Navigating the Malaysian Resale Car Market: A Data-Driven Analysis of Valuation

Deciphering Price Drivers through Machine Learning

1.0 Problem Statement

The Malaysian secondary car market is often opaque, leaving buyers uncertain of a vehicle's "true value" beyond the sticker price. This study investigates three primary sources of pricing confusion:

- There is a lack of clarity regarding whether a larger engine displacement fundamentally justifies a higher price ceiling or if it is merely a proxy for the vehicle's luxury segment.
- It is unclear if purchasing the same vehicle in a different state (e.g., Perak vs. Kuala Lumpur) offers significant cost savings due to local supply-demand dynamics.
- Buyers are unsure if the prestige of a luxury badge is worth the steeper loss in future resale value compared to mass-market brands.

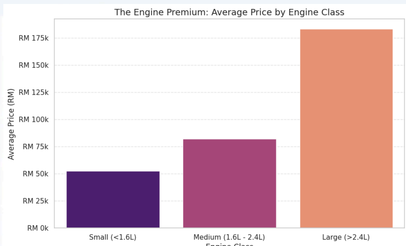
2.0 Dataset

Source: carlist_scraped_data.csv. 4000 entries, 7 columns
Scope: From national staples (Perodua Myvi) to luxury imports (Toyota Alphard, BMW 3 series).
Key Features: Brand, Year, Mileage, Location (State), Engine Capacity (CC),
Preprocessing: Currency parsing, mileage ranges to averages, Car Age calculation (2025 baseline), data extraction from description(year, brand, origin), location generalisation

3.0 Insights & Visuals

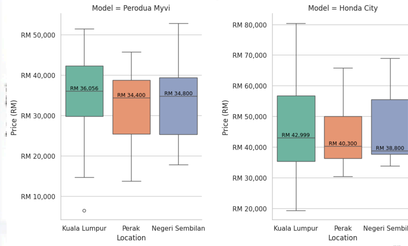
Q: Do cars with larger engine displacements influence car price?

Analysis: Confirmed. Cars with large engine displacements (>2.4L) exhibit a higher median price compared to small engine (<1.6L).



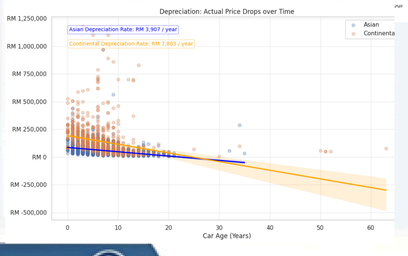
Q: Does location influence car price?

Analysis: Confirmed. Identical car models listed in Kuala Lumpur have higher median price compared to Perak and Negeri Sembilan.



Q: Do continental car depreciate faster?

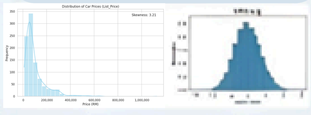
Analysis: Confirmed. Multivariate regression shows Continental cars lose value significantly faster (steeper slope) than Asian cars.



4.0 Machine Learning Models & Methods

To predict **"Fair Market Value"** and isolate price drivers, we implemented:

Target Transformation: Log-Transformation applied to handle high right-skewness (Skew > 3.0).



Random Forest Regressor: Captures non-linear relationships (mileage, age, prestige).

Multiple Linear Regression: Determines specific coefficients of features.

88.26%

5.0 Results & Conclusion

Performance Metrics:

- R' Score: 0.8826(Explains --88.26% of price variance).
- RMSE: RM 43,602 (High precision for mass- market vehicles).

Key Takeaways:

Mileage & Engine Size & Car Age are dominant predictors, impact multiplied by vehicle Origin. Engine Size sets the ceiling, Mileage determines the drop, Brand Origin dictates the speed.