REPORT ON 2nd PROJECT

REQUIREMENTS

To perform this we need a data set and a python libraries such as Pandas, Matplotlib, Seaborn ,Scikit-learn and kaleido.

Importing libraries:

A screenshot of a computer program

Description automatically generated



A close-up of a white background

Description automatically generated

Loading the dataset:

A screenshot of a computer

Description automatically generated



Dataset 1-

A table with numbers and letters

Description automatically generated

Dataset 2-

A screenshot of a data

Description automatically generated

Performing EDA:

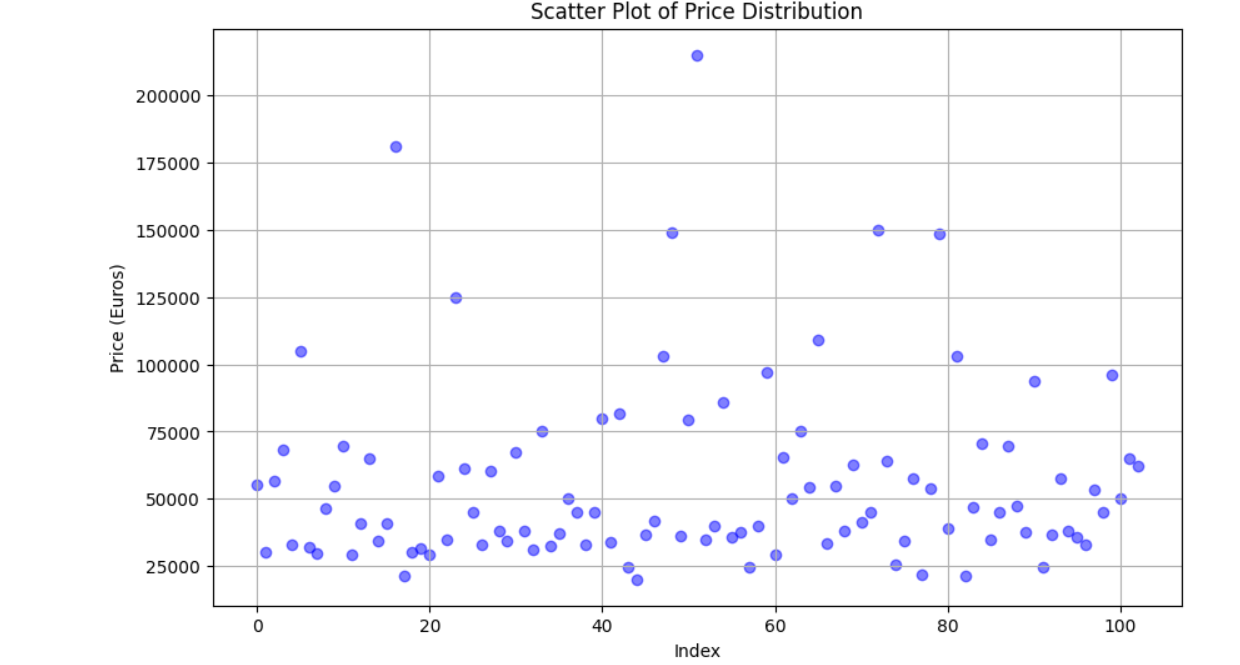
Exploratory Data Analysis (EDA) is a crucial step in data analysis that involves examining and understanding the structure, patterns, and characteristics of a dataset.

Exploratory Data Analysis (EDA) is a crucial step in data analysis that involves examining and understanding the structure, patterns, and characteristics of a dataset. Here's a general outline of how you might approach EDA:

1. **Data Collection**: Gather the dataset you want to analyze. This could be from various sources such as databases, CSV files, APIs, etc.
2. **Initial Inspection**:
   * Check the first few rows of the dataset to understand its structure.
   * Look for missing values, outliers, and inconsistencies in the data.
3. **Summary Statistics**:
   * Calculate basic statistics such as mean, median, mode, standard deviation, minimum, maximum, etc., for numerical columns.
   * For categorical variables, count the frequency of each category.
4. **Data Visualization**:
   * Use plots such as histograms, box plots, scatter plots, and bar charts to visualize the distribution of numerical data, identify outliers, and understand relationships between variables.
   * For categorical data, use bar charts, pie charts, and stacked bar charts to visualize the distribution of categories.
5. **Correlation Analysis**:
   * Calculate correlation coefficients (e.g., Pearson correlation for numerical variables) to understand the linear relationship between variables.
   * Visualize correlations using heatmaps for better interpretation.
6. **Feature Engineering**:
   * Create new features if necessary based on domain knowledge and insights gained during EDA.
   * Transform variables (e.g., log transformation for skewed data) to make them more suitable for modelling.
7. **Handling Missing Values and Outliers**:
   * Decide on strategies to handle missing values (e.g., imputation, deletion) based on the extent of missingness and domain knowledge.
   * Identify and potentially remove outliers that can significantly affect analysis and modelling.
8. **Data Quality Check**:
   * Validate data quality by cross-checking against domain knowledge and business rules.
   * Ensure data consistency and correctness.
9. **Data Segmentation**:
   * If applicable, segment the data based on certain criteria to analyse subsets separately (e.g., segmenting customers based on demographics).
10. **Documentation**:
    * Document your findings, insights, and decisions made during EDA.
    * Summarize key takeaways and prepare

Implementing EDA:

Scatter plot:



Line chart : A line graph with a blue line

Description automatically generated

In this visualize between range (kilometres) and price.

Violin chart:

A diagram of a graph

Description automatically generated with medium confidence

Bar graph:

In this we are going to explore the top speed of each vehicle brand.

A graph of different colored bars

Description automatically generated

Doughnut chart:

In this we are looking at different types of plug time and their percentage of use.

A blue circle with different colored circles

Description automatically generated

Line chart:

A blue line graph with black text

Description automatically generated

A graph of a graph

Description automatically generated

A blue line graph with numbers

Description automatically generated

A graph showing a speed of up to a speed of up to a speed of up to a speed of up to a speed of up to a speed of up to a speed of up to a

Description automatically generated

K-MEANS CLUSTERING

Elbow Method:

A graph with a line

Description automatically generated

Convex Clustering:

A graph with blue dots and a line

Description automatically generated

2ND DATA SET

Plotting the count plot on different martial status.

A graph of a car loan

Description automatically generated

A graph of a car loan

Description automatically generated

Plotting the Doughnut chart car loans:

A pie chart with numbers and text

Description automatically generatedPlotting the frequency of the different customers:

A comparison of graphs with numbers

Description automatically generated with medium confidence

Optimal number of Clusters Using Elbow method:

A graph with a line

Description automatically generated

A chart with green and red dots

Description automatically generated

A chart with green and red dots

Description automatically generated

A graph with green and red dots

Description automatically generated

SEGMENTATION APPROACH

Clustering : Clustering is a technique used in unsupervised machine learning to group similar data points together. In Python, you can perform clustering using libraries such as Scikit-learn or Keras (with TensorFlow backend).

K-means clustering : K-means clustering is a popular unsupervised machine learning algorithm used for clustering data into distinct groups.

Here is the result:

A graph of a number of clusters

Description automatically generated with medium confidence

Elbow Method : The Elbow Method is a technique used to determine the optimal number of clusters (K) in K-means clustering. It involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow" point where the rate of decrease in WCSS slows down significantly. This point indicates the optimal number of clusters.

A line graph with numbers

Description automatically generated

Kaleido: Is a Python library that enables the generation of static image files (such as PNG, JPEG, SVG) directly from Plotly figures without relying on an external rendering engine. This library is particularly useful for creating high-quality static images of interactive Plotly visualizations.

If I had given additional time, I would like to search columns such as range and charging time. I feel this will keep evolving in terms of providing the comfort to the customers. The other column will be the price.

I will try kaleido. Which I just imported the library in this project but couldn’t work on. I was getting the errors and I was unable to make on time. The other model that I would like to work on is Average Silhouette method. This measures how similar a cluster is to its cluster set compared to other cluster.