Final Project Presentation

Nomor Kelompok: 7
Nama Mentor: Aditya Bariq
Nama:

- Mhd. Arsya Fikri
- Rhisa Adika Putri

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Latar Belakang

Latar Belakang Project

Sumber Data: https://www.kaggle.com/datasets/hellbuoy/car-price-prediction

Problem: **Regression**

Tujuan:

- Memprediksi “Harga Mobil” berdasarkan spesifikasi dan *brand* mobil.

Eksplorasi Data dan Visualisasi

Business Understanding

- Harga mobil sangat bervariasi dari yang murah hingga mahal.
- Harga mobil dipengaruhi oleh banyak hal, seperti spesifikasi dan keadaan mobil.
- Calon pelanggan tertarik dengan mobil yang memiliki harga sebanding dengan spesifikasinya.
- Semakin bersaing harga, maka semakin menarik minat calon pelanggan untuk memilih mobil.
- Sehingga, perusahaan mobil perlu melakukan penyesuaian harga jual dan kualitas produksi mobil.

Data Cleansing

- Dataset terdiri dari: 26 kolom (10 kategorikal, 16 numerikal) dan 205 baris, serta tidak ada *missing data*. Dataset perlu dibersihkan karena terdapat beberapa kesalahan data.

Data Cleansing

Kesalahan pada dataset:

- Kolom *symboling* bertipe numerik. Solusinya adalah mengganti tipe data kolom *symboling*.
- Kolom *CarName* memiliki banyak *brand* dan model mobil yang berbeda-beda. Solusinya adalah hanya menggunakan nama *brand* mobil.

Data Cleansing

- Pada data *brand* mobil, terdapat kesalahan penulisan. Sehingga perlu untuk memperbaiki kesalahan penulisan data tersebut.

```
array(['alfa-romero', 'audi', 'bmw', 'chevrolet', 'dodge', 'honda',
       'isuzu', 'jaguar', 'maxda', 'mazda', 'buick', 'mercury',
       'mitsubishi', 'Nissan', 'nissan', 'peugeot', 'plymouth', 'porsche',
       'porcshce', 'renault', 'saab', 'subaru', 'toyota', 'toyouta',
       'vokswagen', 'volkswagen', 'vw', 'volvo'], dtype=object)
```



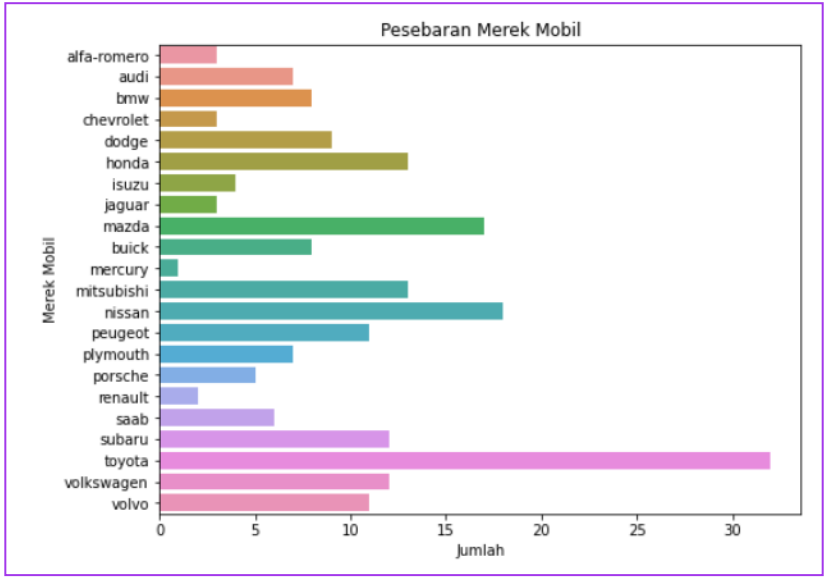
```
array(['alfa-romero', 'audi', 'bmw', 'chevrolet', 'dodge', 'honda',
       'isuzu', 'jaguar', 'mazda', 'buick', 'mercury', 'mitsubishi',
       'nissan', 'peugeot', 'plymouth', 'porsche', 'renault', 'saab',
       'subaru', 'toyota', 'volkswagen', 'volvo'], dtype=object)
```

Data Cleansing

- Kolom *car_ID* tidak digunakan sebagai features karena tidak memiliki pengaruh, sehingga kolom tersebut dihapus.

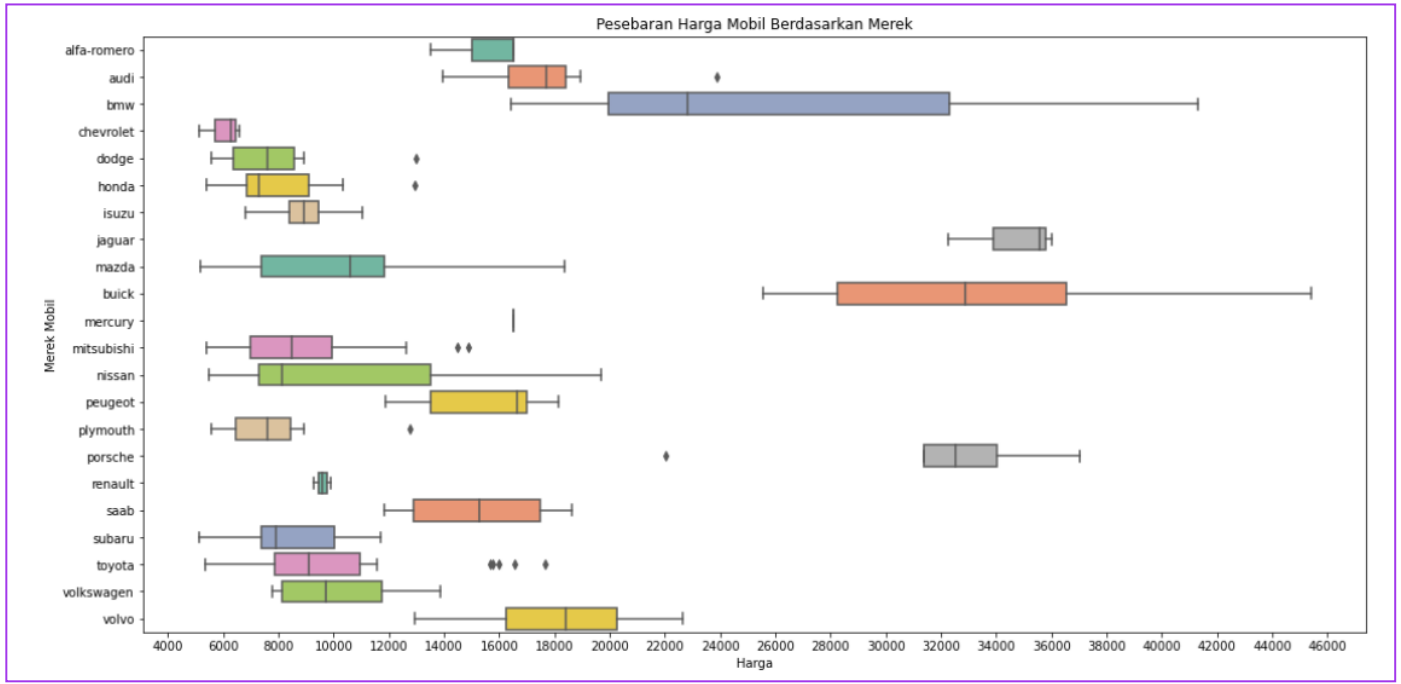
Exploratory Data Analysis

- Data mobil yang paling banyak adalah *brand* Toyota dan yang paling sedikit adalah *brand* Mercury.



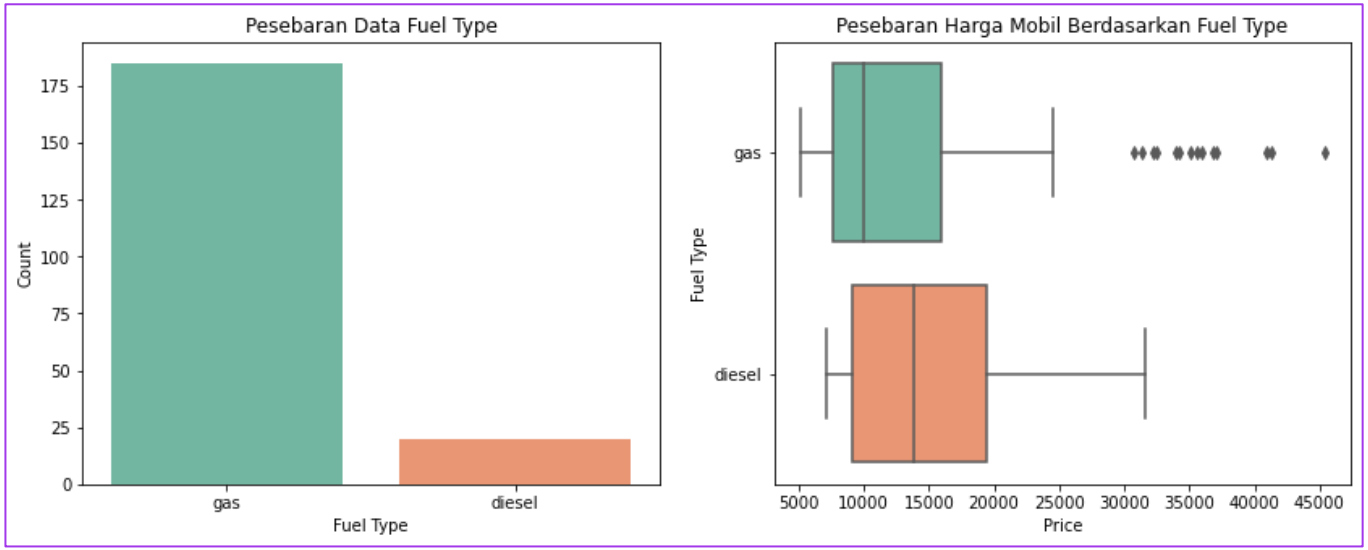
Exploratory Data Analysis

- *Brand* mobil yang memiliki rentang harga paling besar adalah BMW dengan harga terkecil ~16000 dan harga tertinggi ~41000.
- Mobil dengan harga terendah adalah *brand* Subaru dan harga tertinggi adalah *brand* Buick.



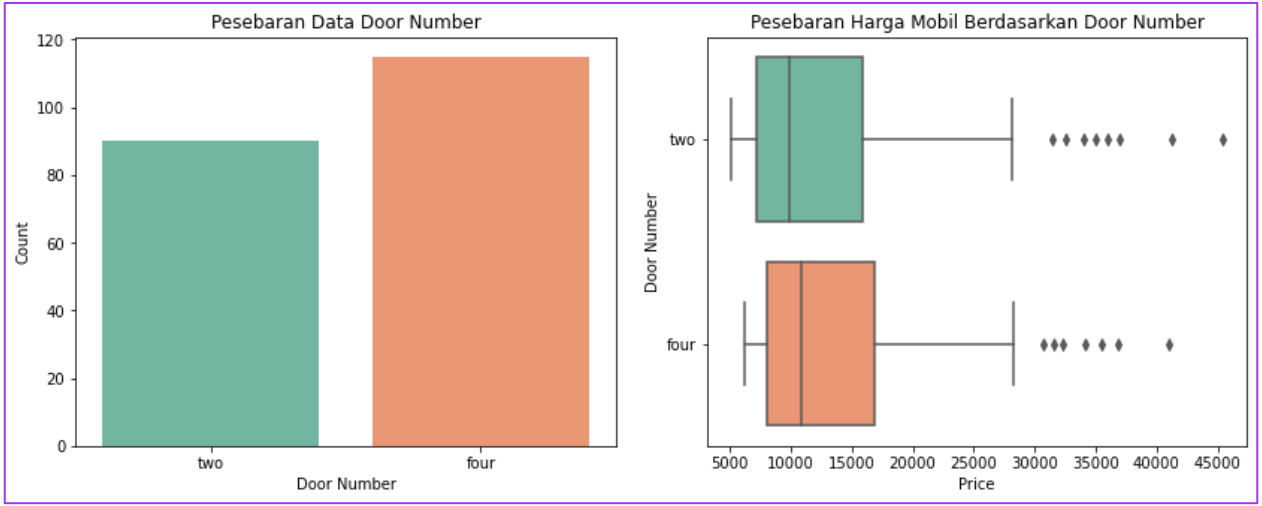
Exploratory Data Analysis

- Kebanyakan mobil bertipe bahan bakar gas.
- Harga mobil dengan tipe bahan bakar diesel cenderung lebih mahal.



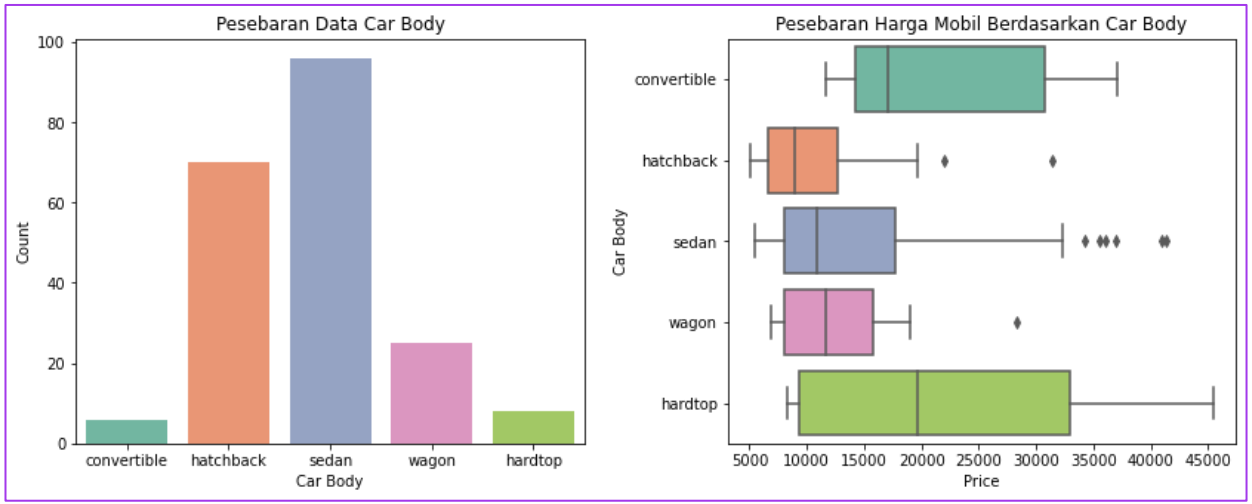
Exploratory Data Analysis

- Kebanyakan mobil memiliki empat pintu, namun selisihnya tidak terlalu jauh dibandingkan dengan mobil dua pintu.
- Ternyata mobil empat pintu sedikit lebih mahal daripada mobil dua pintu



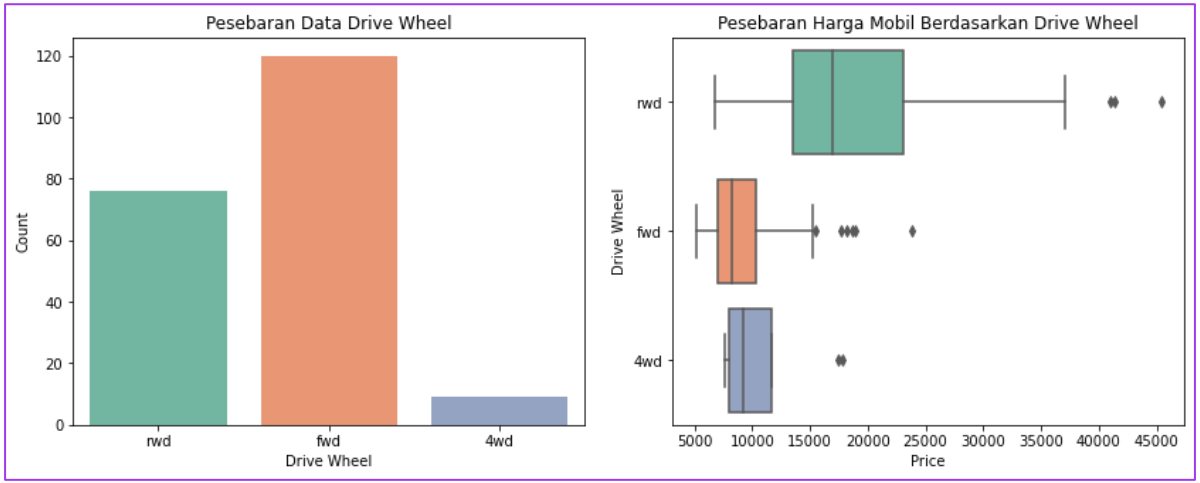
Exploratory Data Analysis

- Kebanyakan mobil memiliki bodi sedan, sedangkan tipe bodi paling sedikit adalah convertible.
- Mobil dengan bodi hardtop memiliki harga paling tinggi dibandingkan jenis bodi lainnya.



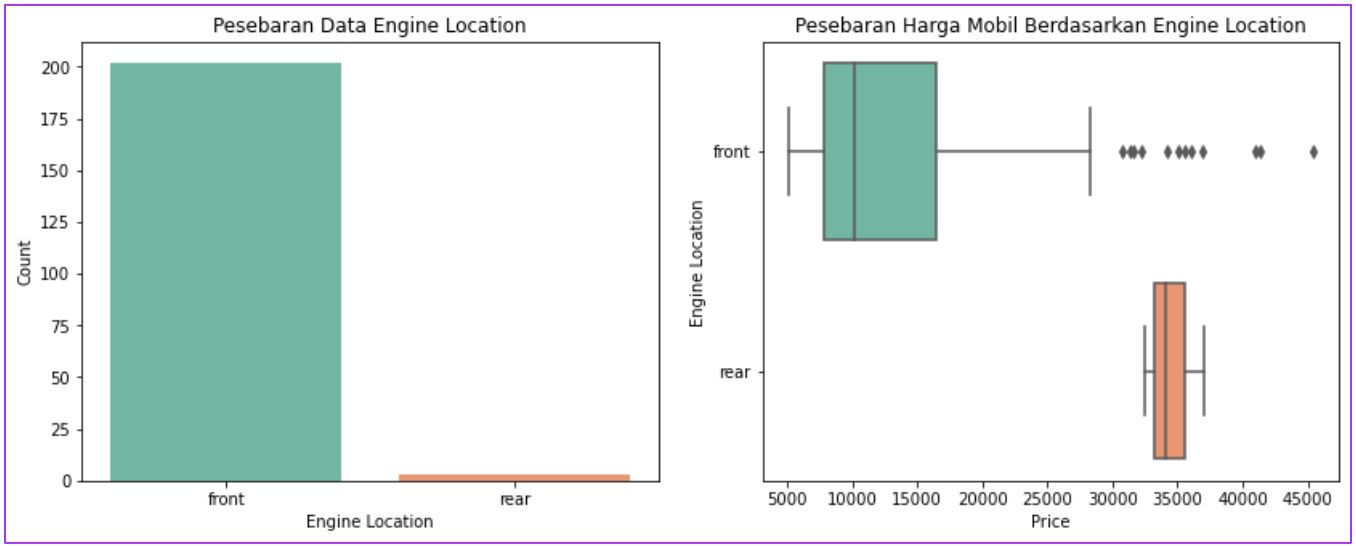
Exploratory Data Analysis

- Kebanyakan mobil memiliki jenis penggerak fwd, kemudian rwd, dan paling sedikit adalah 4wd.
- Mobil dengan jenis penggerak rwd memiliki harga lebih mahal dari jenis penggerak yang lain.



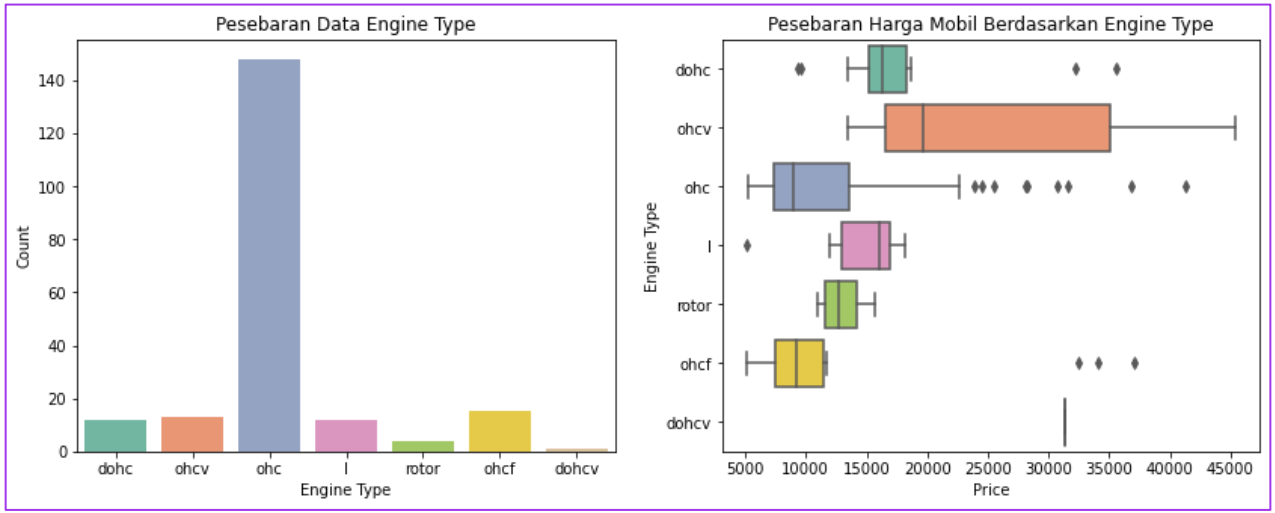
Exploratory Data Analysis

- Hampir seluruh data mobil memiliki lokasi mesin di depan.
- Mobil dengan lokasi mesin di belakang memiliki harga jauh lebih tinggi.



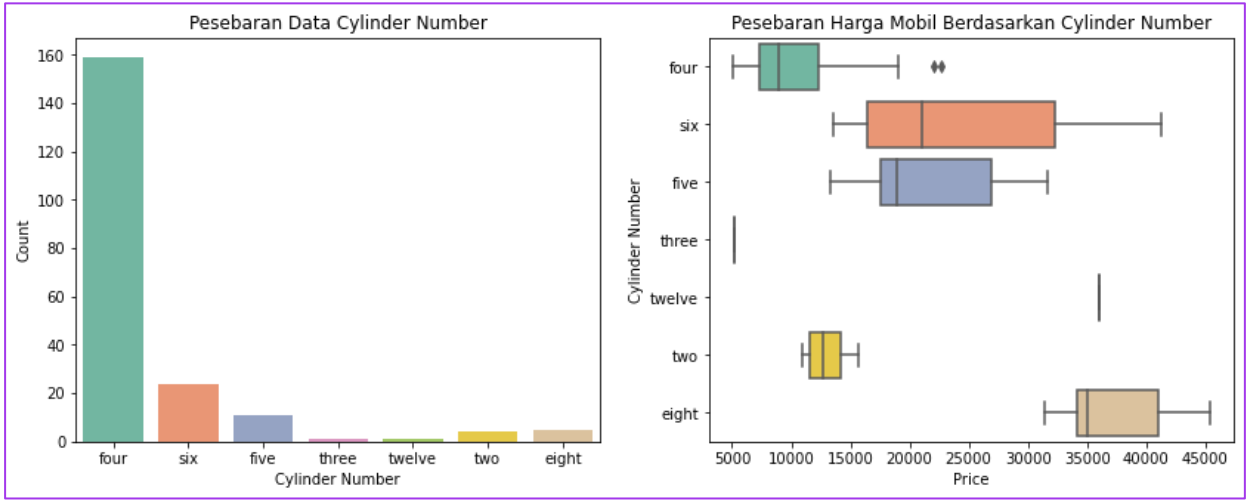
Exploratory Data Analysis

- Kebanyakan mobil memiliki tipe mesin ohc.
- Mobil dengan tipe mesin ohcv memiliki harga lebih tinggi dibandingkan dengan yang lain.



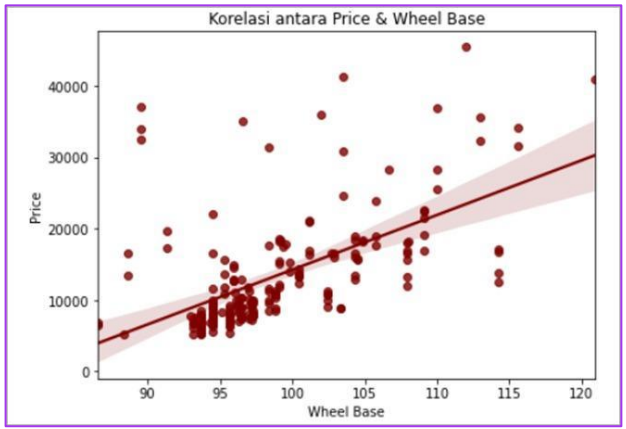
Exploratory Data Analysis

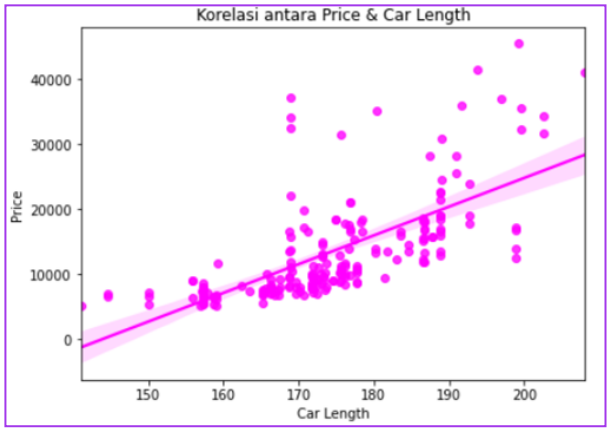
- Kebanyakan mobil memiliki 4 silinder.
- Untuk kebanyakan kasus harga mobil berbanding lurus dengan jumlah silinder.Tetapi mobil dengan 8 silinder memiliki harga tertinggi.

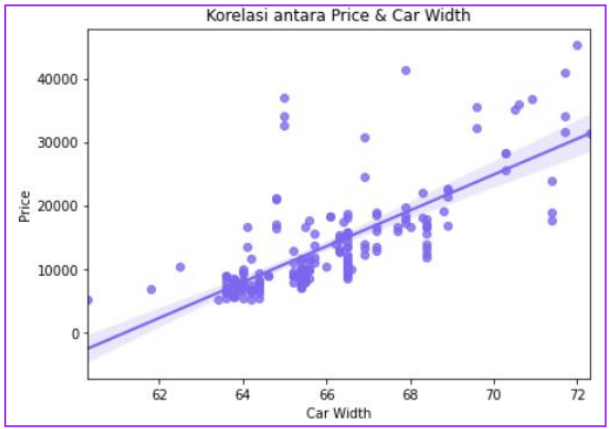


Exploratory Data Analysis

- Harga mobil memiliki korelasi positif dengan WheelBase, CarLength, CarWidth.

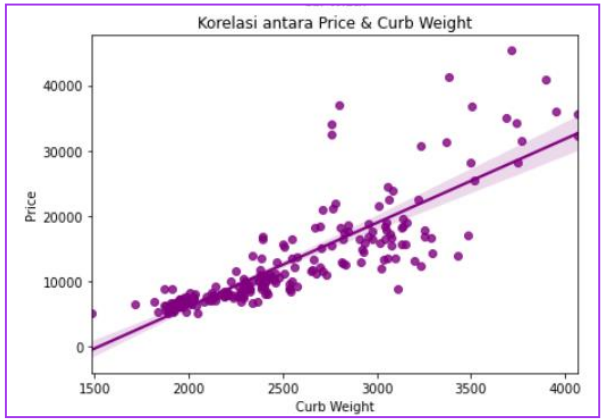


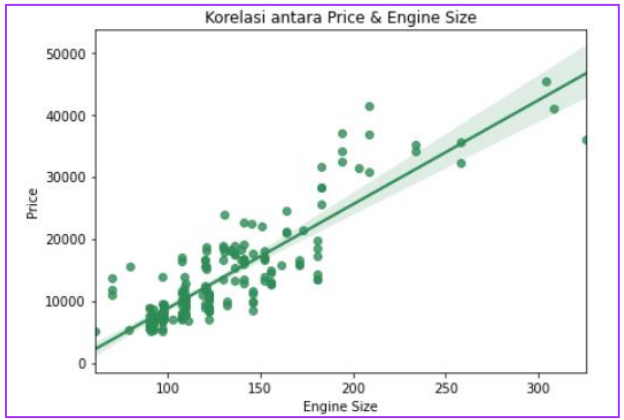


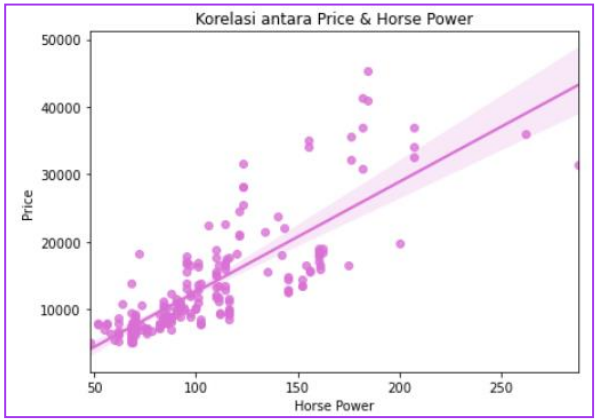


Exploratory Data Analysis

- Hal yang sama juga terjadi dengan CurbWeight, EngineSize, dan HorsePower. Sama-sama memiliki korelasi positif.

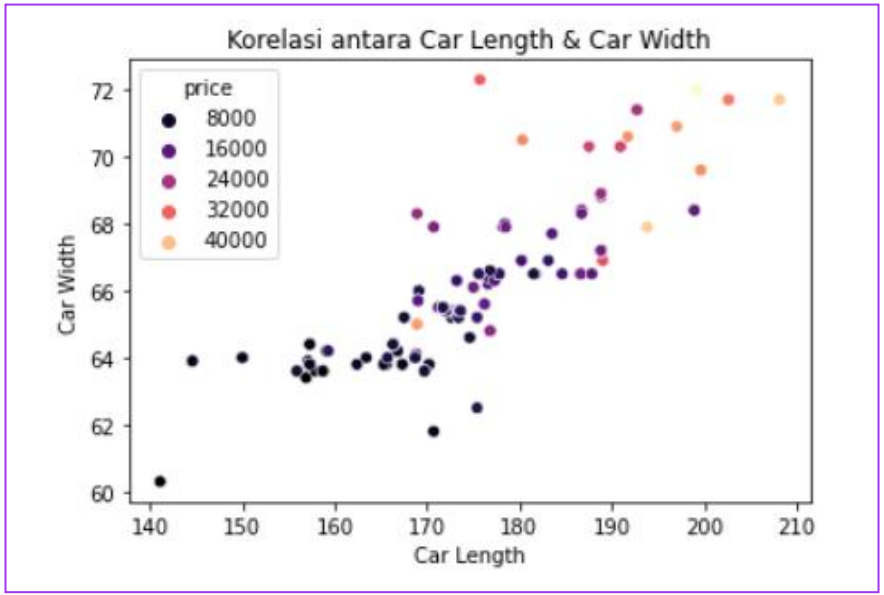






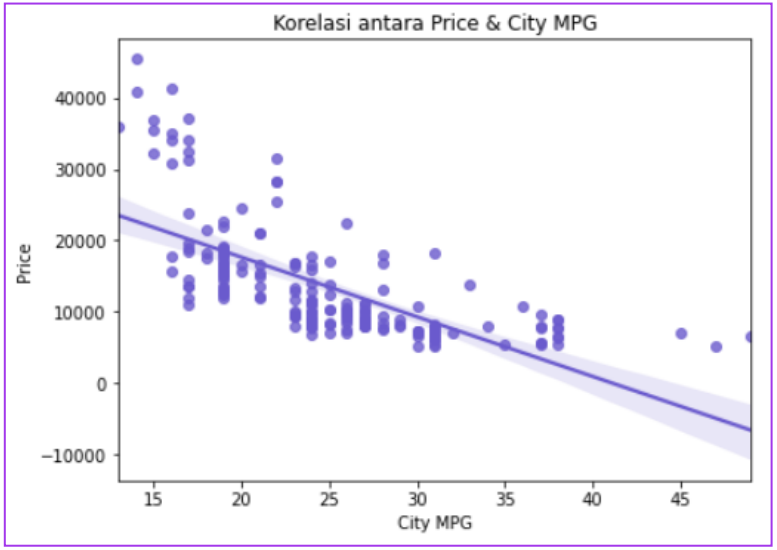
Exploratory Data Analysis

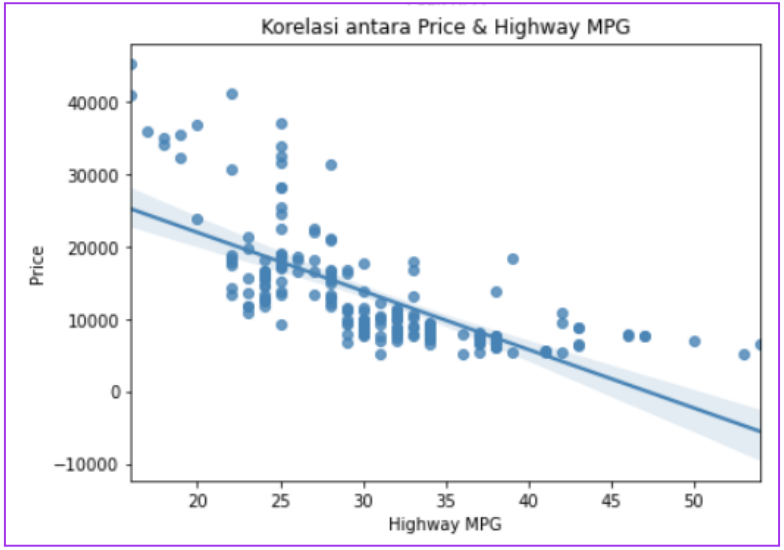
- Relasi antara *CarLength dan CarWidth* kuat



Exploratory Data Analysis

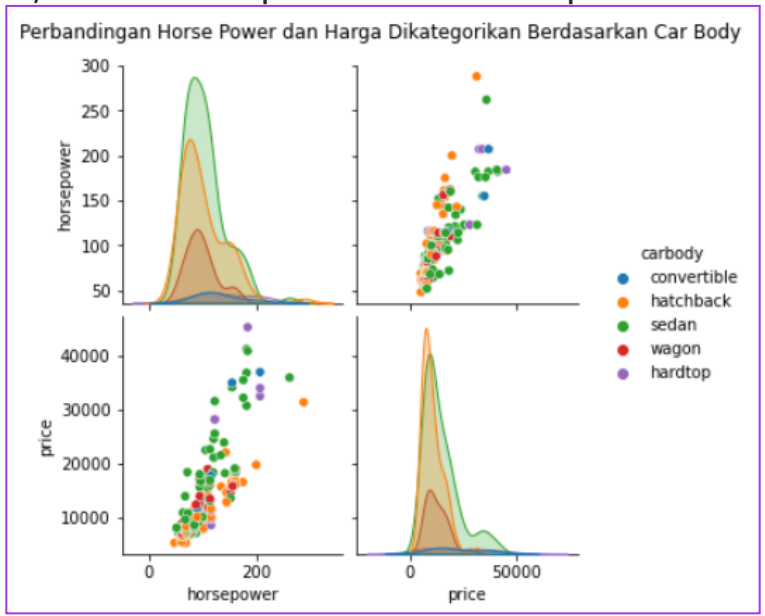
- Sedangkan korelasi negatif didapati antara *CityMPG* dan *HighwayMPG* dengan harga mobil.





Exploratory Data Analysis

- Harga dan horsepower mobil sedan sangat tersebar.
- Mobil sedan, convertible, dan hardtop memiliki horsepower dan harga yang tinggi.



Modelling

Train Test Split

- Menggunakan tiga *train test set* yang berbeda. Ketiga *train test set* memiliki perbandingan 80:20.

- *Train test set* 1 menggunakan semua *features* (kolom) sebagai variabel X.

- *Train test set* 2 hanya menggunakan lima belas *features* yang dipilih menggunakan Recursive Feature Elimination (RFE) sebagai variabel X. Features tersebut adalah: CarName, carbody, wheelbase, carlength, carwidth, carheight, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg.

Train Test Split

- Terakhir, *train test set* 3 sama seperti *train test set* 2, namun tidak menggunakan kolom *CarName* sebagai variabel X. Features tersebut adalah: carbody, wheelbase, carlength, carwidth, carheight, curbweight, enginesize, fuelsystem, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg.

Metrik Evaluasi & Model

Dua metrik yang digunakan untuk melakukan evaluasi model, yaitu:

- R Square
- Root Mean Square Error (RMSE)

Model yang digunakan untuk melakukan prediksi data adalah:

- Linear Regression
- CatBoost Regressor
- Random Forest Regressor

Hyperparameter Tuning

- Untuk model Linear Regression, menggunakan teknik *regularization* dengan model Ridge Regression, Lasso Regression, dan ElasticNet.

- Model CatBoost Regressor menggunakan teknik *randomized search on hyper parameters*. Dengan rentang parameter:
  - 'border_count': [32, 5, 10, 20, 50, 100, 200],
  - 'depth': [3, 1, 2, 6, 4, 5, 7, 8, 9, 10],
  - 'iterations': [250, 100, 150, 300, 200],
  - 'l2_leaf_reg': [3, 1, 5, 10, 100, 25],
  - 'learning_rate': [0.03, 0.001, 0.02, 0.1, 0.2, 0.3]

Hyperparameter Tuning

- Model Random Forest Regressor juga menggunakan teknik *randomized search on hyper parameters*. Dengan rentang parameter:
  - 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
  - 'max_features': ['auto', 'sqrt'],
  - 'min_samples_leaf': [1, 2, 4],
  - 'min_samples_split': [2, 5, 10],
  - 'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]

Evaluasi Model Basic

Model \ Train Test Set	Train Test Set 1	Train Test Set 2	Train Test Set 3
Linear Regression	**R2: 0.888** **RMSE: 2938.658**	R2: 0.776 RMSE: 4158.476	R2: 0.838 RMSE: 3536.155
CatBoost Regressor	R2: 0.891 RMSE: 2896.256	**R2: 0.905** **RMSE: 2706.349**	R2: 0.904 RMSE: 2714.947
Random Forest Regressor	R2: 0.901 RMSE: 2764.136	R2: 0.899 RMSE: 2793.524	**R2: 0.901** **RMSE: 2761.339**

Evaluasi Model Tuning

Model \ Train Test Set	Train Test Set 1	Train Test Set 2	Train Test Set 3
Linear Regression	**(Ridge)** **R2: 0.889** **RMSE: 2919.691**	(Ridge) R2: 0.862 RMSE: 3259.783	(Lasso) R2: 0.837 RMSE: 3550.84
CatBoost Regressor	R2: 0.898 RMSE: 2800.123	**R2: 0.911** **RMSE: 2614.817**	R2: 0.893 RMSE: 2866.955
Random Forest Regressor	R2: 0.902 RMSE: 2754.433	R2: 0.903 RMSE: 2734.886	**R2: 0.903** **RMSE: 2732.574**

Model Final

- Menggunakan Model CatBoost Regressor
- Menggunakan lima belas *features* yang dipilih menggunakan Recursive Feature Elimination (RFE) sebagai variabel X (Train Test Set 2)

  15 features tersebut: ['CarName', 'carbody', 'wheelbase', 'carlength', 'carwidth', 'carheight',
  'curbweight', 'enginesize', 'boreratio', 'stroke', 'compressionratio',
  'horsepower', 'peakrpm', 'citympg', 'highwaympg']

Model Final

- Menggunakan Hyperparameter Tuning dengan parameter:

  {'border_count': 100, 'depth': 3, 'iterations': 200, 'l2_leaf_reg': 10, 'learning_rate': 0.3}

- Hasil:
  - R2: 0.9116818545120756
  - RMSE: 2614.8172380899923

Conclusion

Kesimpulan

- Harga mobil ditentukan berdasarkan spesifikasi dan nama *brand*-nya.
- Semakin bagus spesifikasi suatu mobil, maka akan semakin mahal pula harga mobil tersebut.
- Kebanyakan harga jual mobil di pasaran berkisar antara 9.000 – 15.000.
- Di mayoritas kasus, mobil dengan *brand* yang dikenal mewah memiliki harga lebih mahal.
- Semakin tinggi jarak tempuh suatu mobil, maka akan semakin murah harga mobil tersebut.

Saran

- Perusahaan mobil harus mengutamakan spesifikasi dan kualitas mobil mereka serta memberikan harga jual yang bersaing.
- Perusahaan dapat menentukan harga jual mobil dengan kisaran antara 9.000 – 15.000 sesuai dengan banyaknya harga mobil di pasaran.

Saran

- Faktor-faktor kuat yang dapat digunakan perusahaan dalam penentuan harga mobil di antaranya:
  - *Brand* mobil
  - Dimensi mobil
  - Spesifikasi dan kualitas mesin
  - Jarak tempuh mobil.

Terima kasih!

Ada pertanyaan?