Electrical Vehicle Market in India

Market segmentation

Adesh Mekhe

Problem Statement:

The task is to apply segmentation analysis to study the Indian market for electric vehicles and provide a workable entry strategy that focuses on the demographic, Behavioural, demographic, and geographic groups most likely to use the product.

Understand the Indian electric vehicle market by dividing it into different groups based on demographics, behaviour, geography, cost, vehicle type, retail stores, manufacturers, body style, plug types, safety, and other factors. Focus on the most promising segments to create an entry strategy. Establish a strong retail presence, improve charging infrastructure, and offer excellent after-sales service. Collaborate with local partners and monitor market trends for continuous improvement.

Data Collection:

The data has been collected manually, and the sources used for this process are listed below :

- https://www.kaggle.com/datasets
- https://data.gov.in/
- https://data.worldbank.org/
- https://datasetsearch.research.google.com/

Market Segmentation Target Market:

The target market of Electric Vehicle Market Segmentation can be categorised into Geographic, SocioDemographic, Behavioral, and Psychographic Segmentation

- Behavioural Segmentation: searches directly for similarities in behavior or reported behavior.
- Psychographic Segmentation: grouped based on beliefs, interests, preferences, aspirations, or benefits sought when purchasing a product. Suitable for lifestyle segmentation. Involves many segmentation variables
- Demographic Segmentation: includes age, gender, income and education. Useful in industries.

Implementation Packages/Tools used:

1. Numpy: To calculate various calculations related to arrays.

2. Pandas: To read or load the datasets.

Data-Preprocessing :

Data Cleaning

- The data collected is compact and is partly used for visualisation purposes and partly for clustering. Python libraries such as NumPy, Pandas, Scikit-Learn, and SciPy are used for the workflow, and the results obtained are ensured to be reproducible
- Two datasets are used and achieved the results from both the sets

EDA

- We start the Exploratory Data Analysis with some data Analysis drawn from the data without Principal Component Analysis and with some Principal Component Analysis in the dataset obtained from the combination of all the data we have.
- PCA is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. The process helps in reducing dimensions of the data to make the process of classification/regression or any form of machine learning, cost-effective.

Correlation Matrix

- A correlation matrix is simply a table that displays the correlation. It is best used in variables that demonstrate a linear relationship between each other. Coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values through the heat-map in the below figure.
- The relationship between two variables is usually considered strong when their correlation coefficient value is larger than 0.7

Extracting Segments :

Dendrogram

Agglomerative Hierarchical Clustering:

- Starts with each data point as a separate cluster.
- Joins clusters together in a hierarchical manner based on their distances.

Determining Optimal Number of Clusters:

- Utilizes a dendrogram, which is a tree-like chart.
- Dendrogram shows the sequence of merges or splits of clusters.
- When two clusters are merged, the dendrogram graphically represents this.

- The height of the join in the dendrogram corresponds to the distance between the clusters.
- The optimal number of clusters can be chosen based on the hierarchical structure of the dendrogram.

Cluster Validation Metrics:

- Other cluster validation metrics suggest that four to five clusters are suitable for the agglomerative hierarchical method.
- Please note that the passage is discussing how to use dendrograms in hierarchical clustering to determine the optimal number of clusters, with a suggestion of using four to five clusters based on cluster validation metrics.

Elbow Method

- The Elbow method is a popular method for determining the optimal number of clusters.
- The method is based on calculating the Within-Cluster-Sum of Squared Errors (WSS) for a different number of clusters (k) and selecting the k for which change in WSS first starts to diminish.
- The idea behind the elbow method is that the explained variation changes rapidly for a small number of clusters and then it slows down leading to an elbow formation in the curve. The elbow point is the number of clusters we can use for our clustering algorithm.
- The KElbowVisualizer function fits the KMeans model for a range of clusters values between 2 to 8.

Analysis and Approaches used for Segmentation:

Clustering

Clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean based distance or correlation-based distance.

The decision of which similarity measure to use is application-specific. Clustering analysis can be done on the basis of features where we try to find subgroups of samples based

on features or on the basis of samples where we try to find subgroups of features based on samples.

K-Means Algorithm

K Means algorithm is an iterative algorithm that tries to partition the dataset into pre-defined distinct non-overlapping subgroups (clusters) where each data point be14 longs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The **k-means clustering algorithm** performs the following tasks:

- Specify number of clusters K
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid). • Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing. According to the Elbow method, here we take K=4 clusters to train KMeans model.

Principal component analysis (PCA):

- Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation.
- PCA reduces the dimensionality of a dataset while preserving the maximum amount of information.
- PCA does this by linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data.
- Many studies use the first two principal components in order to plot the data in two dimensions and to visually identify clusters of closely related data points.

- PCA has applications in many fields such as population genetics, microbiome studies, and atmospheric science.

Conclusion:

conclusion for manual process: From the data and Its interpretation it clearly says that

1. for "Everything" usage [Gemopai Ryder,Evolet Polo] bikes are more preferred
2. for 'Daily Commute" usage [Hero Electric NYX HX] bike is more preferred
3. for "Occasional Commute" usage [Bajaj Chetak] bike is more preferred
4. for "Leisure Rides" usage [Yo Drift] bike is more preferred
5. for "Tours" usage [Hero Electric Optima] bike is more preferred

The above conclusions are made after analyzing all the data. i.e based on the ratings

given by each customers for their corresponding bike.

GitHub links:

Adesh Mekhe: https://github.com/adesh-mekhe/EV-Market-Segmentation-Analysis