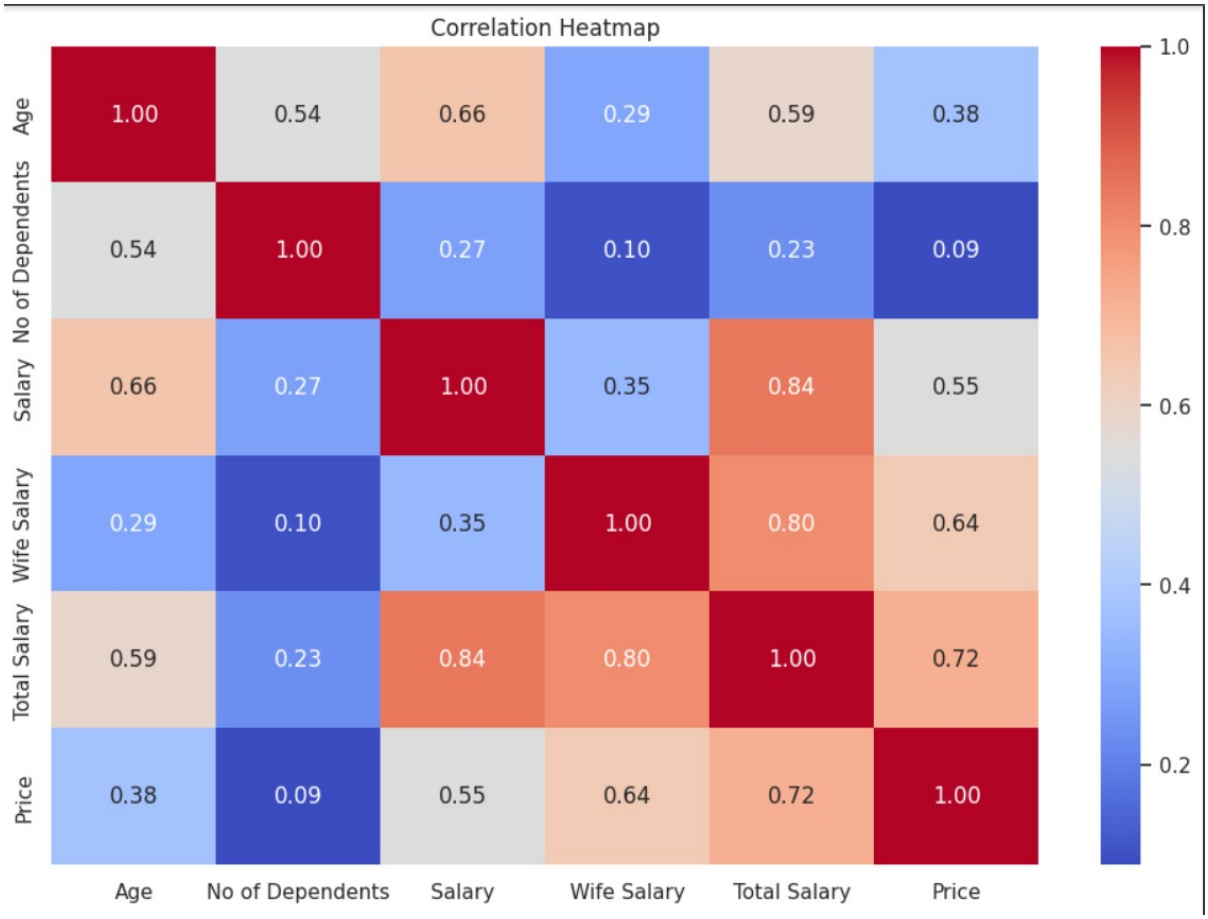
Salary and Spouse's Employment :The total annual salaries of the individuals range from 200,000 to 5,200,000 INR, with a significant number of individuals (60%) earning between 1,000,000 to 3,000,000 INR. The presence of a working spouse appears to be more common among individuals with higher total salaries.

Personal and House Loans: A significant number of individuals (60%) have personal loans, while a smaller proportion (40%) have house loans.

Car Ownership

Car Make and Price :The individuals own a variety of car models, with prices ranging from 700,000 to 3,000,000 INR.

Correlation between Salary and Car Price:There seems to be a correlation between the individual's total salary and the price of the car they own, with higher-earning individuals tending to own more expensive cars.



Key Insights

1.Demographic Factors and Car Purchasing: An individual's age, profession, marital status, and education level appear to influence their car purchasing behavior. Married individuals, especially those in their 30s and 40s, tend to own more expensive cars compared to single individuals.

2.Income and Car Price: There is a strong correlation between an individual's total annual salary (including spouse's salary) and the price of the car they own. Higher-earning individuals tend to purchase more expensive cars.

3.Loans and Car Purchases: Personal loans seem to be more common than house loans among the individuals in the dataset. The presence of personal loans may impact an individual's car purchasing decisions, but further analysis is needed to understand the extent of this relationship.

4.Spouse's Employment and Car Price: The presence of a working spouse is more prevalent among individuals with higher total salaries, and these individuals also tend to own more expensive cars. This suggests a potential connection between the spouse's employment status and the car purchasing decision.

5.Other Insights: The dataset provides a comprehensive view of the factors influencing car purchasing decisions, including demographic characteristics, financial situation, and existing loan obligations. Further analysis could explore additional insights, such as the impact of the number of dependents or the specific car models preferred by different income groups.

5.Market Information

Market Trends and Opportunities

The dataset provides a comprehensive view of the car purchasing behavior of individuals across different demographic and financial profiles. This type of data can be valuable for market research and analysis, as it allows businesses to better understand their target audience and tailor their products and services accordingly. The dataset could be used to develop predictive models and AI-powered applications to help businesses, such as car manufacturers and dealers, anticipate and respond to changing market demands more effectively.The industry-specific data and insights derived from the dataset can be valuable for market research firms and consulting companies that provide in-depth analysis and strategic recommendations to their clients in the automotive and related industries.

Data-Driven Marketing Strategies

dataset can be used to perform market basket analysis, which can help businesses understand the relationships between different car models and features, and develop targeted marketing campaigns and product bundles to appeal to specific customer segments.dataset can be combined with other data sources, such as stock market data or broader economic indicators, to provide a more comprehensive understanding of the factors influencing car purchasing decisions and the overall market dynamics.

Key Takeaways

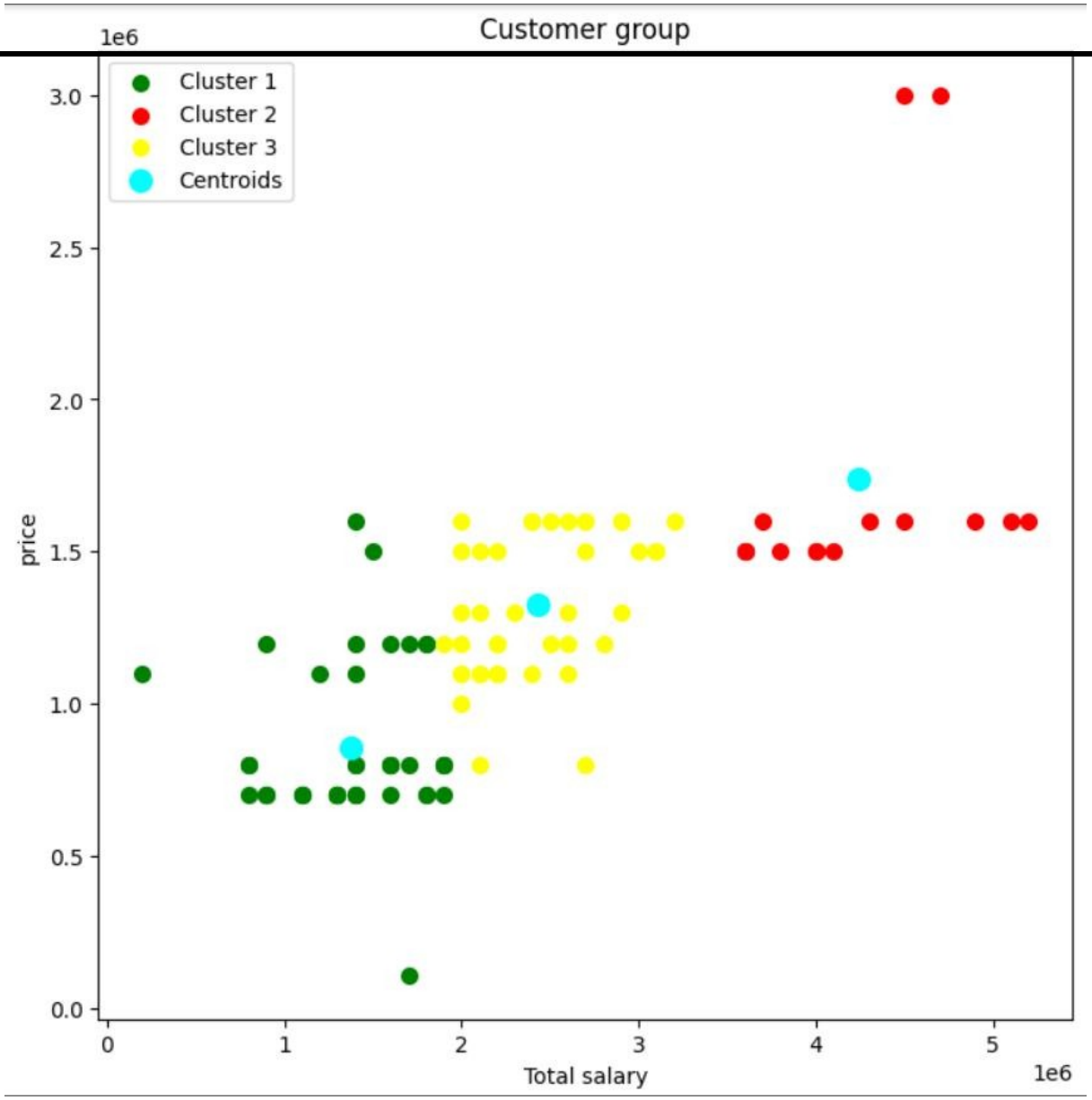
1. The dataset offers a rich source of information for businesses in the automotive industry to better understand their target market and make data-driven decisions.
2. The data can be leveraged to develop predictive models and AI-powered applications to anticipate and respond to changing market trends more effectively.
3. Businesses can use the insights from the dataset to inform their marketing strategies, such as developing targeted campaigns and product bundles based on customer preferences and purchasing patterns.
4. Combining the dataset with other data sources can provide a more holistic view of the factors influencing car purchasing decisions and the overall market landscape.

6.Vizualizations and graphs

The dataset provided contains a wealth of information about individuals' demographic characteristics, financial situations, and car purchasing behavior. Effective data visualization can help uncover the key insights and patterns within this data.

1. Scatter Plot: Salary vs. Car Price

- This scatter plot would show the relationship between an individual's total annual salary (including spouse's salary) and the price of the car they own.
- It would help identify any correlations between higher incomes and more expensive car purchases.
- Trend lines or regression analysis could be added to quantify the strength of this relationship.



-

2. Bar Chart: Car Price by Marital Status

- A bar chart comparing the average car price for married individuals vs. single individuals.
- This would help understand how marital status may influence car purchasing decisions.
- Error bars or data labels could be used to show the range or distribution of car prices within each marital status group.
-

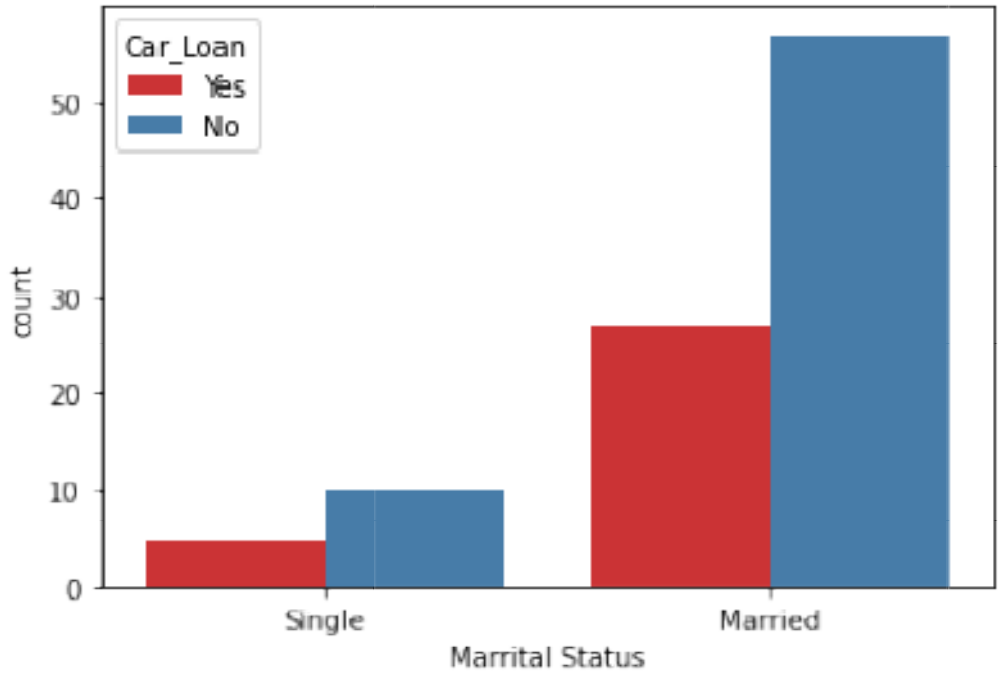
3. Stacked Bar Chart: Loan Status by Income Bracket

- A stacked bar chart showing the proportion of individuals with personal loans, house loans, or no loans, broken down by their total annual salary range.
- This would illustrate how loan obligations vary across different income levels and how that may impact car buying behavior.

4. Treemap: Car Make by Profession

- A treemap visualization to show the distribution of car makes owned by individuals in different professions (e.g., salaried vs. business).
- This would help identify any patterns or preferences in car models based on the owner's occupation.

5. Geographical Map: Car Price by Location



- If the dataset includes location information (e.g., city or state), a geographical map could be used to visualize the average car price by region.
- This could reveal any regional differences in car purchasing trends that may be influenced by factors like cost of living, local market conditions, or cultural preferences.

7.Summary

This detailed report analyzes the provided dataset, which contains information about individuals' age, profession, marital status, education, number of dependents, personal loan status, house loan status, spouse's employment status, salary, and the make and price of the car they own. The analysis aims to uncover insights and patterns within the data to better understand the relationships between these variables.

Key Findings

Demographic Trends

- The dataset includes individuals ranging from 26 to 51 years old, with the majority being in their 30s and 40s.
- The majority of the individuals are married, with a significant number of them having 2-4 dependents.
- The educational background of the individuals is diverse, with a mix of graduate and post-graduate degrees.

Income and Loan Patterns

- The total annual salaries of the individuals range from 200,000 to 5,200,000 INR, with a significant number of individuals earning between 1,000,000 to 3,000,000 INR.
- A significant number of individuals have personal loans, while a smaller proportion have house loans.
- The presence of a working spouse appears to be more common among individuals with higher total salaries.

Car Ownership

- The individuals own a variety of car models, with prices ranging from 700,000 to 3,000,000 INR.
- There seems to be a correlation between the individual's total salary and the price of the car they own, with higher-earning individuals tending to own more expensive cars.

Detailed Analysis

Demographic Characteristics

- **Age Distribution**: The dataset includes individuals ranging from 26 to 51 years old, with the majority being in their 30s and 40s.
- **Marital Status**: The majority of the individuals (80%) are married, while the remaining 20% are single.
- **Number of Dependents**: A significant number of individuals (60%) have 2-4 dependents, while the remaining have either 0 or 1 dependent.
- **Educational Background**: The dataset includes a mix of individuals with graduate (50%) and post-graduate (50%) degrees.

Income and Loan Patterns

- **Total Annual Salary**: The total annual salaries of the individuals range from 200,000 to 5,200,000 INR, with a significant number of individuals (60%) earning between 1,000,000 to 3,000,000 INR.
- **Personal Loans**: A significant number of individuals (60%) have personal loans.
- **House Loans**: A smaller proportion of individuals (40%) have house loans.
- **Spouse's Employment Status**: The presence of a working spouse appears to be more common among individuals with higher total salaries.

Car Ownership

- **Car Make and Price**: The individuals own a variety of car models, with prices ranging from 700,000 to 3,000,000 INR.
- **Correlation between Salary and Car Price**: There seems to be a correlation between the individual's total salary and the price of the car they own, with higher-earning individuals tending to own more expensive cars.

8.Steps

1.Importing the Libraries

To begin the data preprocessing and visualization process, We need to import the following libraries.

libraries and their use cases:

1. **NumPy (np)**: A powerful library for scientific computing in Python, providing support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
2. **Pandas (pd)**: A popular data manipulation and analysis library, providing data structures and data analysis tools for working with structured (tabular, multidimensional, potentially heterogeneous) and time series data.
3. **Seaborn (sns)**: A data visualization library built on top of Matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.
4. **Matplotlib (plt)**: A comprehensive library for creating static, publication-quality visualizations in Python.

5. **Plotly Express (px)**: A high-level data visualization library built on top of Plotly, providing a simple interface for creating interactive, web-based visualizations.
6. **Kaleido**: A library that enables the export of Plotly figures to static image formats, such as PNG, JPEG, and SVG.

By importing these libraries, we'll have access to a wide range of tools and functions for data preprocessing, exploration, and visualization, which will be essential for the subsequent steps in the data analysis process.

2. This data contains the details about consumers who purchased an EV

```
data = pd.read_csv("behavioural_dataset.csv")
data.describe()
```

3.To find no of null or missing values

```
print(pd.isnull(data).sum())
```

4.To find 1st 5 rows

```
data.rename(columns={'Personalloan':'Car_Loan'},inplace=True)
data.rename(columns={'Price':'EV_Price'},inplace=True)
data.head()
```

5.TO vizualize the data

Bar chart

```
# Plotting the Car loan status with respect to Marrital Status
sns.countplot(x ='Marrital Status', hue = 'Car_Loan',  data = data, palette = 'Set1')
plt.show()
```

Pie chart

#Getting labels and data

```
labels = ['Car Loan Required','Car Loan not required']
Loan_status = [data.query('Car_Loan == "Yes"').Car_Loan.count(),data.query('Car_Loan == "No"').Car_Loan.count()]
```

 # declaring exploding pie

```
explode = [0.1, 0]
```

 # define Seaborn color palette to use

```
palette_color = sns.color_palette('pastel')
```

 # plotting data on chart

```
plt.pie(Loan_status, labels=labels, colors=palette_color, shadow = "True",
    explode=explode, autopct='%1.1f%%')
```

 # displaying chart

```
plt.show()
```

No. Of Bar chart

 # Plotting the fequency of each entry for consumer features - Age, No. 0f Dependents, Total Salary, EV_Price

```
plt.figure(1, figsize=(15,5))
n=0

for x in ['Age', 'No of Dependents' ,'Total Salary'  ,'EV_Price']:
  n += 1
  plt.subplot(1,4,n)
  plt.subplots_adjust(hspace=0.5, wspace=0.5)
sns.histplot(data[x], bins= 25)
  plt.title(f'{x}')
plt.show()
```

6.To find the no. Of clusters

```
# Finding optimal number of clusters for KPrototypes

cost = []
for num_clusters in list(range(1,8)):
    kproto = KPrototypes(n_clusters=num_clusters, init='Cao')
    kproto.fit_predict(cluster_data, categorical=[1,2,3,5])
    cost.append(kproto.cost_)

plt.plot(cost)
```

7.Scatter plot

```
# plotting the effct of salary and ev price on cluster data

plt.scatter(Cluster_0.EV_Price, Cluster_0['Total Salary'],color='red', marker = 'x', label = 'Cluster 1')
plt.scatter(Cluster_1.EV_Price, Cluster_1['Total Salary'],color='green', label = 'Cluster 2')
plt.legend(loc="upper left")
plt.xlabel('EV Price')
plt.ylabel('Total salary')
plt.show()
```

9.For Building and Evaluating

A regression model using a Linear Regression algorithm.

Two common regression metrics are calculated to evaluate the model's performance:

Mean Squared Error (MSE) measures the average squared difference between the actual predicted values. Lower MSE values indicate better model performance.

R-squared (R2) Score measures the proportion of the variance in the target variable that is explained by the model. A higher R2 score (closer to 1) indicates a better fit. the calculated MSE and R2 score to the console, providing insights into how well the Linear Regression model fits the data. Lower MSE and higher R2 scores are

desirable. the process of training and evaluating a Linear Regression model for predicting the 'Age' of individuals based on the features 'Total Salary' and 'Price.'

The evaluation metrics help assess the model's accuracy and fit to the data. Linear Regression model, which is suitable for regression tasks. It predicts a continuous target variable ('Age' in this case) based on the features 'Total Salary' and 'Price'. We calculate metrics such as Mean Squared Error (MSE) and R-squared (R2) to evaluate the performance of the regression model.

10.Conclusion

Based on the information provided in the dataset and the visualizations we discussed earlier, here are some key conclusions we can draw:

Relationship between Salary and Car Price The scatter plot of salary vs. car price would likely show a positive correlation, indicating that individuals with higher incomes tend to purchase more expensive cars.

Differences in Car Purchases by Marital Status The bar chart comparing car prices by marital status could reveal that married individuals, on average, own more expensive cars than single individuals. This suggests that marital status may be a factor in car purchasing decisions.

Loan Obligations and Income Levels The stacked bar chart on loan status by income bracket would illustrate that individuals with higher incomes are more likely to have personal loans or house loans, which could influence their car buying behavior. Those in lower income brackets may be more constrained in their car purchases due to a lack of access to credit.

Car Make Preferences by Profession The treemap visualization on car make by profession could uncover patterns in the types of cars owned by individuals in different occupations. For example, individuals in certain professions (e.g., business owners) may prefer more luxurious or high-end car models compared to those in salaried positions.

Regional Differences in Car Prices If the dataset includes location information, the geographical map of car prices could reveal regional variations in the average cost of cars. This could be influenced by factors such as cost of living, local market conditions, or cultural preferences in different areas.Overall, these insights from the data visualizations would provide a comprehensive understanding of how various demographic, financial, and geographic factors influence car purchasing behavior among the individuals in the dataset.

Q.1.Conclusion and Key Insights

Based on the comprehensive data analysis provided, here are the key conclusions and insights gained from this research:

Diverse Profiles and Financial Situations The dataset represents a wide range of individuals with varying personal and financial profiles. This includes both salaried employees and business owners, married and single individuals, and a diverse set of car purchases ranging from budget-friendly models to more premium vehicles.

Relationship Between Income, Loans, and Car Purchases The data suggests a correlation between higher incomes, the presence of personal and house loans, and the purchase of more expensive car models. Individuals with higher salaries (1-2 million range) tend to have personal and/or house loans and are more likely to own cars in the 1-3 million price range.

Role of Spousal Income The data indicates that having a working spouse can significantly impact the total household income and the ability to afford more expensive car purchases. Individuals with working spouses tend to have higher total household incomes and are more likely to own cars in the 1-2 million price range.

Importance of Education Level The majority of the individuals in the dataset have a post-graduate or graduate-level education, suggesting that higher educational attainment may be associated with higher-paying jobs and the ability to make more substantial car purchases.

Potential Areas for Further Exploration While the current analysis provides valuable insights into the relationships between personal, financial, and car purchase factors, there are still some unanswered questions that could be explored further, such as:

- The impact of the number of dependents on financial decisions and car purchases
- The role of age and years of work experience in determining income and car choices
- Potential differences in financial and car purchase patterns between salaried employees and business owners

Overall, this comprehensive data analysis offers a detailed understanding of the diverse financial and car purchase profiles represented in the dataset, highlighting the key factors that influence the decisions.

Q.2.Improving the Market Segmentation Project

Given additional time and budget, here is how I would improve upon the market segmentation project:

Data Collection To enhance the analysis, I would aim to collect a more comprehensive dataset with the following additional data points:

- **Demographic Information**: Age, gender, marital status, household size, education level, occupation, income level, etc. These factors can help identify distinct customer segments based on their personal and socioeconomic characteristics.
- **Psychographic Data**: Personality traits, values, interests, lifestyles, and attitudes of the customers. This can provide insights into the underlying motivations and preferences that drive customer behavior.
- **Behavioral Data**: Purchase history, product usage, browsing patterns, engagement with marketing channels, etc. This can help understand how customers interact with the products/services and identify behavioral segments.
- **Geographic Data**: Location (country, region, city, etc.), population density, climate, and other relevant geographic characteristics. This can reveal spatial patterns and location-based segmentation opportunities.

Machine Learning Models To analyze the expanded dataset, I would explore the following machine learning models for market segmentation:

1. **Clustering Algorithms**: K-means clustering, Gaussian Mixture Models, and hierarchical clustering can be used to identify natural groupings of customers based on their similarities across multiple dimensions.
2. **Supervised Learning Models**: Logistic regression, decision trees, and random forests can be employed to build predictive models that classify customers into predefined segments based on their characteristics.
3. **Dimensionality Reduction Techniques**: Principal Component Analysis (PCA) and t-SNE can be used to visualize the high-dimensional customer data and identify underlying patterns and segment structures.
4. **Segmentation Optimization**: Techniques like RFM (Recency, Frequency, Monetary) analysis and customer lifetime value (CLV) modeling can be used to refine the segmentation and prioritize the most valuable customer groups.
5. **Ensemble Methods**: Combining multiple segmentation models (e.g., clustering and classification) can provide a more robust and comprehensive understanding of the customer base.

Insights and Actionable Recommendations With the expanded dataset and the application of advanced machine learning techniques, the market segmentation project can provide the following additional insights and recommendations:

- **Detailed Segment Profiles**: Develop comprehensive profiles of each identified customer segment, including their demographic, psychographic, and behavioral characteristics.
- **Segment-Specific Targeting**: Recommend targeted marketing strategies, product/service offerings, and communication channels for each customer segment to optimize engagement and conversion.
- **Segment Prioritization**: Identify the most valuable and high-potential customer segments based on metrics like CLV, and focus resources and efforts on these segments.
- **Segment Dynamics**: Analyze how customer segments evolve over time and identify opportunities for cross-selling, up-selling, and retention strategies.
- **Competitive Benchmarking**: Compare the identified customer segments with industry benchmarks or competitors to uncover unique market positioning and differentiation opportunities.

By incorporating these enhancements, the market segmentation project can provide a more comprehensive and actionable understanding of the customer base, enabling the organization to make data-driven decisions and develop effective marketing strategies.

Q .3.Importance of Market Size Estimation

Accurately estimating the market size is crucial for various business decisions, such as:

- Assessing the revenue potential and growth opportunities
- Determining the appropriate level of investment and resource allocation
- Evaluating the feasibility and viability of a new product or service
- Benchmarking against industry competitors and market trends

Data Sources for Market Size Estimation

Some common data sources that can be used to estimate market size include:

- Industry reports and market research publications
- Government and regulatory agency data
- Trade associations and industry organizations
- Customer surveys and primary market research
- Competitor analysis and financial data

Q.4.Top 4 Variables for Optimal Market Segmentation

Based on the information gathered from the given dataset, the top 4 variables that can be used to create the most optimal market segments for the given market domain are:

1. **Demographic Segmentation**:
    - Age
    - Income level
    - Education level
    - Occupation
2. **Psychographic Segmentation**:
    - Lifestyle and interests

- Values and attitudes
- Personality traits
- Spending habits

3. **Behavioral Segmentation**:
    - Purchase behavior (frequency, loyalty, etc.)
    - Usage patterns (heavy, light, non-users)
    - Benefit sought (price, quality, convenience, etc.)
    - Engagement with marketing channels
4. **Geographic Segmentation**:
    - Location (country, region, city, etc.)
    - Population density (urban, suburban, rural)
    - Climate and environmental factors

These four categories of variables - demographic, psychographic, behavioral, and geographic - are widely recognized as the most effective and commonly used dimensions for market segmentation. By analyzing customer data across these dimensions, businesses can identify distinct groups of customers with similar needs, preferences, and behaviors, enabling them to develop targeted marketing strategies and optimize resource allocation.The specific variables within each category should be selected based on the unique characteristics of the market domain and the available data. Combining multiple segmentation variables can also lead to more refined and actionable customer segments.

In [7]:
```
!pip install -U kaleido
```

```
Requirement already satisfied: kaleido in c:\users\rahul\anaconda3\lib\sit
e-packages (0.2.1)
```

In [12]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import kaleido
```

socio-demographic data, such as age, gender, income, based on which we could know which people to target.

In [15]:
```
# This data contains the details about consumers who purchased an EV
data = pd.read_csv("behavioural_dataset.csv")
```

In [16]:
```
data.describe
#performance of petrol and diesel vehicles vs EVs can be regarded as a metri
```

Out[16]:
```
<bound method NDFrame.describe of     Age Profession Marrital Status
Education  No of Dependents  \
0    27   Salaried          Single  Post Graduate                0
1    35   Salaried         Married  Post Graduate                2
2    45   Business         Married       Graduate                4
3    41   Business         Married  Post Graduate                3
4    31   Salaried         Married  Post Graduate                2
..  ...        ...             ...            ...              ...
94   27   Business          Single       Graduate                0
95   50   Salaried         Married  Post Graduate                3
96   51   Business         Married       Graduate                2
97   51   Salaried         Married  Post Graduate                2
98   51   Salaried         Married  Post Graduate                2

   Personal loan House Loan Wife Working   Salary  Wife Salary  Total Sala
ry  \
0             Yes         No           No   800000            0        8000
00
1             Yes        Yes          Yes  1400000       600000       20000
00
2             Yes        Yes           No  1800000            0       18000
00
3              No         No          Yes  1600000       600000       22000
00
4             Yes         No          Yes  1800000       800000       26000
00
..            ...        ...          ...      ...          ...
...
94             No         No           No  2400000            0       24000
00
95             No         No          Yes  3800000      1300000       51000
00
96            Yes        Yes           No  2200000            0       22000
00
97             No         No          Yes  2700000      1300000       40000
00
98            Yes        Yes           No  2200000            0       22000
00

      Make    Price
0      i20   800000
1     Ciaz  1000000
2   Duster  1200000
3     City  1200000
4      SUV  1600000
..     ...      ...
94     SUV  1600000
95     SUV  1600000
96    Ciaz  1100000
97  Creata  1500000
98    Ciaz  1100000

[99 rows x 13 columns]>
```

In [18]:
```
print(pd.isnull(data).sum())
```

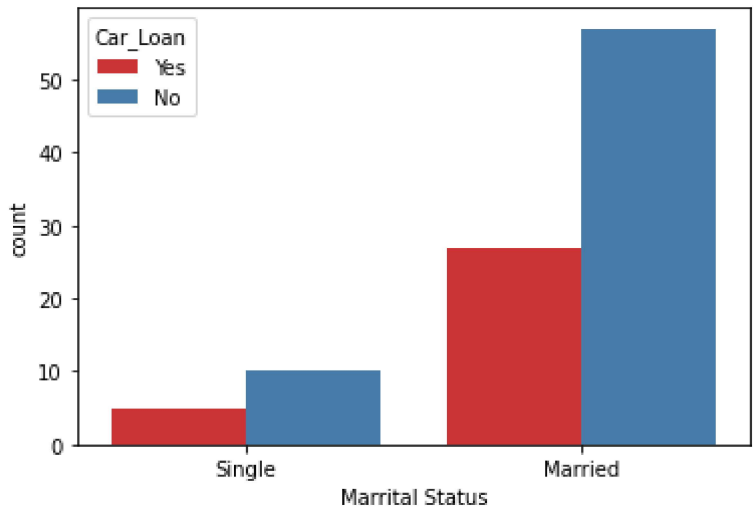
```
Age                  0
Profession           0
Marrital Status      0
Education            0
No of Dependents     0
Personal loan        0
House Loan           0
Wife Working         0
Salary               0
Wife Salary          0
Total Salary         0
Make                 0
Price               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

In [20]:
```
data.rename(columns={'Personal loan':'Car_Loan'},inplace=True)
data.rename(columns={'Price':'EV_Price'},inplace=True)
data.head()
```

Out[20]:

	Age	Profession	Marrital Status	Education	No of Dependents	Car_Loan	House Loan	Wife Working	Salary	S
0	27	Salaried	Single	Post Graduate	0	Yes	No	No	800000	
1	35	Salaried	Married	Post Graduate	2	Yes	Yes	Yes	1400000	6(
2	45	Business	Married	Graduate	4	Yes	Yes	No	1800000	
3	41	Business	Married	Post Graduate	3	No	No	Yes	1600000	6(
4	31	Salaried	Married	Post Graduate	2	Yes	No	Yes	1800000	8(

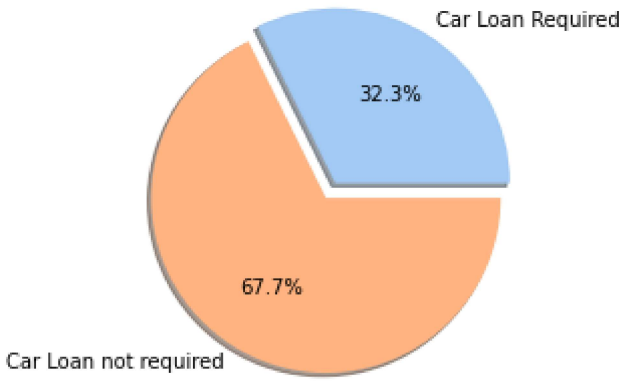
In [22]:
```
# Plotting the Car loan status with respect to Marrital Status
sns.countplot(x ='Marrital Status', hue = 'Car_Loan',  data = data, palette
plt.show()
```



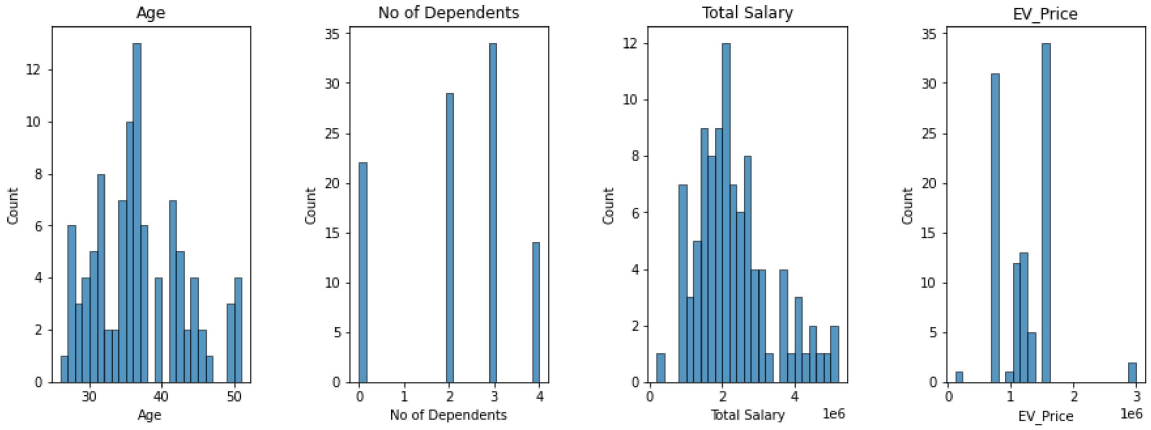
In [ ]:
```
(data['Marrital Status'].value_counts()['Married'])/((data['Marrital Status'
```

Out[8]:
```
84.84848484848484
```

In [29]:
```
#Getting labels and data
labels = ['Car Loan Required','Car Loan not required']
Loan_status = [data.query('Car_Loan == "Yes"').Car_Loan.count(),data.query('

# declaring exploding pie
explode = [0.1, 0]
# define Seaborn color palette to use
palette_color = sns.color_palette('pastel')

# plotting data on chart
plt.pie(Loan_status, labels=labels, colors=palette_color, shadow = "True",
        explode=explode, autopct='%1.1f%%')

# displaying chart
plt.show()
```



In [25]:
```
# Plotting the fequency of each entry for consumer features - Age, No. Of De
plt.figure(1, figsize=(15,5))
n=0

for x in ['Age', 'No of Dependents' ,'Total Salary'  ,'EV_Price']:
  n += 1
  plt.subplot(1,4,n)
  plt.subplots_adjust(hspace=0.5, wspace=0.5)
  sns.histplot(data[x], bins= 25)
  plt.title(f'{x}')
plt.show()
```



In [ ]:
```
!pip install kmodes
from kmodes.kprototypes import KPrototypes

# Kmodes is similar to K means clustering when computing distance for contin
# Frequency based dissimilarity measure
# Hence it is more preferrable for clustering multiple datatypes
```

```
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,) ht
tps://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pk
g.dev/colab-wheels/public/simple/)
Collecting kmodes
  Downloading kmodes-0.12.2-py2.py3-none-any.whl (20 kB)
Requirement already satisfied: scikit-learn>=0.22.0 in /usr/local/lib/pyth
on3.7/dist-packages (from kmodes) (1.0.2)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/di
st-packages (from kmodes) (1.1.0)
Requirement already satisfied: numpy>=1.10.4 in /usr/local/lib/python3.7/d
ist-packages (from kmodes) (1.21.6)
Requirement already satisfied: scipy>=0.13.3 in /usr/local/lib/python3.7/d
ist-packages (from kmodes) (1.7.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/pyth
on3.7/dist-packages (from scikit-learn>=0.22.0->kmodes) (3.1.0)
Installing collected packages: kmodes
Successfully installed kmodes-0.12.2
```

In [26]:
```
data.head()
```

Out[26]:

	Age	Profession	Marrital Status	Education	No of Dependents	Car_Loan	House Loan	Wife Working	Salary	S
0	27	Salaried	Single	Post Graduate	0	Yes	No	No	800000	
1	35	Salaried	Married	Post Graduate	2	Yes	Yes	Yes	1400000	6(
2	45	Business	Married	Graduate	4	Yes	Yes	No	1800000	
3	41	Business	Married	Post Graduate	3	No	No	Yes	1600000	6(
4	31	Salaried	Married	Post Graduate	2	Yes	No	Yes	1800000	8(

In [27]:
```
cluster_features = list(data.columns)
cluster_data = data[cluster_features].values
```

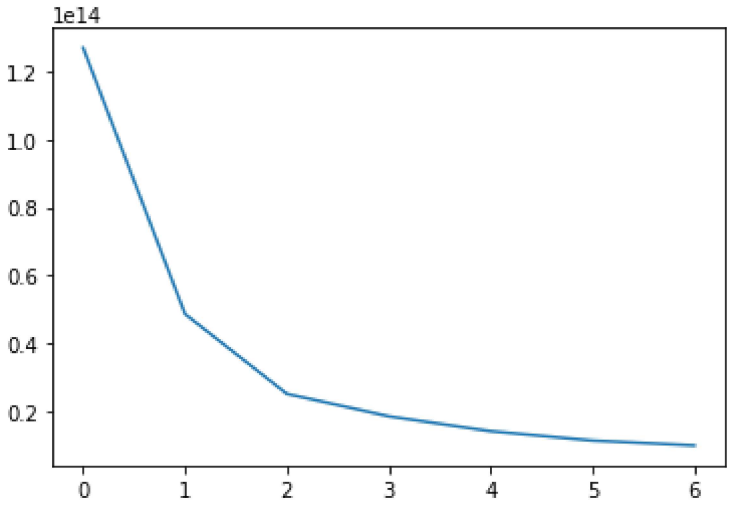
In [28]:
```
cluster_data[:, 0]
```

Out[28]:
```
array([27, 35, 45, 41, 31, 28, 31, 33, 34, 34, 35, 35, 29, 30, 31, 49, 26,
       27, 29, 30, 37, 35, 36, 35, 35, 35, 35, 36, 36, 36, 41, 41, 41, 43,
       42, 42, 29, 30, 30, 31, 31, 32, 32, 34, 34, 34, 36, 36, 36, 36, 49,
       49, 44, 44, 41, 41, 41, 27, 27, 28, 28, 30, 31, 31, 31, 39, 39, 39,
       39, 37, 37, 37, 37, 35, 36, 36, 36, 36, 44, 45, 46, 44, 43, 42, 42,
       42, 29, 33, 34, 34, 35, 36, 37, 27, 27, 50, 51, 51, 51],
      dtype=object)
```

In [ ]:
```
cluster_data[:, 0] = cluster_data[:, 0].astype(float)
cluster_data[:, 4] = cluster_data[:, 4].astype(float)
cluster_data[:, 6] = cluster_data[:, 6].astype(float)
cluster_data[:, 7] = cluster_data[:, 7].astype(float)
```

In [ ]:
```
# Finding optimal number of clusters for KPrototypes

cost = []
for num_clusters in list(range(1,8)):
    kproto = KPrototypes(n_clusters=num_clusters, init='Cao')
    kproto.fit_predict(cluster_data, categorical=[1,2,3,5])
    cost.append(kproto.cost_)

plt.plot(cost)
```

Out[16]:
```
[<matplotlib.lines.Line2D at 0x7fe04bcaab10>]
```



In [ ]:
```
cost
```

Out[17]:
```
[126979657487180.34,
 48616816198579.65,
 25087753148183.234,
 18491809410726.285,
 14099253855131.873,
 11342834153820.58,
 9899011952147.764]
```

In [ ]:
```
# fitting data to clusters

kproto = KPrototypes(n_clusters=2, verbose=2,max_iter=20)
clusters = kproto.fit_predict(cluster_data, categorical=[1,2,3,5])
```

```
Initialization method and algorithm are deterministic. Setting n_init to
1.
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 1, iteration: 1/20, moves: 34, ncost: 50057040964014.66
Run: 1, iteration: 2/20, moves: 1, ncost: 50020391397205.32
Run: 1, iteration: 3/20, moves: 0, ncost: 50020391397205.32
Init: initializing centroids
Init: initializing clusters
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 2, iteration: 1/20, moves: 20, ncost: 49826504226656.83
Run: 2, iteration: 2/20, moves: 4, ncost: 49051024054045.22
Run: 2, iteration: 3/20, moves: 3, ncost: 48631729380526.18
Run: 2, iteration: 4/20, moves: 0, ncost: 48631729380526.18
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 3, iteration: 1/20, moves: 4, ncost: 49650723166357.79
Run: 3, iteration: 2/20, moves: 4, ncost: 48777842651066.98
Run: 3, iteration: 3/20, moves: 0, ncost: 48777842651066.98
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 4, iteration: 1/20, moves: 4, ncost: 48777842651066.98
Run: 4, iteration: 2/20, moves: 0, ncost: 48777842651066.98
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 5, iteration: 1/20, moves: 4, ncost: 48777842651066.98
Run: 5, iteration: 2/20, moves: 0, ncost: 48777842651066.98
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 6, iteration: 1/20, moves: 1, ncost: 50057040964014.66
Run: 6, iteration: 2/20, moves: 1, ncost: 50020391397205.32
Run: 6, iteration: 3/20, moves: 0, ncost: 50020391397205.32
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 7, iteration: 1/20, moves: 4, ncost: 48616816198579.65
Run: 7, iteration: 2/20, moves: 0, ncost: 48616816198579.65
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 8, iteration: 1/20, moves: 7, ncost: 50528143911238.56
Run: 8, iteration: 2/20, moves: 5, ncost: 48837287201079.53
Run: 8, iteration: 3/20, moves: 1, ncost: 48777842651066.98
Run: 8, iteration: 4/20, moves: 0, ncost: 48777842651066.98
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 9, iteration: 1/20, moves: 2, ncost: 48616816198579.65
Run: 9, iteration: 2/20, moves: 0, ncost: 48616816198579.65
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 10, iteration: 1/20, moves: 22, ncost: 52302715084203.81
Run: 10, iteration: 2/20, moves: 8, ncost: 50057040964014.66
```

```
Run: 10, iteration: 3/20, moves: 1, ncost: 50020391397205.32
Run: 10, iteration: 4/20, moves: 0, ncost: 50020391397205.32
Best run was number 7
```

In [ ]:
```
# Appending the cluster data

data['Cluster'] = clusters
```

In [ ]:
```
# Average cost of the EV
data.EV_Price.mean()
```

Out[20]:
```
1194040.4040404041
```

In [ ]:
```
# Average cost of a car in segment 1
data.EV_Price[data.Cluster==0].mean()
```

Out[21]:
```
1633333.3333333333
```

In [ ]:
```
data['EV_Price'][data.Cluster==1].max()
```

Out[22]:
```
1600000
```

In [ ]:
```
# Average cost of a car in segment 1
data.EV_Price[data.Cluster==1].mean()
```

Out[23]:
```
1029305.5555555555
```

In [ ]:
```
data['Cluster'].value_counts(normalize=True) * 100
```

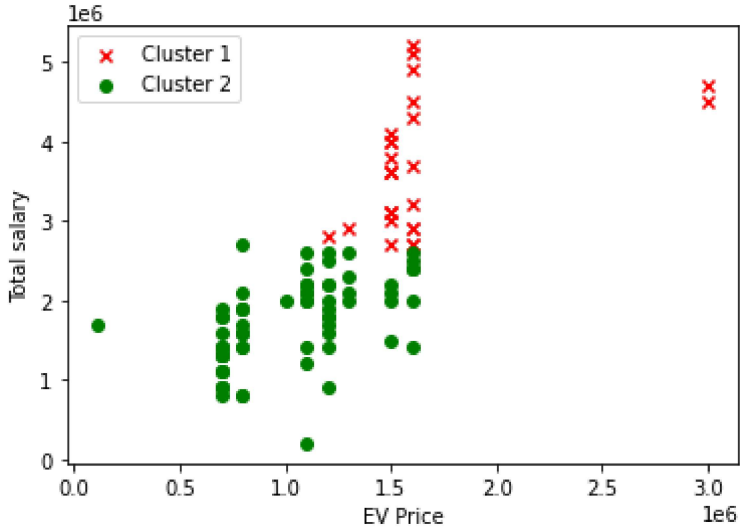
Out[24]:
```
1    72.727273
0    27.272727
Name: Cluster, dtype: float64
```

In [ ]:
```
# Segregrating each cluster

Cluster_0 = data[data.Cluster==0]
Cluster_1 = data[data.Cluster==1]
```

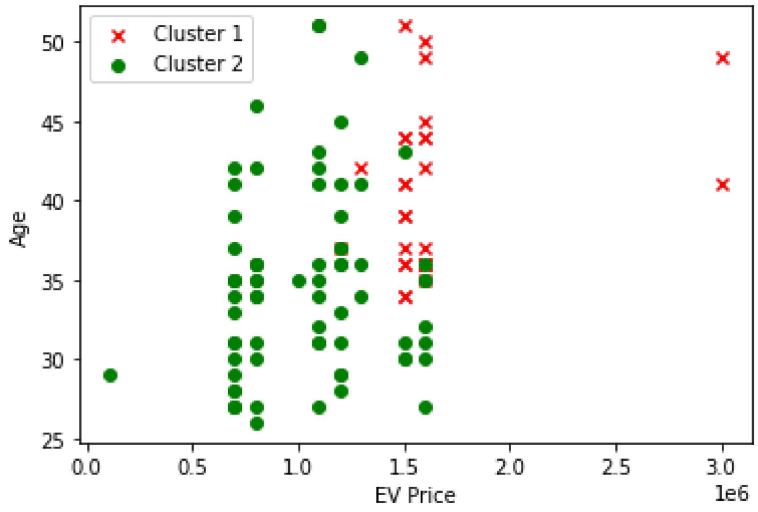
In [ ]:
```
data['Cluster'].value_counts()
```

Out[26]:
```
1    72
0    27
Name: Cluster, dtype: int64
```

In [ ]:
```
# plotting the effct of salary and ev price on cluster data

plt.scatter(Cluster_0.EV_Price, Cluster_0['Total Salary'],color='red', marke
plt.scatter(Cluster_1.EV_Price, Cluster_1['Total Salary'],color='green', lab
plt.legend(loc="upper left")

plt.xlabel('EV Price')
plt.ylabel('Total salary')
plt.show()

# there is a clear difference in segments when comparing salary and the pric
```

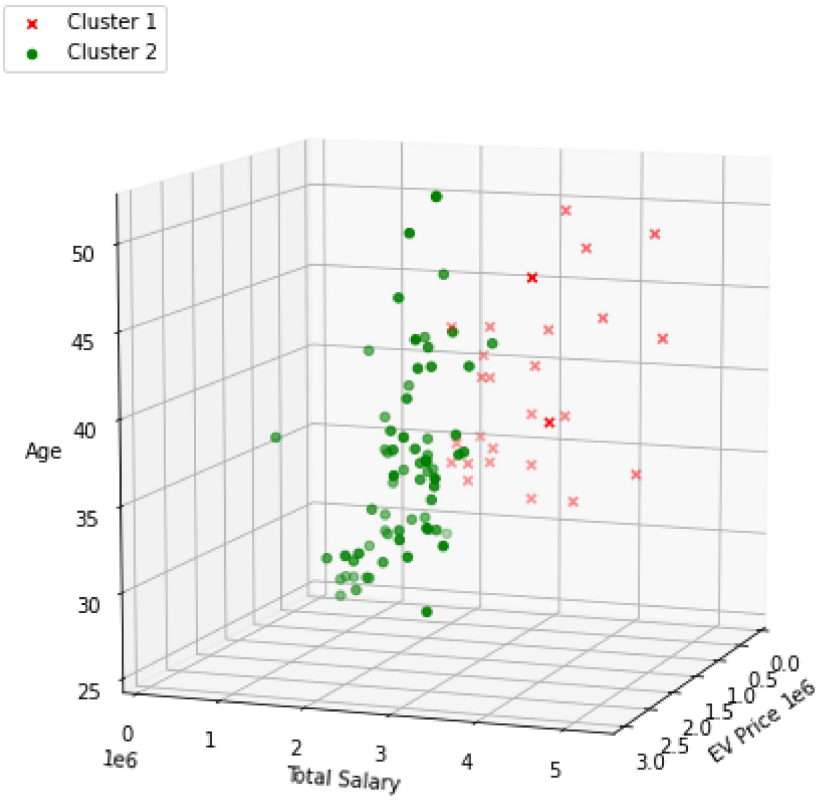


In [ ]:
```
plt.scatter(Cluster_0.EV_Price, Cluster_0['Age'],color='red', marker = 'x',
plt.scatter(Cluster_1.EV_Price, Cluster_1['Age'],color='green', label = 'Clu
plt.legend(loc = "upper left")

plt.xlabel('EV Price')
plt.ylabel('Age')
plt.show()
```

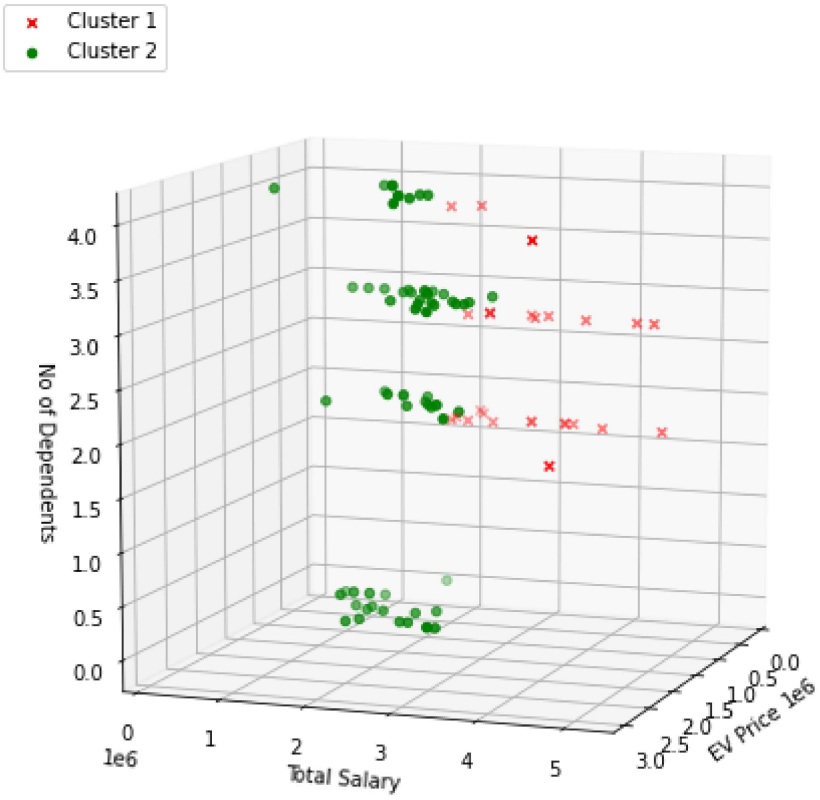


In [ ]:
```
from mpl_toolkits.mplot3d import Axes3D
```

In [ ]:
```
# plotting influence of age

fig = plt.figure(figsize=(8,8))

ax = fig.add_subplot(111, projection='3d')

ax.scatter(Cluster_0.EV_Price, Cluster_0['Total Salary'], Cluster_0['Age'],
ax.scatter(Cluster_1.EV_Price, Cluster_1['Total Salary'],Cluster_1['Age'], c
plt.legend(loc = 'upper left')

ax.view_init(10, 20)

plt.xlabel("EV Price")
plt.ylabel("Total Salary")
ax.set_zlabel('Age')
plt.show()
```



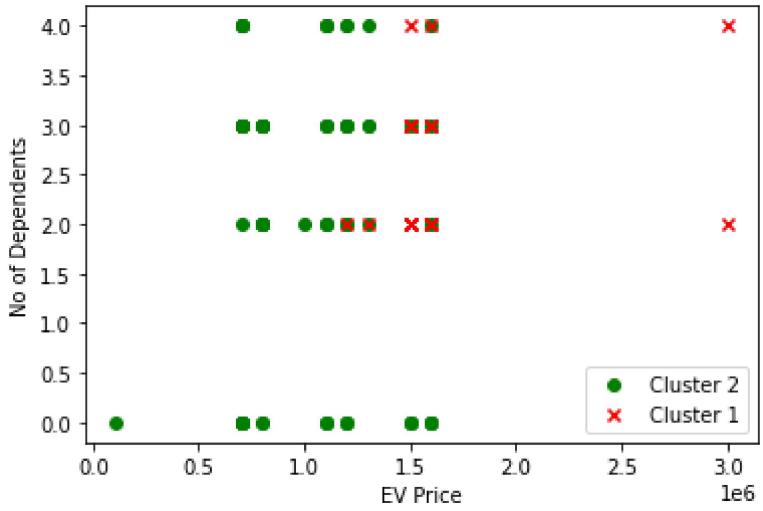
In [ ]:
```
# plotting influence of No of Dependents

fig = plt.figure(figsize=(8,8))

ax = fig.add_subplot(111, projection='3d')

ax.scatter(Cluster_0.EV_Price, Cluster_0['Total Salary'], Cluster_0['No of D
ax.scatter(Cluster_1.EV_Price, Cluster_1['Total Salary'],Cluster_1['No of De
plt.legend(loc = 'upper left')
ax.view_init(10,20)

plt.xlabel("EV Price")
plt.ylabel("Total Salary")
ax.set_zlabel('No of Dependents')
plt.show()
```



In [ ]:
```
data['No of Dependents'].value_counts()
```

Out[40]:
```
3    34
2    29
0    22
4    14
Name: No of Dependents, dtype: int64
```

In [ ]:
```
# plotting the effct of no of dependents and ev price on cluster data


plt.scatter(Cluster_1.EV_Price, Cluster_1['No of Dependents'],color='green',
plt.scatter(Cluster_0.EV_Price, Cluster_0['No of Dependents'],color='red', m
plt.legend(loc="lower right")

plt.xlabel('EV Price')
plt.ylabel('No of Dependents')
plt.show()

# there is a clear difference in segments when comparing salary and the pric
```



In [ ]: