**DSA210 - FINAL REPORT**

**Emre Pelit**

**32307**

**Price Prediction Using Car Attributes (Hyundai & Mercedes)**

**Supervised by Berke Odacı**

**SABANCI UNIVERSITY**

**Introduction to Report**

This project aimed to predict the prices of used Hyundai and Mercedes vehicles using various car attributes such as engine size, mileage, age, fuel type, and model. The goal was to build a regression model that captures the relationship between these features and car price, providing insights into which attributes are most influential in pricing.

**Dataset Description**

Columns used:

* model: Categorical variable that represents the model of the car.
* transmission: Categorical variable that represents the transmission type of the car.
* fuelType: Categorical variable that represents the fuel type of the car.
* price: Numeric variable corresponding to the price of the car.
* year: Numeric variable that represents the production year of the car.
* mileage: Numeric variable that represents the amount of miles the car is used.
* mpg: Numeric variable corresponding to the miles per gallon.
* engineSize: Numeric variable corresponding to the engine capacity of the car.

Feature Engineering:

* Created new column:

age = 2025 - year

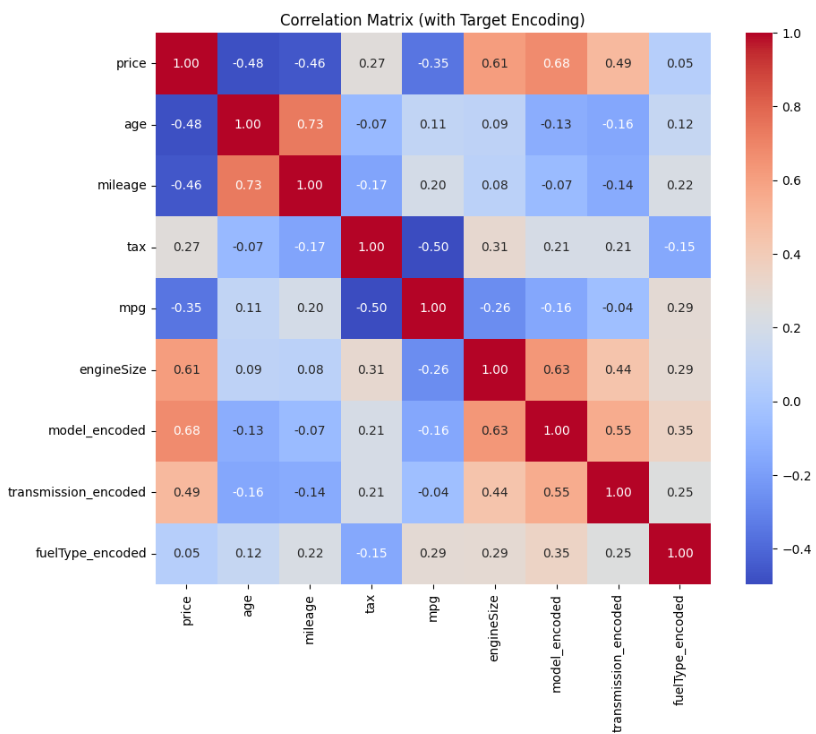
* Removed outliers:

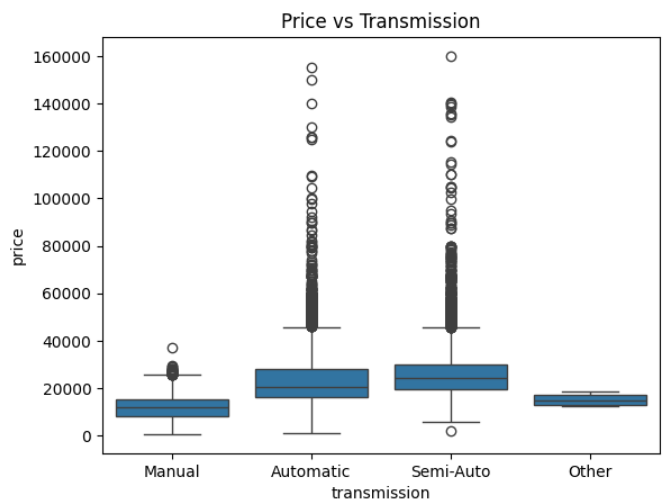
engineSize = 0 rows

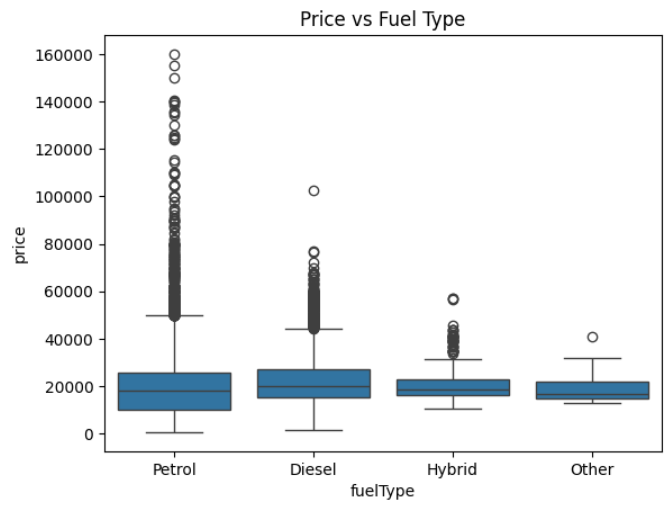
Price outside [1000, 100000]

“Other” categories in fuelType and transmission

**Exploratory Data Analysis (EDA)**





,

Observations:

* Histograms and boxplots were used to understand the distribution of price, mileage, mpg etc.
* Correlation analysis showed:
* Negative correlation between price and age/mileage.
* Positive correlation between price and engine size.
* Price distributions across different models, transmissions, and fuel types were visualized using boxplots.

**Hypothesis Testing**

* Transmission vs Price: T-test showed a statistically significant difference (p<0.05) between manual and automatic transmissions.
* Fuel type vs Price: T-test showed petrol and diesel cars differ significantly in average price.
* Brand vs Price: Mercedes cars have significantly higher average prices than Hyundai.

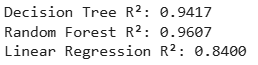
**Model Building and Evaluation**

* Linear Regression:

Using 4 numeric variables (mileage, engineSize, age, mpg) and 3 categorical variables (model, transmission, fuelType), linear regression model is created.

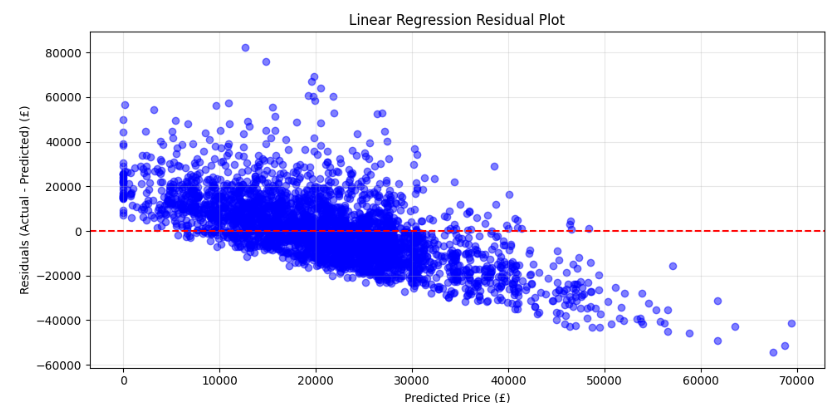
* Decision Tree
* Random Forest

**Comparison**

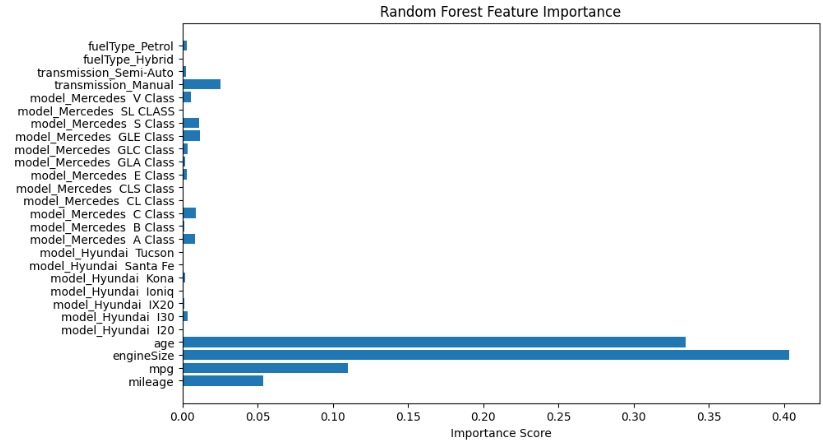
****

* Random forest had the best accuracy, but Linear Regression offered easier interpretation.

**Visualizations**



* The residual plot above visually supports the assumption that price has a linear relationship with the predictors.



* The bar chart derived from the Random Forest model confirms the dominance of engine size, age, mpg, mileage.

**Conclusion**

The project successfully achieved its goal of predicting car prices using both numerical and categorical features. The data cleaning and filtering processes eliminated noise and rare categories. While Random Features achieved the highest R square value, Linear Regression offered strong interpretability, allowing us to clearly understand the impact of each variable.

Key insights:

* Larger engine sizes and Mercedes models drive higher prices.
* Older cars and higher mileage reduce predicted value.
* Specific models (e.g., S Class, GLE Class) contribute significantly.

**Future Work**

To improve the project further:

* Try more advanced models like XGBoost, which might increase accuracy.
* Try using combinations of features together, or adjust their scale to improve the performance of some models.
* Use methods that explain how each feature affects the prediction.