DSA 210 - Introduction to Data Science

The Impact of Mileage across Different Car Brands

Cem Kaya - 31957

My Dataset:

- Car brand and model
- Year of manufacture
- Mileage (km)
- Fuel type (gasoline, diesel, electric)
- Price (feature to be analyzed)
- Location

Future Engineering and Transformation

By Feature Transformation and Enrichment:

- car_age. *(By current year - car model year)*
- brand_classification (luxury,economy..) *(by classification)*
- log_price *(by mathematical transformation)*
- log_mileage *(by mathematical transformation)*
- price_per_km *(by mathematical transformation)*

-> data points are added to the dataset to enrich the dataset.

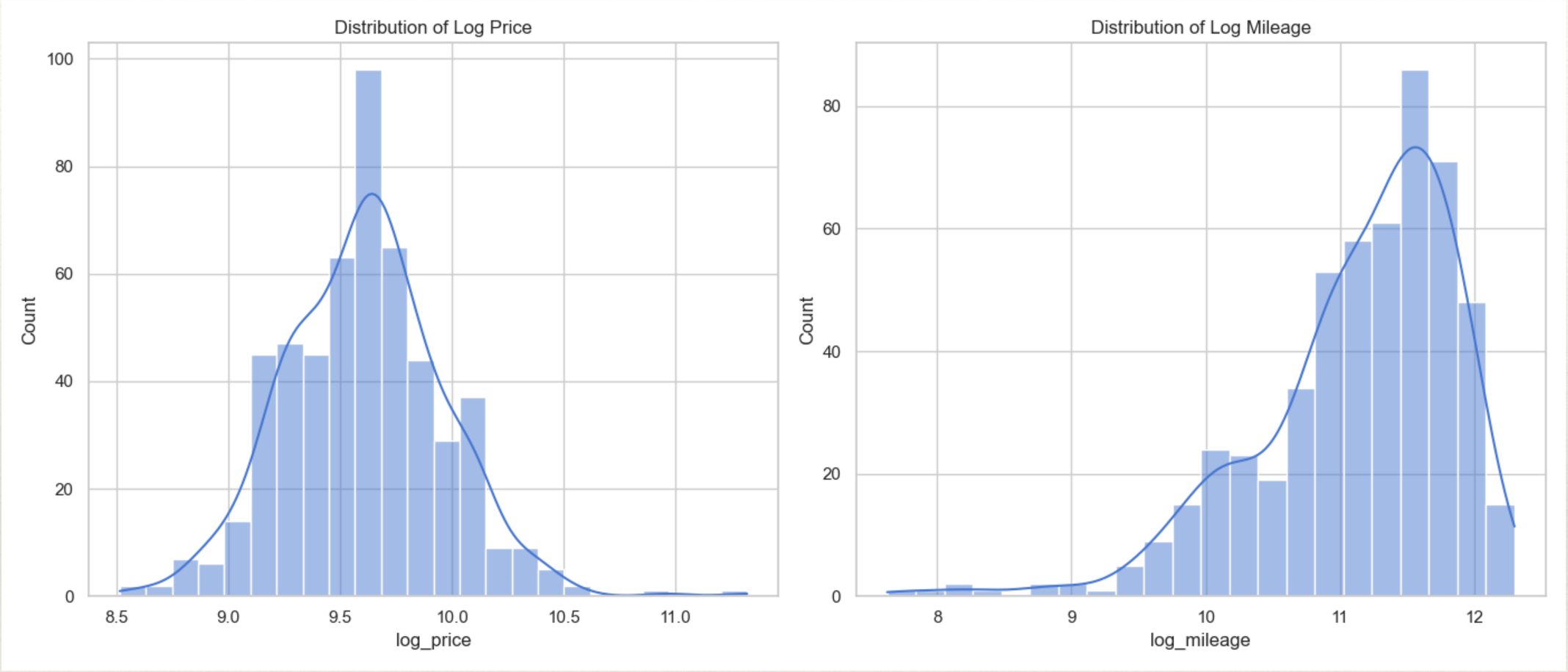
Hypothesis

Null Hypothesis (H0): "The price sensitivity to mileage does not have a different relationship across different car brands regardless of them being a luxury brand or an economy brand."

Alternative Hypothesis (H1): "The price sensitivity to mileage changes across different car brands, with luxury brands showing higher decrease per kilometer compared to economy brands."

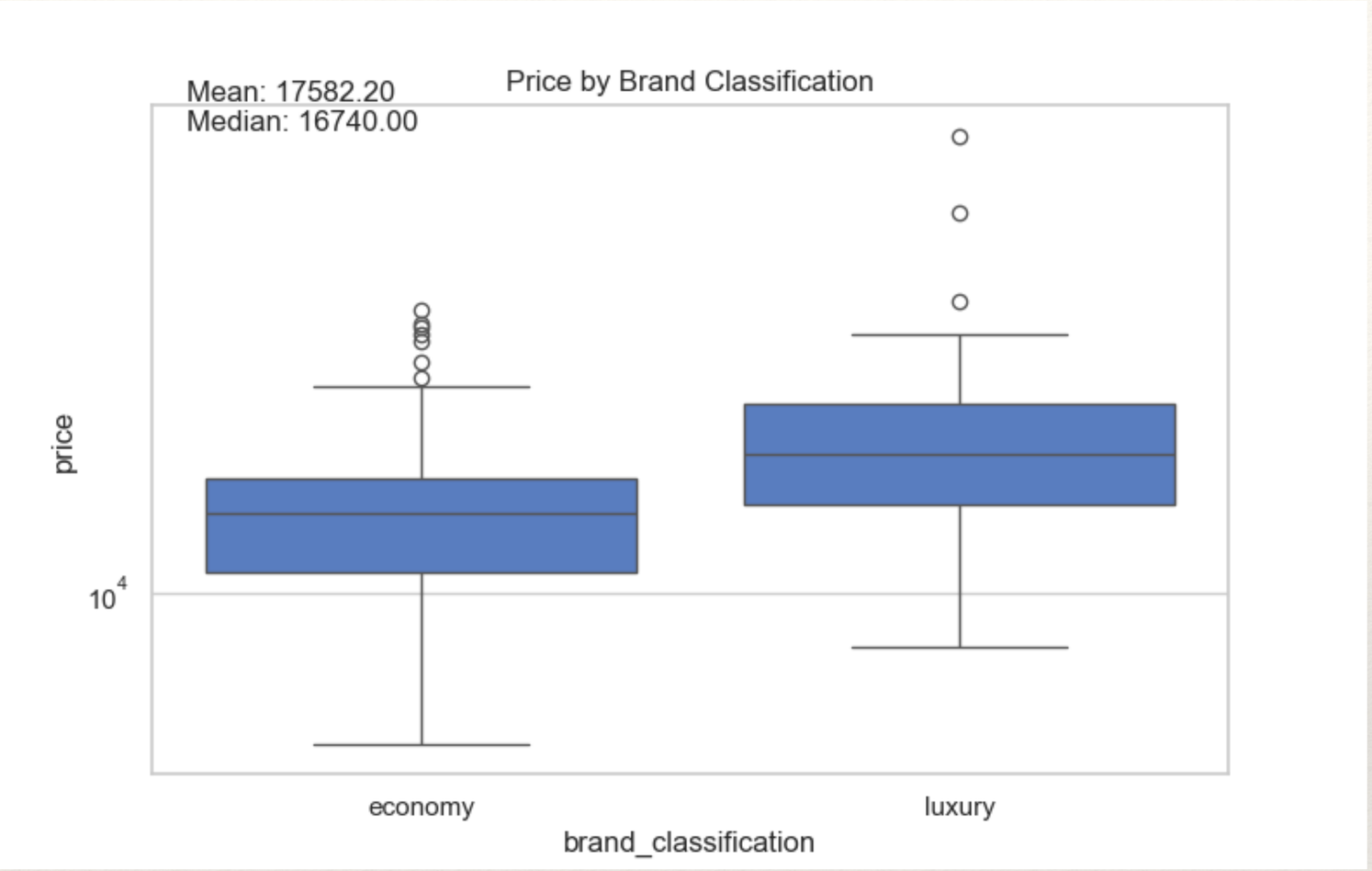
Exploratory Questions:

- Which brands hold their value the longest?
- Can I estimate a fair market price for a car based on key attributes?
- Are some car listings significantly overpriced or underpriced compared to the estimated market price?



Mean: 9.61
Median: 9.62

Mean: 11.13
Median: 11.32



From this graph we can conclude luxury car brands have higher prices than economy car brands in general.

Economy car brands:

Pearson Coefficient: -0.38

P-value: 2.90e-16

Luxury car brands:

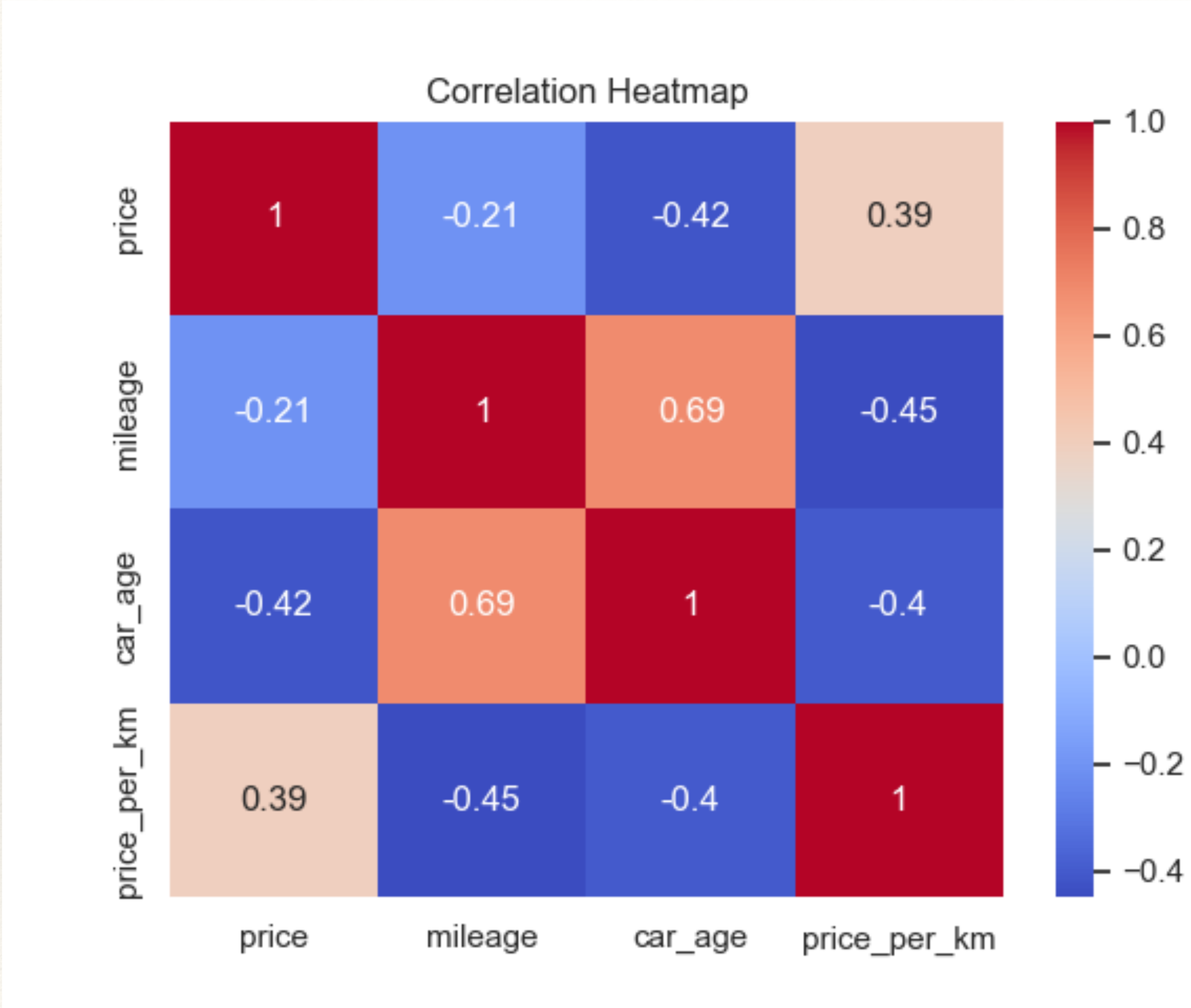
Pearson Coefficient: -0.34

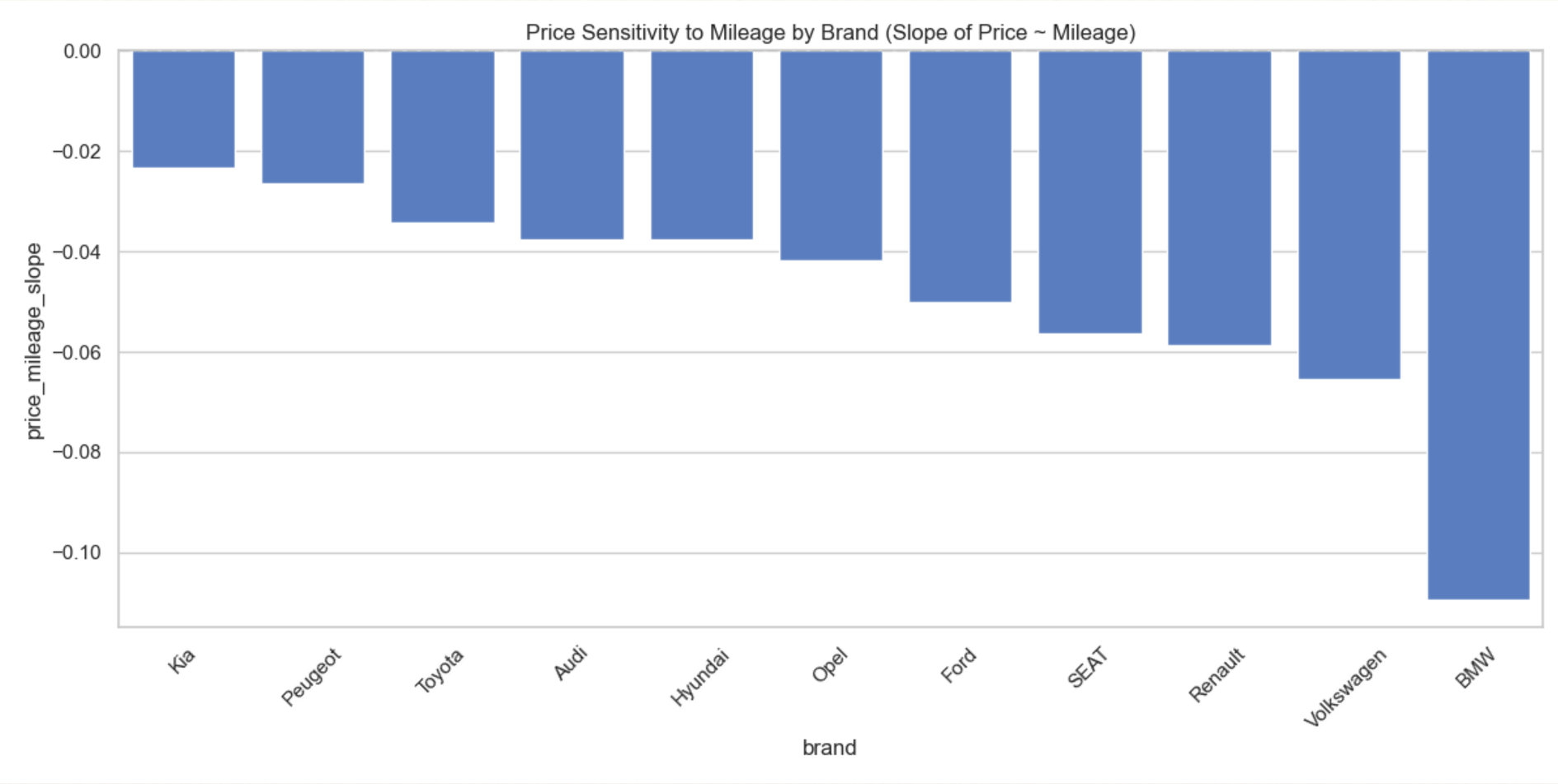
P-value: 2.62e-04

When individually examined Economy car brands have higher correlation coefficient and lower p value than luxury car brands. It means their price have higher correlation with mileage changes; however, it doesn't mean that their price drops more than luxury cars by mileage. We will examine that in the following slides.

Multivariate Analysis: Correlation Heatmap

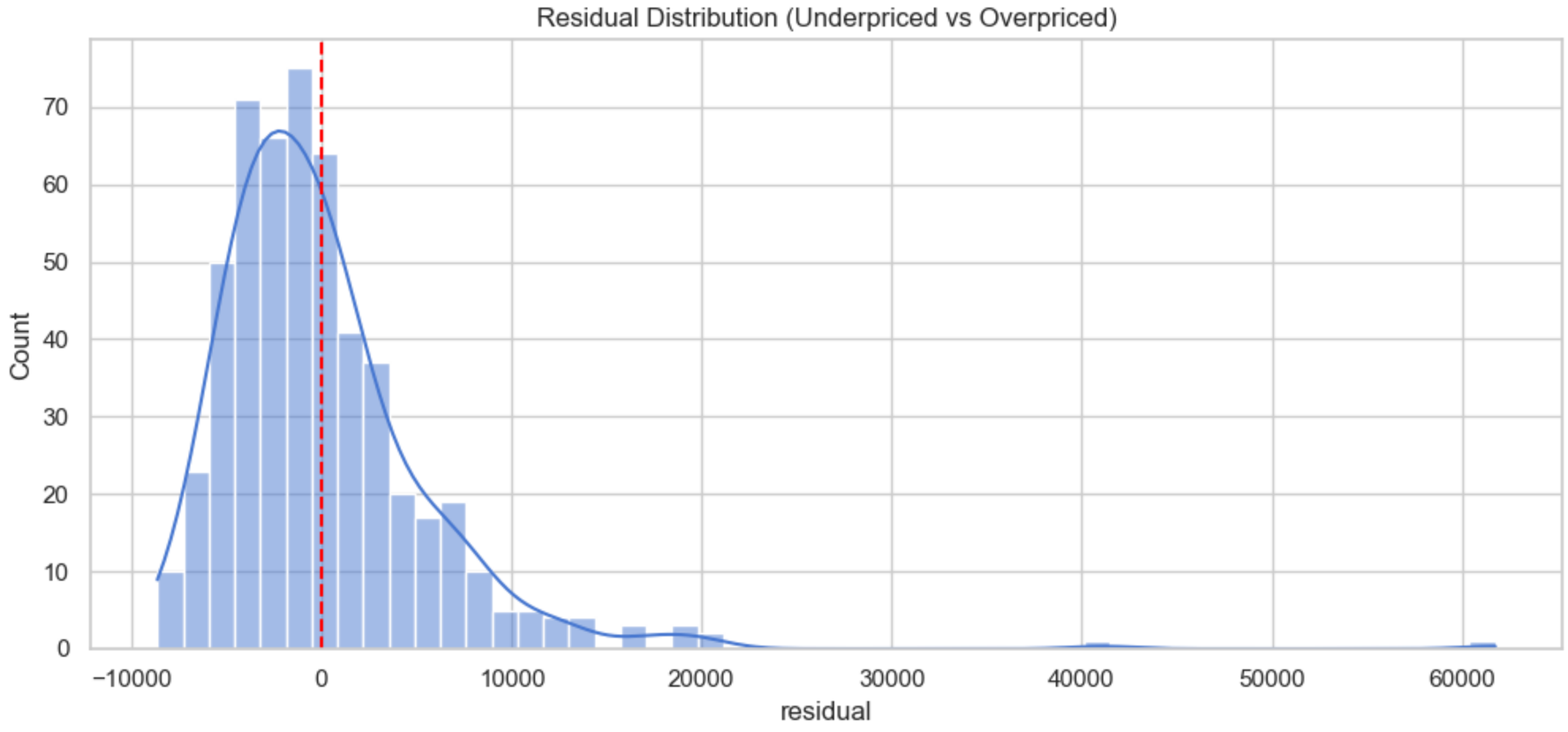
- The highest correlation with the price is car_age
- The car_age has also high correlation with the mileage of the car.



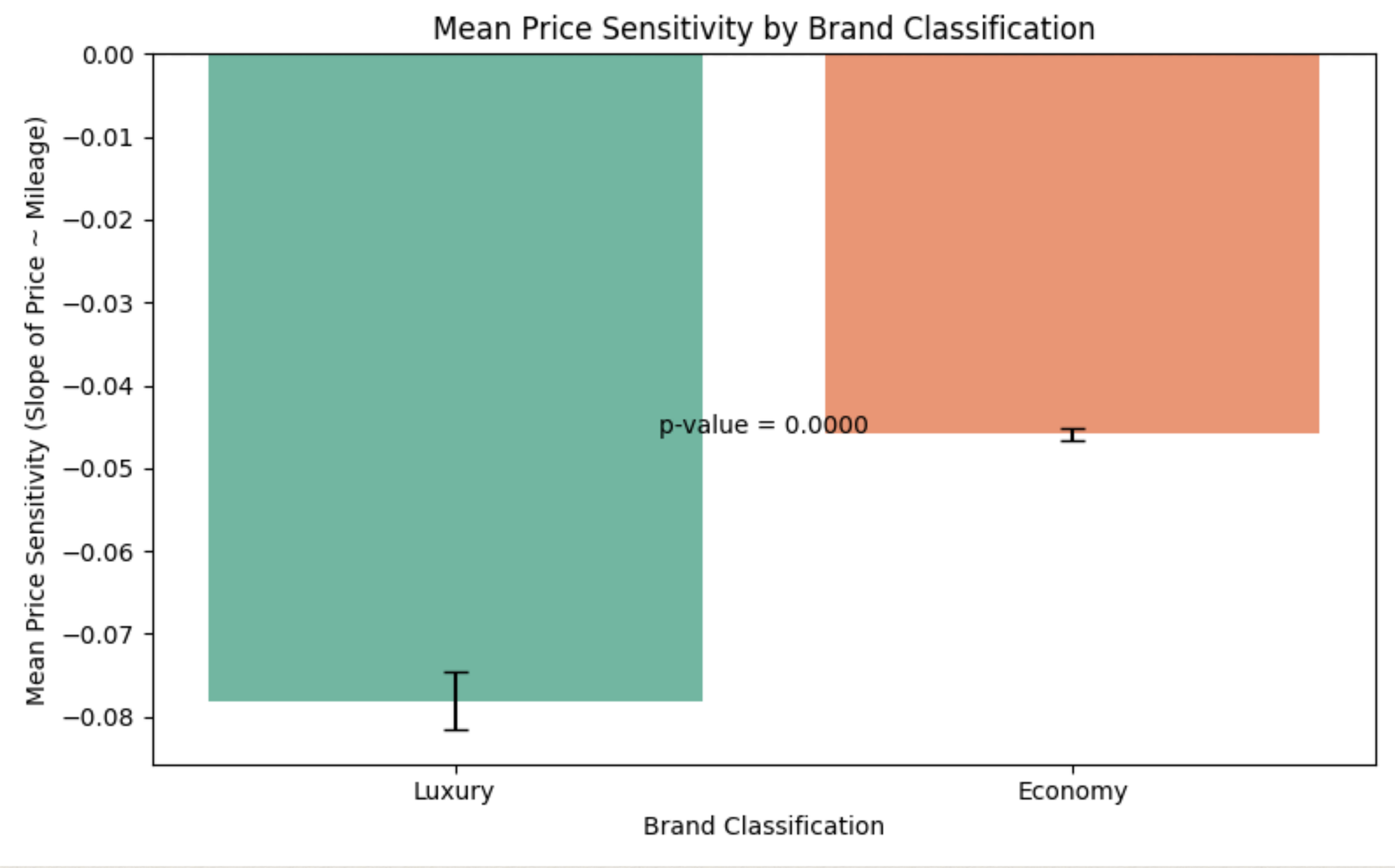


While BMW (a luxury car brand) is the car that has the most value loss per mileage kia (an economy car brand) has the lowest value loss per mileage.

This chart tells us that whether the car is underpriced or overpriced In terms of its mileage without taking anything else into account.



The residual distribution shows that most used car listings have small prediction errors centered near zero, but some listings are underpriced, which causes the distribution to lean a little to the right.



The chance of luxury car brands being more sensitive to mileage in terms of price being just out of luck is 0%. (Calculated by the p-value shown on the graph)

Conclusion & Findings:

- There is an actual and significant correlation between a car brand being luxury or economy to its price sensitivity to increased mileage **we reject the Null Hypothesis** since the p-value from that analysis came out to be 0.000… which is way lower than 0.05 needed to reject the Null Hypothesis.

- The feature that effects the price most is **car_age**.

- Luxury car brands are effected by higher mileage more in terms of value loss than economy car brands.

- Kia (an economy brand) holds their value the most while BMW (a luxury brand) unable to hold its value against increasing mileage.

Conclusion & Findings:

- What we discovered from our **exploratory analysis** is that car prices are effected by mileage, car age etc. ; however this is **different in every car brand** and we found out exactly *how much* difference it makes for each car brand.

- Our goal is to **predict the car prices accurately** using the given data and the data we feature engineered. We will also keep in mind what we found in our explatory analysis that **car brands are also an important factor** in measuring the effect of mileage, car age etc. on a cars price.