Applied Data Science – Assignment on Clustering and Fitting

Student Name: Vaishali

Student ID: 23038567

GitHub Link: https://github.com/Vaishali8567/Applied-Data-Science---Clustering-and-Fitting

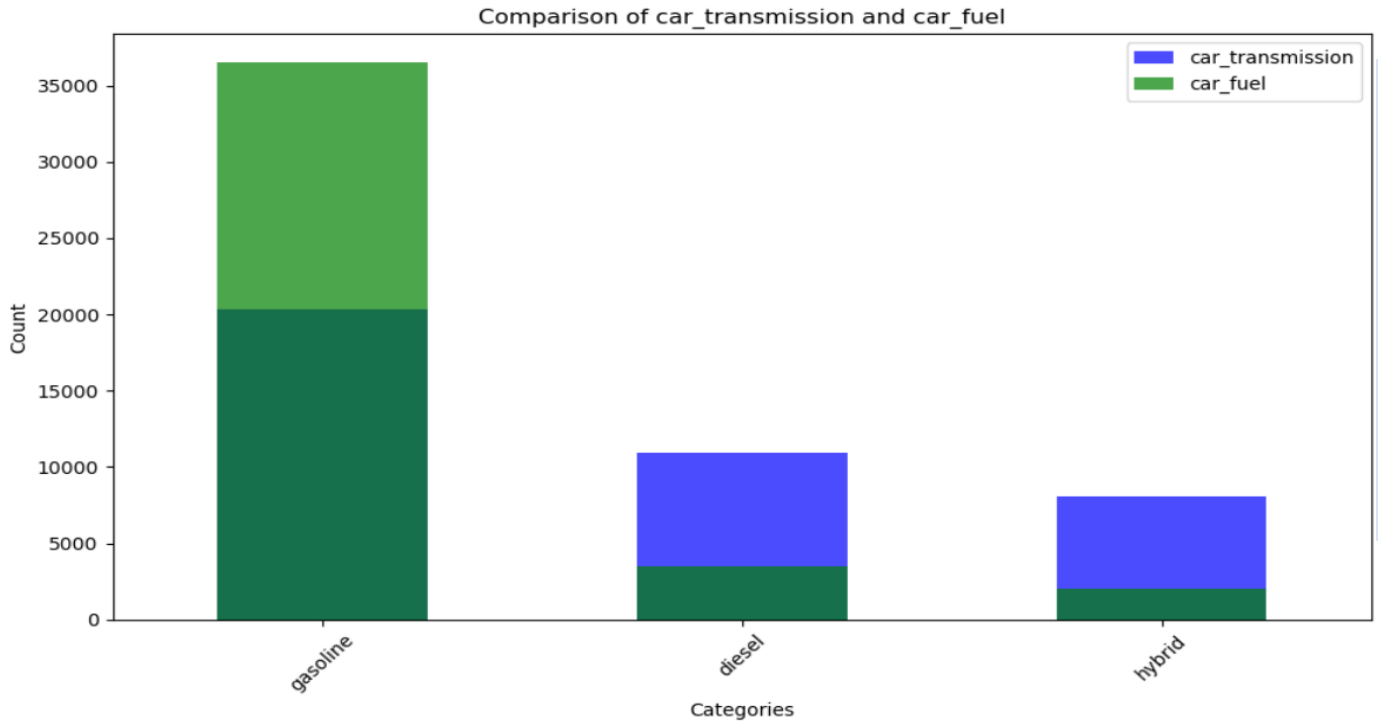
Dataset Link: https://www.kaggle.com/datasets/volkanastasia/dataset-of-used-cars

I. Introduction:

The "used cars" dataset contains information about various attributes of used cars such as brand, model, price, mileage, age, fuel type, and transmission type. Various visualizations and statistical techniques are used to gain insights into the relationships between different car attributes.

II. Findings:

- **Categorical Graph: *Bar Plot***



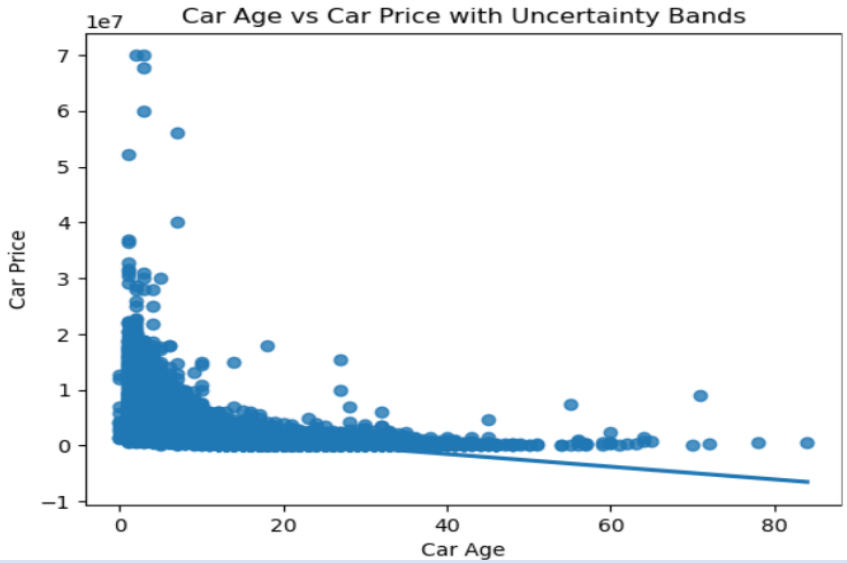
Interpretation:

The comparison aids in understanding the distribution of various transmission and fuel types among the vehicles in the dataset.

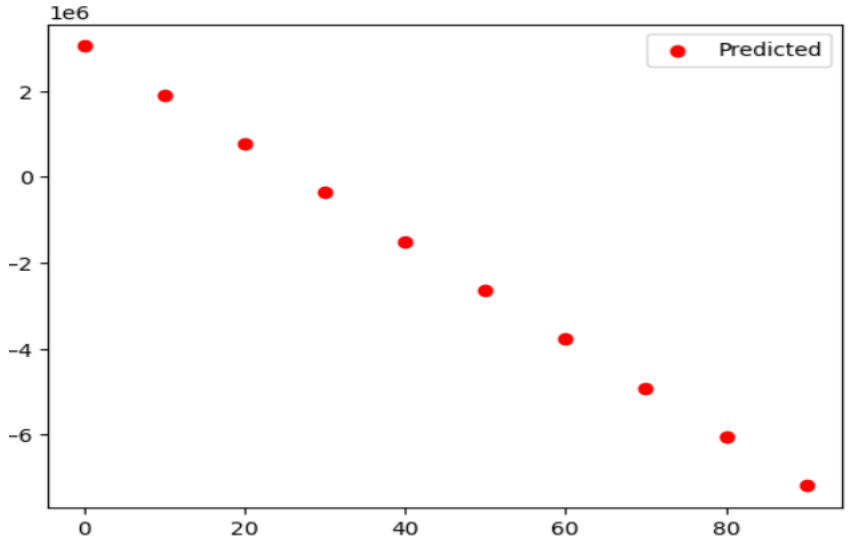
The relative heights of bars for each category can be used to estimate the popularity or prevalence of certain transmission and fuel types in the dataset.

Significant discrepancies in category counts between the two columns could reveal potential linkages or dependencies between transmission type and fuel type.

Relational Graph: ***Scatter Plot with uncertainty bands and predicted points***

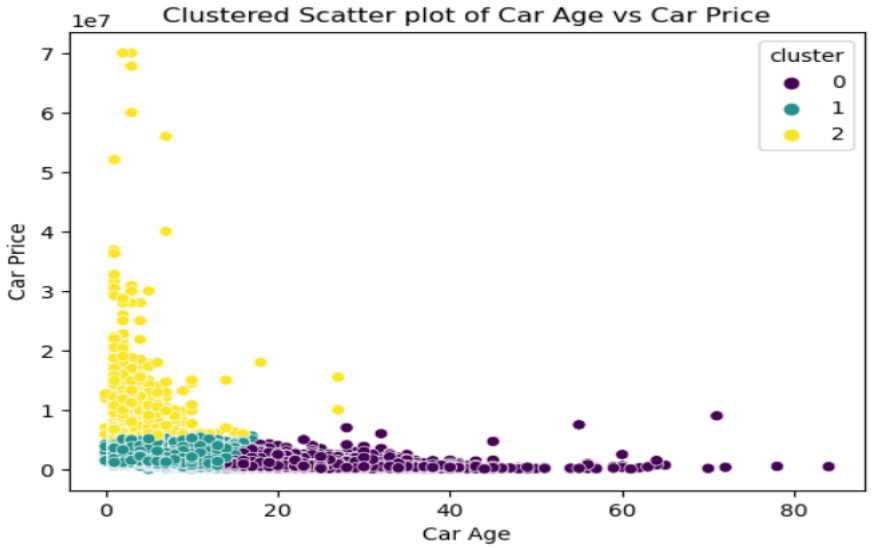


Interpretation: The downward slope of the line indicates a negative correlation between car age and price. A wider band suggests higher variabilities. the relationship between car age and price is less predictable within that range.

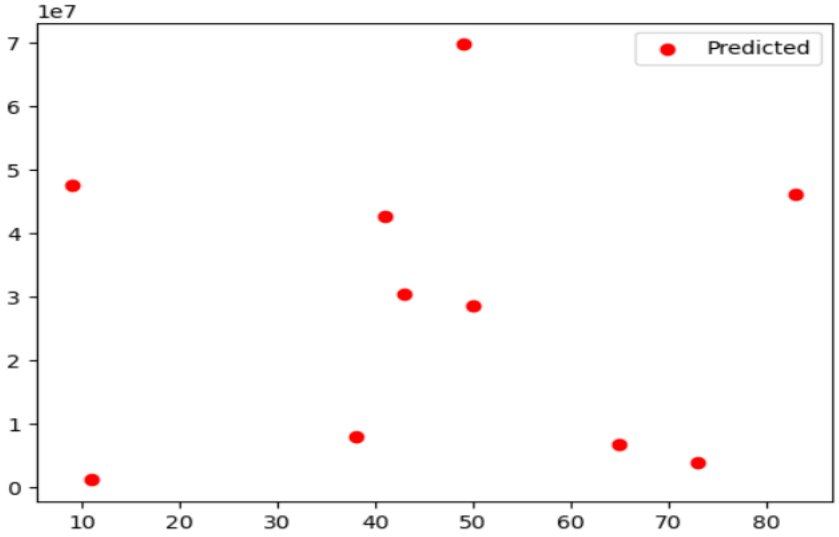


Interpretation: A pricing estimate for cars older than any observed data point can be found in these anticipated points. They assist predict future price trends based on the proven association between automobile age and price

Relational Graph: ***Clustered Scatter Plot with predicted points***



Interpretation: Different colors indicate the various clusters discovered by the KMeans clustering method. It enables us to find patterns or trends in the relationship between vehicle age and pricing within each cluster.

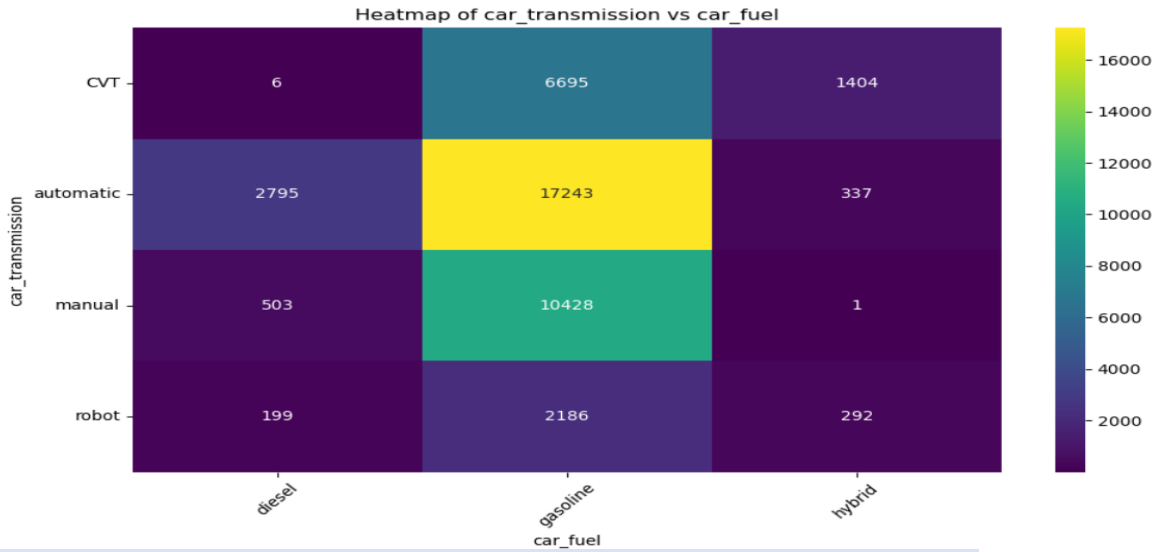


Interpretation: Based on the clustering patterns in the dataset, the anticipated values indicate where new data points may lie within the age-price space. A comparison of anticipated values and actual data points can assist in determining the clustering model's accuracy and predictive capabilities.

Statistical Depth

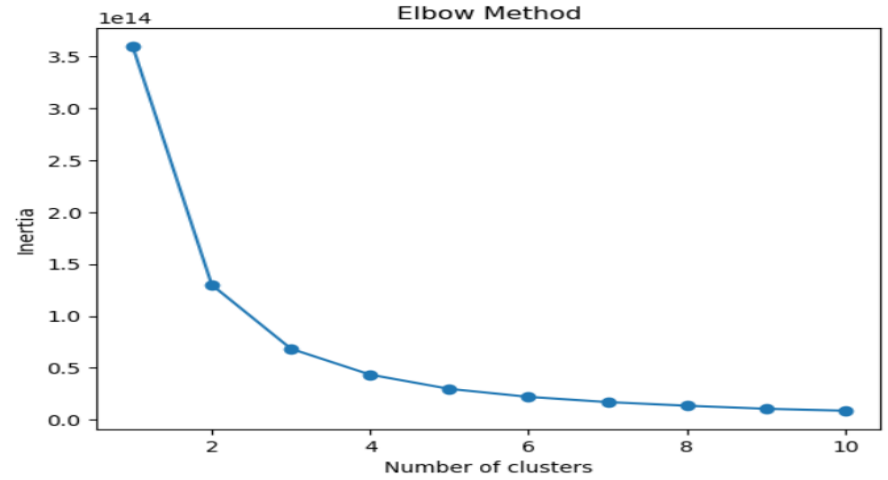
```
Mean: 1712716.6593409204
Median: 1250000.0
Standard Deviation: 1976669.3965921274
Skewness: 8.385723462913422
Kurtosis: 173.94108094393107
```

Statistical Graph: *Heat Map*



Interpretation: Darker tones represent higher frequencies, implying that certain transmission and fuel types are commonly used. In contrast, lighter tones suggest fewer common combinations. The diagonal cells show examples where the transmission and fuel types are the same, providing information about the frequency of homogenous pairings.

Clustering and Fitting




```
Line equation: y = -113683.73731538343x + 3042504.956937287
```

Interpretation: The plot shows the link between the number of clusters (k) and the inertia, which is the sum of the squared distances between samples and their closest cluster center.
The plot often shows a decrease in inertia as the number of clusters grows. However, adding more clusters has diminishing returns in terms of reducing inertia.
The resulting slope and intercept coefficients define the fitted line's equation (y = mx + b), where 'm' represents the slope and 'b' represents the intercept.

Conclusion:

This analysis outlines the important findings and implications of the exploratory data analysis, emphasizing the importance of understanding the dataset's characteristics and variable relationships within the context of the used automobile market.