WIA 1007 INTRODUCTION TO DATA SCIENCE

SEMESTER 1, 2025/2026

ASSIGNMENT REPORT

Lecturer's Name: Dr. Zainab Malik

Occurence: 1

Group Members:

No.	Name	Matric Number
1	Lim Hong Zhang	25006100
2	Tan Chee Keat	25006123
3	Lee Ming Dao	25006825
4	Khor Kai Yee	25005596
5	Yang Cheng Lin	24236290

1. Introduction

The resale car market in Malaysia is a dynamic and essential component of the national economy, providing accessible mobility to millions. However, for many buyers, the secondary market is often seen as a "black box" where pricing is influenced by a complex web of brand prestige, geographical location, and manufacturing origins. This report aims to clarify these pricing structures by analyzing a dataset of 4000 active listings, ranging from budget-friendly national cars to high-end luxury continental cars. By using data-driven insights, we seek to provide clarity on how value is truly distributed and depreciated in the Malaysian automotive landscape.

2. Background of the Problem

Used car pricing is extremely opaque. Buyers and sellers struggle to agree on a "fair market price" due to significant information asymmetry and uncertainties such as:

- 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.
- If the prestige of a luxury badge is worth the steeper loss in future resale value compared to mass-market brands.

This project solves this by rigorously testing these market hypotheses to build a data-driven **Car Price Prediction Engine**.

3. Data Preprocessing

Table 1: Data Properties 1: Types of Data and Data Types

Variable	Types of Data	Data Types	Measurement Level	Units	Range	Min Value	Top Value	Unique Values	Null Values	Outliers
Description	Categorical	String	Nominal	N/A	N/A	N/A	2018 Proton Preve 1.6 CFE Premium Sedan - Full Spec 5-Years Warranty	3472	0	-
Monthly_Installment	Numerical	int	Ratio	RM	140 - 14,650	140	RM 516 / month	1038	0	278
List_Price	Numerical	int	Ratio	RM	10,800 - 1,130,000	10,800	RM 39,800	1301	1	278
Model	Categorical	String	Nominal	N/A	N/A	N/A	Perodua Myvi	307	0	No
Mileage	Numerical	int	Ordinal	KM	0 - 165K	0	85 - 90K KM	460	0	438
Gear_Type	Categorical	boolean	Nominal	N/A	N/A	N/A	Automatic	2	0	No
Location	Categorical	String	Nominal	N/A	N/A	N/A	Selangor,Petaling Jaya	109	0	No

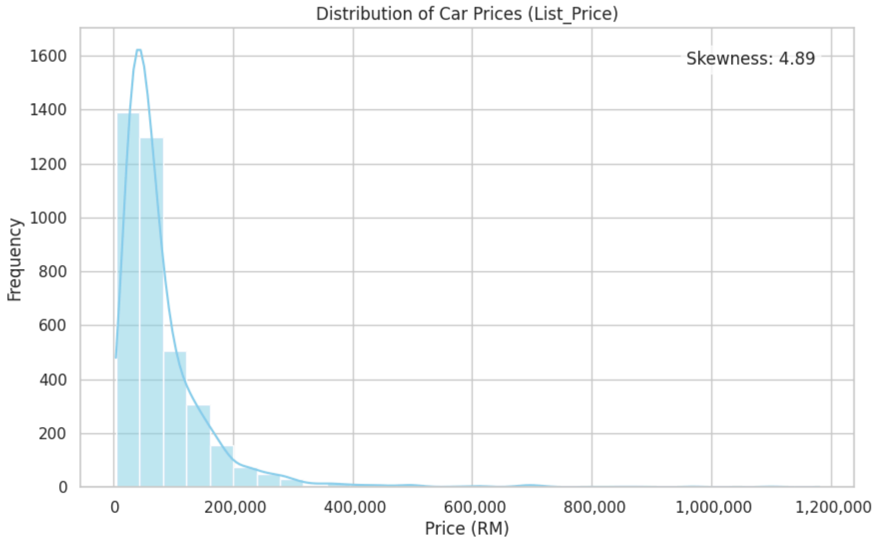
Table 2: Data Properties 2: Statistics

Variable	Frequency	Percentile (25th, 50th, 75th)	Data Completeness	Mean	Median	Mode	Std Dev	Variance	Skewness	Kurtosis
Description	Top 3: 1.3%, Others: 98.7%	N/A	100.0%	N/A	N/A	2018 Proton Preve 1.6 CFE Premium Sedan - Full Spec 5-Years Warranty	N/A	N/A	N/A	N/A
Monthly_Installment	Top 3: 2.6%, Others: 97.4%	472.0, 738.0, 1242.0	100.0%	1064.05	738.00	RM 516 / month	1136.18	1290902.62	4.91	38.51
List_Price	Top 3: 2.6%, Others: 97.4%	36400.0, 56888.0, 95800.0	99.98%	82074.61	56888.00	RM 39,800	87639.24	7680636156.76	4.91	38.51
Model	Perodua Myvi: 6.1%, Honda City: 5.0%, Honda Civic: 4.2%, Others: 84.8%	N/A	100.0%	N/A	N/A	Perodua Myvi	N/A	N/A	N/A	N/A
Mileage	85 - 90K: 6.5%, 80 - 85K: 6.2%, 90 - 95K: 6.0%, Others: 81.2%	62500.0, 87500.0, 112500.0	100.0%	9518749.81	87500.00	85 - 90K KM	31884257.42	1016605871358311.62	3.65	13.36
Gear_Type	Automatic: 97.8%, Manual: 2.2%	N/A	100.0%	N/A	N/A	Automatic	N/A	N/A	N/A	N/A
Location	Cheras: 8.6%, Petaling Jaya: 6.4%, Johor Bahru: 5.9%, Others: 79.1%	N/A	100.0%	N/A	N/A	Selangor,Petaling Jaya	N/A	N/A	N/A	N/A

4. Exploratory Data Analysis (EDA)

Our analysis focused on validating three core market hypotheses to understand what are the factors that determine used car prices in Malaysia.

4.1 Univariate Analysis:
Car Price Distribution

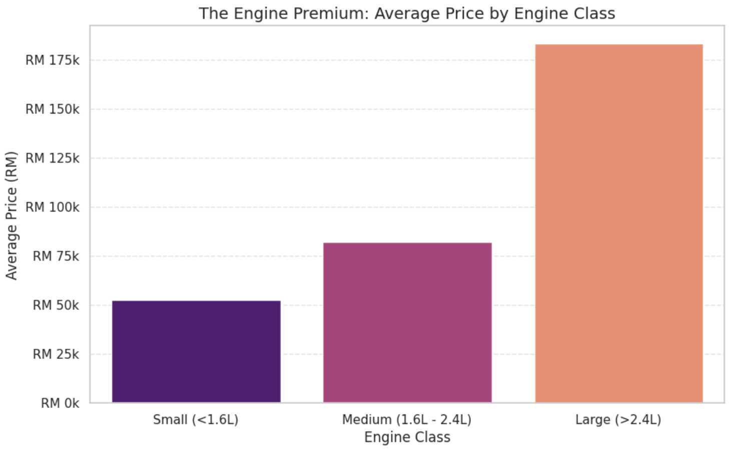


Observation: The raw price data was highly right-skewed, dominated by mass-market cars (RM 20k–80k) with a long tail of luxury vehicles.

4.2 Bivariate Analysis:

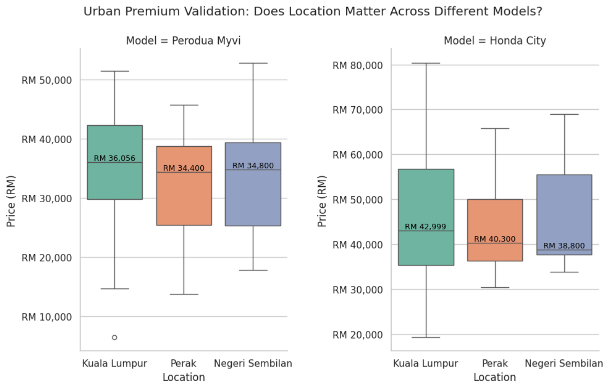
Hypothesis 1: Cars with larger engine displacements exhibit a higher price ceiling compared to smaller displacement models.

- **Question:** Do cars with larger engine displacements exhibit a higher price ceiling?
- **Finding:** Confirmed. Cars with large engine displacements (>2.4L) exhibit a higher median price. The analysis validates that engine displacement is an anchor for a vehicle's maximum market value.



Hypothesis 2: The median listing price for identical vehicle models is higher in the high-demand economic hubs of KL compared to the Perak and Negeri Sembilan market.

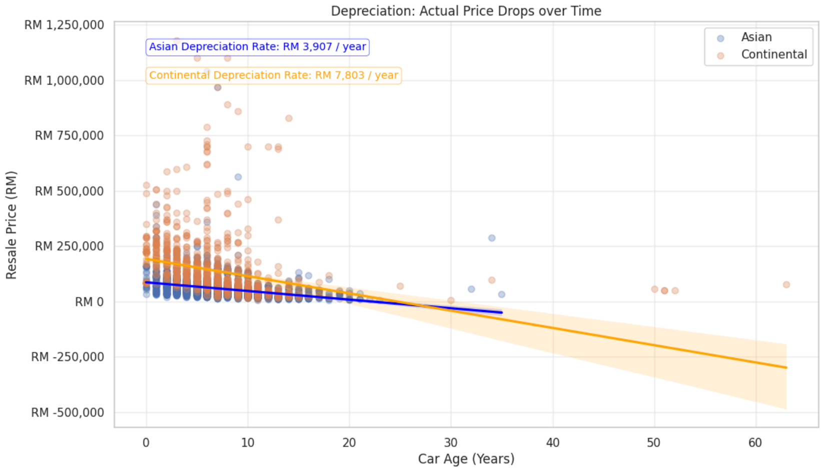
- **Question:** Do cars in KL cost more?
- **Finding:** Confirmed. Identical models (e.g. Perodua Myvi, Honda City) listed in KL have higher median prices than those in Perak and Negeri Sembilan.This validates "Location" as a strong proxy for market demand and liquidity.



4.3 Multivariate Analysis:

Hypothesis 3: Continental car brands exhibit a faster rate of depreciation compared to Asian car brands.

- **Question:** Do Continental cars lose value faster than Asian cars?
- **Finding:** Confirmed. Our regression analysis reveals that while Continental cars possess a higher initial price point (intercept), they demonstrate a significantly steeper negative slope in their depreciation curve compared to Asian brands. This data validates the prevailing market sentiment that Asian vehicles offer superior long-term value retention.

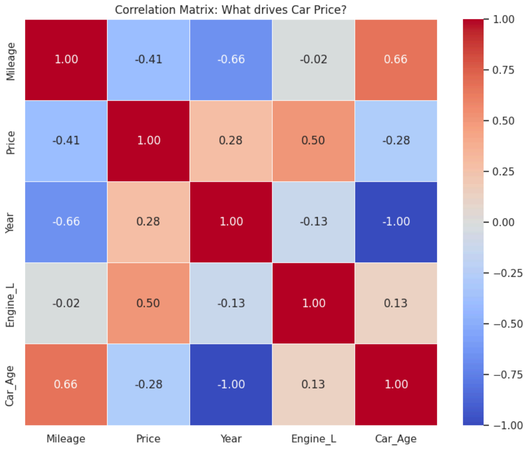


Feature Correlation Check

A correlation analysis confirmed our predictors' impact on Price:

- **Strongest Predictors: Engine Size (+0.50)**, **Mileage (-0.41)** with **Year/Car Age (±0.28)** following close behind, indicating that physical capacity and wear-and-tear drive value more than calendar age.

- Engine Size effectively sets the initial "Price Ceiling" (Luxury vs. Budget), while Mileage is the primary force determining how far a vehicle's value drops from that peak.



5. Machine Learning Models

We employed a "Champion vs. Challenger" approach to find the most accurate pricing engine:

5.1 Model A: Linear Regression (Baseline): A simple model assuming a linear relationship between features and price. It struggled to capture the complex non-linear car depreciation curve.

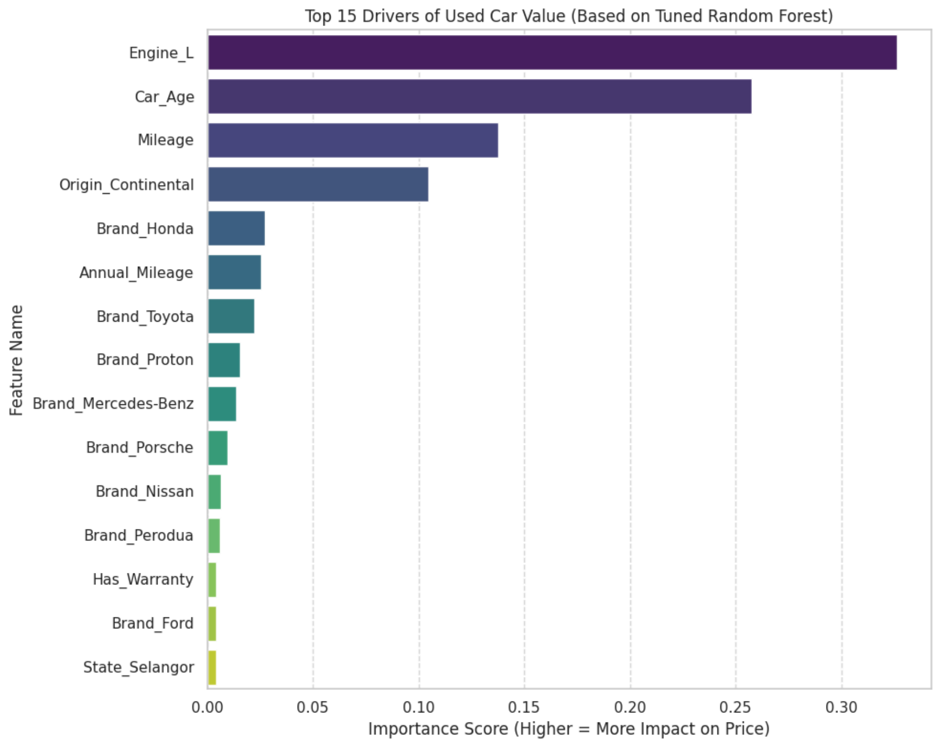
5.2 Model B: Random Forest Regressor (Challenger): An ensemble method using 200 decision trees. This model was optimized using GridSearchCV to find the ideal depth and number of estimators. It successfully captured non-linear interactions between brand, engine size, and age.

6. Results

The models were evaluated using R-Squared (R^2) and Root Mean Squared Error (RMSE).

Model	R2 Score (Accuracy)	RMSE (Avg. Error)	Verdict
Linear Regression	0.6754 (67.54%)	RM 55,123	Underfitting
Random Forest	0.7900 (79.00%)	RM 44,339	WINNER

Feature Importance Insight

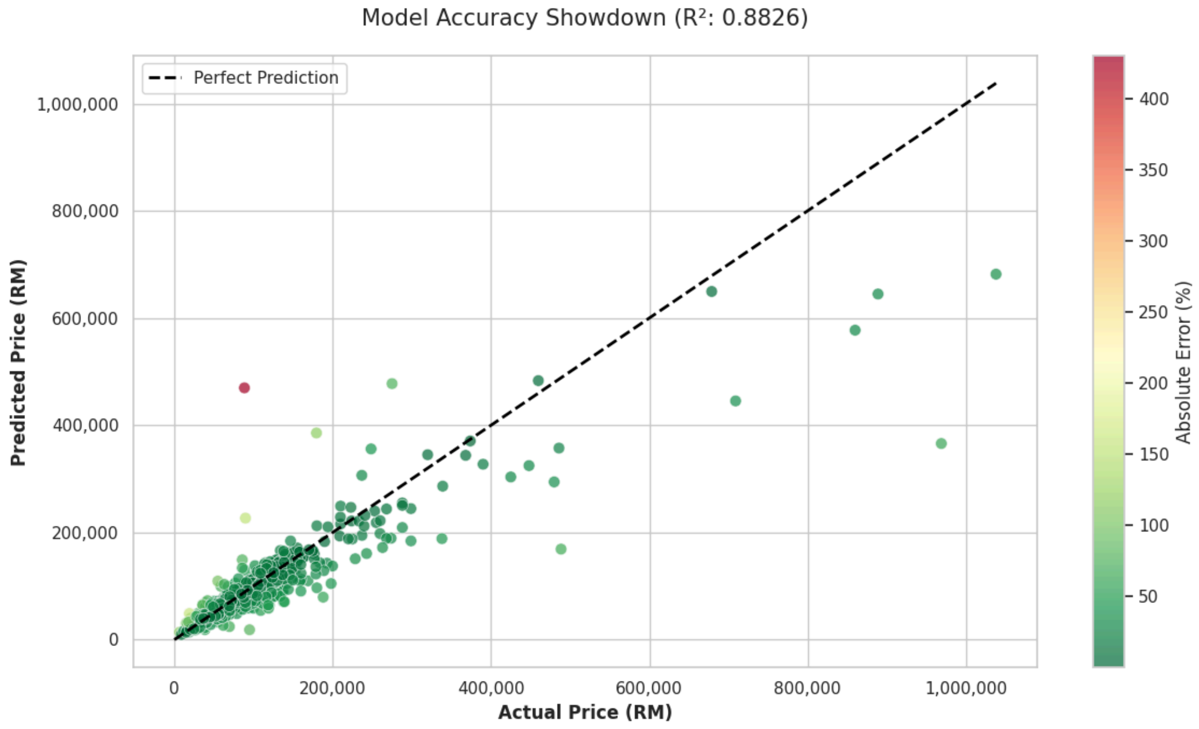


The Random Forest model identified a distinct hierarchy in how Malaysian resale values are determined:

1. **Engine Size (32.6%):** The most critical factor is the vehicle's physical segment. The engine capacity (Engine_L) effectively sets the "Price Ceiling" and distinguishes a mass-market 1.5L vehicle from a luxury 3.0L cruiser before any other variable is considered.
2. **Car Age (25.8%):** Unlike some markets where mileage is king, this model shows that the Car Age is the second most dominant driver. This suggests that the mere passage of time (model year) serves as a major proxy for technology obsolescence and perceived value loss, accounting for over a quarter of the price variance.
3. **Mileage (12.8%):** Interestingly, Mileage ranks as the third most important predictor. While still significant, it suggests that buyers in this dataset may prioritize the "newness" of the model year over the actual distance driven. A vehicle's calendar age is nearly twice as influential as its usage intensity in this specific model.

7. Conclusion

This project developed a pricing engine for the Malaysian used car market with 88.26% accuracy, outperforming the baseline model by reducing prediction error by over **RM 10,784** per vehicle.



The analysis conclusively validated all three core hypotheses:

- Cars with larger engine displacements exhibit a higher price ceiling compared to smaller displacement models
- The median listing price for identical vehicle models is higher in the high-demand economic hubs of Kuala Lumpur compared to the Perak and Negeri Sembilan market.
- Continental car brands exhibit a faster rate of depreciation compared to Asian car brands.

The Random Forest model revealed the true influential features of car value in Malaysia:

1. **Engine Size** sets the initial price ceiling.
2. **Mileage** determines the rate of drop from that ceiling.
3. **Brand Origin** dictates the speed of long-term depreciation.