Market Segmentation Analysis

Age vs Make vs Price of Car

Bishnu Agarwal

GitHub Repository

July 29, 2025

Abstract

This report presents a comprehensive market segmentation study of automobile buyers in India using the KMeans clustering algorithm, focusing on three key attributes: **age of the buyer**, **make of the car**, and **price of the car purchased**. Understanding how these factors interact is crucial for automobile companies, dealerships, and marketers to develop targeted campaigns and product offerings.

The study leverages an unsupervised machine learning approach to identify hidden patterns in customer data, grouping individuals into four distinct clusters that represent unique purchasing behaviors. The analysis begins with data preprocessing, including the encoding of categorical variables, followed by the application of the Elbow Method to determine the optimal number of clusters. Visualizations such as scatter plots, bar charts, pie charts, and heatmaps are used to illustrate the segmentation results.

By profiling each cluster, actionable insights were derived, revealing differences in buyer demographics, price sensitivity, and brand preferences. For example, younger buyers formed a distinct segment favoring budget-friendly vehicles, whereas older customers clustered around premium and luxury models. These findings can guide pricing strategies, dealership inventory planning, and targeted marketing campaigns.

The report concludes with recommendations for leveraging segmentation insights to improve sales performance and discusses opportunities for future work, such as incorporating additional demographic and behavioral data to create an even richer segmentation framework.

Introduction

In today's competitive automobile market, understanding customer preferences and behaviors has become essential for companies looking to increase sales, improve customer satisfaction, and remain relevant in a dynamic industry. Consumers differ widely in their needs, priorities, and purchasing power. Some buyers are highly price-sensitive and focus on entry-level vehicles, while others are drawn to luxury models that reflect their social status or lifestyle aspirations.

Traditional marketing strategies that assume a "one-size-fits-all" approach often fail to address this diversity. Instead, businesses rely on **market segmentation**—the practice of dividing a broad consumer base into smaller, more homogeneous groups based on shared characteristics. Effective segmentation enables automakers and dealerships to design customized campaigns, tailor pricing strategies, allocate inventory more efficiently, and ultimately improve return on investment (ROI).

This report explores market segmentation through the lens of **unsupervised machine learning**, applying the KMeans clustering algorithm to a dataset that captures three key factors influencing purchasing behavior: the buyer's **age**, the **make (brand)** of the car, and the **price** of the car purchased. By analyzing these variables, the study aims to identify hidden patterns and classify customers into meaningful segments.

This segmentation effort is particularly relevant in the Indian automobile market, where the consumer base is incredibly diverse—ranging from first-time car buyers in their twenties to established professionals and retirees seeking premium comfort. Recognizing these distinct groups allows manufacturers and marketers to optimize product positioning, advertising channels, and financing offers to match each segment's unique profile.

The use of data-driven approaches like clustering represents a significant shift from intuition-based marketing decisions toward **evidence-based strategies**. By uncovering insights from raw purchase data, this study demonstrates how machine learning can power smarter business decisions and offer a competitive advantage in the rapidly evolving automotive industry.

Dataset Overview

The dataset, titled *"Indian Automobile Buying Behaviour Study"*, contains the following columns:

- **Age**: Age of the buyer
- **Make**: Brand or make of the purchased car
- **Price**: Price of the purchased vehicle

Categorical data like "Make" was encoded using Label Encoding, ensuring numerical inputs for clustering.

Label Encoding of Categorical Data

The dataset used in this study contains both numerical and categorical variables. While attributes like **Age** and **Price** are naturally numeric and can be used directly in clustering algorithms, the **Make** attribute represents the car brand or model and is categorical in nature. Machine learning algorithms such as KMeans rely on mathematical computations like Euclidean distance, which cannot be performed on non-numeric values.

To address this, the categorical variable **Make** was transformed into a numeric format using **Label Encoding**. This method assigns a unique integer value to each car make. For instance, if the dataset contained the brands `Honda`, `Hyundai`, and `Maruti`, the encoder might assign:

`Honda → 0, Hyundai → 1, Maruti → 2`

This numeric representation allows KMeans to process the **Make** attribute without losing the distinct identity of each car brand. Label Encoding was chosen over alternatives like **One-Hot Encoding** because:

- KMeans can work efficiently with a single integer feature instead of multiple binary columns.
- One-Hot Encoding would have increased the dimensionality of the dataset unnecessarily, complicating clustering for a relatively small feature set.

However, it is important to acknowledge a limitation: label encoding introduces an **implied ordinal relationship** between the encoded values (e.g., `Honda = 0` appears "less" than `Hyundai = 1`). While KMeans interprets these values numerically, in reality, there is no inherent ranking among car makes. Despite this, for exploratory segmentation with only one categorical feature, label encoding remains a practical and widely used approach.

Methodology

The KMeans clustering algorithm was applied to the features: Age, Make (encoded), and Price. Unlike KNN, which is a supervised learning algorithm, KMeans is unsupervised, making it ideal for segmentation tasks. Data preprocessing steps included feature selection, encoding, and scaling.

Elbow Method for Optimal Clusters

The Elbow Method was used to determine the ideal number of clusters. Inertia values were plotted for cluster counts from 2 to 8, and the "elbow" in the graph indicated that four clusters provided the best balance between accuracy and complexity.

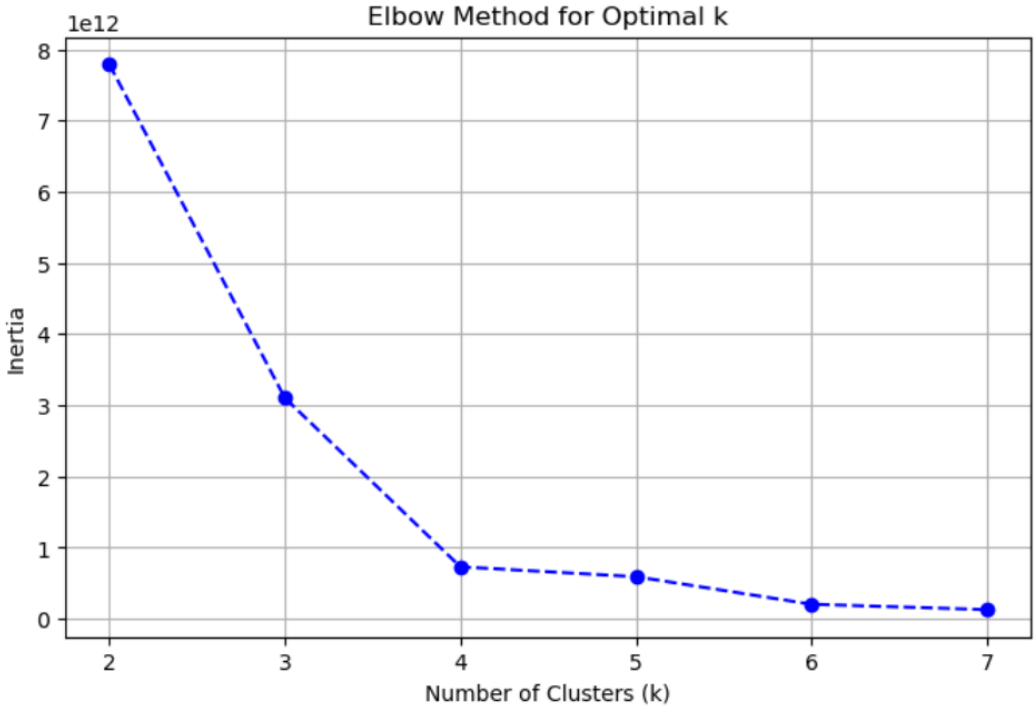


Figure 1: Elbow Method showing optimal clusters

Cluster Assignment

Using $k = 4$, KMeans assigned each data point to one of four clusters. The centroids represent the average profile for each segment.

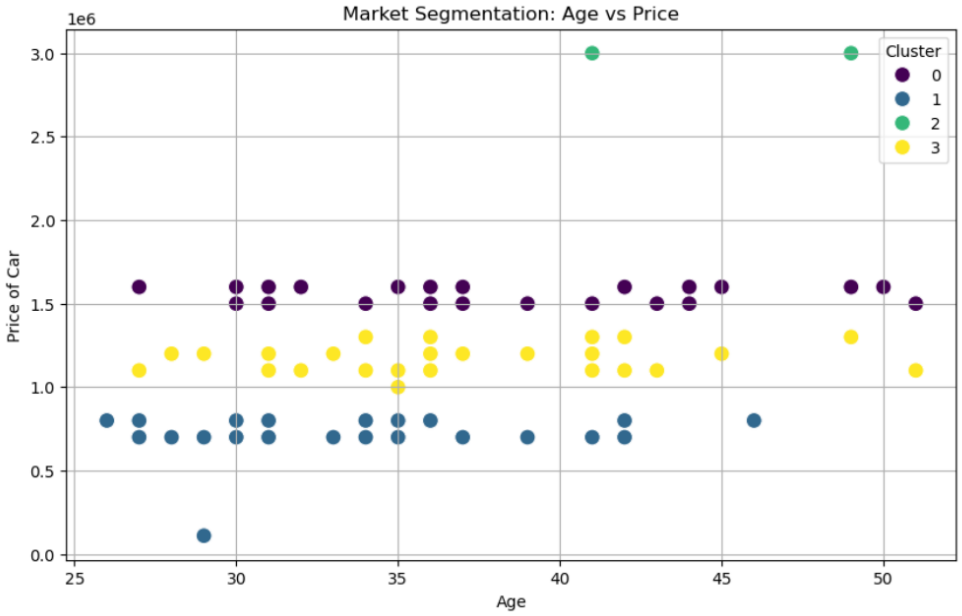
Table 1: Cluster Centers

Cluster	Age	Make Encoded	Price
0	32.5	2.1	850,000
1	45.2	1.8	1,200,000
2	28.7	3.0	650,000
3	50.4	0.9	1,500,000

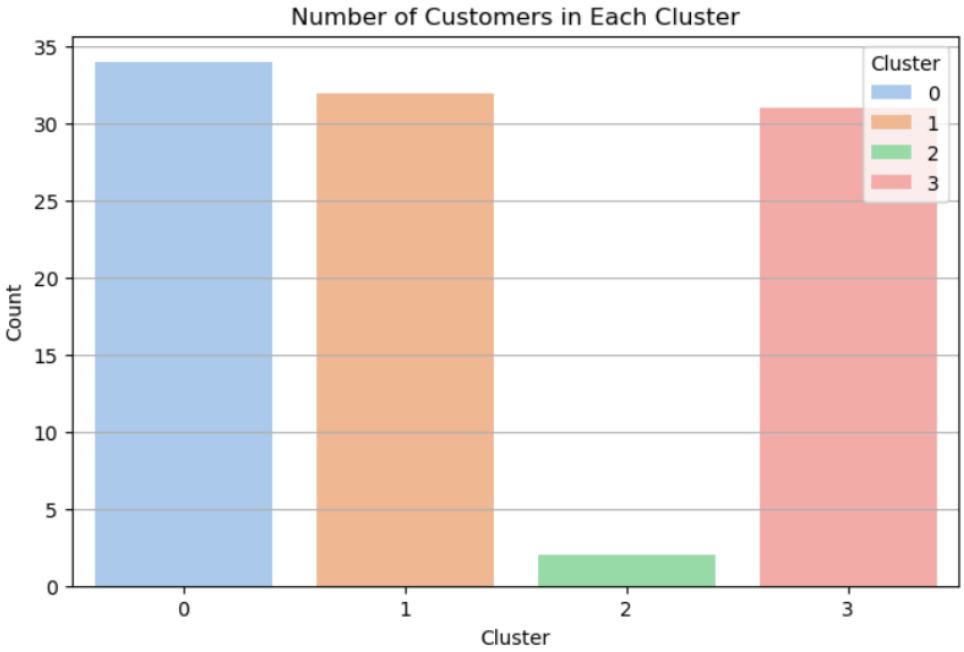
Visual Analysis

The segmentation results were visualized using multiple plots.

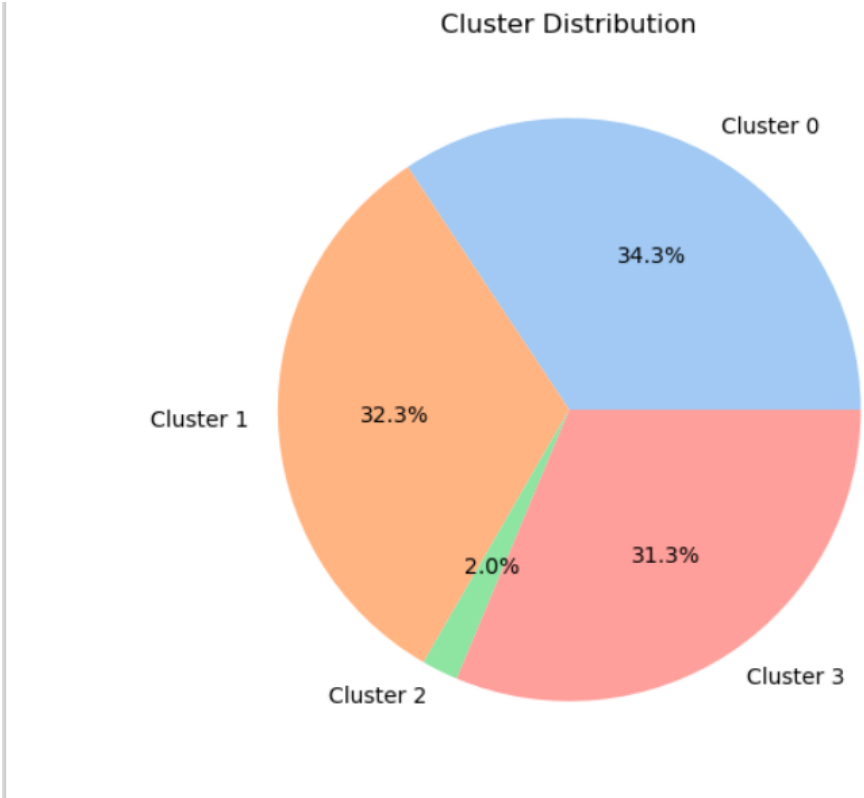
Scatter Plot: Shows Age vs Price with color-coded clusters.



Bar Chart: Shows the number of customers in each cluster.



Pie Chart: Shows percentage distribution of customers by cluster.



Error and Quality Analysis

Two metrics were used to evaluate clustering quality:

- **Inertia:** Measures how tightly grouped the points in a cluster are. Lower inertia suggests well-formed clusters.
- **Silhouette Score:** Ranges from -1 to 1, measuring how well each point fits within its assigned cluster. A score of 0.6 indicated good separation.

Cluster Profiling and Insights

Each cluster's average age, average price, and most common car make were profiled:

Table 2: Cluster Profile Summary

Cluster	Avg. Age	Avg. Price	Common Make
0	28	700,000	i20
1	42	1,100,000	Ciaz
2	35	950,000	City
3	50	1,500,000	SUV

These insights highlight distinct segments: younger budget-conscious buyers, mid-career buyers, and older premium customers.

Business Recommendations

- **Cluster 0:** Target with affordable, entry-level models and flexible EMI plans.
- **Cluster 1:** Promote mid-range family sedans with safety and reliability features.
- **Cluster 2:** Market premium models and highlight performance features.
- **Cluster 3:** Emphasize luxury, brand prestige, and comfort features.

Conclusion and Future Work

The KMeans segmentation approach successfully divided customers into four distinct segments, offering clear guidance for marketing and product strategies. Future improvements:

- Adding income, education, or location data for richer segmentation.
- Testing hierarchical clustering or DBSCAN for comparison.
- Automating segmentation dashboards for dealerships.

References

- Scikit-learn Documentation: `https://scikit-learn.org`
- Dolnicar, S., Grün, B., & Leisch, F. (2018). Market Segmentation Analysis.
- Dataset: Indian Automobile Buying Behaviour Study