What do Indonesian look for in a used car?

Aldo Gadra Paulus Simatupang

Problem Statement

Indonesia used car market have been growing fast in the last few years. In 2025, used car sold twice as much as new car [1]. This new market shift is caused by the rise of new online used vehicle-oriented marketplace and economic downturn [2] .

The main challenge in buying used or secondhand item is determining the best price of an item. Luckily, there have been a lot of car expert giving online guides on it based on their experience. But what does the market tell us? **What determined used car selling price?**

[1] https://www.cnnindonesia.com/otomotif/20250523173354-579-1232563/mobil-bekas-lebih-laku-dari-mobil-baru-di-indonesia
[2] https://otomotif.kompas.com/read/2024/08/07/181526015/daya-beli-lemah-masyarakat-pilih-mobil-bekas

Goal

- Identify variables that significantly contribute to the used car's and its relevant contribution.
- Furthermore, to make machine learning model to predict used car price.

Dataset Introduction

- The dataset is uploaded by Indra in Kaggle (https://www.kaggle.com/datasets/indraputra21/used-car-listings-in-indonesia/data).
- The dataset made up of various listings scraped from carsome.id
- Initial dataset have 620 rows x 20 columns.

Data Dictionary

- **car name:** The name or model of the car.
- **brand:** The brand or manufacturer of the car.
- **year:** The year the car was manufactured.
- **mileage (km):** The mileage or distance traveled by the car in kilometers (km).
- **location:** The location where the car is listed for sale.
- **transmission:** The transmission type, such as "Manual" or "Automatic."
- **plate type:** The type of license plate, which can be an even plate or an odd plate.
- **rear camera:** Indicates whether the car has a rear camera (0 for no, 1 for yes).
- **sun roof:** Indicates whether the car has a sunroof (0 for no, 1 for yes).
- **auto retract mirror:** Indicates whether the car has auto-retracting mirrors (0 for no, 1 for yes).

Data Dictionary

- **electric parking brake:** Indicates whether the car has an electric parking brake (0 for no, 1 for yes).
- **map navigator:** Indicates whether the car has a built-in map navigator (0 for no, 1 for yes).
- **vehicle stability control:** Indicates whether the car has vehicle stability control (0 for no, 1 for yes).
- **keyless push start:** Indicates whether the car has a keyless push start (0 for no, 1 for yes).
- **sports mode:** Indicates whether the car has a sports mode (0 for no, 1 for yes).
- **360 camera view:** Indicates whether the car has a 360-degree camera view (0 for no, 1 for yes).
- **power sliding door:** Indicates whether the car has a power sliding door (0 for no, 1 for yes).
- **auto cruise control:** Indicates whether the car has auto cruise control (0 for no, 1 for yes).
- **price (Rp):** The price of the car in Indonesian Rupiah (Rp).
- **instalment (Rp|Monthly):** The monthly installment amount for the car, in Indonesian Rupiah (Rp).

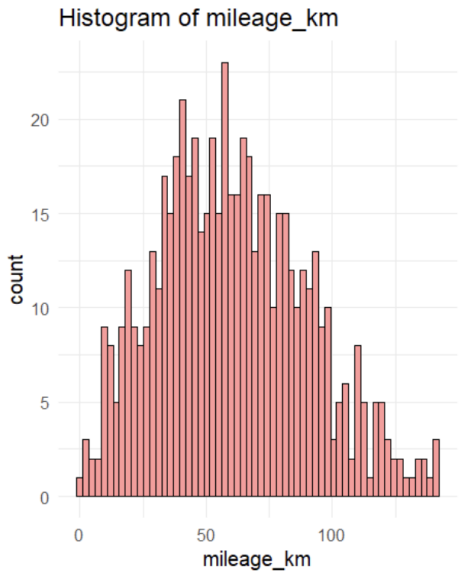
Data Cleaning & Handling

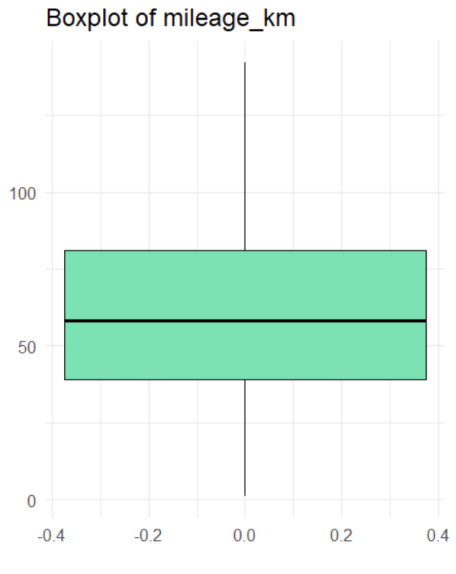
Subject	Report	Handling
Data Transformation		
Convert categorical characters	“car _brand” has important information but is in string or categorical format.	Generate “brand_popularity” column by calculating (brand_counts / total rows) * 100% “Brand popularity” represent the popularity of a single brand in the dataset and is aimed to proxied for brand awareness
Convert binary characters columns to dummy	“transmission” & “plate.type” is still in character format	Converted both columns to dummy variables: is.manual & is.odd.plate.
Generate “car’s age” from “year” column.		Calculate each car’s age from 2024 - “year” column.

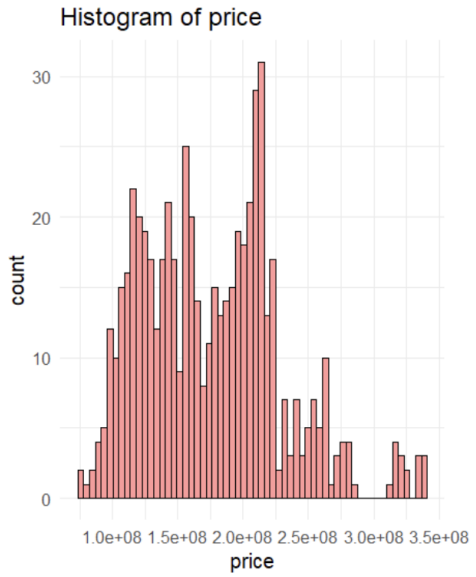
Data Cleaning & Handling

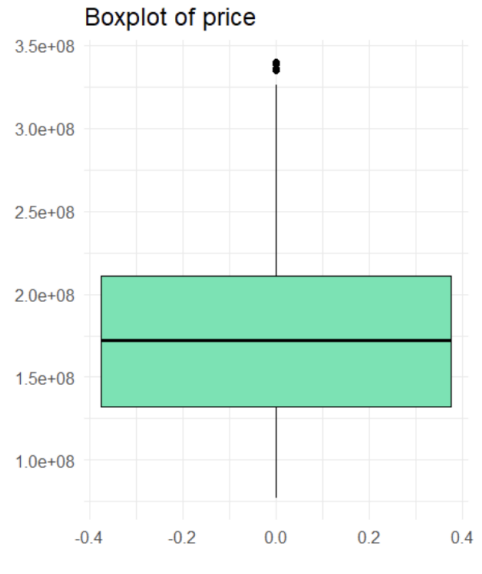
Data Cleaning		
Duplicates row	No duplicates found	
Missing data (NA, empty string “”, and “NULL” entry)	No Missing data found	
Extreme outliers	There are 12 observation that lies outside of 3 standard deviation range from the mean	Imputed the aforementioned 12 observations.

EDA: Continuous Variables





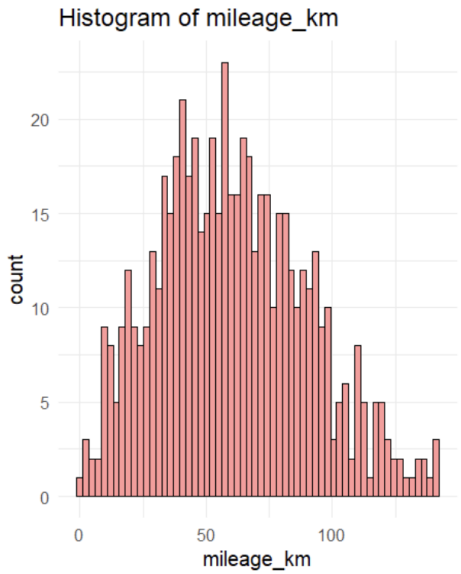


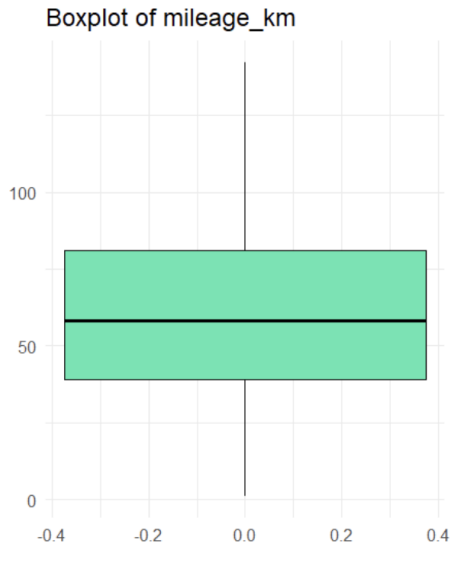


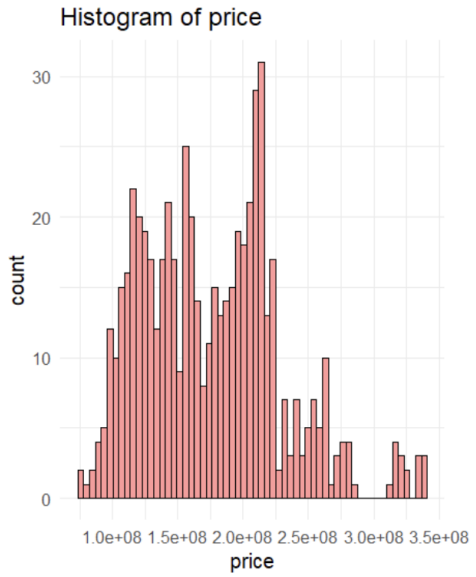
Data Distribution

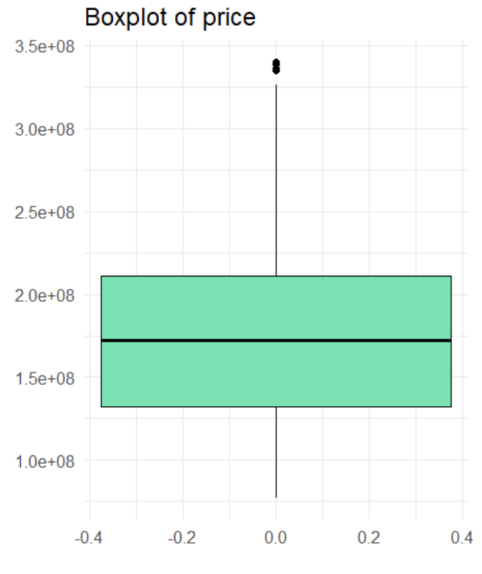
- All continuous variables are positively skewed.
- Furthermore, boxplot reveals that the outliers are in the right side.
- This might indicate a need for variable transformation.

EDA: Continuous Variables









Data Distribution

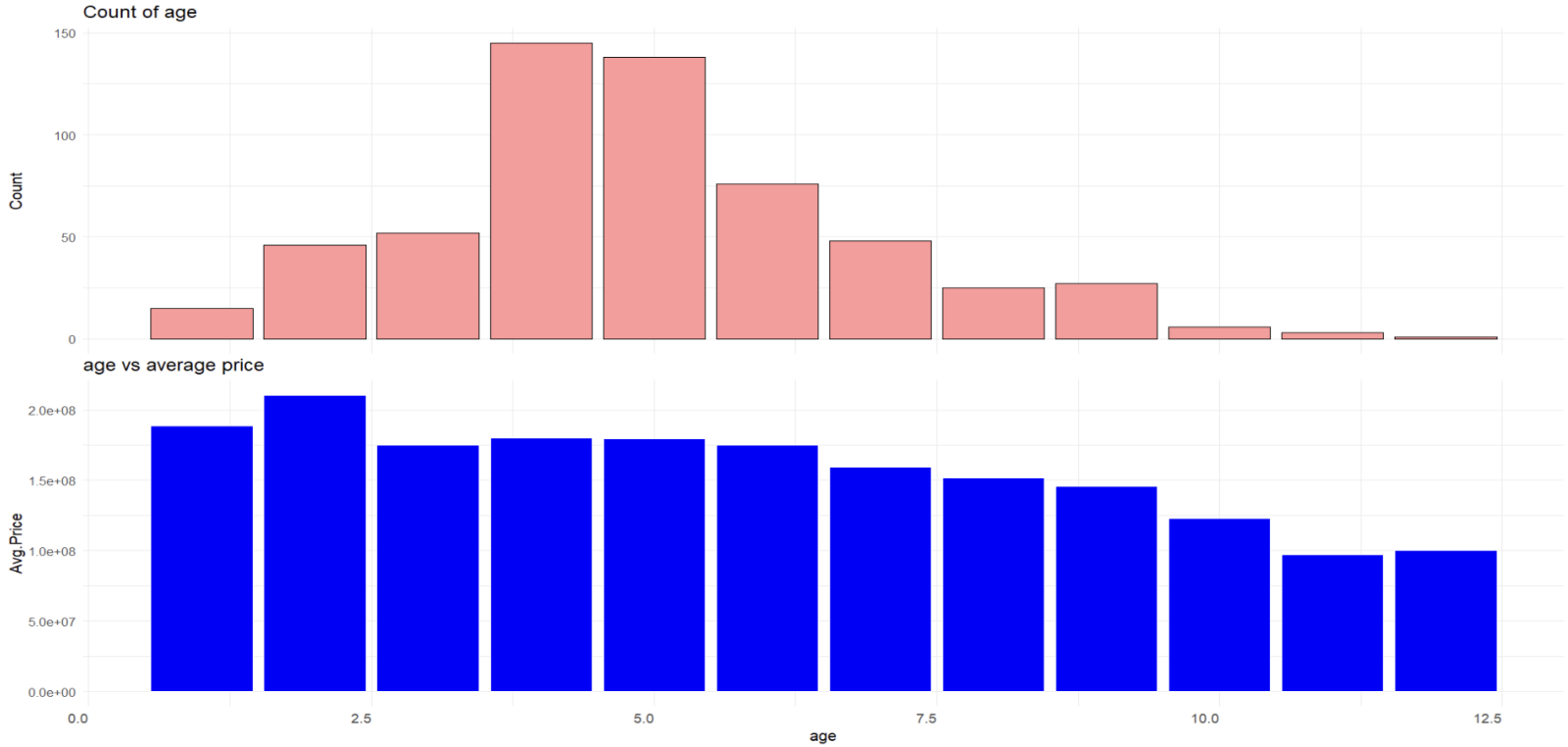
- All continuous variables are positively skewed.
- Furthermore, boxplot reveals that the outliers are in the right side.
- This might indicate a need for variable transformation.

EDA: Discrete Variables

Insights

- Most of the car listed have engine capacity 1.500cc and lower.
- As the engine capacity gets higher, the price tends to increase.

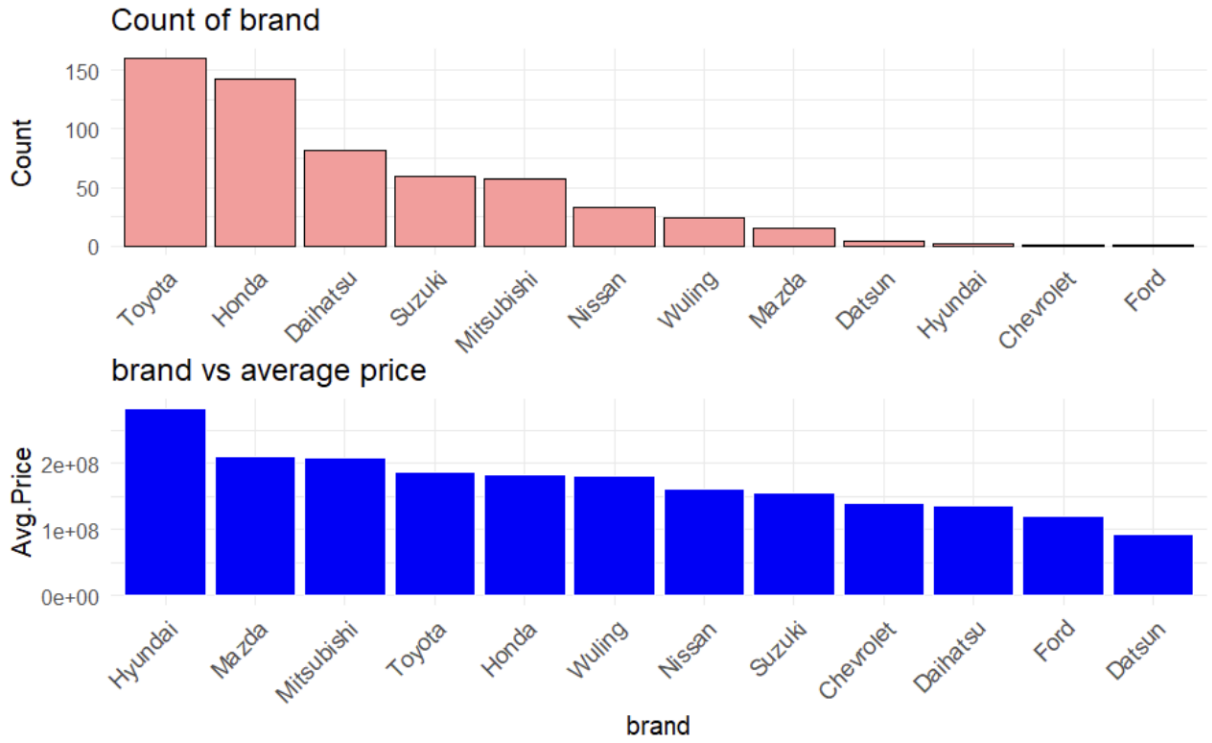
EDA: Discrete Variables



Insights

- Most of the car in the dataframe are 4 tol 5 years old.
- Furthermore, the average price tends to increase as it gets older

EDA: Categorical Variables



Insights

- Toyota, Daihatsu, and Hyundai are the most common brand listed on the website. But interestingly, they don't hold the top highest average price.
- Hyundai, Mazda, and Mitsubishi are the top 3 brand with the highest average price despite their relatively low listing popularity. This can indicate that they are the luxury brands.

Hedonic Model: Variable Selection

Dependent Variable: “Price” column

Multicollinearity Check

	Variable1 <fctr>	Variable2 <fctr>	Correlation <dbl>
14	age	mileage_km	0.6346377
222	mileage_km	age	0.6346377

- The age and mileage variable are highly correlated, which is understandable as both variable can serve as a proxy for car use.
- I choose to keep the mileage variable as it is a more accurate variable to reflect car use.

Best variable subset (based on R2)

Using exhaustive search, the best performing model is made up from 12 from the 16 independent variables. Variables that are excluded from the best model are: “keyless.push.start”, “sports.mode”, “x360.camera.view”, and “is.odd.plate”.

Hedonic Model: BLUE assumption check

Assumption	Test	Result
Linearity	Remsey RESET test	The p-value is < 0.05, which indicates that the dependent and independent variable relationship are not linear. But this can be caused by other BLUE assumption violations.
No Multicollinearity	VIF (Variance Inflation Factors)	There is no variable with VIF value > 5, so it can be concluded there are no to little multicollinearity.
Homoscedasticity	Breusch-Pagan test	Overall model p-value is < 5, so it can be concluded that the model have a heteroscedasticity problem.

The model have:

- Non-linear relationship
- Heteroscedasticity

Hedonic Model: BLUE assumption check

Model	Best variable subset	Log transformed	R2	RMSE	Violation
OLS	-	-	0.702	29034580.3	- Non-Linear Relationship - Heteroskedasticity
OLS (Best subset)	yes	-	0.703	28974723.6	- Non-Linear Relationship - Heteroskedasticity
Double Log OLS	yes	yes	0.714	0.162	- Non-Linear Relationship - Heteroskedasticity
WLS (Weighted Least Squares)	yes	yes	0.724	0.163	- Non-Linear Relationship - Heteroskedasticity

Machine Learning: Regression Tree

Pre-Pruning Parameter

Starting CP	Minsplit	Maxdepth
0.01	20	30

Post-Pruning Parameter

CP
0.0108673

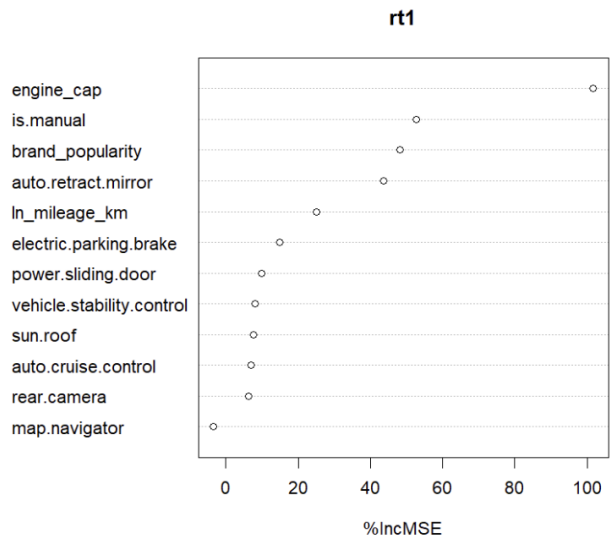
Variable Importance

Engine_cap	Is.manual	auto.retract.mirror	brand_popularity	ln_mileage_km
42	19	15	13	11

Machine Learning: Random Forest

Parameter

Random Subspace Dimension	N_tree	K-fold's fold number
6	501	10



Variable Importance

The top 3 variables are

- engine_cap
- is.manual
- brand_popularity

Prediction Evaluation

Model	RMSE
OLS	29034580.3
OLS (Best subset)	28974723.6
Double Log OLS	0.162
WLS (Weighted Least Squares)	0.163
Regression Tree (Pre-purning)	0.219
Regression Tree (Post-purning)	0.229
Random Forest	0.153

Conclusion

Variable Importance

- Engine capacity, manual/automatic transmission, and brand popularity are the top significant and important variables in predicting used car price.
- Hence, Indonesian considers those variables to be top characteristic to keep in mind when deciding used car price.

Model performance

- Random Forest, WLS model, and double log model are the top model to predict used car price.
- Surprisingly, WLS and double log model have higher prediction performance than regression tree. This is unusual, considering previous test have shown that the model might has non-linear relationship that hinders regression performance.
    - But this might show that the Remsey RESET test is biased due to heteroscedasticity.