Feynn Labs: Project 2

Link: <https://github.com/pankajsingh016/Feynn-Labs-Internship>

**Market Segmentation Analysis of Electric Vehicles Market in India**

**Pankaj Singh Kanyal**



*Fig1. EV Vehicles*

**Introduction**

The electric vehicle (EV) market has witnessed exponential growth in recent years, driven by environmental concerns, technological advancements, and supportive government policies. As the global focus shifts towards reducing carbon emissions and fostering sustainable development, EVs have emerged as a key solution to mitigate the impacts of climate change. This report explores the segmentation of the electric vehicle market, breaking down the diverse landscape of consumer demands, regional influences, vehicle types, and technological advancements.

Market segmentation within the EV industry offers valuable insights into different consumer groups, preferences, and purchasing behaviors, enabling companies to tailor strategies to meet specific needs and optimize their market presence. The primary segments analyzed include vehicle type (passenger cars, commercial vehicles, and two-wheelers), battery technology (lithium-ion, solid-state, and lead-acid), charging infrastructure, and geographic region.

This segmentation analysis provides a comprehensive understanding of market dynamics and forecasts, highlighting opportunities and challenges across various categories. By leveraging this information, stakeholders—including manufacturers, policymakers, and investors—can make informed decisions to drive further growth and innovation within the electric vehicle market.

**Datasets**

In this project, I took 4 dataset for market analysis

1. European EV Market Dataset
2. European EV Sales Dataset
3. Indian Market EV Dataset
4. Open Government Dataset of Electric Vehicles Sales

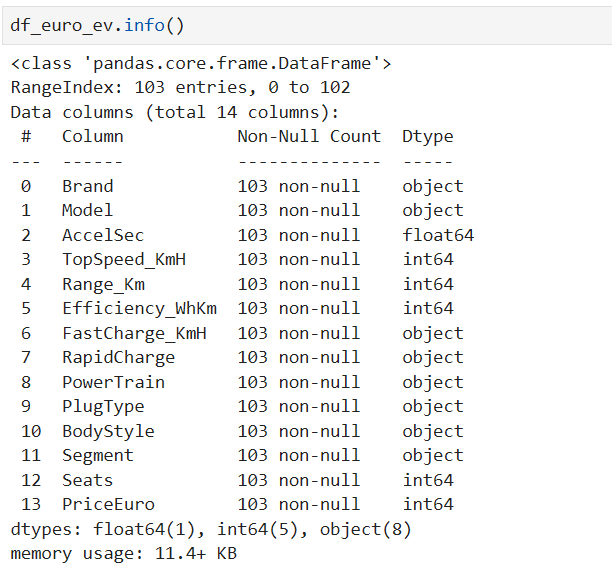
**Problem Statement**

1. **Definition:** Looking into alternate segment like electric vehicles to generate the early foot in the market.
2. **Data Collection:** The data for this analysis is collected mainly through three different sources
   1. Kaggle
   2. Open Government Dataset (India)
   3. European EV Market Dataset
3. **Identify key metrics:** Factors like Customer\_convinence, EV\_Stations, regions, geography, customer\_habits, rangeofvehicle
4. **Feature Engineering:** The primary goal here to clean and alter data values according to our needs, so that the data can become valuable.
5. **EDA (Exploratory Data Analysis):** Exploring data using various visualization tool, like seaborn and matplotlib to get a visual representation.
6. **Algorithm Selection:** Selecting best algorithm which describes and learn on data very well, It also depend upon the input and output of the data.
7. **Model Training:** Model is being trained to make
8. **Hyperparameter Tuning**
9. **Model Evaluation/Validation**

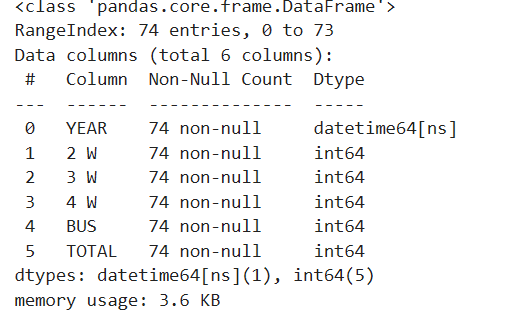
**Feature Engineering and Changes**

The Feature Engineering is Specific to dataset and end goal to derive from each dataset. I used four datasets, below I have explained about the feature engineering in all the datasets

1. **European EV Market Dataset**

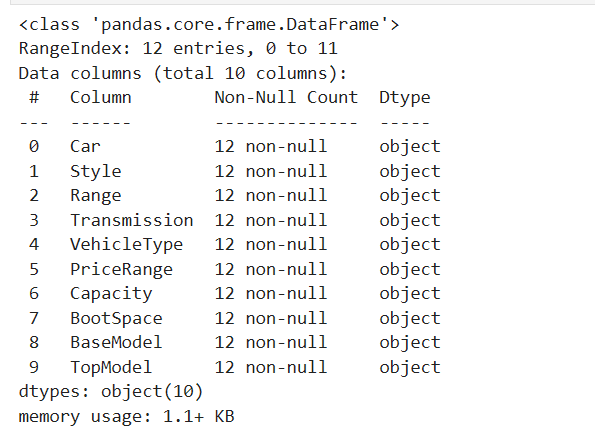


1. The Dataset contains 8 Objects, 4 int64 and 1 float64 type datatype features
2. Objects value like **FastCharge\_KmH, RapidCharge** were converted easily into the int64 type data
3. The remaining dataset can be changed according to the need of the end model algorithm.
4. **European EV Sales Dataset**
   1. The dataset is all about the sales done by companies in USA of electric vehicles.

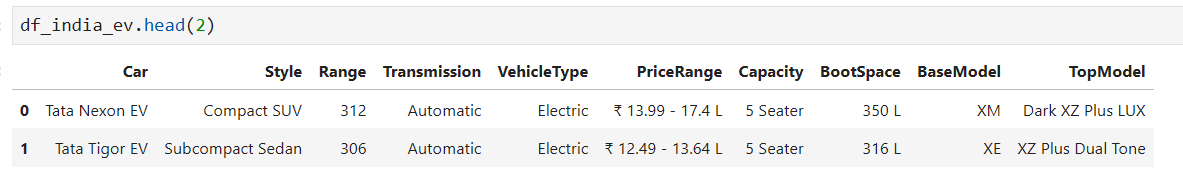
****

* 1. The dataset contains all int64 type dataset, with features representing yearly sales in the market

1. **Indian Market EV Dataset**

****

* 1. The dataset contains all the object type of datatype.
  2. Features like range, PriceRange, BootSpace are converted into the integer type.

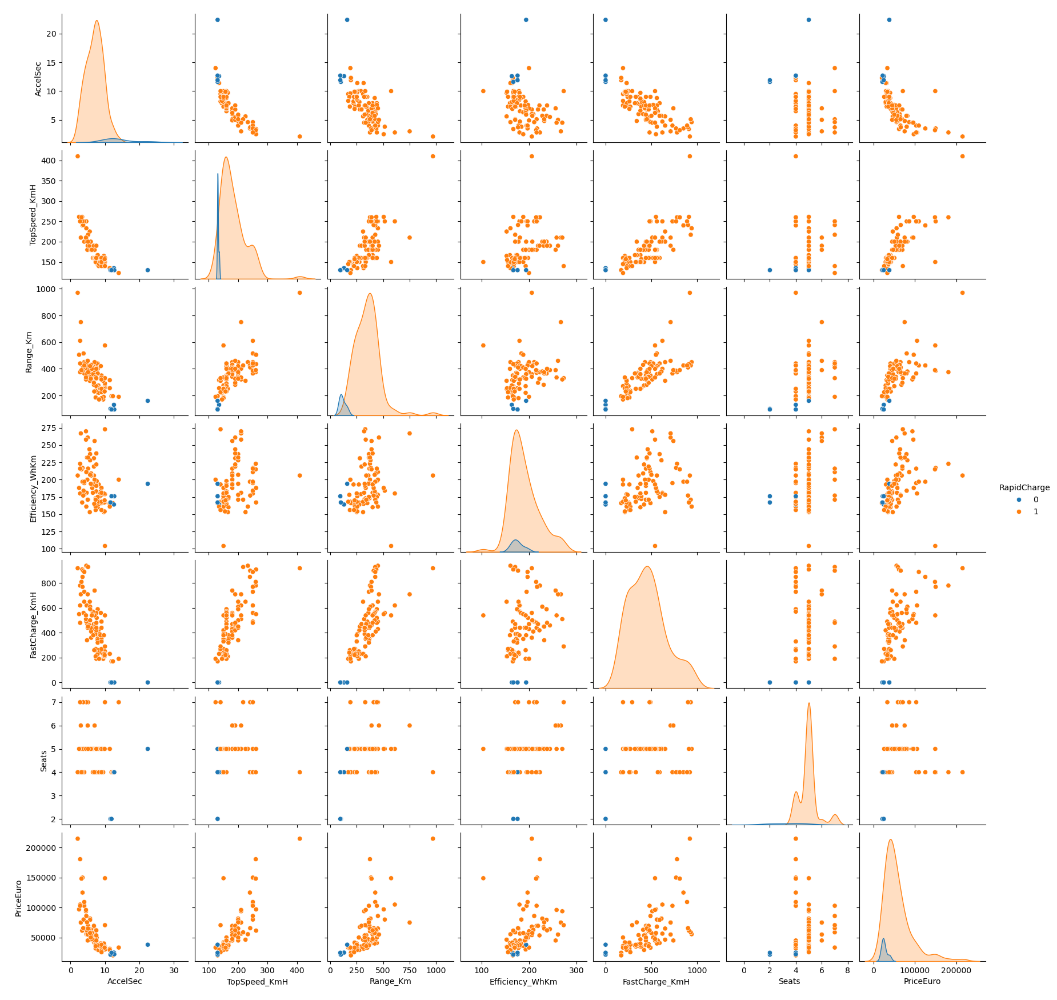


* 1. The sample dataset is shown, the remaining features are not changed as, we needed them to perform the EDA.

1. **Open Government Dataset of Electric Vehicles Sales**
   1. The dataset contains all the sales of the Electric Vehicles in India
   2. It also have the data of the total Electric Vehicles per State in India
2. **Data Type Adjustments**: Converts object types to integers where appropriate, especially for features like FastCharge\_KmH that should contain numerical data but have text values, such as '-', which are replaced with 0.
3. **Binary Encoding**:
4. The RapidCharge feature, indicating whether rapid charging is available, is converted to binary values: 1 for "Yes" and 0 for "No".
5. **Unique Value Checks**: Runs .unique() on features like PowerTrain, PlugType, BodyStyle, Segment, and Seats to review categories within each feature, which may inform the segmentation process

**EDA (Exploratory Data Analysis)**

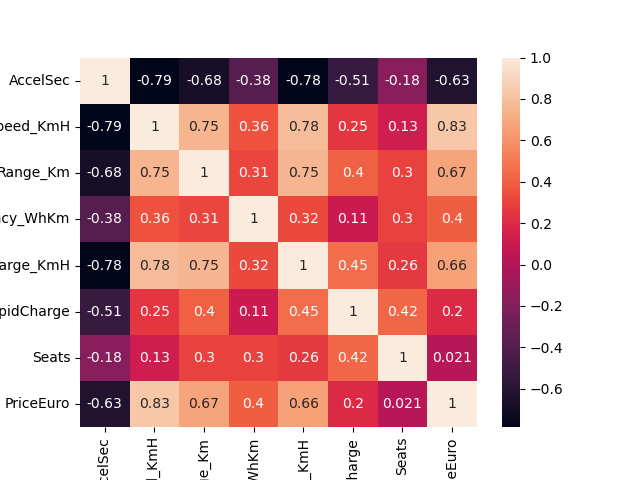
1. Pair Plot



A **pair plot** is a created to explore the relationship between the variables present in the dataset, especially when dealing with numerical data. The plot displays scatter plots of variable pairs in a grid format, with each subplot representing a relationship between two variables. Diagonal plots shows the distribution (histograms or KDE plots) for each individual variable.

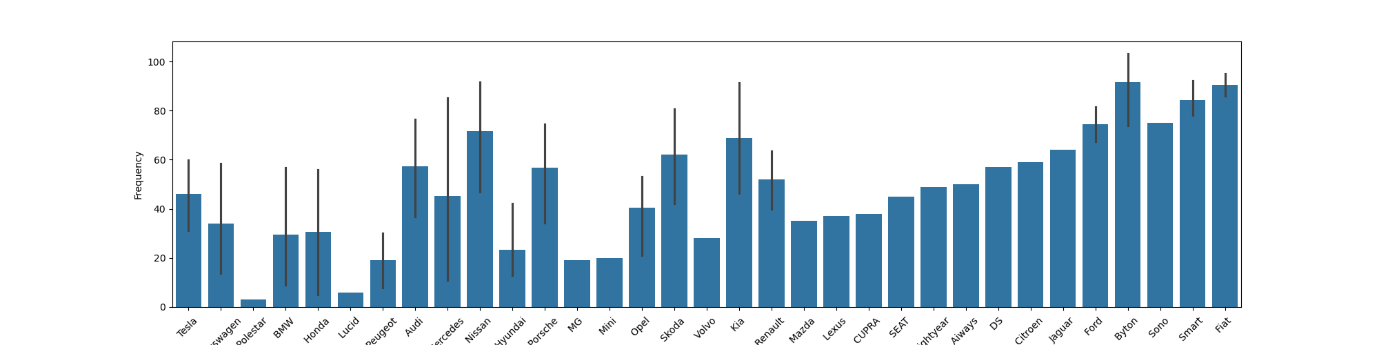
On Observing above pair pltos we can see that correlated pair of variables are appear to form a single line, horizontal or vertical created lines represent there are not correlated at all.

1. Correlation heatmap to see how features are related

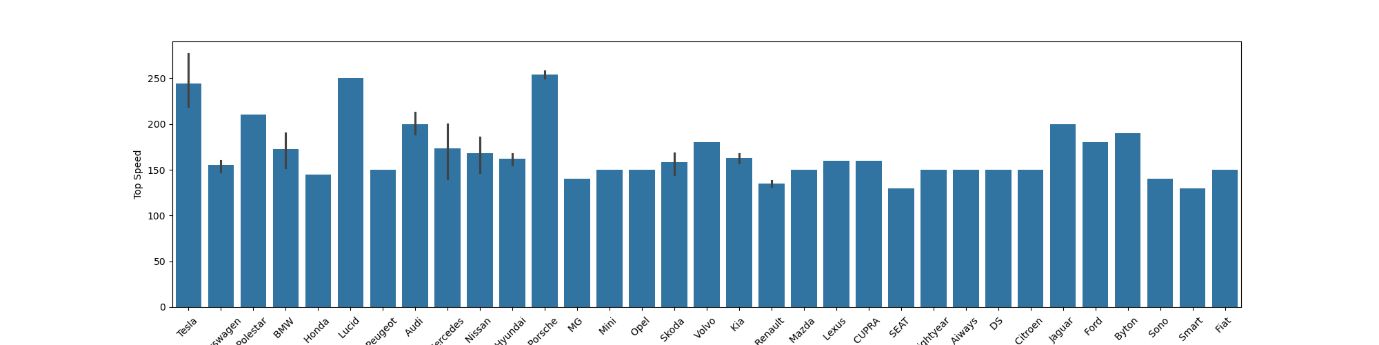


The heat map shows the correlation between the variables,

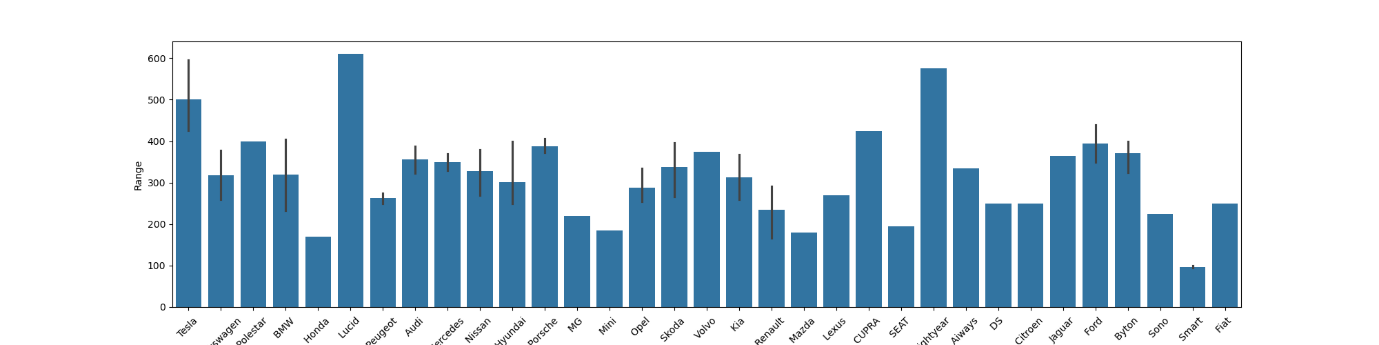
1. **1** indicates a perfect positive correlation (as one variable increases, so does the other).
2. **-1** indicates a perfect negative correlation (as one variable increases, the other decreases).
3. **0** indicates no linear correlation.
4. Frequency vs Brand



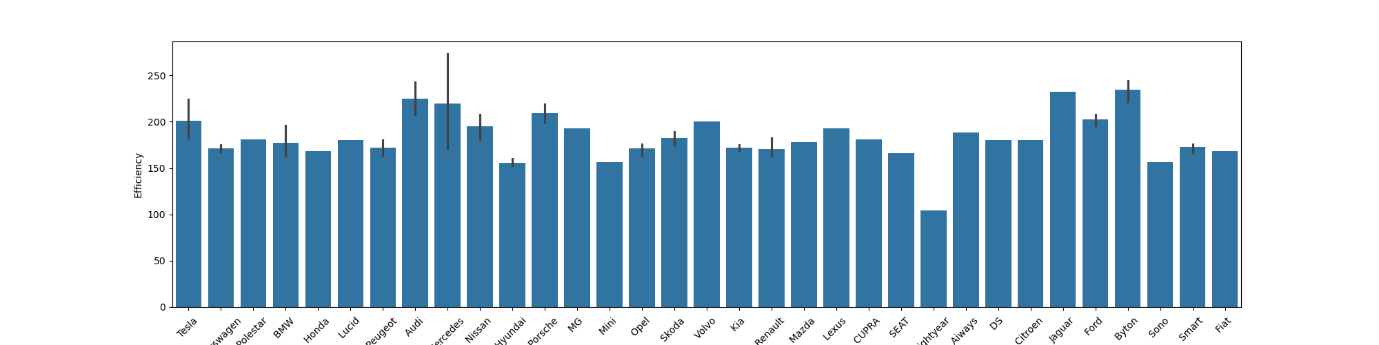
1. Top Speed Vs Brand



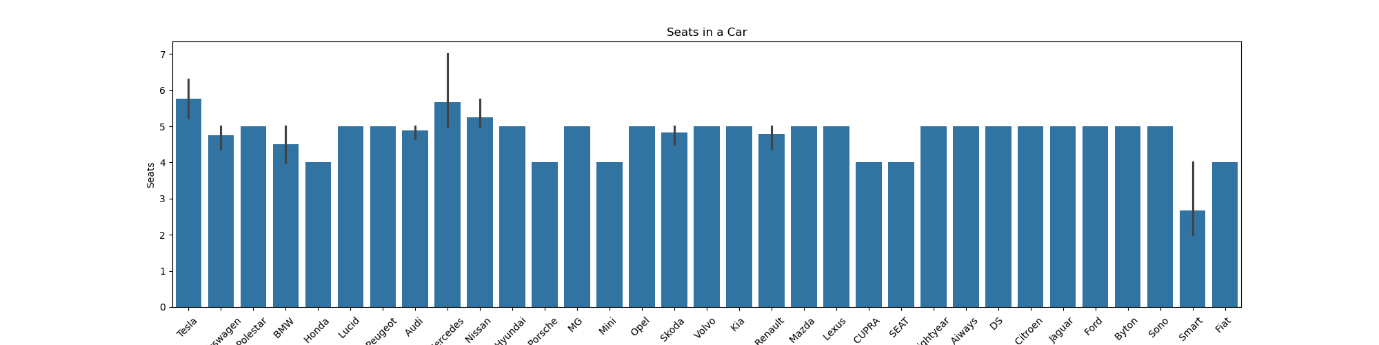
1. Brand Vs Range Km

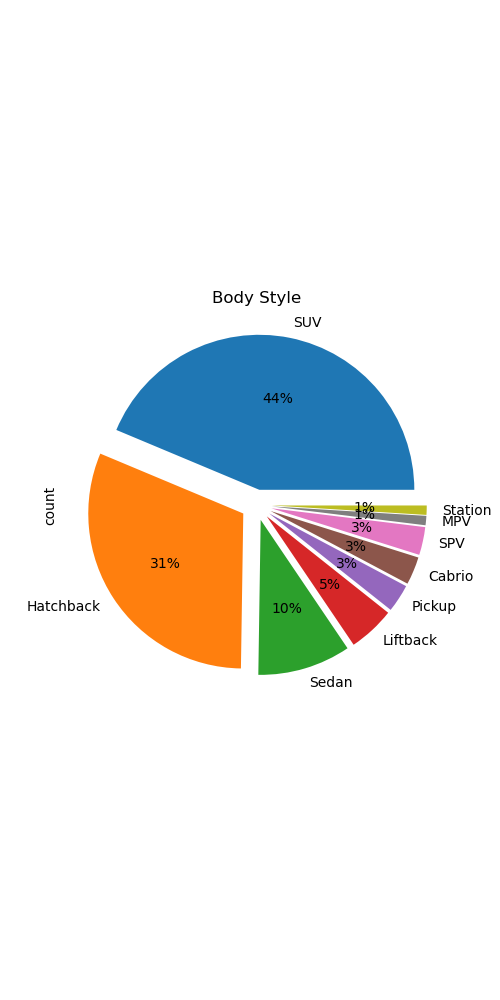


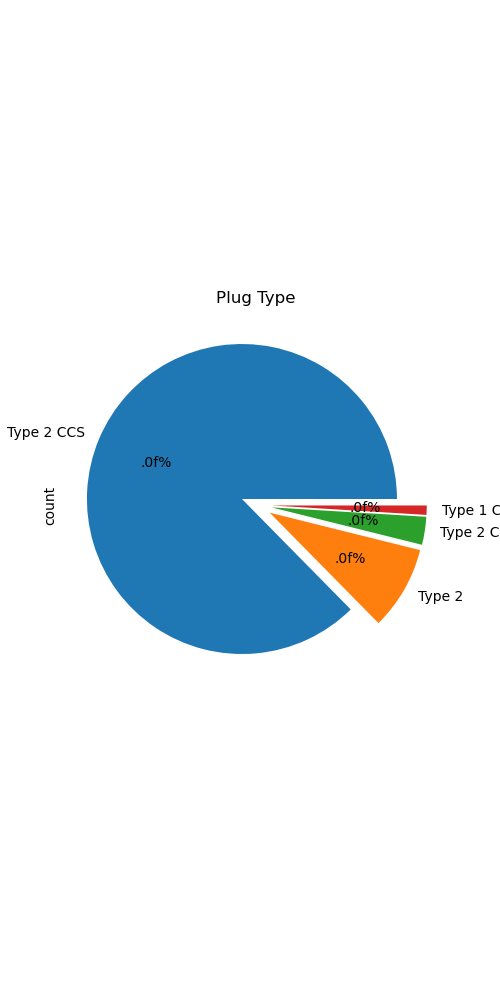
1. Brand Vs Efficiency

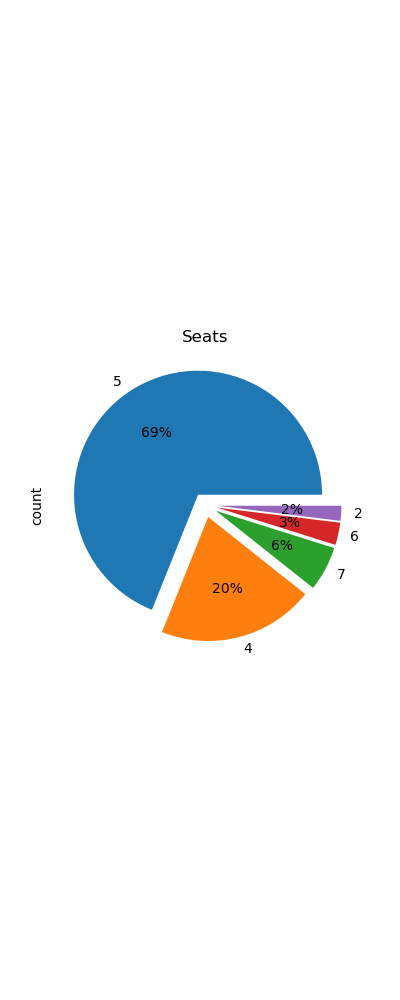


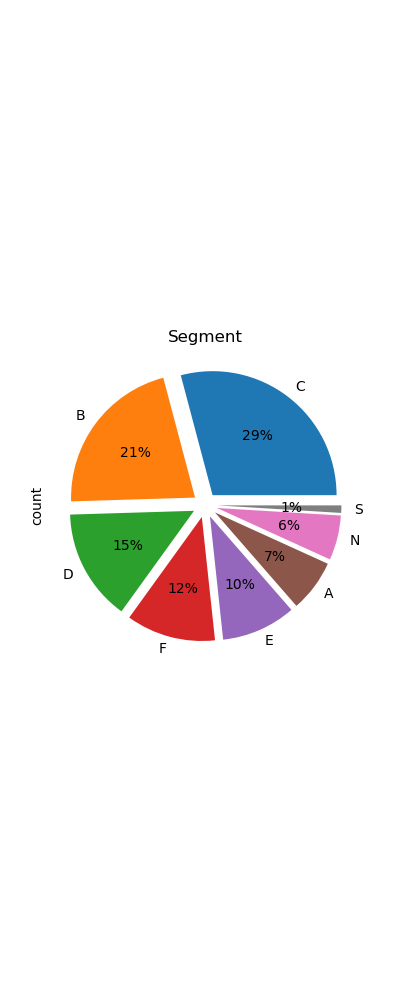
1. Brand Vs Seats



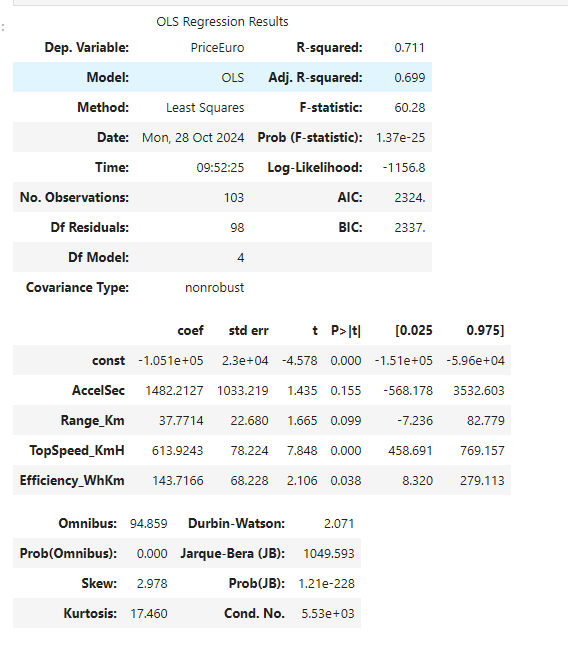








**Algorithm Used for Training**

****

**Steps in Applying Linear Regression on Your Dataset:**

1. **Feature Selection:** Identify key features like battery capacity, fast charge rate, and price that could influence the outcome.
2. **Data Preparation:** Clean and preprocess data, handling any missing values, standardizing units, and encoding categorical variables if needed.
3. **Model Fitting:** Fit a linear regression model to the dataset. For example, price might be predicted based on battery capacity, charging rate, and range.
4. **Evaluation**: Use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared to assess how well the model fits the data.

**Interpreting Results:**

* **Coefficients:** Each coefficient shows the impact of a feature on the outcome. For instance, a positive coefficient for BatteryCapacity would suggest that higher battery capacity correlates with a higher price.
* **R-squared Value:** This metric indicates the percentage of the variance in the dependent variable that the independent variables explain. A higher R-squared means a better fit.

**Use Case in EV Market Analysis:**

* Predicting EV Prices: Based on features like battery size, range, and fast-charging capability.
* Sales Forecasting: Estimating future sales based on market trends and features.

**Key Metrics and Interpretation:**

1. **R-squared (0.711):**
   * This value indicates that approximately 71.1% of the variance in the dependent variable (PriceEuro) is explained by the independent variables in the model.
   * This suggests a moderately strong model, though there is still unexplained variance (about 28.9%).
2. **Adjusted R-squared (0.699):**
   * The adjusted R-squared accounts for the number of predictors and sample size, adjusting the R-squared downward if additional variables do not improve the model fit.
   * Here, the adjusted R-squared is slightly lower than the R-squared, suggesting that the model complexity is appropriate.
3. **Coefficients (coef):**
   * **Intercept (const):** The constant term is -105,100. This would be the predicted PriceEuro if all other features were zero, though it has limited interpretative value alone.
   * **AccelSec:** The coefficient of 1482.21 suggests that, all else equal, a one-second increase in acceleration time (0-100 km/h) is associated with an increase of €1482 in the price. However, the p-value (0.155) indicates that this effect is not statistically significant at the 0.05 level.
   * **Range\_Km:** A coefficient of 37.77 implies that each additional kilometer in range is associated with an increase of €37.77 in price. The p-value (0.099) suggests this effect is marginally significant.
   * **TopSpeed\_KmH:** The coefficient of 613.92 shows that for every additional km/h in top speed, the price increases by approximately €614, which is statistically significant (p-value < 0.001).
   * **Efficiency\_WhKm:** With a coefficient of 143.72, this suggests that better efficiency (lower Wh/km) increases the price by €143.72 per unit, and this is significant (p-value = 0.038).
4. **Significance (P>|t|):**
   * The p-values indicate statistical significance for TopSpeed\_KmH and Efficiency\_WhKm (p < 0.05), while AccelSec and Range\_Km are not statistically significant predictors in this model.
5. **F-statistic (60.28) and Prob(F-statistic):**
   * The F-statistic tests whether at least one predictor variable has a non-zero coefficient. The Prob(F-statistic) is extremely low (1.37e-25), suggesting that the model overall is statistically significant.
6. **Model Diagnostics:**
   * Omnibus, Prob(Omnibus), Skew, and Kurtosis: These tests assess the normality of residuals. Here, the skewness and kurtosis values indicate some non-normality, which may affect the model's assumptions.
   * Durbin-Watson (2.071): This value tests for autocorrelation in the residuals. A value around 2 suggests little to no autocorrelation, which is good for this model.

**Conclusion and Recommendation**

This linear regression model explains a significant portion of the variance in EV prices based on features such as top speed and efficiency, though it could be improved further. Significant predictors (TopSpeed\_KmH and Efficiency\_WhKm) provide valuable insights into factors that influence the price, whereas other variables (like AccelSec and Range\_Km) may need further investigation or additional data to establish significance.

This analysis can help in understanding which features drive pricing in the EV market and potentially inform market segmentation and pricing strategies.