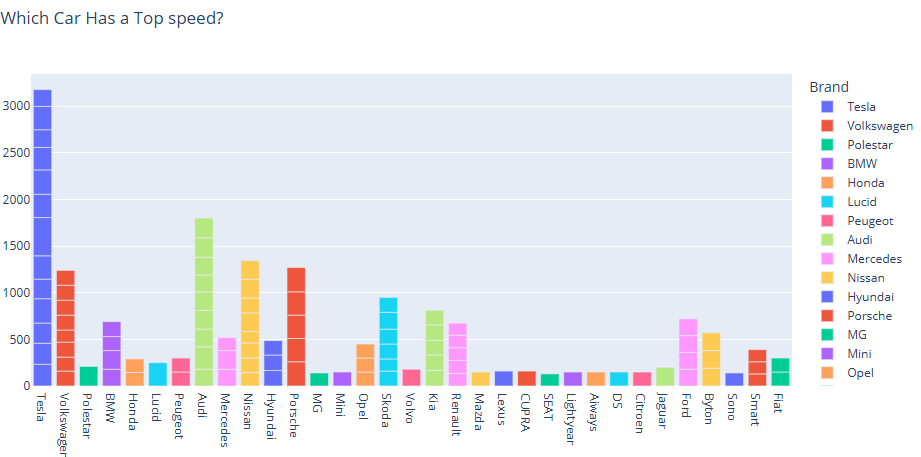
EV Cars Market Segmentation And Customer Segmentation

Amol Sahebrao Kasbe

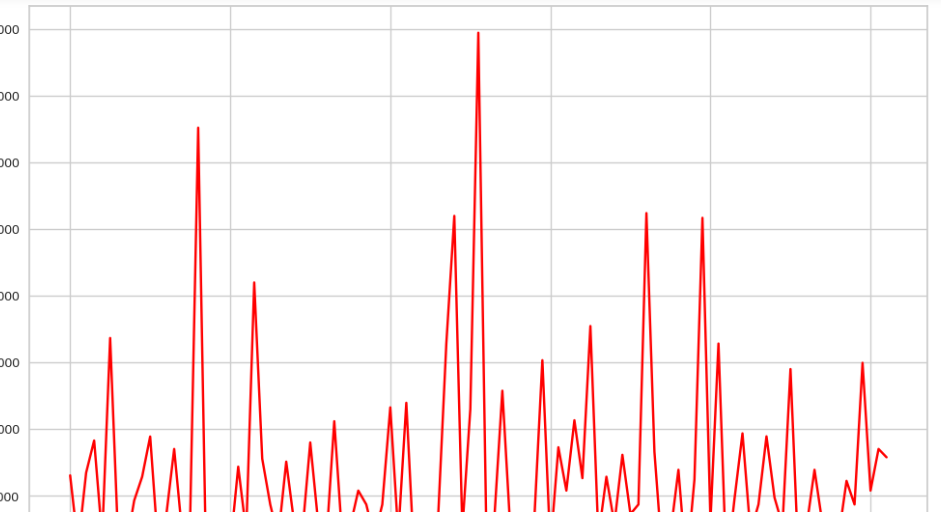
Github:

Objective:

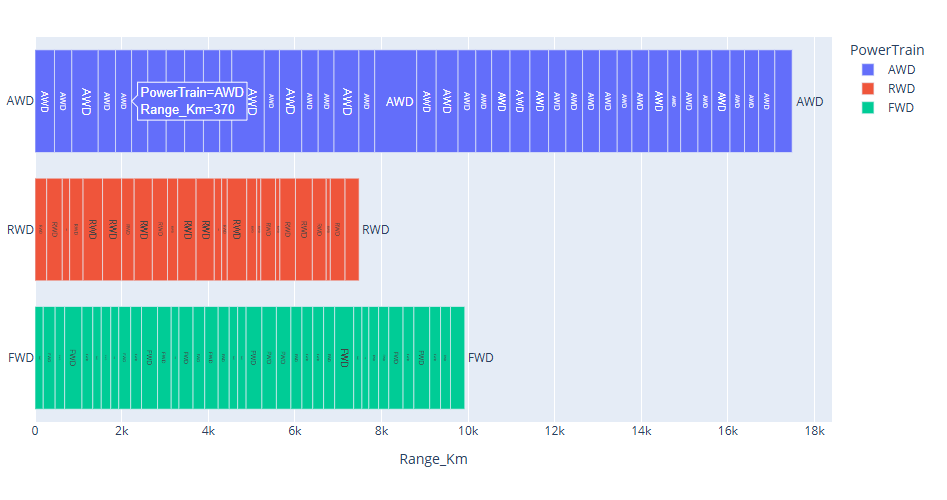
This dataset contains electric vehicle (EV) specifications, including acceleration, top speed, range, efficiency, charging speed, drivetrain, body style, segment, seating capacity, and price. Based on this data, here are some potential business objectives:



This graph helps in **benchmarking EV brands based on performance**, showing that Tesla leads in top speed, while other premium brands like Porsche and Audi also compete strongly. Meanwhile, budget-friendly EV brands focus more on efficiency and urban usability rather than top speed.

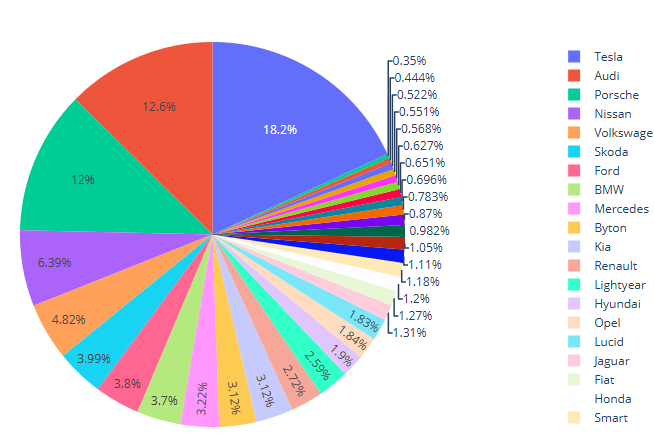


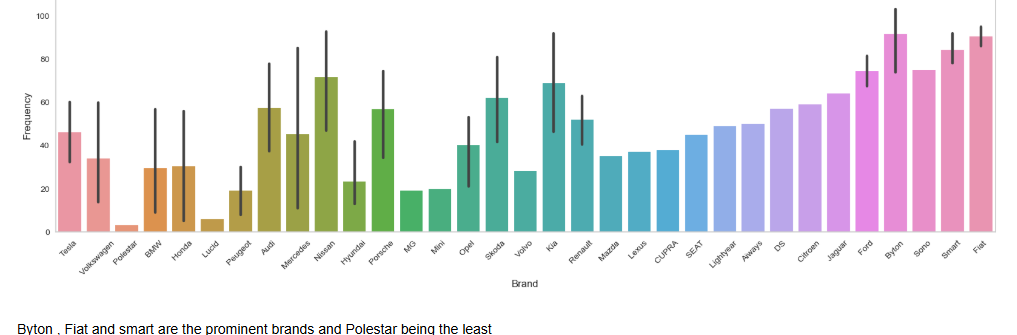
This graph suggests a **highly fluctuating dataset** where values rise and drop significantly. To provide a more precise interpretation, I'd need details on the **X-axis (time, categories, models?) and Y-axis (speed, efficiency, sales, price?).**



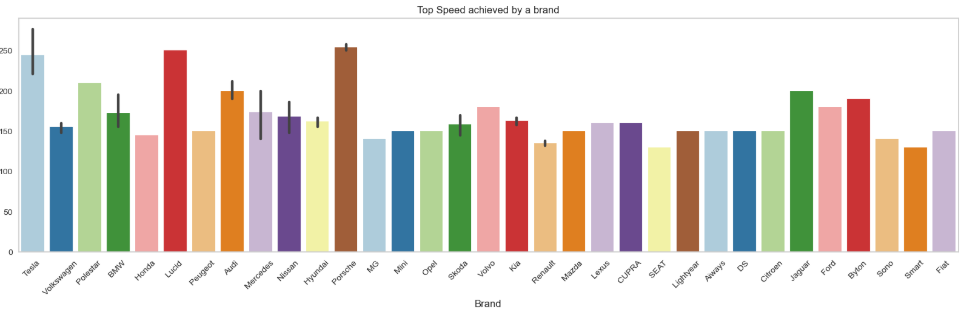
**Business Insights:**

* **Luxury & Performance Vehicles (AWD)** tend to have higher range, making them a premium choice.
* **Mass-market or affordable EVs (FWD, RWD)** focus on moderate range, targeting urban or mid-range customers.
* **Manufacturers might focus on AWD models for long-range efficiency** and RWD/FWD for cost-effective urban EVs.

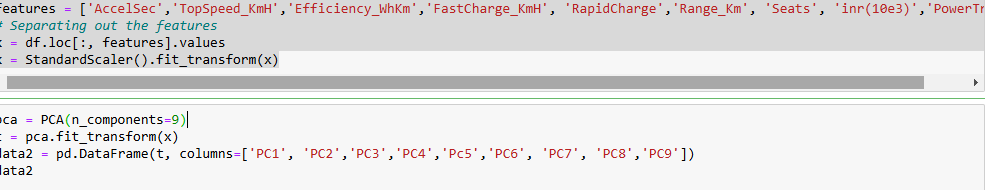




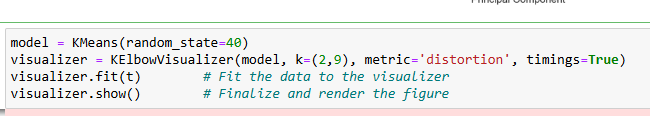
Byton , Fiat and smart are the prominent brands and Polestar being the least



Porsche, Lucid and Tesla produce the fastest cars and Smart the lowest



data2 now contains the transformed dataset with reduced dimensions. The new features (PCs) capture the most important information from the original features while reducing redundancy.



* The **elbow point** is where adding more clusters **doesn’t significantly reduce the distortion**.
* This point is the **ideal number of clusters** for **K-Means**.

📌 **Example Interpretation**:

* If the elbow occurs at k=4, it means that **4 clusters** is the best choice for K-Means.

#Customer Segmentation on age dataset:

### ****Business Objective for the Dataset****

This dataset appears to contain customer demographic and financial details related to purchasing decisions, possibly for **loan approval, credit risk assessment, or pricing strategy optimization**. Based on its structure, the **business objectives** could be:

**1. Loan Eligibility & Credit Risk Assessment**

* **Objective**: Develop a model to predict whether a customer should be granted a personal loan based on their age, profession, marital status, dependents, salary, and education.
* **Use Case**: Banks and financial institutions can assess the risk of default before approving a loan.

**2. Customer Segmentation for Financial Services**

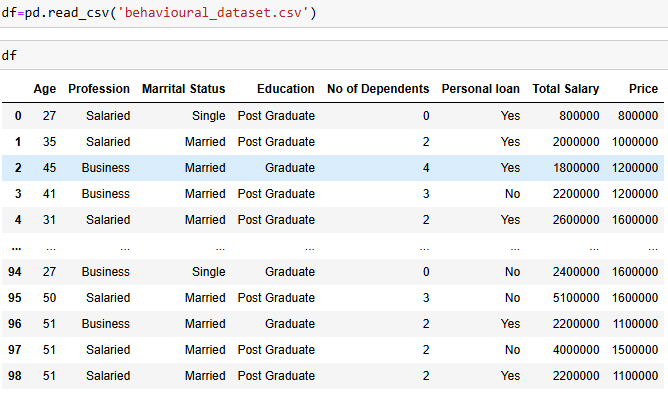
* **Objective**: Identify different customer segments based on salary, profession, and loan-taking behavior.
* **Use Case**:
  + Offer **personalized financial products** (e.g., higher credit limits for high-income salaried individuals).
  + Target specific customer groups for **loan promotions**.

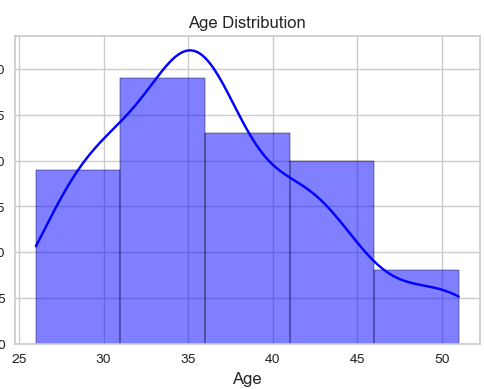
**4. Predicting Customer Purchase Behavior**

* **Objective**: Build a predictive model to estimate the price a customer is likely to pay based on their demographic and financial details.
* **Use Case**:
  + Helps **real estate companies** or **automobile dealers** in **price recommendations**.
  + Helps lenders assess the **maximum loan amount** that a customer can afford.

**5. Evaluating Financial Stability of Customers**

* **Objective**: Determine if a customer's **salary, dependents, and marital status** influence their ability to take and repay loans.
* **Use Case**: Banks can use this to set **credit limits** and determine eligibility for **higher-value loans**.





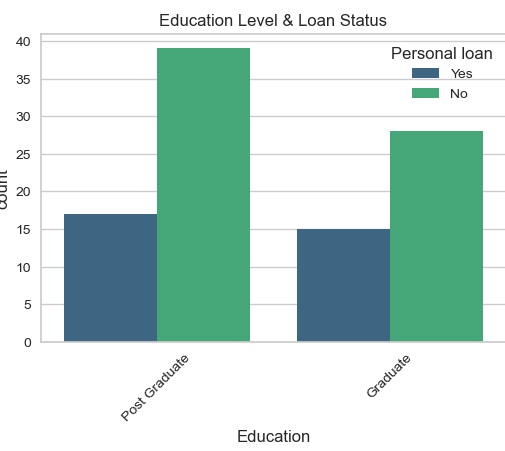
This **histogram with KDE (Kernel Density Estimation)** represents the **distribution of ages** in the dataset. Let’s break down each part of the code and what it visually represents:



 If **most customers have taken personal loans**, financial institutions might **target them with additional credit offers**.

 If **few customers take loans**, banks may need to **adjust interest rates** or **offer better terms** to attract more borrowers.

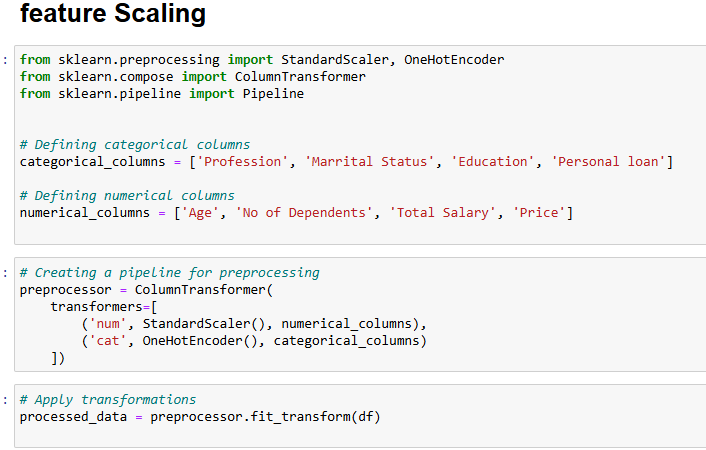
 Can be **cross-analyzed** with income, age, or profession to see which groups prefer taking loans.



 If **Post Graduates have more "Yes" bars**, they may be more financially capable or **more willing to take loans**.

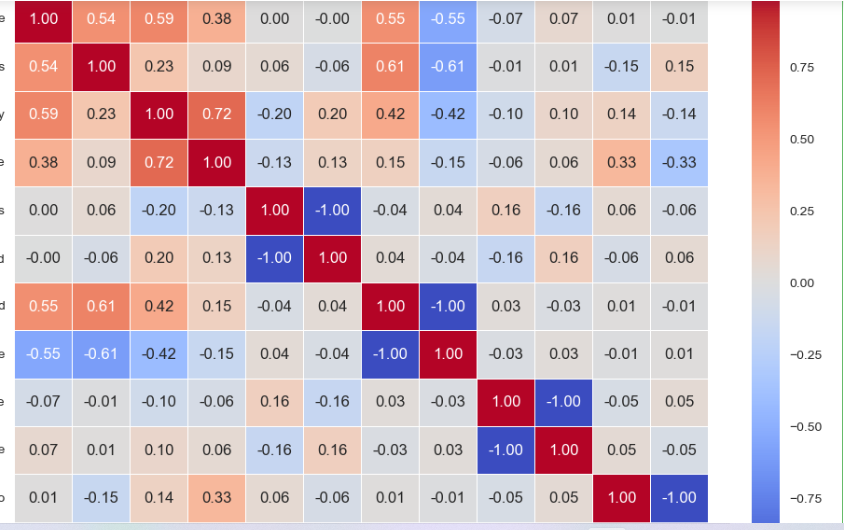
 If **Graduates have fewer loans**, they may be **less eligible** or **less inclined** to take loans.

 If there is **no significant difference**, it suggests **education level alone may not be a strong factor** in loan-taking behavior.



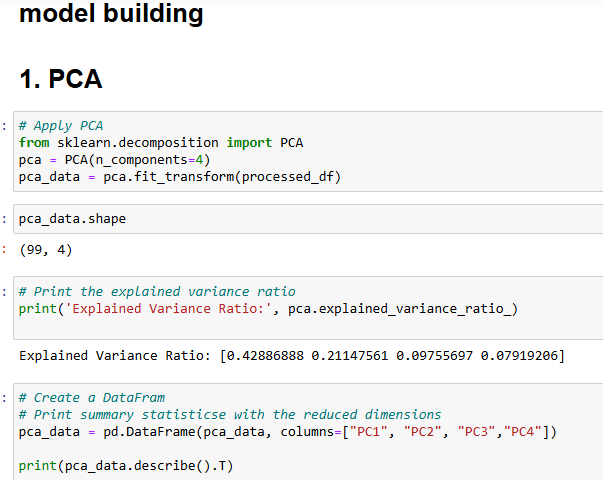
 The pipeline prepares your data for machine learning models by scaling numerical data and encoding categorical data in a structured way.

 This is essential when applying models that are sensitive to the scale of features



This code generates a **correlation heatmap** that visualizes the strength and direction of the relationships between numerical variables in your dataset (processed\_df).

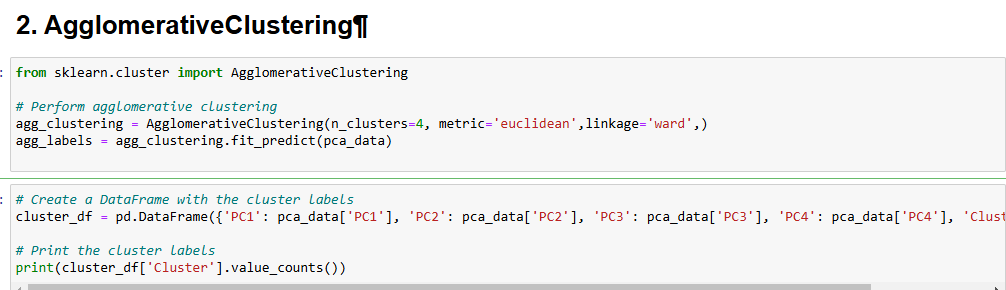
* **Red areas** (positive correlation) mean that as one variable increases, the other tends to increase as well.
* **Blue areas** (negative correlation) mean that as one variable increases, the other tends to decrease.
* The correlation coefficient values are annotated on the heatmap to make it easier to interpret.



This code performs PCA on a dataset, reducing its dimensionality from the original number of features to 4 principal components. It then:

* Prints the proportion of variance each component explains.
* Creates a new DataFrame with the transformed data.
* Displays summary statistics (like mean, standard deviation) for the 4 principal components.

PCA is useful for understanding the underlying structure of the data, visualizing it in lower dimensions, and potentially reducing noise or irrelevant features for further analysis or modeling.

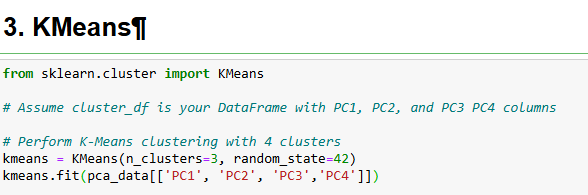


This code performs **Agglomerative Clustering** to group data into 4 clusters based on the 4 principal components obtained from PCA. After clustering:

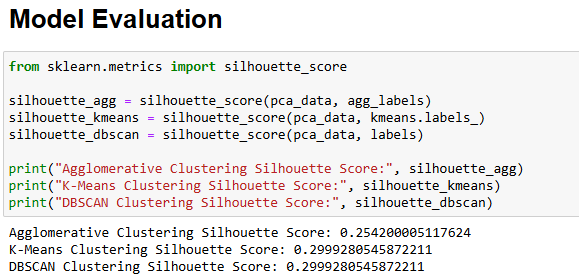
* A new DataFrame cluster\_df is created that contains the original PCA components along with the cluster labels.
* The cluster distribution is printed using value\_counts() to show how many data points belong to each cluster.

Agglomerative Clustering is useful when you want to discover hierarchical relationships between data points. This method doesn’t assume any prior knowledge of the number of clusters (but here, it is specified as 4) and allows for an interpretation of the structure of the data in a nested way.

4o mini



This code applies **K-Means clustering** with 3 clusters on the dataset pca\_data, which contains the first 4 principal components from PCA. The fit() method assigns each data point to one of the 3 clusters based on their proximity to the cluster centroids. The results will be stored in the kmeans object and can be accessed for further analysis, such as viewing cluster labels (kmeans.labels\_) or centroids (kmeans.cluster\_centers\_).



The Silhouette Score helps you understand how good the clustering result is:

* A higher score means better-defined and well-separated clusters.
* A lower or negative score suggests that the clusters might not be well-separated or that the algorithm misclassified data points.

Comparing the silhouette scores allows you to assess which clustering algorithm worked best for your dataset in terms of forming clear, well-separated clusters.