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
    a. Kaggle
    b. Open Government Dataset (India)
    c. 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**

```
df_euro_ev.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103 entries, 0 to 102
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Brand            103 non-null    object
 1   Model            103 non-null    object
 2   AccelSec         103 non-null    float64
 3   TopSpeed_KmH     103 non-null    int64
 4   Range_Km         103 non-null    int64
 5   Efficiency_WhKm  103 non-null    int64
 6   FastCharge_KmH   103 non-null    object
 7   RapidCharge      103 non-null    object
 8   PowerTrain       103 non-null    object
 9   PlugType         103 non-null    object
 10  BodyStyle        103 non-null    object
 11  Segment          103 non-null    object
 12  Seats            103 non-null    int64
 13  PriceEuro        103 non-null    int64
dtypes: float64(1), int64(5), object(8)
memory usage: 11.4+ KB
```

   a. The Dataset contains 8 Objects, 4 int64 and 1 float64 type datatype features
   b. Objects value like **FastCharge_KmH, RapidCharge** were converted easily into the int64 type data
   c. The remaining dataset can be changed according to the need of the end model algorithm.

2. **European EV Sales Dataset**
   a. The dataset is all about the sales done by companies in USA of electric vehicles.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74 entries, 0 to 73
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   YEAR    74 non-null     datetime64[ns]
 1   2 W     74 non-null     int64
 2   3 W     74 non-null     int64
 3   4 W     74 non-null     int64
 4   BUS     74 non-null     int64
 5   TOTAL   74 non-null     int64
dtypes: datetime64[ns](1), int64(5)
memory usage: 3.6 KB
```

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

3. Indian Market EV Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12 entries, 0 to 11
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Car           12 non-null     object
 1   Style         12 non-null     object
 2   Range         12 non-null     object
 3   Transmission  12 non-null     object
 4   VehicleType   12 non-null     object
 5   PriceRange    12 non-null     object
 6   Capacity      12 non-null     object
 7   BootSpace     12 non-null     object
 8   BaseModel     12 non-null     object
 9   TopModel      12 non-null     object
dtypes: object(10)
memory usage: 1.1+ KB
```

a. The dataset contains all the object type of datatype.

b. Features like range, PriceRange, BootSpace are converted into the integer type.

```
df_india_ev.head(2)
```

	Car	Style	Range	Transmission	VehicleType	PriceRange	Capacity	BootSpace	BaseModel	TopModel
0	Tata Nexon EV	Compact SUV	312	Automatic	Electric	₹ 13.99 - 17.4 L	5 Seater	350 L	XM	Dark XZ Plus LUX
1	Tata Tigor EV	Subcompact Sedan	306	Automatic	Electric	₹ 12.49 - 13.64 L	5 Seater	316 L	XE	XZ Plus Dual Tone

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

4. Open Government Dataset of Electric Vehicles Sales

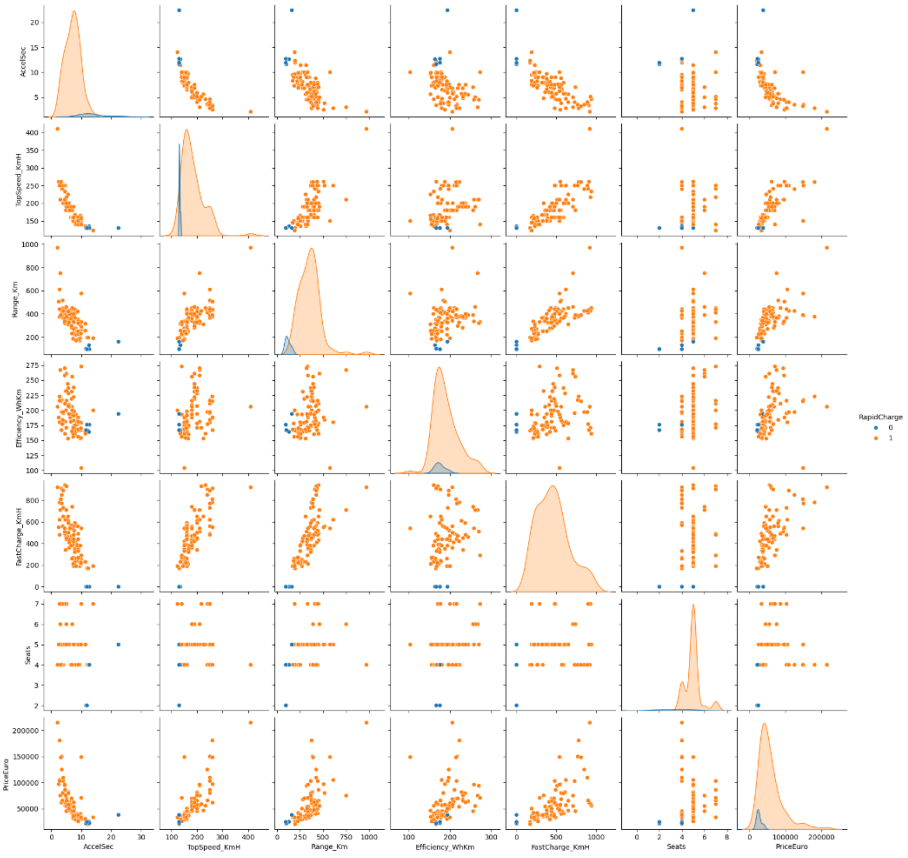
a. The dataset contains all the sales of the Electric Vehicles in India

b. It also have the data of the total Electric Vehicles per State in India

1. **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.
2. **Binary Encoding**:
3. The RapidCharge feature, indicating whether rapid charging is available, is converted to binary values: 1 for "Yes" and 0 for "No".
4. **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)

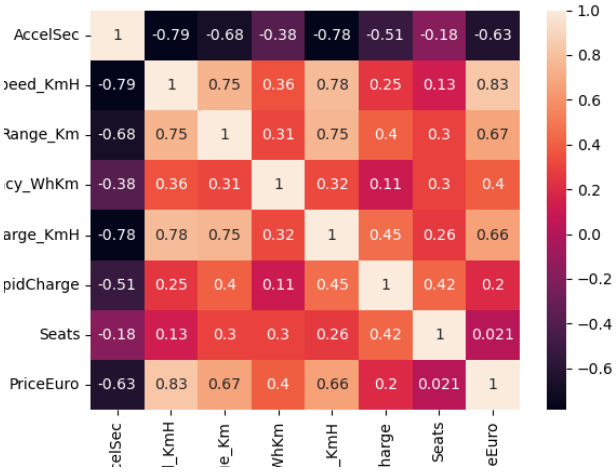
1. Pair Plot



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.

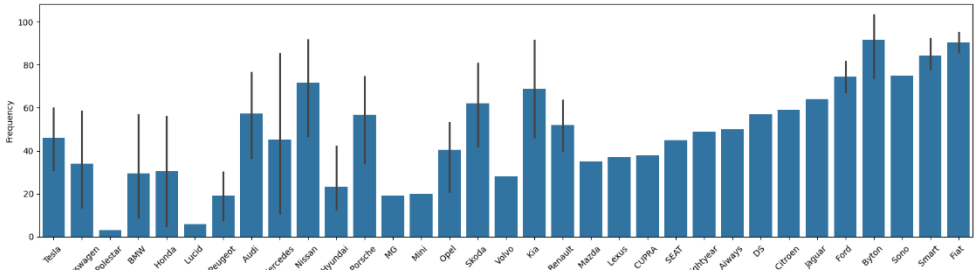
2. Correlation heatmap to see how features are related



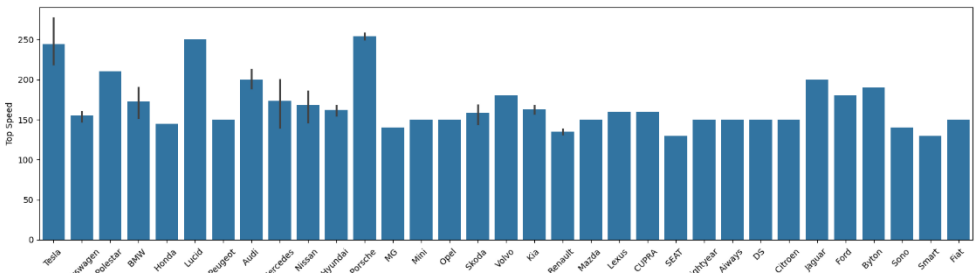
The heat map shows the correlation between the variables,

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.

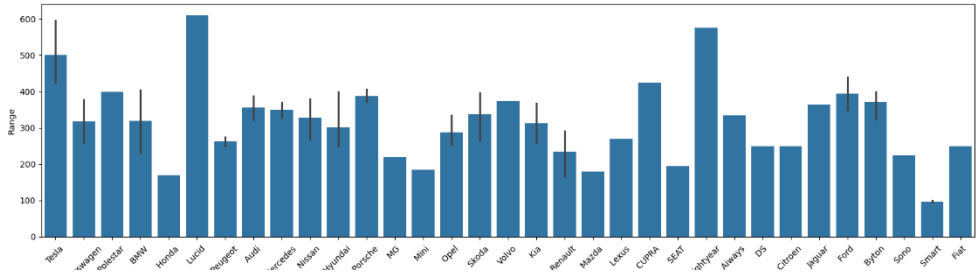
4. Frequency vs Brand



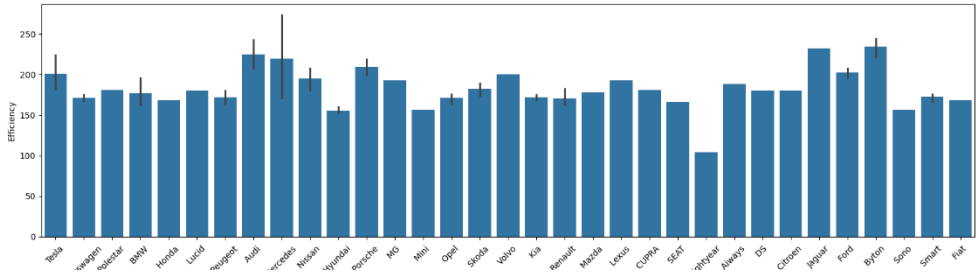
5. Top Speed Vs Brand



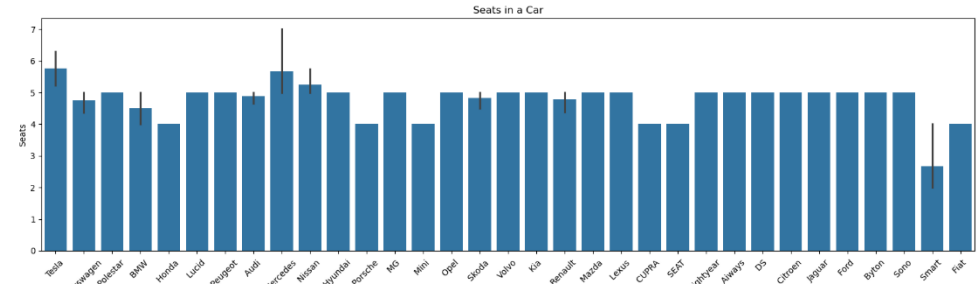
6. Brand Vs Range Km

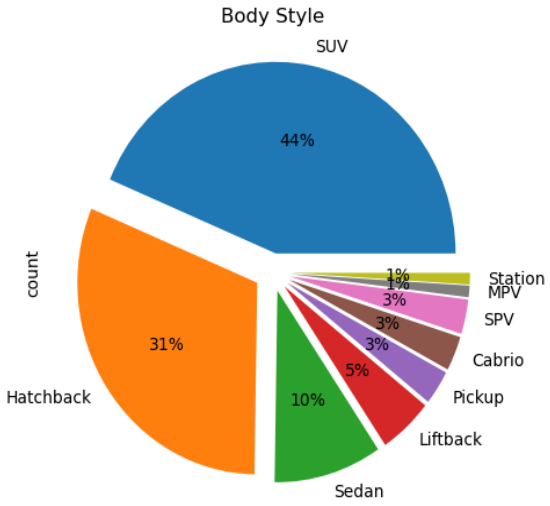


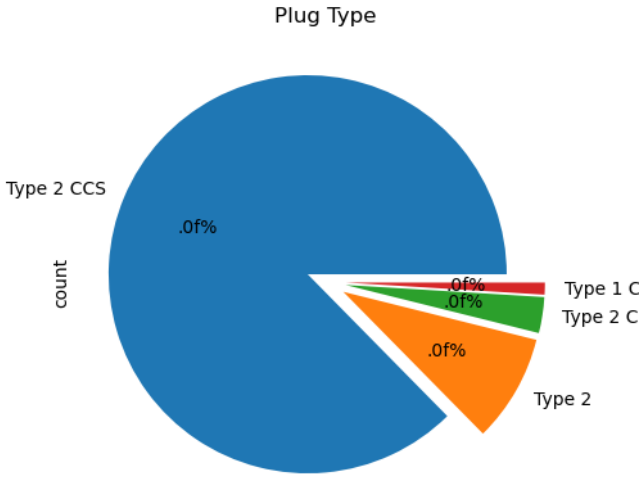
7. Brand Vs Efficiency

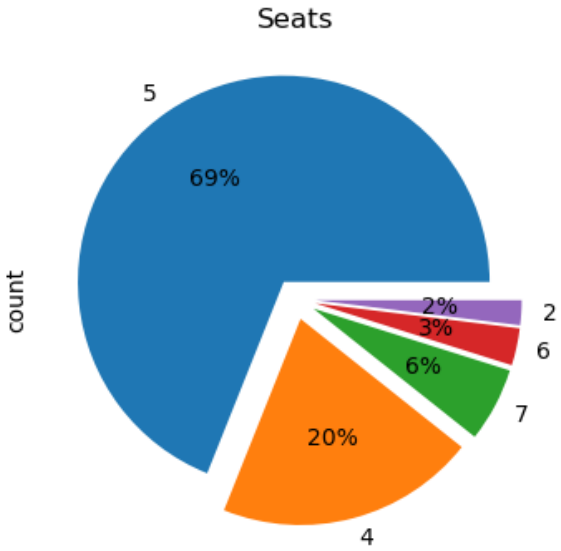


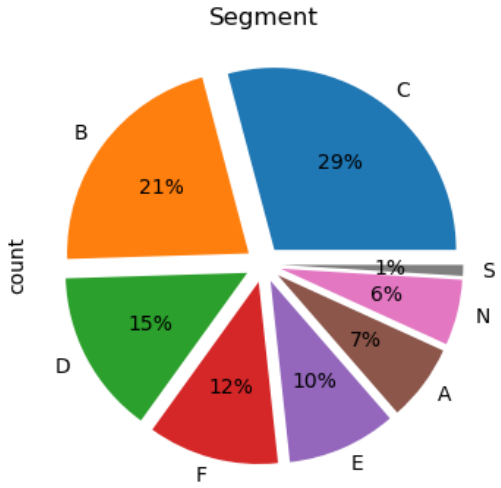
8. Brand Vs Seats






Plug Type
Type 2 CCS
.0f%
count
.0f%
Type 1 C
.0f%
Type 2 C
.0f%
Type 2


Seats
5
69%
count
2%
2
3%
6
6%
7
20%
4


Segment
C
B
29%
21%
count
1%
S
6%
N
15%
7%
D
A
12%
10%
F
E

Algorithm Used for Training

OLS Regression Results

Dep. Variable:	PriceEuro	**R-squared:**	0.711
Model:	OLS	**Adj. R-squared:**	0.699
Method:	Least Squares	**F-statistic:**	60.28
Date:	Mon, 28 Oct 2024	**Prob (F-statistic):**	1.37e-25
Time:	09:52:25	**Log-Likelihood:**	-1156.8
No. Observations:	103	**AIC:**	2324.
Df Residuals:	98	**BIC:**	2337.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.051e+05	2.3e+04	-4.578	0.000	-1.51e+05	-5.96e+04
AccelSec	1482.2127	1033.219	1.435	0.155	-568.178	3532.603
Range_Km	37.7714	22.680	1.665	0.099	-7.236	82.779
TopSpeed_KmH	613.9243	78.224	7.848	0.000	458.691	769.157
Efficiency_WhKm	143.7166	68.228	2.106	0.038	8.320	279.113

Omnibus:	94.859	**Durbin-Watson:**	2.071
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1049.593
Skew:	2.978	**Prob(JB):**	1.21e-228
Kurtosis:	17.460	**Cond. No.**	5.53e+03

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.