Indian Population Buying Behavior Study

Author: Nikhil Mane

To successfully enter a competitive market, it is crucial to have an in-depth understanding of the psychology and behavior of the target end users. This comprehensive market research is imperative for multiple reasons, such as setting optimal pricing strategies, analyzing consumer spending habits, and identifying the specific products that dominate their usage patterns. For instance, in the automobile sector, it is vital to study whether consumers prefer 4-wheel diesel or petrol vehicles, and to understand the price range they are willing to consider. Additionally, it is essential to determine the specific requirements and features they seek in an automobile, such as fuel efficiency, brand reputation, safety features, and technology integration.

Understanding automobile buying behavior in India is critical for manufacturers and marketers aiming to cater to diverse consumer needs. The automotive market in India is influenced by a range of factors, including income levels, family size, and educational background.

In the following series of visualizations, we delve into this niche market by conducting a thorough requirement analysis. These visualizations will provide insights into consumer preferences and behaviors, enabling us to make informed decisions that align with the needs and expectations of our target audience. By carefully analyzing these factors, we aim to position our product effectively within the market and maximize its appeal to potential buyers.

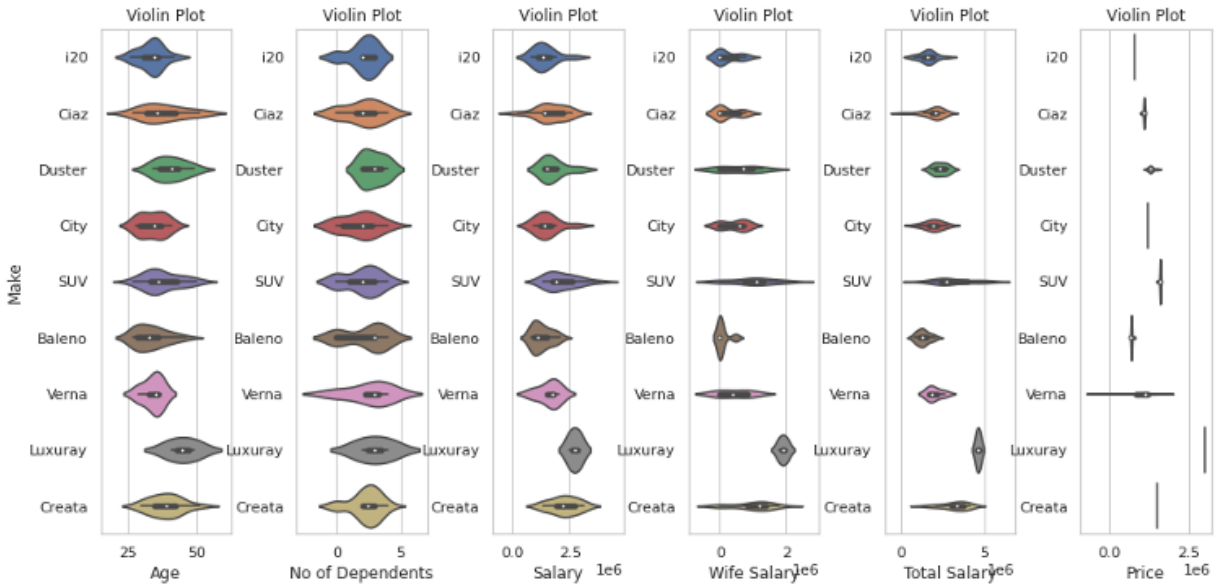
Below is a look at the dataset used:

	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
5	28	Salaried	Married	Graduate	3	Yes	Yes	No	900000	0	900000	Baleno	700000
6	31	Salaried	Married	Graduate	4	No	No	Yes	1200000	600000	1800000	City	1200000
7	33	Business	Married	Post Graduate	4	No	No	No	1400000	0	1400000	Baleno	700000
8	34	Business	Married	Post Graduate	4	No	No	No	2000000	0	2000000	Verna	1100000
9	34	Salaried	Married	Graduate	3	Yes	Yes	Yes	1200000	700000	1900000	i20	800000

The dataset used for this analysis contains several demographic and economic attributes, including:

- **Demographic Variables**: Profession, Marital Status, Education, Number of Dependents.
- **Economic Variables**: Total Salary, Price of the Vehicle.
- **Ownership Variables**: Make, Model of the vehicle.

Data Pre-processing: Missing values were handled by dropping rows where necessary. Categorical variables were encoded for clustering and trend analysis.



The series of violin plots offers a comprehensive view of how various demographic and economic factors influence the choice of vehicle models among buyers in India. Each plot reveals the distribution of a specific variable across different car models, allowing us to identify key trends and preferences within the market.

Starting with **age**, the data shows a clear preference for "Luxury" vehicles among older buyers, while younger consumers are more inclined towards models like the "i20" and "Ciaz." This suggests that as individuals age and potentially gain more financial stability, they may be more likely to invest in higher-end vehicles. Conversely, younger buyers, possibly at earlier stages of their careers, opt for more affordable, practical options.

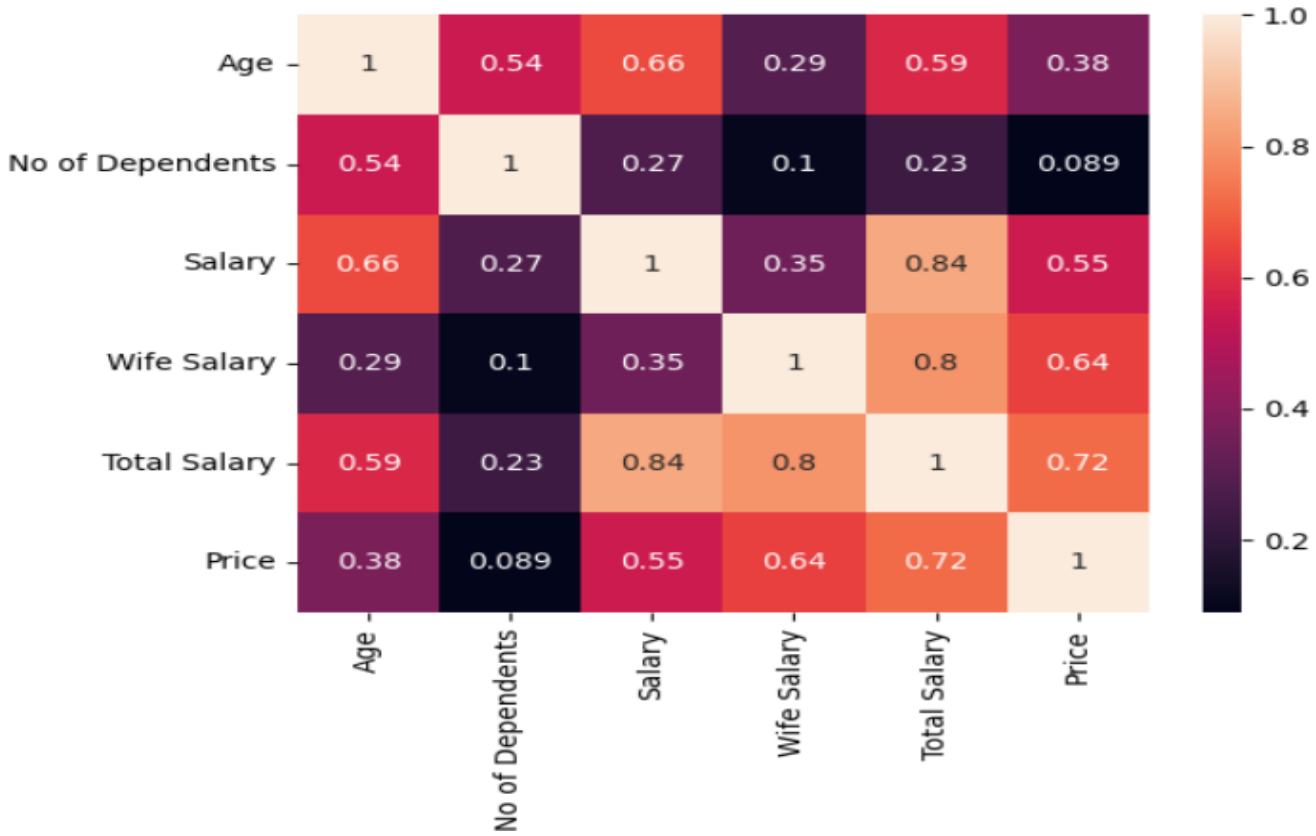
The analysis of **dependents** highlights a distinct pattern where cars like the "Creta" and "City" are favoured by buyers with larger families. This could be due to the spaciousness and practicality these models offer, making them more suitable for family needs. On the other hand, models categorized as "Luxury" are more commonly purchased by individuals with fewer dependents, indicating that these buyers may prioritize personal luxury and status over family-oriented practicality.

When examining **salary**, a strong correlation emerges between income levels and vehicle choice. High earners predominantly choose "Luxury" vehicles, reflecting their ability to afford premium models. Meanwhile, cars such as the "Duster" and "Verna" display a broader salary range among their buyers, indicating their appeal to a wider economic demographic. This trend is consistent across the

plots for both **wife's salary** and **total household income**, reinforcing the idea that household financial capacity significantly impacts vehicle preferences.

Finally, the **price** violin plot further corroborates the income-related trends observed earlier. As expected, "Luxury" vehicles are positioned at the higher end of the price spectrum, aligning with their association with higher-income buyers. In contrast, other models like the "i20" and "Baleno" cater to those with more moderate financial means, showing a more even distribution of vehicle prices across a range of buyers.

In conclusion, the violin plots clearly illustrate the impact of age, family size, and income on automobile purchasing behavior in India. Older, wealthier individuals with fewer dependents tend to gravitate towards luxury cars, while wenger, middle-income buyers with larger families prefer more affordable, practical vehicles. These insights can be instrumental in shaping targeted marketing strategies, enabling manufacturers and marketers to better address the needs and preferences of distinct consumer segments.



In the above heatmap, starting with **age**, we observe a moderate positive correlation with both **salary** (0.66) and **total household salary** (0.59). This indicates that as individuals' age, their income tends to increase, reflecting typical career advancement and greater financial stability over time. Interestingly, there is also a positive correlation between **age** and the **number of dependents** (0.54),

suggesting that older individuals may have larger families. However, the correlation between **age** and **vehicle price** is weaker (0.38), indicating that while older individuals may earn more, this doesn't strongly influence their choice to purchase more expensive vehicles.

When examining the relationship between **salary** and other variables, we see a strong positive correlation with **total household salary** (0.84), as expected, since total income is a combination of individual salaries. There is also a notable correlation between **salary** and **vehicle price** (0.55), implying that higher earners tend to purchase more expensive cars. This trend is further supported by the correlation between **wife's salary** and **vehicle price** (0.64), indicating that households where both partners have higher incomes are more likely to invest in pricier vehicles.

The **total household salary** exhibits strong correlations with both **salary** (0.84) and **wife's salary** (0.80), underscoring the importance of combined household income in determining purchasing power. Additionally, the correlation between **total household salary** and **vehicle price** (0.72) is significant, suggesting that households with higher combined incomes are more likely to afford luxury or higher-end vehicles.

Lastly, the **number of dependents** shows a weak correlation with most variables, except for a moderate correlation with **age** (0.54). This suggests that while family size might be related to the age of the buyer, it doesn't have a strong direct influence on income levels or vehicle price.

KMeans Clustering & Algorithms

Clustering is one of the foundational techniques in exploratory data analysis (EDA), widely used to gain insights into the underlying structure of a dataset. This method involves grouping data points into distinct subgroups, known as clusters, where each cluster contains data points that exhibit a high degree of similarity to one another. Conversely, data points from different clusters are notably different from each other. The primary goal of clustering is to uncover these homogeneous subgroups within the dataset, ensuring that the data points within a single cluster are as similar as possible based on a chosen similarity measure. Common measures of similarity include Euclidean distance, which considers the straight-line distance between points in a multi-dimensional space, and correlation-based distance, which assesses how closely data points follow the same pattern.

The choice of similarity measure is critical and should be tailored to the specific application at hand. For instance, Euclidean distance might be more appropriate in scenarios where the physical distance between data points is relevant, while correlation-based measures might be better suited for identifying patterns in data where the direction of change is more important than the magnitude.

Clustering analysis can be approached in two primary ways: feature-based clustering and sample-based clustering. In feature-based clustering, the focus is on finding subgroups of data samples that are similar based on a set of features or attributes. This approach is useful when the goal is to segment

the data into groups of similar observations. On the other hand, sample-based clustering involves identifying subgroups of features that are similar based on the samples. This method is particularly useful in scenarios where the relationships between variables or attributes are of primary interest, allowing for the identification of patterns across different features rather than across data points.

The **K-Means algorithm** is an iterative method designed to divide a dataset into a specified number of distinct, non-overlapping subgroups known as clusters, where each data point is assigned exclusively to one cluster. The core objective of the algorithm is to create clusters in which the data points within each cluster are as similar to each other as possible, while simultaneously ensuring that the different clusters are as distinct or far apart from each other as possible.

The algorithm achieves this by assigning data points to clusters in a way that minimizes the sum of the squared distances between each data point and the centroid of its assigned cluster. The centroid of a cluster is calculated as the arithmetic mean of all data points within that cluster, serving as a central point that represents the cluster. The goal is to reduce the variation within each cluster, so that the data points in a cluster are more homogeneous or similar to each other.

To elaborate, during the K-Means clustering process, the algorithm begins by randomly initializing the centroids of the clusters. It then iteratively refines the clusters by assigning each data point to the nearest centroid and recalculating the centroids based on the new cluster assignments. This process repeats until the cluster assignments stabilize, meaning that data points no longer switch between clusters, or until a predefined number of iterations is reached.

The effectiveness of the K-Means algorithm largely depends on how well it minimizes the within-cluster variance. When the variance within clusters is minimized, the algorithm successfully groups similar data points together, resulting in well-defined, homogeneous clusters. This makes K-Means a powerful tool for tasks such as market segmentation, image compression, and pattern recognition, where the goal is to uncover meaningful subgroups within the data

---

Clustering Vehicle Data Based on Price: A Detailed Analysis Using K-Means

The first step in the clustering workflow involved data pre-processing, specifically utilizing a label encoder. Label encoding is a crucial technique used to convert categorical variables into numerical values, which are more suitable for machine learning algorithms. By applying label encoding, we transformed the categorical data within the dataset into a numerical format. This step is particularly important for algorithms like K-Means, which rely on numerical inputs to calculate distances between data points. Ensuring that all categorical variables were appropriately encoded allowed for a smoother clustering process later on.

Price, being a continuous and numerical feature, is a significant factor that can reveal natural groupings within the dataset, such as differentiating between budget, mid-range, and luxury vehicles. By focusing on the price as the primary feature for clustering, we aimed to uncover distinct groups of vehicles that share similar pricing characteristics. This approach can be particularly insightful for market segmentation, helping to identify different price tiers within the vehicle market.

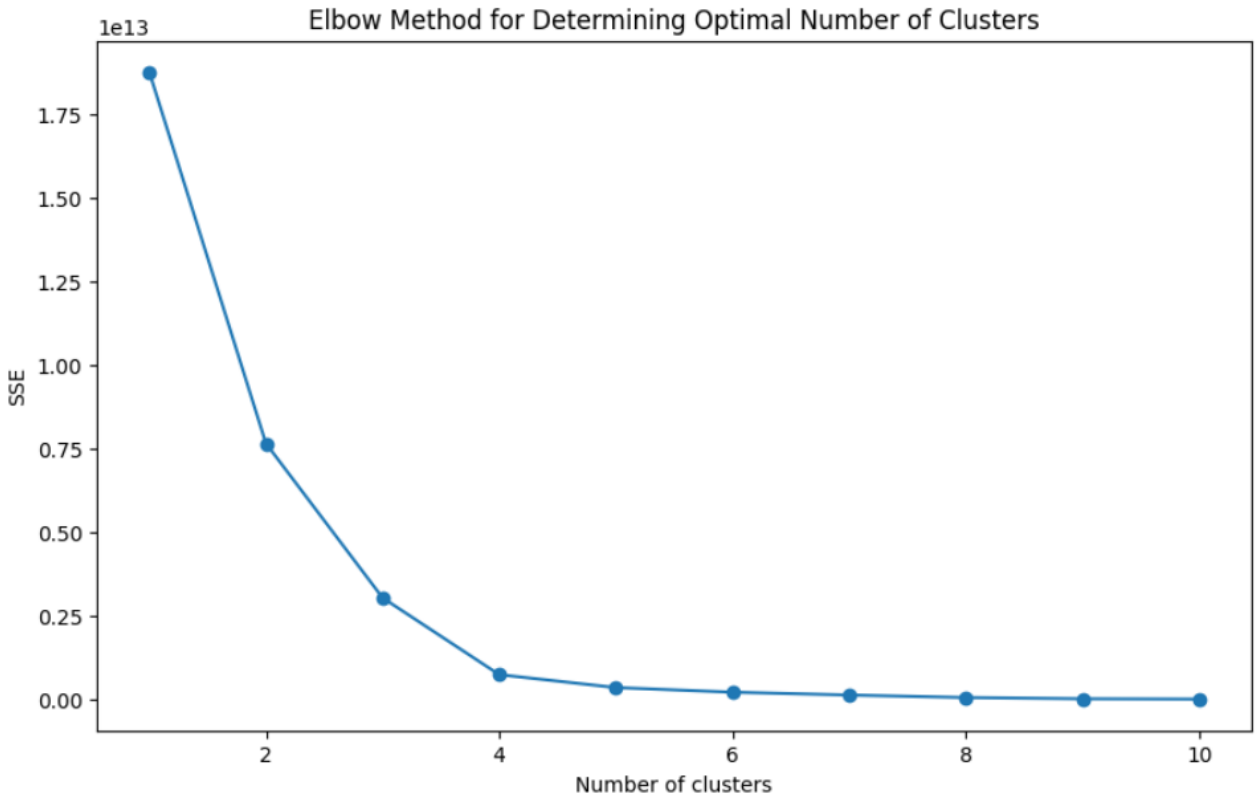
	Age	Profession	Marrital Status	Education	No of Dependents	Personal loan	House Loan	Wife Working	Salary	Wife Salary	Total Salary	Price
0	27	1	1	1	0	1	0	0	800000	0	800000	800000
1	35	1	0	1	2	1	1	1	1400000	600000	2000000	1000000
2	45	0	0	0	4	1	1	0	1800000	0	1800000	1200000
3	41	0	0	1	3	0	0	1	1600000	600000	2200000	1200000
4	31	1	0	1	2	1	0	1	1800000	800000	2600000	1600000
...	...	...	...	...	...	...	...	...	...	...	...	...
94	27	0	1	0	0	0	0	0	2400000	0	2400000	1600000
95	50	1	0	1	3	0	0	1	3800000	1300000	5100000	1600000
96	51	0	0	0	2	1	1	0	2200000	0	2200000	1100000
97	51	1	0	1	2	0	0	1	2700000	1300000	4000000	1500000
98	51	1	0	1	2	1	1	0	2200000	0	2200000	1100000

99 rows × 12 columns

After performing the initial clustering based on price, we decided to drop the 'Make' column from the dataset. The 'Make' column, which likely represented the manufacturer or brand of the vehicles, was removed to eliminate any potential bias or redundancy that could interfere with the clustering results. By excluding this categorical feature, we allowed the clustering algorithm to focus solely on the numerical attributes of the vehicles, such as price, leading to more meaningful and unbiased clusters. We then assigned the modified dataset to a new variable, preparing it for further analysis.

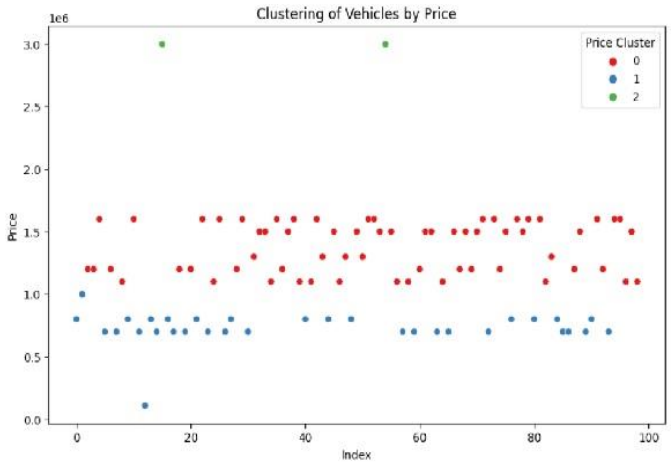
With the data prepared and cleaned, we implemented the K-Means clustering algorithm. K-Means is a powerful method for partitioning data into distinct clusters by minimizing the variance within each cluster. During this step, we assigned each vehicle in the dataset to one of the clusters based on its price, iteratively refining the cluster centroids to achieve the best possible partitioning. The K-Means algorithm repeatedly assigned data points to the nearest centroid and recalculated the centroids until the clusters stabilized, ensuring that each cluster contained vehicles with similar price points.

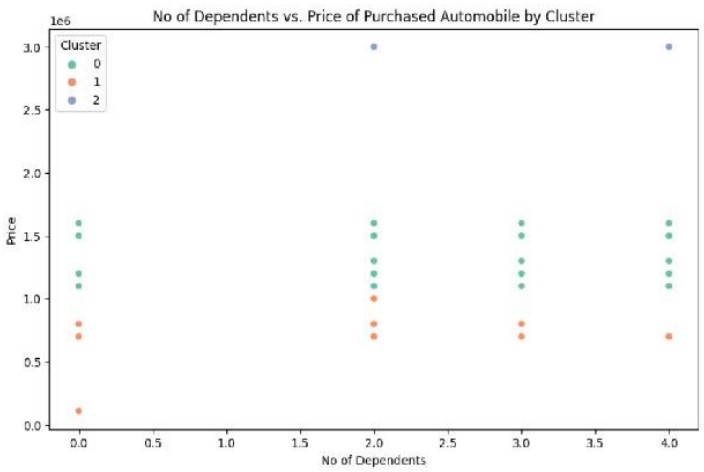
To identify the optimal number of clusters for the dataset, we employed the elbow method. The elbow method is a graphical tool that helps determine the point at which adding more clusters no longer significantly reduces the within-cluster variance. By plotting the sum of squared distances between data points and their respective cluster centroids against the number of clusters, we observed the characteristic "elbow" in the plot. This elbow indicates the optimal number of clusters, where further increases in the number of clusters result in diminishing returns. In the case, the optimal number of clusters was identified as three, meaning that the data naturally grouped into three distinct clusters based on vehicle price

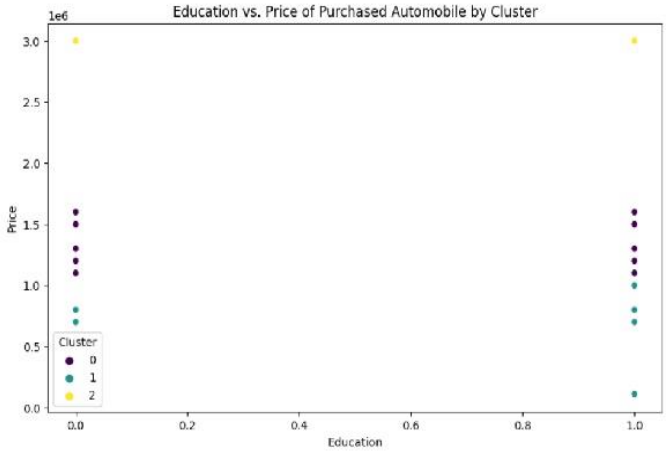


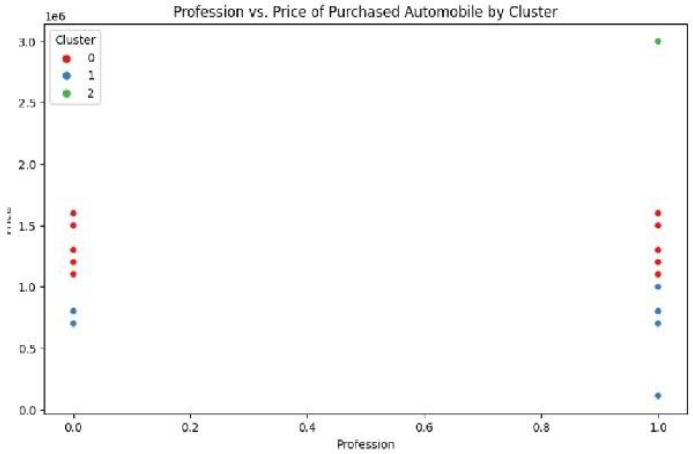
.

To better understand and visualize the clusters identified by the K-Means algorithm, we created scatterplots with the clusters as the hue. Scatterplots are an effective way to visualize the distribution of data points across different clusters, especially when dealing with two or three key features. By adding the clusters as the hue, we could easily differentiate between the clusters in the visual representation. This allowed us to observe how well the clustering algorithm separated the vehicles based on price and provided insights into the characteristics of each cluster. The scatterplots likely revealed the distinct groups within the data, highlighting any patterns or trends that emerged from the clustering process.









The visualizations provided depict the results of a clustering analysis based on the price of vehicles, offering valuable insights into how different vehicle price segments are grouped and the potential influence of demographic factors on vehicle purchasing decisions.

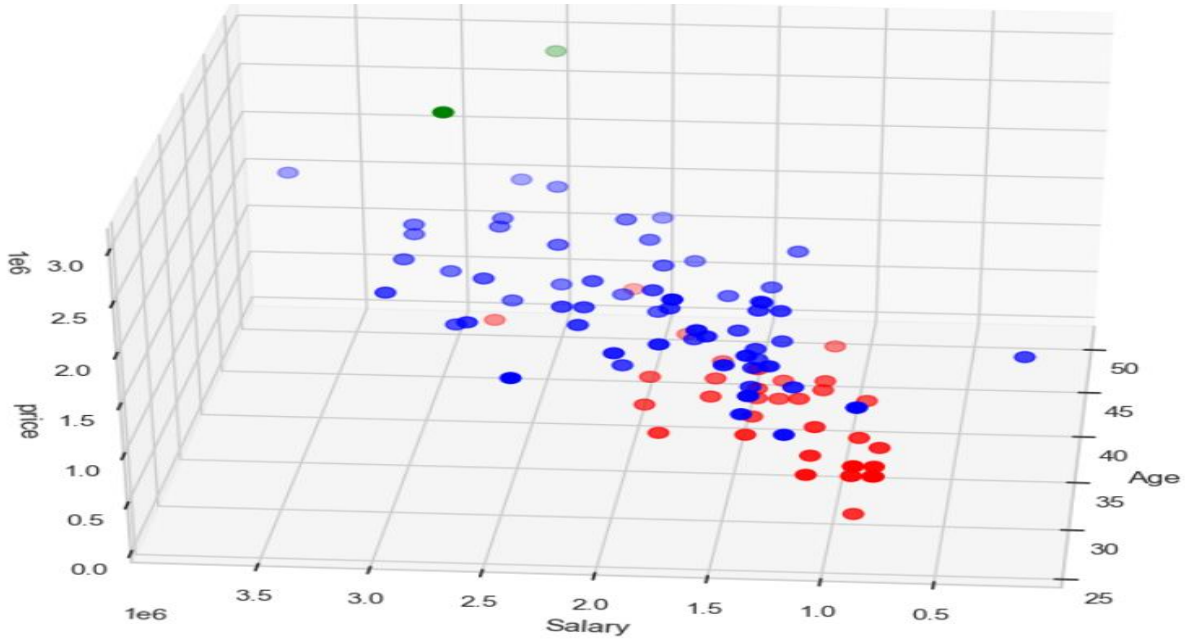
The first scatterplot, titled "Clustering of Vehicles by Price," visually separates the vehicles into three distinct clusters. Cluster 1, represented in blue, generally includes lower-priced vehicles, Cluster 0, shown in red, and captures mid-priced vehicles, while Cluster 2, marked in green, comprises higher-priced vehicles. This clear segmentation reveals that the clustering algorithm has successfully identified groups of vehicles with similar price ranges, emphasizing the presence of distinct market segments. Interestingly, Cluster 2 contains significantly fewer data points, indicating that high-priced vehicles are less common, likely representing luxury or premium models. The greater density of vehicles in Clusters 0 and 1 suggests that the majority of vehicles fall within the low to mid-price range, which could be indicative of broader market demand in these segments.

The second scatterplot, which explores the relationship between the number of dependents and the price of the purchased automobile, provides further insights into consumer behavior. Despite the

variation in the number of dependents, the price of vehicles purchased appears to be consistent across all clusters, suggesting that the number of dependents does not significantly influence the price range of vehicles that buyers select. The distribution of vehicles in Cluster 2 (higher-priced vehicles) across all dependent categories implies that individuals purchasing luxury vehicles do so regardless of family size. This lack of a strong correlation between dependents and vehicle price suggests that other factors, such as income level, lifestyle, or personal preferences, may play a more decisive role in the purchasing decision.

The third scatterplot examines the relationship between education level and vehicle price across the identified clusters. Here again, the data indicates that there is no strong correlation between education level and the price of the purchased vehicle, except for a slight tendency for higher-educated individuals to purchase vehicles from Cluster 2 (the highest-priced segment). This pattern suggests that while education may have some influence on the decision to purchase higher-priced vehicles, it is not a definitive factor, as individuals with varying education levels are found in both lower and mid-priced segments. The diversity in the education levels of buyers across Clusters 0 and 1 further emphasizes that vehicle purchasing decisions in these segments are likely influenced by a variety of factors, not just education.

Overall, these visualizations provide a comprehensive view of how vehicles are clustered by price and the extent to which demographic factors like the number of dependents and education level influence purchasing decisions. The clear segmentation of vehicles into distinct price clusters highlights potential market segments, while the lack of strong correlations between demographic factors and vehicle price suggests that other variables, such as income, brand loyalty, or specific consumer preferences, might be more influential in determining the vehicles consumers choose. These insights can be instrumental in shaping marketing strategies, product development, and targeting efforts within the automotive industry.



This 3D scatterplot provides valuable insights into how vehicle prices are influenced by the buyer's salary and age. The clear segmentation of clusters highlights distinct market segments, with younger, lower-salary individuals tending to purchase more affordable vehicles, and older, higher-salary individuals gravitating towards more expensive options. The outlier cluster, marked in green, suggests a small but significant market for luxury vehicles among high-income individuals.

These insights can be critical for automotive businesses in tailoring their marketing strategies and product offerings. For instance, targeting luxury vehicles towards older, wealthier consumers, while offering more affordable options to younger buyers, could align with the purchasing behaviors illustrated in this scatterplot. Additionally, understanding the correlation between salary and vehicle price can help in creating financing options or packages that cater to different income segments within the market.

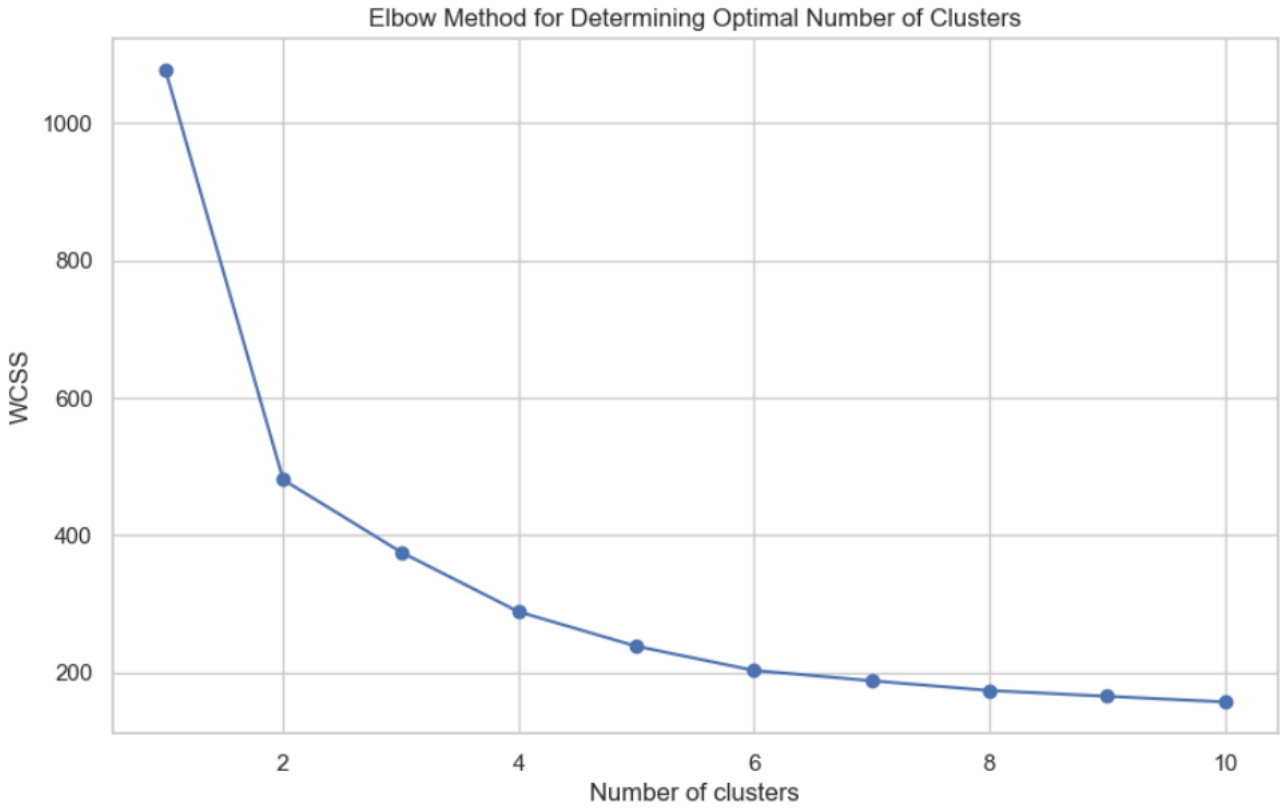
Clustering vehicle data based on Make

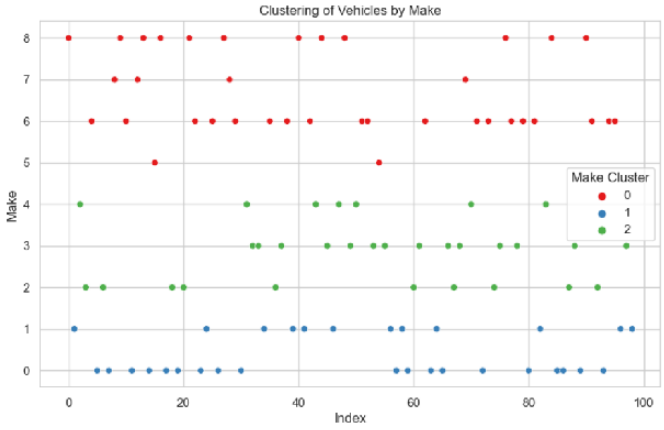
To gain additional insights into the data, we performed a second round of clustering analysis, this time focusing on the **make** of the vehicles. Instead of clustering based on the price, as in the previous analysis, we dropped the price column and used the make of the vehicle as the key variable for clustering. This approach allowed for the identification of patterns and similarities between different vehicle makes, independent of their price tags.

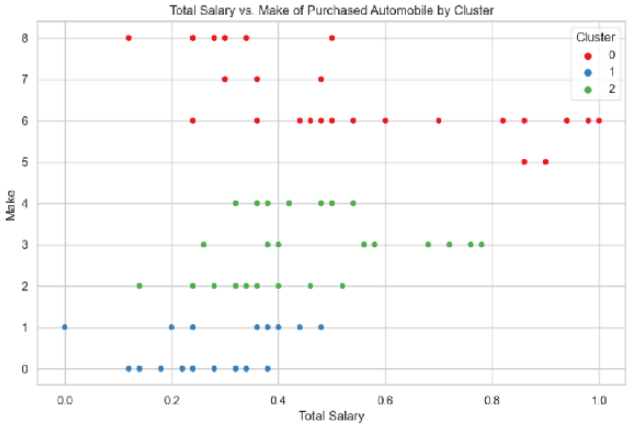
The clustering process revealed how various vehicle makes grouped together, providing a clearer understanding of how different manufacturers or brands may share certain characteristics or appeal to similar segments of buyers. By clustering on the make of the vehicle, we were able to identify distinct market segments that may not have been as apparent in the price-based analysis.

	Age	Profession	Marrital Status	Education	No of Dependents	Personal loan	House Loan	Wife Working	Salary	Wife Salary	Total Salary	Make
0	27	1	1	1	0	1	0	0	800000	0	800000	8
1	35	1	0	1	2	1	1	1	1400000	600000	2000000	1
2	45	0	0	0	4	1	1	0	1800000	0	1800000	4
3	41	0	0	1	3	0	0	1	1600000	600000	2200000	2
4	31	1	0	1	2	1	0	1	1800000	800000	2600000	6
...	...	...	...	...	...	...	...	...	...	...	...	...
94	27	0	1	0	0	0	0	0	2400000	0	2400000	6
95	50	1	0	1	3	0	0	1	3800000	1300000	5100000	6
96	51	0	0	0	2	1	1	0	2200000	0	2200000	1
97	51	1	0	1	2	0	0	1	2700000	1300000	4000000	3
98	51	1	0	1	2	1	1	0	2200000	0	2200000	1

99 rows × 12 columns







After performing clustering based on the make of the vehicle, several insights can be drawn. By focusing on the make instead of price, we can explore the purchasing patterns in relation to the specific brands chosen by customers. The resulting visualizations demonstrate how different makes cluster together based on various factors like salary and other demographic data.

In the first scatter plot, where clustering is based solely on the make of the vehicle, it is evident that distinct clusters form along specific make categories. This suggests a strong brand preference among different customer segments, possibly influenced by other variables such as income, age, or other demographic factors.

The second plot, which examines total salary against the make of the purchased automobile by cluster, reveals patterns in brand preference relative to the buyer's income. For instance, certain clusters are dominated by high-salary individuals purchasing specific makes, indicating a correlation between income levels and brand choice. The spread of clusters across various salary ranges also points to diverse market segments, each with unique preferences and affordability criteria.

Overall, clustering based on the make of the vehicle rather than price offers a nuanced perspective on consumer behavior. It highlights how brand loyalty, influenced by factors like salary, can significantly impact purchasing decisions, providing valuable insights for targeted marketing and product positioning strategies.

Target Segments

High-Income, Brand-Conscious Consumers: The visualizations reveal a distinct cluster of high-income individuals who show a strong preference for premium vehicle brands. These consumers tend to gravitate towards well-known, luxurious makes, as indicated by the high concentration of high-salary individuals in certain clusters. These customers are less sensitive to price and more focused on brand reputation, quality, and the status that comes with owning a high-end vehicle. Their brand loyalty suggests they value not just the vehicle's functionality but also the prestige associated with specific makes.

Mid-Income, Value-Oriented Consumers: Another significant cluster consists of mid-income individuals who exhibit a balanced approach to their vehicle purchases. They lean towards brands that offer a good mix of quality, reliability, and affordability. This segment is more price-sensitive compared to the high-income group but still places a strong emphasis on brand reputation. They are likely to choose brands known for their long-term value, good resale potential, and solid customer service.

Low-Income, Practical Buyers: The final cluster consists of lower-income individuals who prioritize practicality and affordability in their vehicle choices. These consumers are most likely to be influenced by the initial purchase price, maintenance costs, and fuel efficiency. They prefer brands that offer budget-friendly models with essential features, reliability, and low running costs.

Marketing EVs to the Target Segments

Promoting EVs to different income segments requires a nuanced approach that speaks directly to the values and concerns of each group. High-income consumers are driven by luxury, innovation, and exclusivity, so the marketing strategy should focus on premium features and a bespoke ownership

experience. Mid-income consumers value practicality, reliability, and long-term cost savings, so messaging should emphasize the economic and functional benefits of EVs. Low-income consumers prioritize affordability and ease of use, so the strategy should focus on making EVs accessible and highlighting the day-to-day cost savings. By tailoring the promotion of EVs to these specific segments, brands can more effectively reach and convert potential buyers across the income spectrum, driving broader adoption of electric vehicles.

GITHUB REPO

Buying Behaviour