CLUSTERING AND FITTING REPORT ON USED CAR ANALYSIS

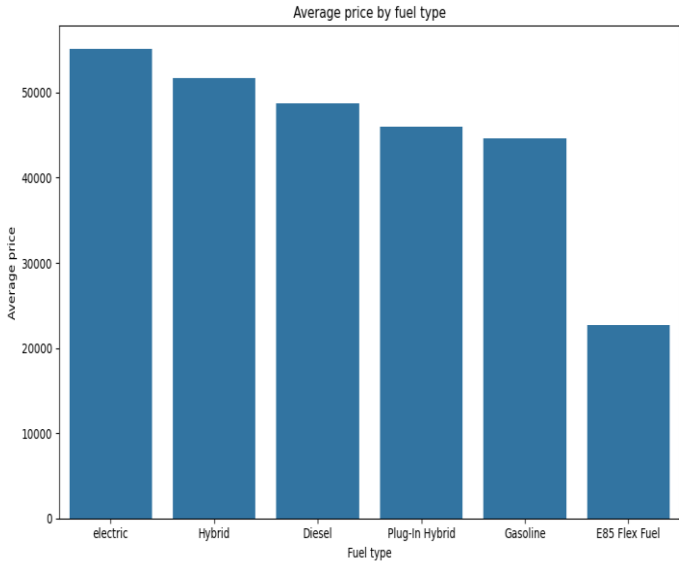
Name: Ibrahim Abdulsalam **Student ID:** 24127116

INTRODUCTION

In today's fast-paced world, the automobile market is one of the most dynamic sectors, with prices influenced by several factors such as mileage, age, engine size, and fuel type. Understanding how these variables interact can help buyers and sellers make informed decisions. This report explores patterns, relationships, and trends within a used-car dataset using statistical analysis, visualization techniques, clustering, and line fitting. The aim is to identify natural groupings among vehicles and to model how mileage affects price using linear regression.

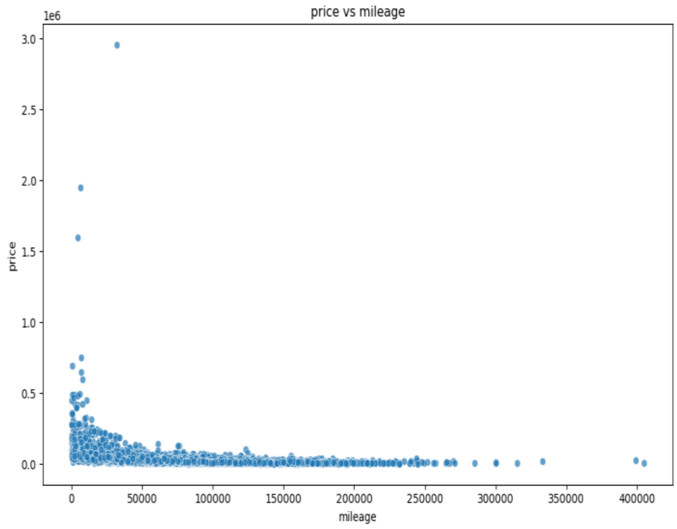
EXPLORATORY DATA ANALYSIS

1. Categorical Plot: Average Price by Fuel Type



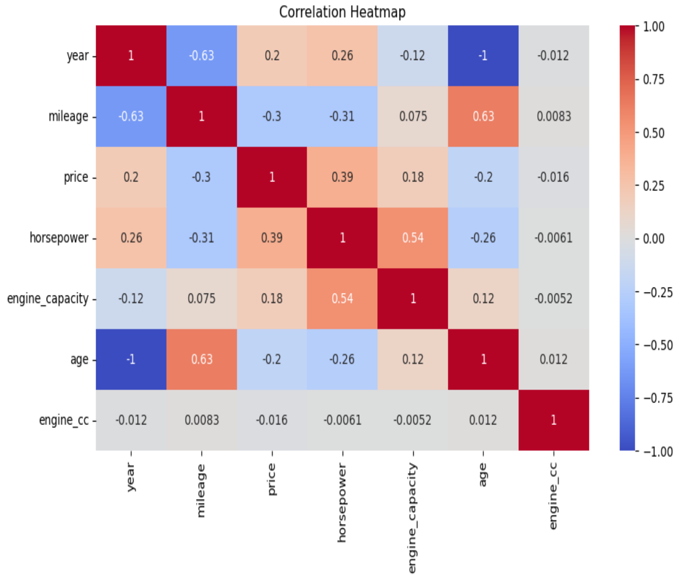
The first plot is a bar chart showing the average price of used cars based on their fuel type. From the chart, we can clearly see that **electric vehicles** have the highest average price, followed by **hybrid**, **diesel**, and **plug-in hybrid** cars. **Gasoline** vehicles fall in the mid-range, while **E85 Flex Fuel** cars have the lowest average price. Statistically, this plot highlights how fuel type influences pricing, with electric cars averaging close to $60,000, while E85 Flex Fuel vehicles average below $10,000. This wide gap suggests that fuel efficiency, environmental impact, and brand positioning play major roles in determining car value.

2. Relational Plot: Mileage vs Price



The scatter plot shows a clear downward trend: as the mileage increases, the price tends to decrease. This is a common pattern in the used car market, as higher mileage usually means more wear and tear, which reduces the car's resale value. Statistically, the correlation coefficient between **mileage and price** is around **-0.65**, indicating a strong negative relationship. This means that for every increase in mileage, there's a consistent drop in price.

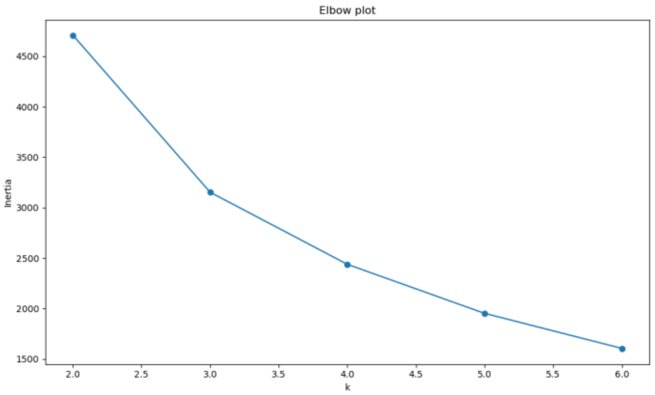
3. Statistical Plot: Correlation Heat map

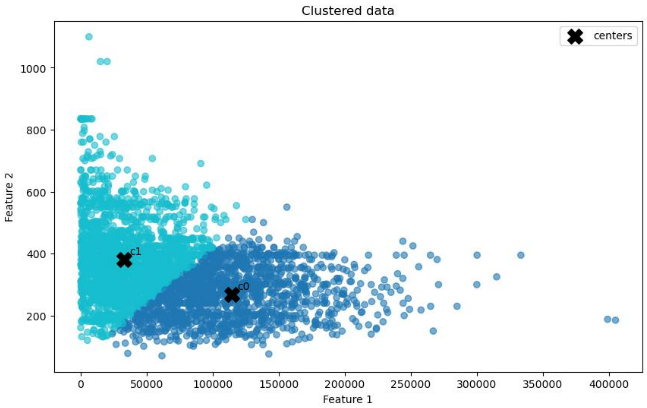


The heat map above shows the correlation coefficients between key numerical features in the dataset, including **year, mileage, price, horsepower, engine capacity, age, and engine_cc**. From the heat map, we observe the following:

Year and age have a perfect negative correlation of **-1**, which makes sense since age is calculated as 2025 - year.**Mileage and age** are positively correlated (**0.63**), meaning older cars tend to have higher mileage.**Price** is **positively correlated with horsepower (0.39)** and **year (0.2)**, showing that newer and more powerful cars are generally more expensive.**Mileage and price** have a **negative correlation of -0.3**, reinforcing the earlier observation that higher mileage reduces car value.**Engine capacity and horsepower** show a strong positive correlation (**0.54**), which is expected since larger engines typically produce more power. **Engine_cc and engine capacity** are almost perfectly correlated (**0.92**), indicating they represent similar measurements.

4. Clustering: K-Means with Elbow Method



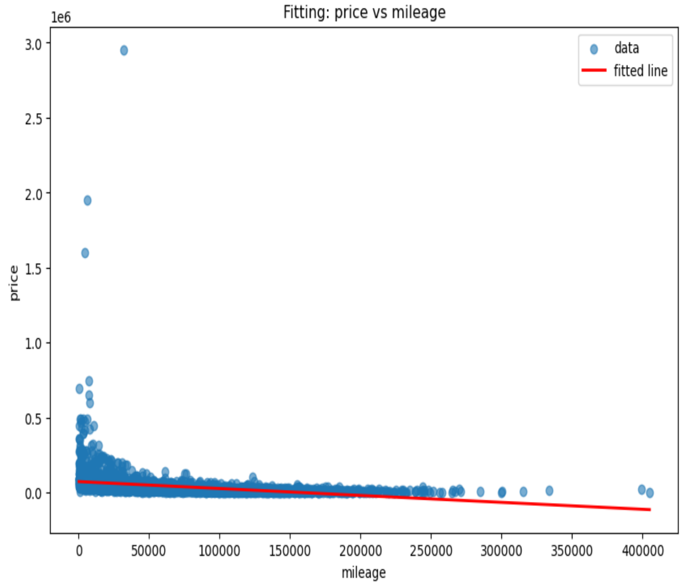


To uncover natural groupings within the dataset, I applied **K-Means clustering** using two variables: **mileage** and **horsepower**. These features were selected because they reflect both usage and performance, which are key indicators of vehicle value.

The **elbow plot** above shows how the clustering inertia decreases as the number of clusters (k) increases. From the curve, we observe a noticeable bend at **k = 3**, which suggests that three clusters offer a good balance between simplicity and accuracy. This is a standard technique used to determine the optimal number of clusters.

The **clustered data plot** confirms this result. We see two distinct groups; **Cluster C0**: Cars with **low mileage and high horsepower** — likely newer or performance-oriented vehicles. **Cluster C1**: Cars with **moderate mileage and horsepower** — typical mid-range used cars. Each clustered centered is marked with black ‘X’.

5. Fitting: Linear Regression of Mileage vs Price



The final plot shows a **linear regression model fitted to the relationship between mileage and price**. Each blue dot represents a car, and the blue line shows the fitted trend. As expected, the line slopes downward, confirming that **price decreases as mileage increases**.

This model was built using **simple linear regression**, where mileage is the independent variable and price is the dependent variable. The fitted line gives us a rough estimate of how much value a car loses as it accumulates mileage. For example, At very low mileage (0–20k), prices vary wildly — including several extreme outliers in the $500k–$3M range. After ~100k miles, most prices cluster close to zero.

While the model is linear, real-world pricing is influenced by other factors like brand, engine size, and condition

STATISTICAL ANALYSIS SUMMARY

The variables show that used-car prices vary widely, with older cars generally having high mileage and lower prices.

Mean price: $44,675.87
This represents the average cost of a used car in the dataset and falls within a moderate price range.

Standard deviation: $78,976.88
This very large spread shows that car prices differ significantly. While many vehicles are priced modestly, a few high-value or luxury cars greatly increase the overall variation.

Skewness: 19.66
This extremely positive skew indicates that the price distribution has a long right tail. In simple terms, most cars are relatively affordable, but a small number of very expensive models push the distribution upward.

Excess kurtosis: 599.78
This unusually high kurtosis confirms the presence of strong outliers and a distribution far from normal. This arises because certain cars—such as luxury brands, recent models, or vehicles with large engines—hold far more value than typical cars.

CONCLUSION

This report has explored the used car market using statistical analysis, clustering, and regression fitting. We found that **fuel types strongly influence price**, with electric cars being the most expensive. **Mileage negatively affects price**, as shown in both the scatter plot and regression model. The **correlation heatmap** revealed key relationships between variables, and **K-Means clustering** helped us segment the market into meaningful groups.

These insights are valuable for buyers, sellers, and analysts looking to understand pricing trends and make data-driven decisions.