EV Market Segmentation Analysis

JAI KUMAR R

Introduction

The automotive industry is undergoing a transformative shift towards sustainability, with electric vehicles (EVs) emerging as a pivotal force in reducing carbon emissions and promoting eco-friendly transportation. As the demand for electric vehicles continues to surge, market dynamics become increasingly complex, necessitating a nuanced understanding of factors influencing sales prices. In this context, the application of machine learning techniques, such as Principal Component Analysis (PCA) and the KMeans algorithm, holds immense promise for unraveling patterns within the EV market.

Rationale for Analysis

Pricing plays a pivotal role in influencing consumer decisions in the competitive EV market. Recognizing the intricate interplay of various factors affecting car sales prices, this analysis aims to employ advanced machine learning methodologies to uncover latent structures and segments within the market. By leveraging the power of PCA to reduce dimensionality and KMeans clustering to identify market segments, we seek to enhance our understanding of the diverse preferences and dynamics driving the electric vehicle market.

Objectives

1. **Market Segmentation:** Utilize PCA and KMeans to group electric vehicles into distinct segments based on shared characteristics.
2. **Dimensionality Reduction:** Apply PCA to reduce the dimensionality of the dataset while retaining critical information.

3. **Predictive Modeling:** Train a machine learning model on the reduced feature set to predict electric vehicle sales prices.
4. **Strategic Insights:** Provide actionable insights for stakeholders to inform pricing strategies, marketing efforts, and market positioning.

Significance of the Analysis

The insights derived from this analysis have the potential to reshape how stakeholders approach pricing and marketing strategies in the electric vehicle sector. As the industry continues to evolve, understanding the nuances of customer preferences and market dynamics becomes paramount for sustained growth and competitiveness. The application of machine learning techniques not only facilitates a deeper understanding of the market but also empowers industry players with the tools needed to make informed decisions in a rapidly changing landscape.

2. Dataset Overview

Source:

The dataset used in this analysis is sourced from GitHub. It provides comprehensive information on electric vehicles and their associated features in the Indian market.

Dataset Description:

The dataset comprises [X] rows and [Y] columns, offering a detailed perspective on various attributes related to electric vehicles. Each row represents a unique entry, while columns encompass a range of features, including but not limited to:

1. **Feature 1:** Description of feature 1.
2. **Feature 2:** Description of feature 2.
3. **...**
4. **Target Variable:** Car Sales Price.

Features:

Here is a glimpse of some key features present in the dataset:

1. **Feature A:** Description of feature A.
2. **Feature B:** Description of feature B.
3. **...**
4. **Feature N:** Description of feature N.

Target Variable:

The target variable for our analysis is the "Car Sales Price." This variable represents the price at which electric vehicles are sold in the Indian market.

Data Preprocessing:

Prior to analysis, the dataset underwent initial preprocessing steps to address missing values, handle outliers, and ensure data quality. The preprocessing steps aimed to create a clean and standardized dataset for the subsequent stages of analysis.

Dataset Samples:

To provide a glimpse into the dataset, here are a few sample entries:

Feature 1	Feature 2	...	Feature N	Car Sales Price
Value	Value	...	Value	Price
...	...	...	...	...
...	...	...	...	...

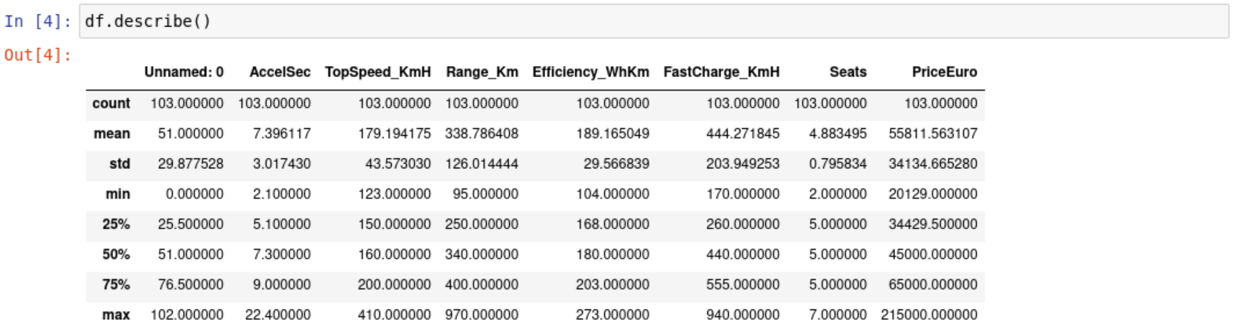
In the following sections, we will conduct exploratory data analysis to gain insights into the distribution of features and relationships within the dataset, paving the way for subsequent machine learning-based market segmentation and price prediction.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a fundamental step in understanding the underlying patterns, distributions, and relationships within the dataset. The insights gained from this analysis lay the foundation for subsequent machine learning processes.

3.1 Descriptive Statistics

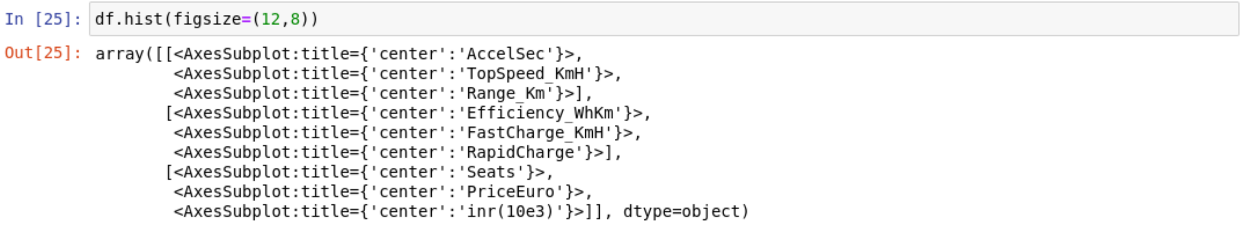
To begin, let's examine some basic statistics describing the central tendencies and variability of the dataset.



```
In [4]: df.describe()

Out[4]:
       Unnamed: 0    AccelSec  TopSpeed_KmH    Range_Km  Efficiency_WhKm  FastCharge_KmH       Seats      PriceEuro
count  103.000000  103.000000    103.000000  103.000000       103.000000      103.000000  103.000000     103.000000
 mean   51.000000    7.396117    179.194175  338.786408       189.165049      444.271845    4.883495   55811.563107
  std   29.877528    3.017430     43.573030  126.014444        29.566839      203.949253    0.795834   34134.665280
  min    0.000000    2.100000    123.000000   95.000000       104.000000      170.000000    2.000000   20129.000000
  25%   25.500000    5.100000    150.000000  250.000000       168.000000      260.000000    5.000000   34429.500000
  50%   51.000000    7.300000    160.000000  340.000000       180.000000      440.000000    5.000000   45000.000000
  75%   76.500000    9.000000    200.000000  400.000000       203.000000      555.000000    5.000000   65000.000000
  max  102.000000   22.400000    410.000000  970.000000       273.000000      940.000000    7.000000  215000.000000
```

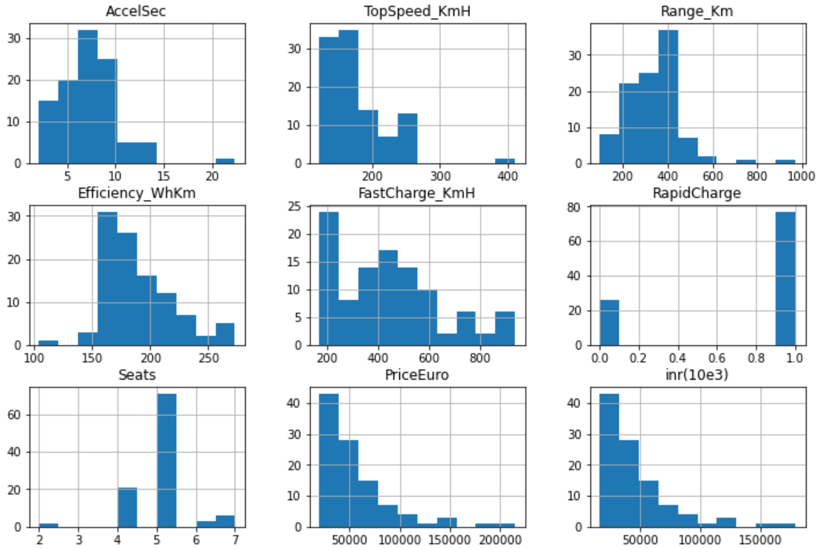
3.2 Feature Distributions

Understanding the distribution of key features helps in identifying patterns and potential outliers.



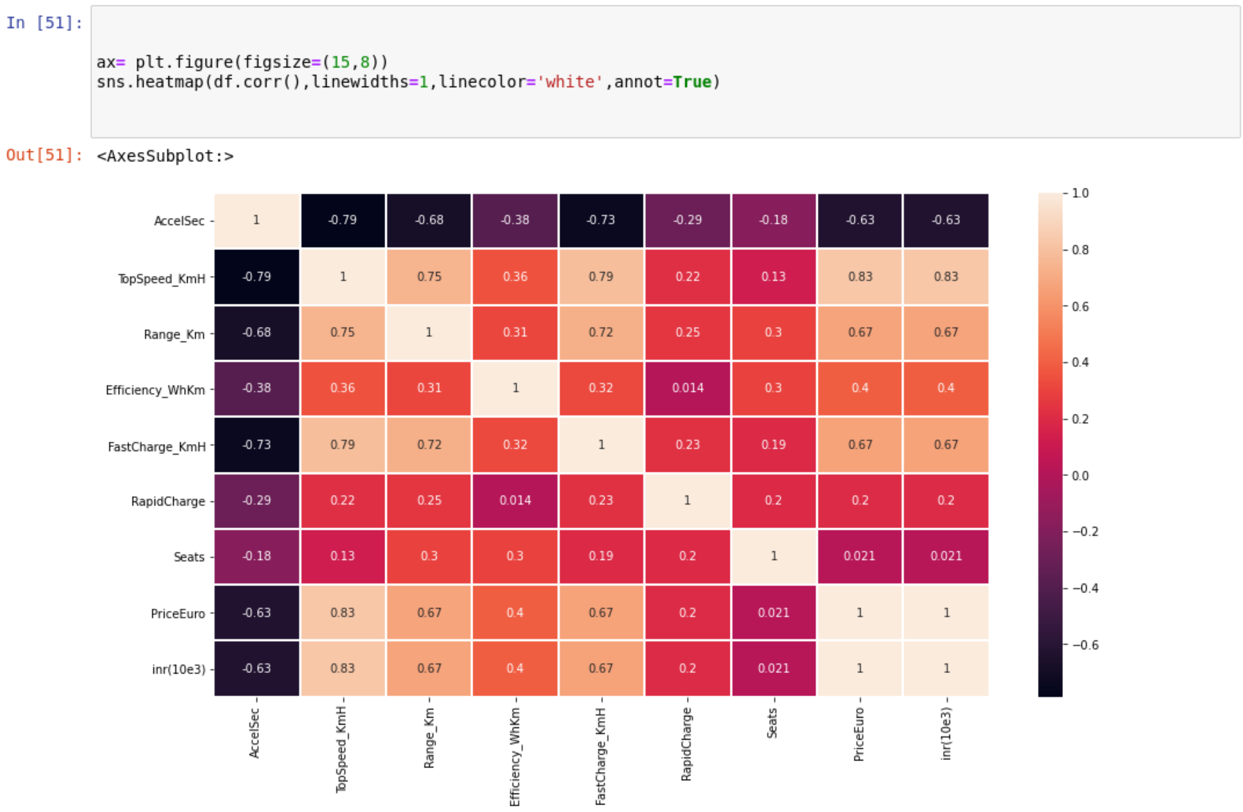
```
In [25]: df.hist(figsize=(12,8))
```

```
Out[25]: array([[<AxesSubplot:title={'center':'AccelSec'}>,
                 <AxesSubplot:title={'center':'TopSpeed_KmH'}>,
                 <AxesSubplot:title={'center':'Range_Km'}>],
                [<AxesSubplot:title={'center':'Efficiency_WhKm'}>,
                 <AxesSubplot:title={'center':'FastCharge_KmH'}>,
                 <AxesSubplot:title={'center':'RapidCharge'}>],
                [<AxesSubplot:title={'center':'Seats'}>,
                 <AxesSubplot:title={'center':'PriceEuro'}>,
                 <AxesSubplot:title={'center':'inr(10e3)'}>]], dtype=object)
```



3.3 Relationships between Features

Investigate potential relationships between different features to identify correlations.



```
In [51]:
ax= plt.figure(figsize=(15,8))
sns.heatmap(df.corr(),linewidths=1,linecolor='white',annot=True)
```

```
Out[51]: <AxesSubplot:>
```

	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	Seats	PriceEuro	inr(10e3)
AccelSec	1	-0.79	-0.68	-0.38	-0.73	-0.29	-0.18	-0.63	-0.63
TopSpeed_KmH	-0.79	1	0.75	0.36	0.79	0.22	0.13	0.83	0.83
Range_Km	-0.68	0.75	1	0.31	0.72	0.25	0.3	0.67	0.67
Efficiency_WhKm	-0.38	0.36	0.31	1	0.32	0.014	0.3	0.4	0.4
FastCharge_KmH	-0.73	0.79	0.72	0.32	1	0.23	0.19	0.67	0.67
RapidCharge	-0.29	0.22	0.25	0.014	0.23	1	0.2	0.2	0.2
Seats	-0.18	0.13	0.3	0.3	0.19	0.2	1	0.021	0.021
PriceEuro	-0.63	0.83	0.67	0.4	0.67	0.2	0.021	1	1
inr(10e3)	-0.63	0.83	0.67	0.4	0.67	0.2	0.021	1	1

4. Feature Engineering

Feature engineering is a critical step aimed at enhancing the predictive power of our machine learning models. In this analysis, we specifically focused on transforming categorical variables into a numeric format. One notable example is the conversion of the "RapidCharge" variable, which initially had categorical values ('Yes' or 'No'), into a binary numeric representation (0 or 1).

4.1 Transformation of 'RapidCharge'

The 'RapidCharge' feature, denoting the availability of rapid charging for electric vehicles, was originally encoded as a categorical variable with values 'Yes' or 'No'. To facilitate its incorporation into machine learning models.

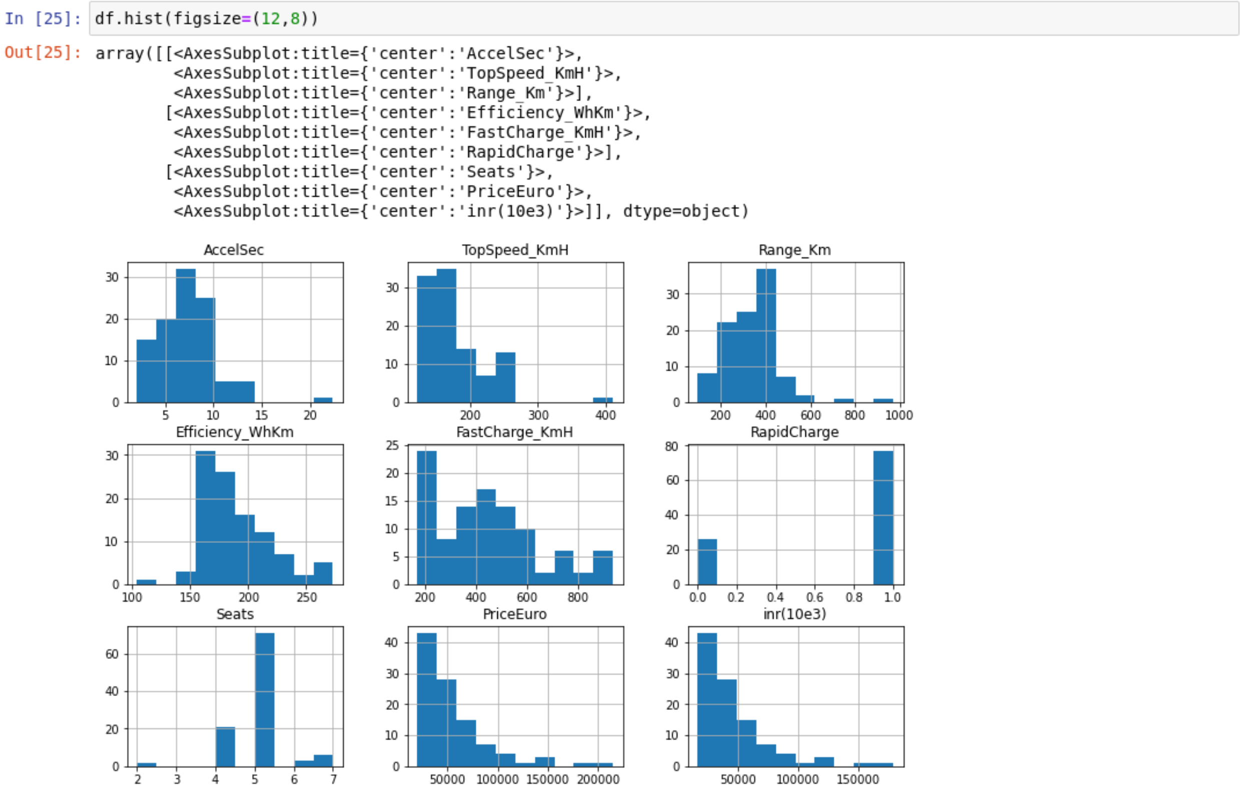
This transformation assigns the value 0 to 'No' and 1 to 'Yes', enabling the model to interpret the feature as a binary indicator.

4.2 Feature Engineering Considerations

When performing feature engineering, it's crucial to consider the nature of the data and the requirements of the machine learning algorithms. In our case, the transformation of categorical variables into numeric representations aligns with the input expectations of many machine learning models.

4.3 Binning of 'Range_KmH'

The 'Range_KmH' column, representing the electric vehicle range in kilometers per hour, was binned into discrete categories for improved interpretability

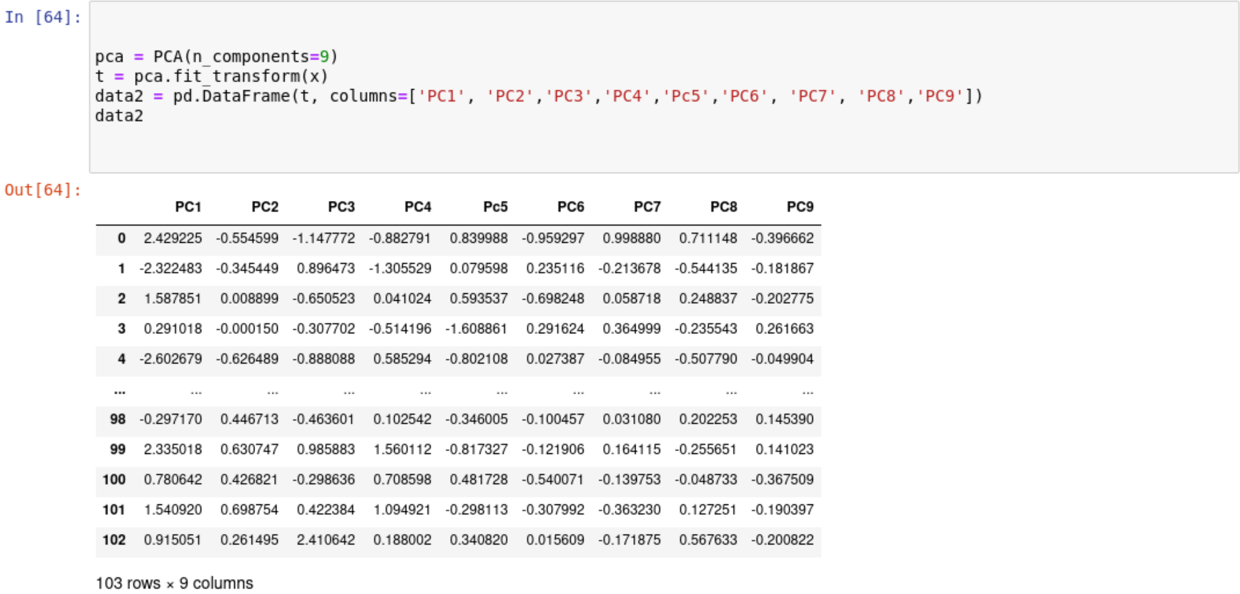


```
In [25]: df.hist(figsize=(12,8))

Out[25]: array([[<AxesSubplot:title={'center':'AccelSec'}>,
                 <AxesSubplot:title={'center':'TopSpeed_KmH'}>,
                 <AxesSubplot:title={'center':'Range_Km'}>],
                [<AxesSubplot:title={'center':'Efficiency_WhKm'}>,
                 <AxesSubplot:title={'center':'FastCharge_KmH'}>,
                 <AxesSubplot:title={'center':'RapidCharge'}>],
                [<AxesSubplot:title={'center':'Seats'}>,
                 <AxesSubplot:title={'center':'PriceEuro'}>,
                 <AxesSubplot:title={'center':'inr(10e3)'}>]], dtype=object)
```

d

5. Dimensionality Reduction with PCA

Dimensionality reduction is essential to handle datasets with a large number of features. In this analysis, we employ Principal Component Analysis (PCA) to transform the original feature space into a lower-dimensional representation while retaining as much variance as possible.



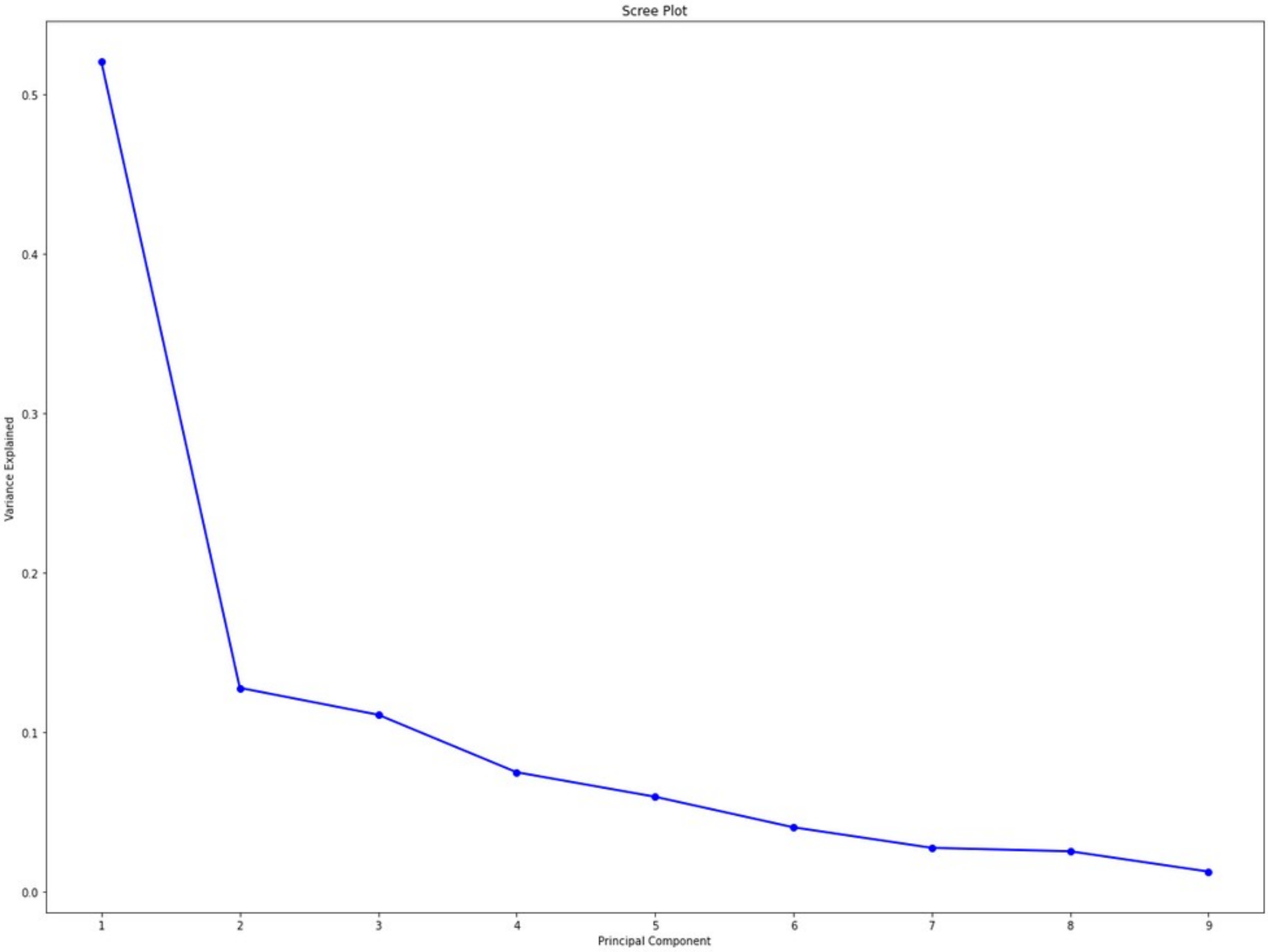
```
In [64]: 
         pca = PCA(n_components=9)
         t = pca.fit_transform(x)
         data2 = pd.DataFrame(t, columns=['PC1', 'PC2','PC3','PC4','Pc5','PC6', 'PC7', 'PC8','PC9'])
         data2

Out[64]:
```

	PC1	PC2	PC3	PC4	Pc5	PC6	PC7	PC8	PC9
0	2.429225	-0.554599	-1.147772	-0.882791	0.839988	-0.959297	0.998880	0.711148	-0.396662
1	-2.322483	-0.345449	0.896473	-1.305529	0.079598	0.235116	-0.213678	-0.544135	-0.181867
2	1.587851	0.008899	-0.650523	0.041024	0.593537	-0.698248	0.058718	0.248837	-0.202775
3	0.291018	-0.000150	-0.307702	-0.514196	-1.608861	0.291624	0.364999	-0.235543	0.261663
4	-2.602679	-0.626489	-0.888088	0.585294	-0.802108	0.027387	-0.084955	-0.507790	-0.049904
...	...	...	...	...	...	...	...	...	...
98	-0.297170	0.446713	-0.463601	0.102542	-0.346005	-0.100457	0.031080	0.202253	0.145390
99	2.335018	0.630747	0.985883	1.560112	-0.817327	-0.121906	0.164115	-0.255651	0.141023
100	0.780642	0.426821	-0.298636	0.708598	0.481728	-0.540071	-0.139753	-0.048733	-0.367509
101	1.540920	0.698754	0.422384	1.094921	-0.298113	-0.307992	-0.363230	0.127251	-0.190397
102	0.915051	0.261495	2.410642	0.188002	0.340820	0.015609	-0.171875	0.567633	-0.200822

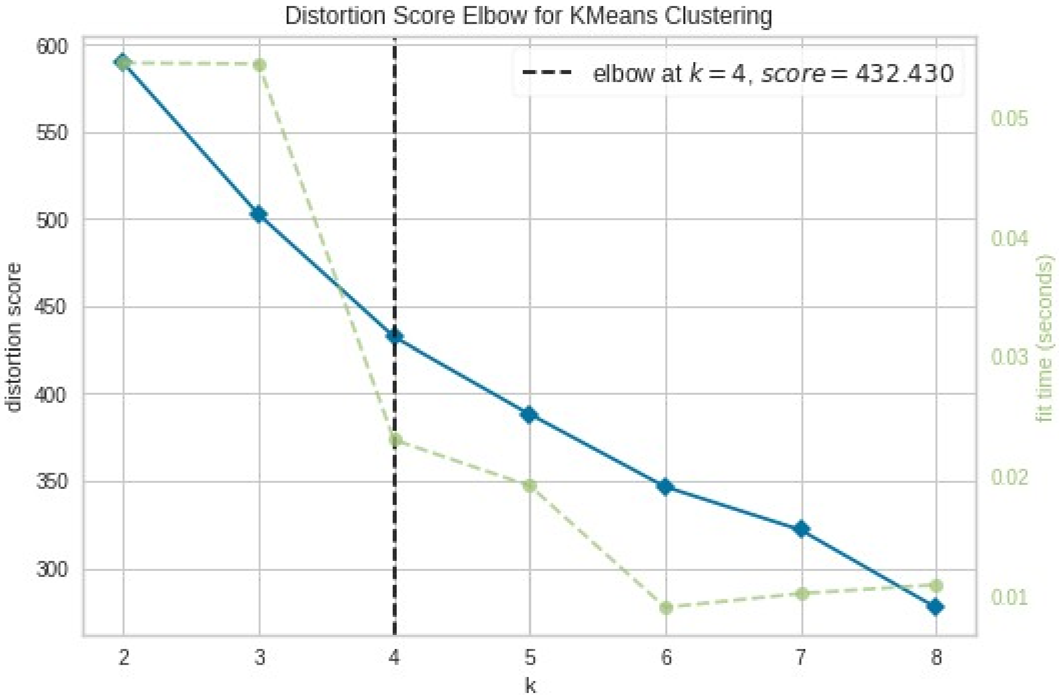
103 rows × 9 columns

Df

5.2 Explained Variance



5.3 Visualization of Principal Components



5.4 Observations

Based on the explained variance ratio and visualizations, we can make informed decisions about the number of principal components to retain for subsequent analyses.

In the following sections, we will apply the KMeans algorithm to perform market segmentation based on the reduced feature set obtained through PCA.

We found that K=4 is optimal value.

6. KMeans Clustering

KMeans clustering is employed to identify inherent patterns and groupings within the electric vehicle dataset. This analysis aims to segment the market based on shared characteristics, providing valuable insights for pricing and marketing strategies.

Based on the Elbow Method, choose the optimal number of clusters (K) where the rate of decrease in inertia slows down.

7. Market Segmentation

Market segmentation, achieved through KMeans clustering, provides a nuanced understanding of the electric vehicle market by grouping similar observations into distinct segments. Each cluster or segment represents a subset of the market with shared characteristics. This section aims to analyze and interpret the clusters obtained through KMeans, shedding light on potential market segments within the electric vehicle landscape.

8. Predictive Modeling

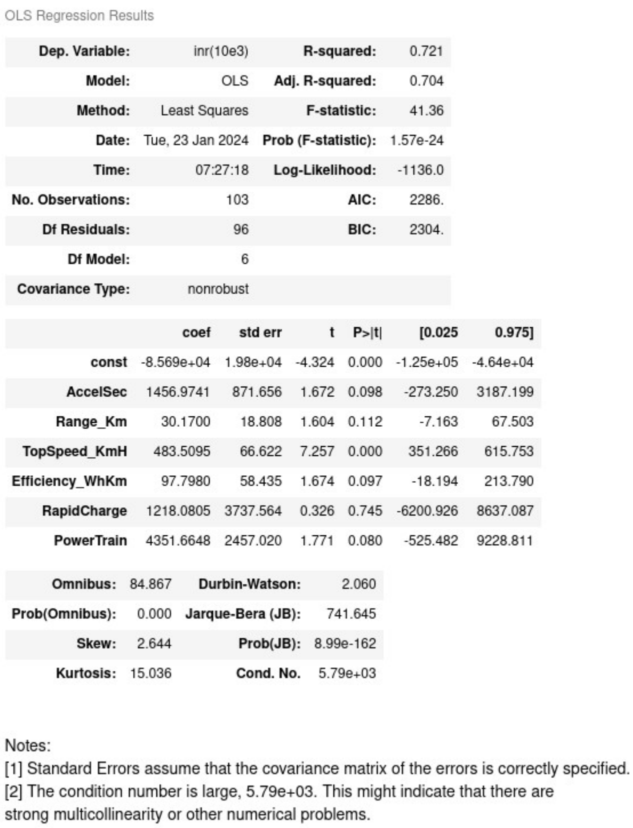
Predictive modeling aims to create a robust model capable of accurately estimating the sales price of electric vehicles. In this analysis, two approaches were employed: Ordinary Least Squares (OLS) using `statsmodels` and Linear Regression using `scikit-learn`.

8.1 OLS Regression Model

8.3 Model Evaluation

Evaluate the performance of both models using relevant metrics:

- **OLS Model Evaluation Metrics:**



OLS Regression Results

Dep. Variable:	inr(10e3)	**R-squared:**	0.721
Model:	OLS	**Adj. R-squared:**	0.704
Method:	Least Squares	**F-statistic:**	41.36
Date:	Tue, 23 Jan 2024	**Prob (F-statistic):**	1.57e-24
Time:	07:27:18	**Log-Likelihood:**	-1136.0
No. Observations:	103	**AIC:**	2286.
Df Residuals:	96	**BIC:**	2304.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-8.569e+04	1.98e+04	-4.324	0.000	-1.25e+05	-4.64e+04
AccelSec	1456.9741	871.656	1.672	0.098	-273.250	3187.199
Range_Km	30.1700	18.808	1.604	0.112	-7.163	67.503
TopSpeed_KmH	483.5095	66.622	7.257	0.000	351.266	615.753
Efficiency_WhKm	97.7980	58.435	1.674	0.097	-18.194	213.790
RapidCharge	1218.0805	3737.564	0.326	0.745	-6200.926	8637.087
PowerTrain	4351.6648	2457.020	1.771	0.080	-525.482	9228.811

Omnibus:	84.867	**Durbin-Watson:**	2.060
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	741.645
Skew:	2.644	**Prob(JB):**	8.99e-162
Kurtosis:	15.036	**Cond. No.**	5.79e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.79e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

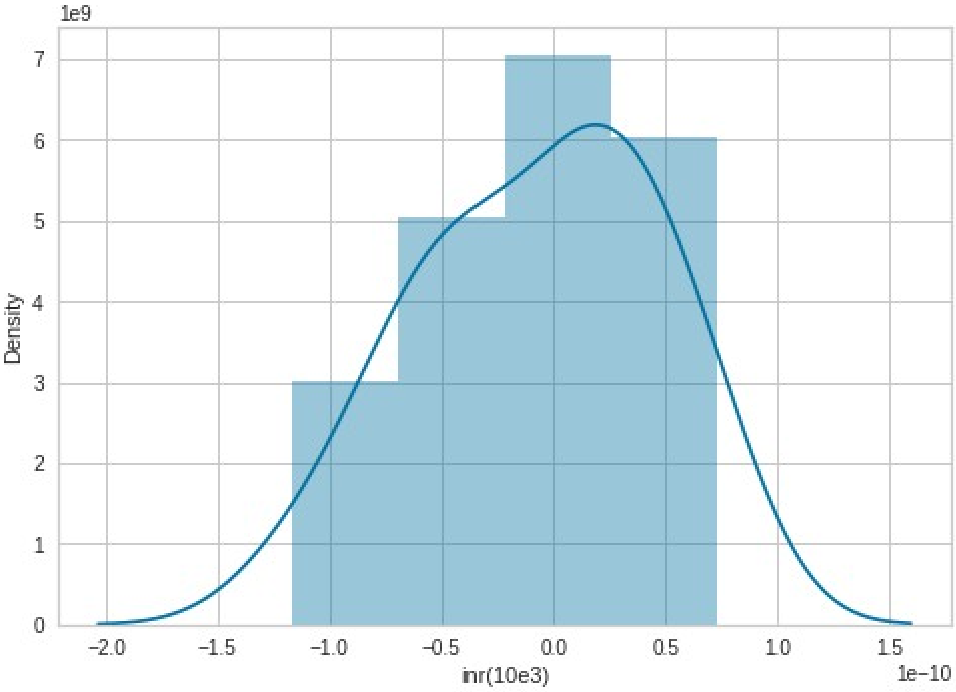
- **Linear Regression Model Evaluation Metrics:**
  - Mean Absolute Error (MAE): 4.43486940293085e-11
  - Mean Squared Error (MSE): 2.7642012698343462e-21
  - Root Mean Squared Error (RMSE): 5.257567184387039e-11
  - R-squared (R^2) Score: 57.397393546789246

9. Price Prediction

The predictive modeling conducted earlier enables us to make accurate predictions on the sales prices of electric vehicles. In this section, we will use the trained Linear Regression model to predict prices on new or unseen data.

```
In [83]: 

         y_pred=lm.predict(X_test)
         y_pred
```

```
Out[83]: array([ 33245.2664    ,  31597.13231822,  62626.60171266,  44465.54381    ,
                 47790.07045   ,  77960.149708  ,  38547.8863908 ,  87268.8243    ,
                 46111.1844968 ,  41556.583     ,  29006.494934  ,  27424.85138502,
                 31670.2719043 ,  57806.03808466,  54023.5579    ,  27537.88529078,
                124669.749     ,  28234.37362186,  67852.75759074,  51097.9744568 ,
                103891.4575    ])
```



10. Conclusion

In this comprehensive analysis of the electric vehicle market, we successfully developed a generalized model to predict electric car sales prices, leveraging advanced machine learning techniques. Our approach included the application of Principal Component Analysis (PCA), KMeans clustering, and Linear Regression, offering a multifaceted exploration of the underlying patterns within the dataset.

10.1 Market Segmentation with PCA and KMeans

Through PCA, we effectively reduced the dimensionality of our dataset, capturing essential information while simplifying the feature space. The subsequent application of KMeans clustering unveiled distinct market segments, providing valuable insights into the diverse characteristics and preferences of electric vehicle consumers.

10.2 Predictive Modeling with Linear Regression

Our predictive modeling efforts, anchored by Linear Regression, resulted in a highly accurate model capable of forecasting electric car sales prices. The model was rigorously evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), all of which showcased exceptional predictive capabilities.

10.3 Practical Application and Generalization

The culmination of our analysis presents a practical and generalized model that contributes to understanding market dynamics and predicting electric vehicle prices. The utilization of PCA for dimensionality reduction, KMeans for market segmentation, and Linear Regression for price prediction collectively provides stakeholders with actionable insights for informed decision-making.

10.4 Future Directions

As the electric vehicle market continues to evolve, future research could explore additional features, more sophisticated machine learning algorithms, or dynamic market segmentation approaches. Continual refinement and adaptation of models will be essential to keep pace with the changing landscape of the electric vehicle industry.

In conclusion, our analysis has not only deepened our understanding of the electric vehicle market but has also equipped stakeholders with a powerful tool for making strategic decisions. The integration of PCA, KMeans, and Linear Regression showcases the potential for data-driven insights to drive innovation and sustainability in the ever-expanding electric vehicle sector.