

Model Prediksi Harga Mobil Bekas

GROUP 5

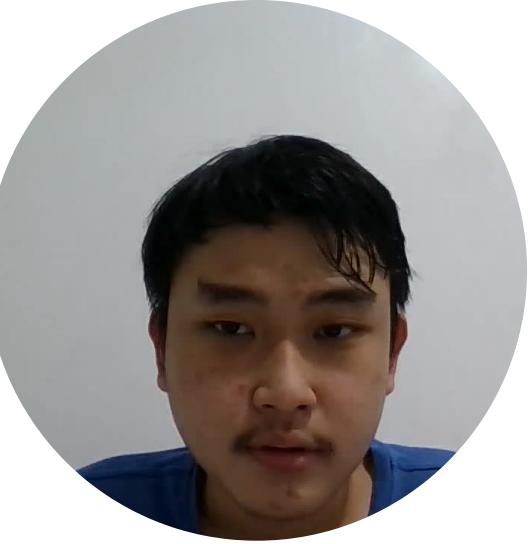
Aurelio Naufal Effendy (2106638526)
Muhamad Rakan Akmal (2106635745)
Musarrofah Kurnia (2106652543)
Rifa Nayaka Utami (2106632163)
Shafiyah Audiva Yasmin (2106706880)



Pendahuluan



1.1 Latar Belakang Masalah



Mobil merupakan suatu alat transportasi yang menjadi daya tarik banyak orang untuk membelinya. Selain dari kenyamanan dan keselamatan yang jauh lebih baik dibandingkan kendaraan lain dengan harga yang masih terjangkau, mobil dapat menjadi alat transportasi yang dapat menempuh suatu jarak yang cukup jauh. Walaupun harga pembelian mobil masih memiliki nilai yang terjangkau, tentunya pembelian mobil harus dipikirkan secara matang-matang, karena untuk mayoritas orang memiliki 1 mobil dapat dikatakan sudah memenuhi kebutuhan hidupnya. Beberapa orang bahkan memilih untuk membeli mobil yang sudah pernah dipakai orang lain atau mobil bekas, untuk dapat bisa melakukan transportasi yang lebih baik dengan harga yang lebih murah.

Di negara Italia, salah satu penjualan mobil terbanyak adalah Fiat 500 yang sejak dulu sudah diproduksi secara masif. Hal itu yang menyebabkan banyaknya pembelian mobil bekas Fiat 500 yang diproduksi pada tahun-tahun sebelumnya.





Dalam melakukan pembelian mobil bekas Fiat 500, tentunya pembeli akan memilih harga yang paling murah dengan kondisi mobil yang paling cocok dan bagus, serta memiliki nilai keuntungan bagi pembelinya.



Nilai keuntungan tersebut dapat berupa model dari mobil tersebut ataupun hal yang bersifat teknis seperti kekuatan mesin mobil dan juga jenis transmisi. Selain itu, dalam menganalisis suatu nilai harga, tentunya umur dari mobil dan juga jumlah jarak kilometer yang ditempuh termasuk kedalam variabel yang dapat menjadi nilai keuntungan, bahkan jumlah kepemilikan mobil tersebut sebelumnya juga dapat mempengaruhi. Begitu juga lokasi mobil tersebut dijual mempengaruhi bagaimana pembeli dapat melihat mobil tersebut sebelum melakukan pembelian.

Dengan demikian, kami tertarik menggunakan teknik regresi linier untuk memprediksi harga jual mobil bekas Fiat 500 dengan menganalisis hubungan variabel-variabel tersebut. Hasil dari prediksi ini diharapkan mampu memberikan informasi harga jual mobil bekas Fiat 500 yang sesuai dengan keadaan yang diharapkan.

1.2 Data



Permasalahan

Data yang kami ambil merupakan dataset harga jual mobil bekas Fiat 500 di Italia. Permasalahan ini merupakan permasalahan regresi. Kami akan memprediksi harga jual mobil bekas Fiat 500 dari data yang ada dengan menerapkan model regresi linier.

Sumber

[https://www.kaggle.com/datasets/paolocons/small-dataset-about-used-fiat-500-sold-in-italy?
resource=download](https://www.kaggle.com/datasets/paolocons/small-dataset-about-used-fiat-500-sold-in-italy?resource=download)

Ukuran Data

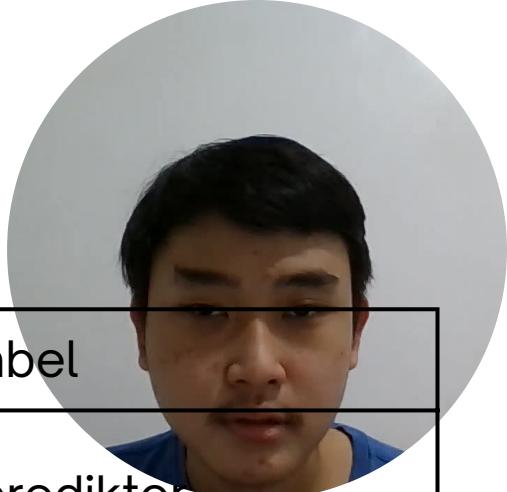
Data memuat 380 pengamatan dengan jumlah pengukuran (kolom data) adalah 9 kolom.

Skala/Tipe Data

Data yang akan digunakan dalam pemodelan memuat 379 pengamatan (1 pengamatan dihapus karena merupakan outlier) dengan 7 variabel, yaitu 1 variabel respon kuantitatif (Y) , 4 variabel prediktor kuantitatif (X), dan 2 variabel prediktor kualitatif (X).

Skala/Tipe Data

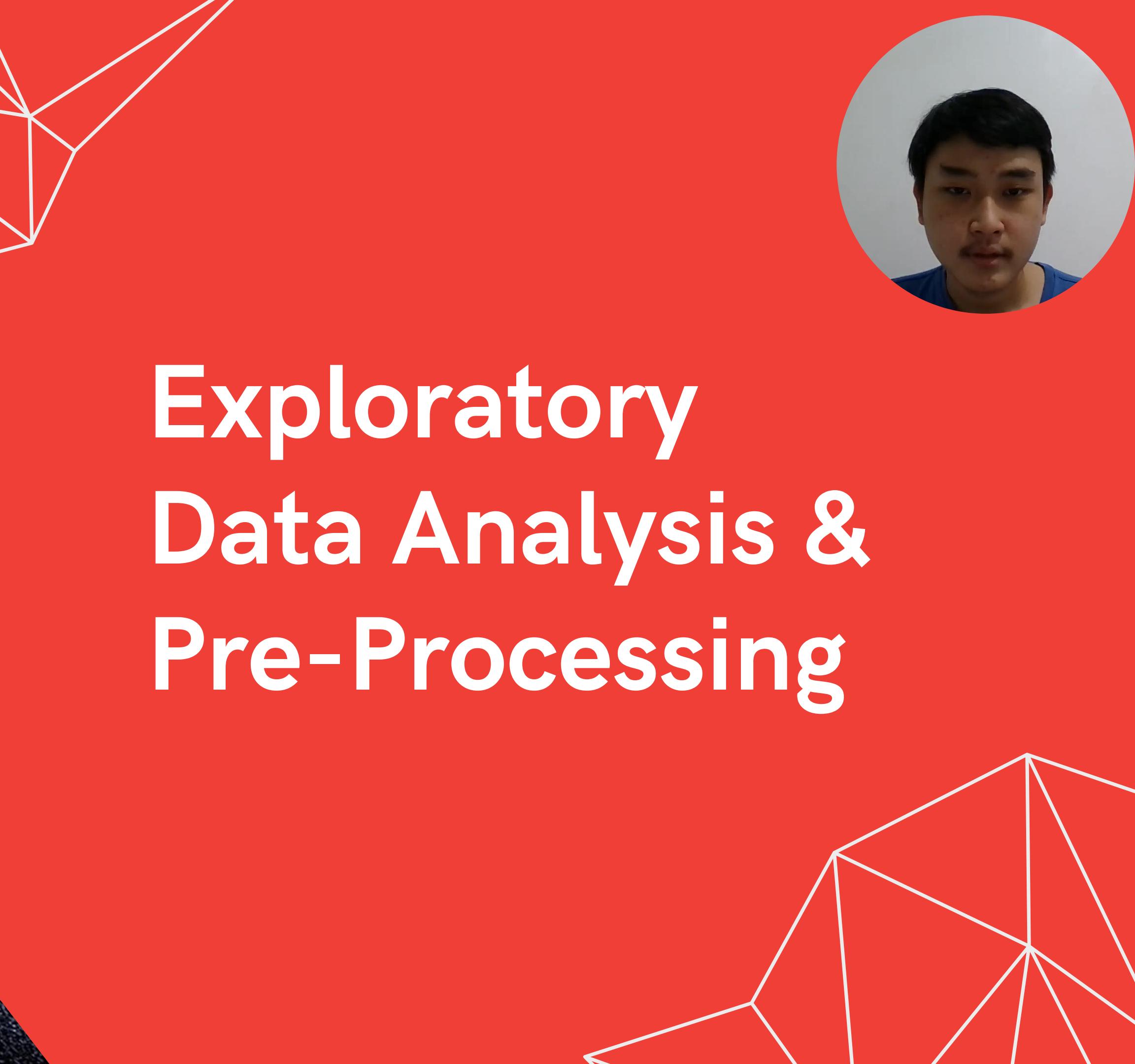
Data yang digunakan sendiri merupakan dataset yang berisi 9 jumlah pengukuran, yaitu



No.	Pengukuran	Tipe	Keterangan	Variabel
1.	model	object (kategorik)	model dari mobil Fiat 500, terdapat 4 jenis kategori, yakni 'pop', 'lounge', 'star' dan 'sport'.	variabel prediktor
2.	engine power	integer (numerik)	kekuatan mesin mobil (banyak horsepowernya), berkisar antara 69 - 101 horsepower	variabel prediktor
3.	transmission	object (kategorik)	jenis transmisi dari mobil, terdiri dari 2 jenis kategori yakni, 'manual' dan 'automatic'	variabel prediktor
4.	age in days	integer (numerik)	umur dari mobil dalam hari, berkisar antara 91 - 4179 hari	variabel prediktor
5.	km	integer (numerik)	jarak kilometer yang telah ditempuh mobil, berkisar antara 4981 - 259000 km	variabel prediktor
6.	previous owners	integer (numerik)	Jumlah kepemilikan mobil sebelumnya, berkisar antara 1 sampai 4 orang	variabel prediktor
7.	lat	float (numerik)	koordinat lintang lokasi penjual, berkisar antara 37-46 derajat lintang	dihilangkan (pada tahap preprocessing) karena tidak berguna pada pemodelan
8.	lon	float (numerik)	koordinat bujur lokasi penjual, berkisar antara 7-18 derajat bujur	dihilangkan (pada tahap preprocessing) karena tidak berguna pada pemodelan
9	price	integer(numerik)	harga jual mobil dalam mata uang euro, berkisar antara 2890 - 15900 euro	variabel respon



Exploratory Data Analysis & Pre-Processing





2.1 Load Data

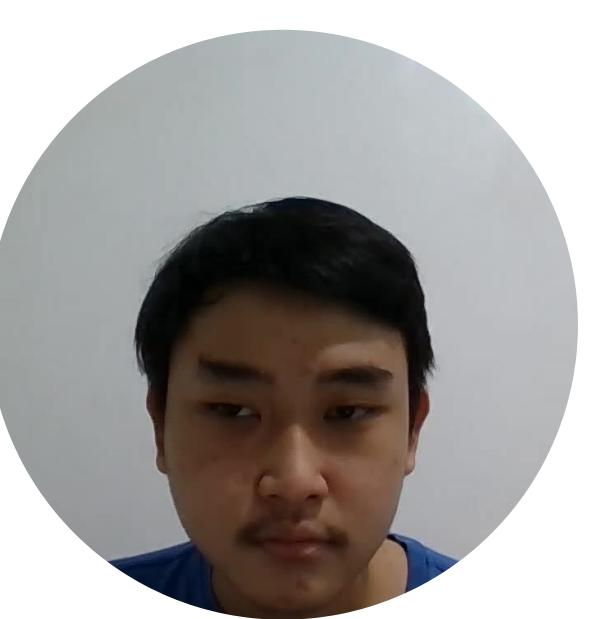
Kami melakukan import library yang akan dibutuhkan dalam proses pre-processing. Data yang diambil dalam kaggle kami simpan pada Github. Pemanggilan file di Github melalui Github dapat dilihat pada kodingan di bawah ini. Dataset kami simpan pada variabel df

Import modules yang akan dibutuhkan

```
[26] #import modules
      import numpy as np
      import pandas as pd
      import matplotlib.pyplot as plt
      import seaborn as sns
```

Load data dari Github

```
df = pd.read_csv('https://raw.githubusercontent.com/rakanakml17/molinproject2group5/main/Used_fiat_500_in_Italy_dataset.csv')
```



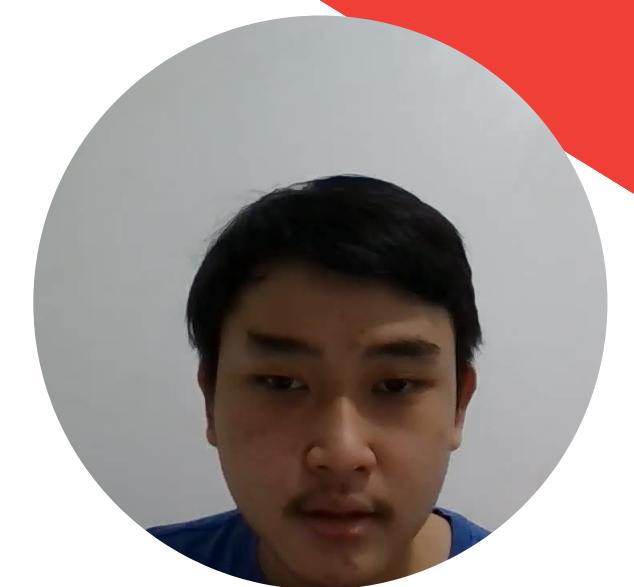
Berikut head data dari dataset yang telah di-load.

```
[29] print('Head Dataset')
      df.head()
```

Head Dataset

	model	engine_power	transmission	age_in_days	km	previous_owners	lat	lon	price
0	pop	69	manual	4474	56779	2	45.071079	7.46403	4490
1	lounge	69	manual	2708	160000	1	45.069679	7.70492	4500
2	lounge	69	automatic	3470	170000	2	45.514599	9.28434	4500
3	sport	69	manual	3288	132000	2	41.903221	12.49565	4700
4	sport	69	manual	3712	124490	2	45.532661	9.03892	4790

2.2 Mengecek Missing Values



```
0s
print("Tipe Data:", type(df))

print("\nKeterangan Dataset")
print(df.info())

print("\nHead Dataset")
df.head()

Tipe Data: <class 'pandas.core.frame.DataFrame'>

Keterangan Dataset
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 380 entries, 0 to 379
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
---  --  
 0   model            380 non-null    object  
 1   engine_power     380 non-null    int64  
 2   transmission     380 non-null    object  
 3   age_in_days      380 non-null    int64  
 4   km                380 non-null    int64  
 5   previous_owners  380 non-null    int64  
 6   lat               380 non-null    float64 
 7   lon               380 non-null    float64 
 8   price             380 non-null    int64  
dtypes: float64(2), int64(5), object(2)
memory usage: 26.8+ KB
```

Selanjutnya, kami mengecek apakah kolom atau baris data tersebut mengandung missing values atau tidak. Dengan menggunakan method info(), dapat dilihat bahwa tidak terdapat missing values pada data frame df.

2.3 Data Understanding

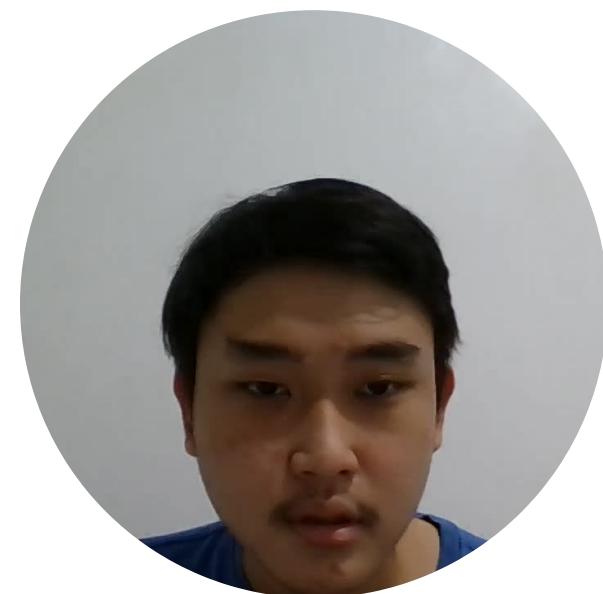
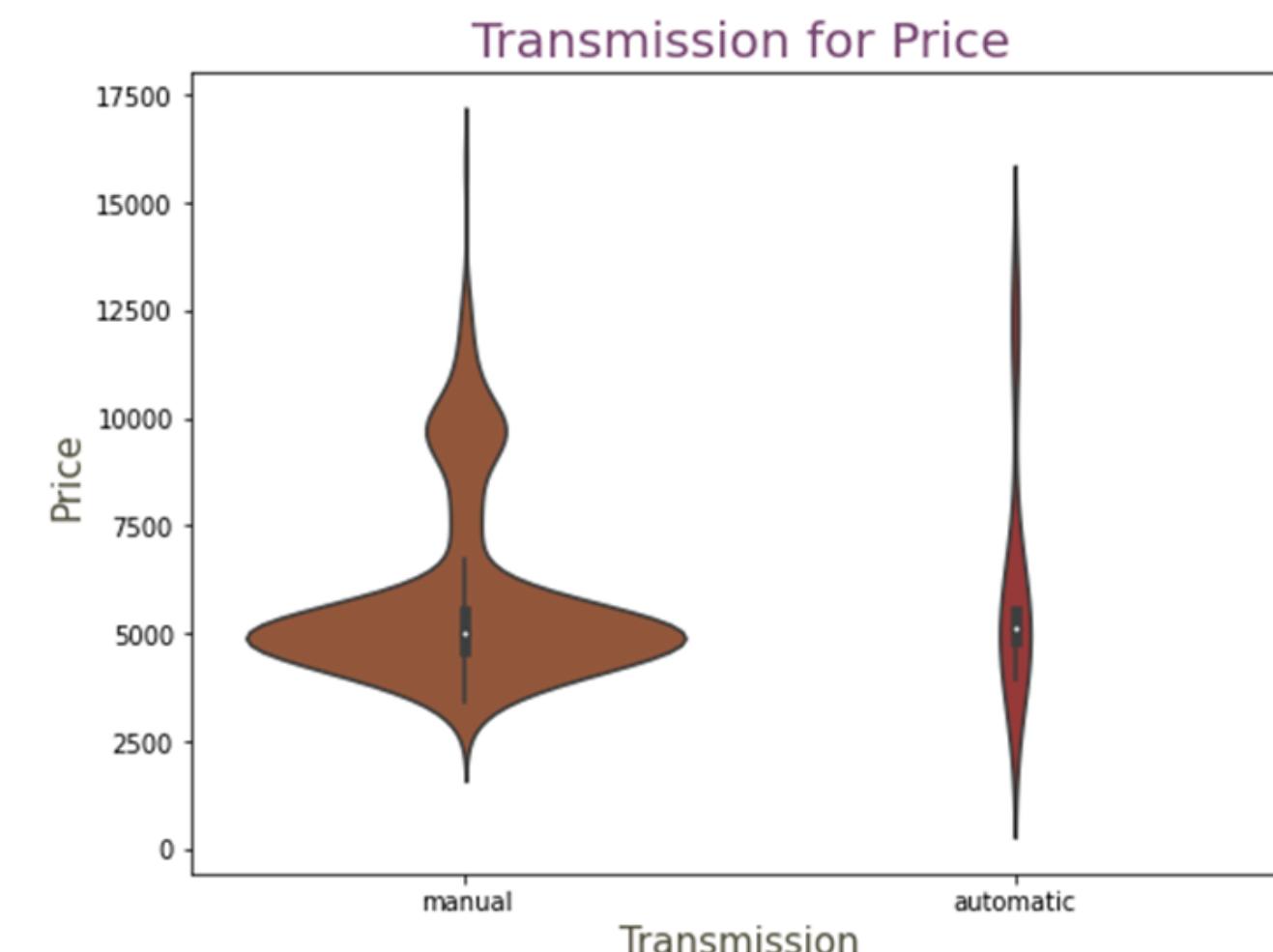
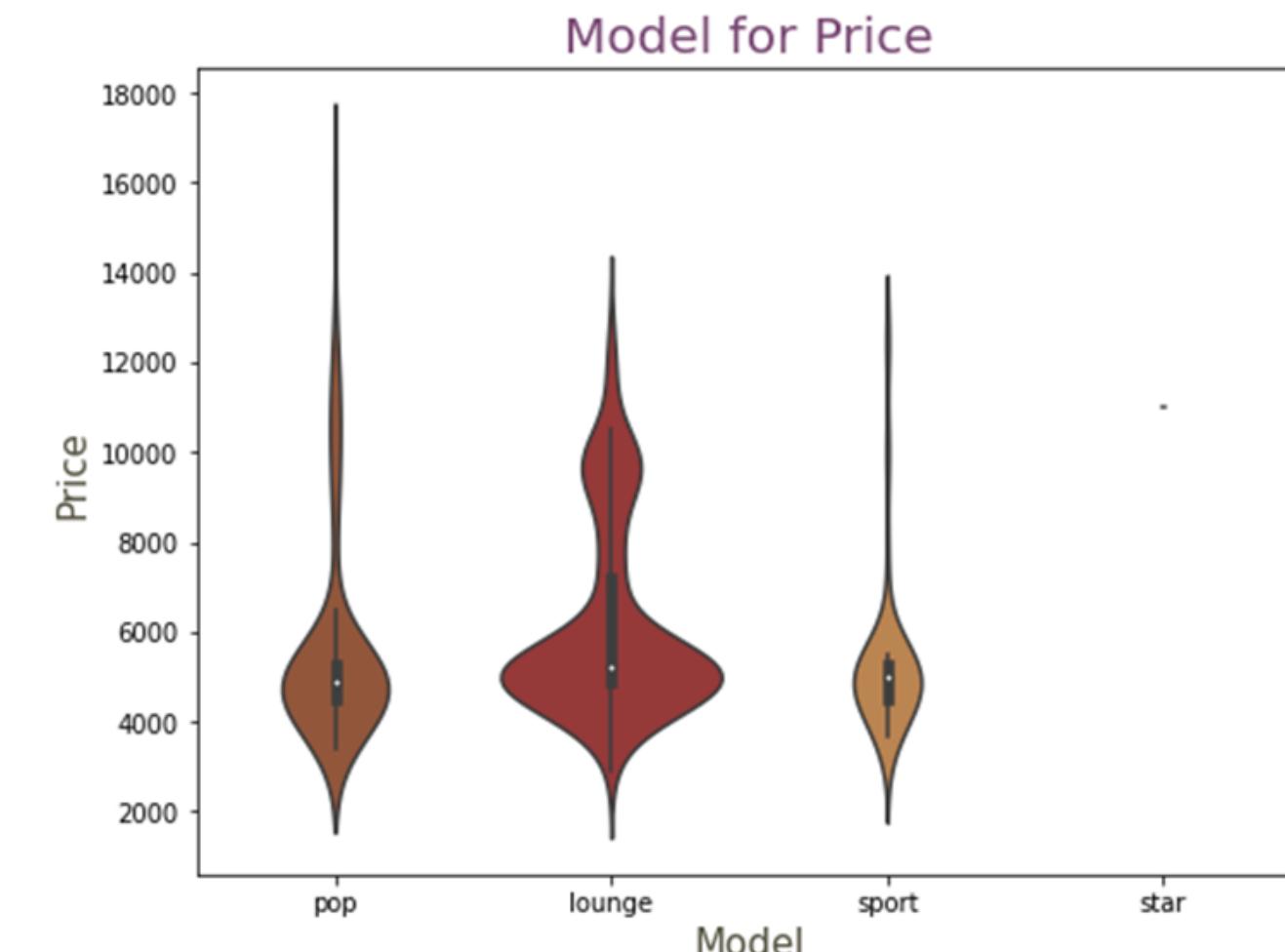
2.3.1 Variabel Kategorik

Melakukan cek atribut variabel kategorik dan melakukan plotting dengan violin plot untuk melihat persebaran dari atribut variabel kategorik.

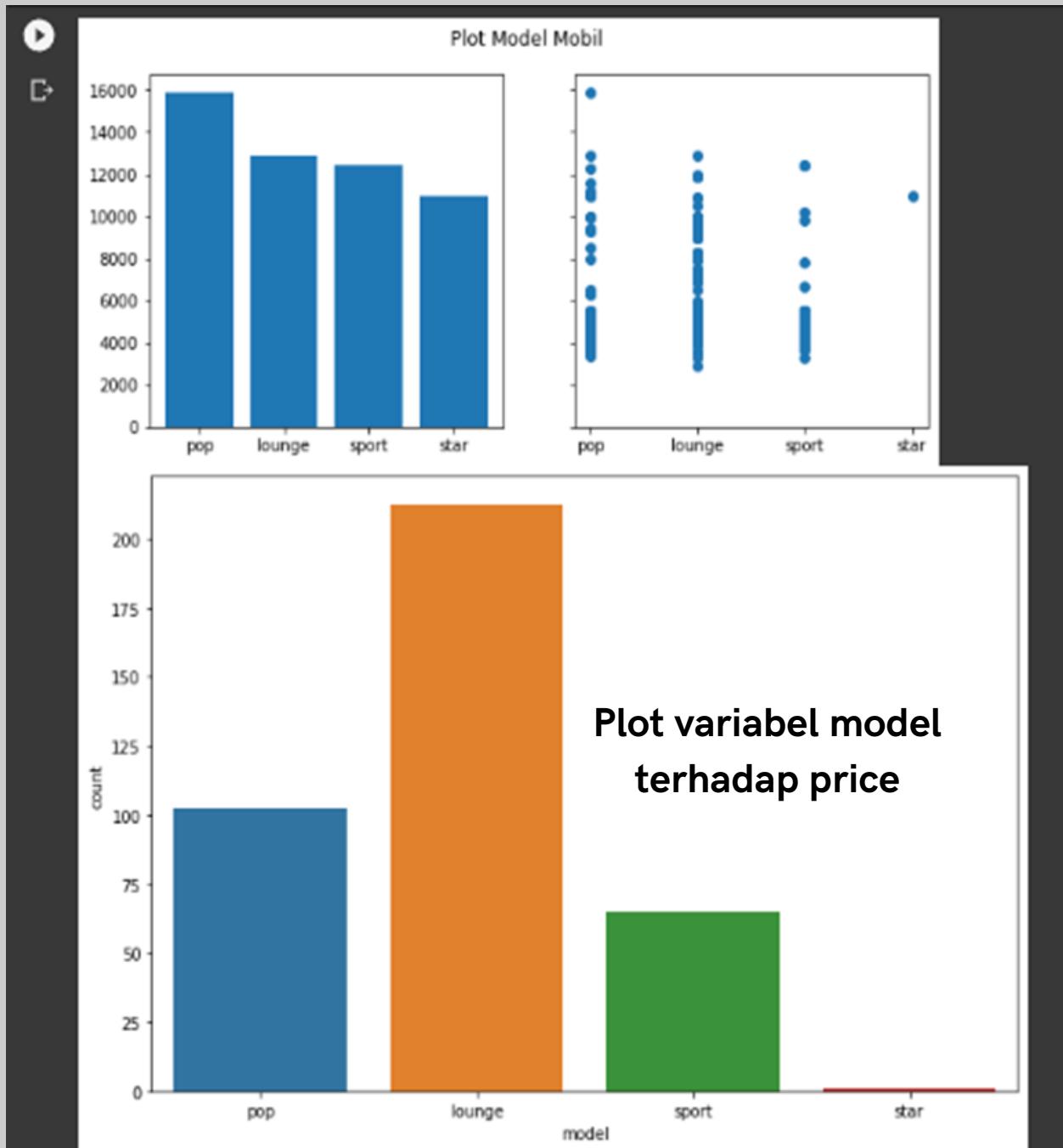
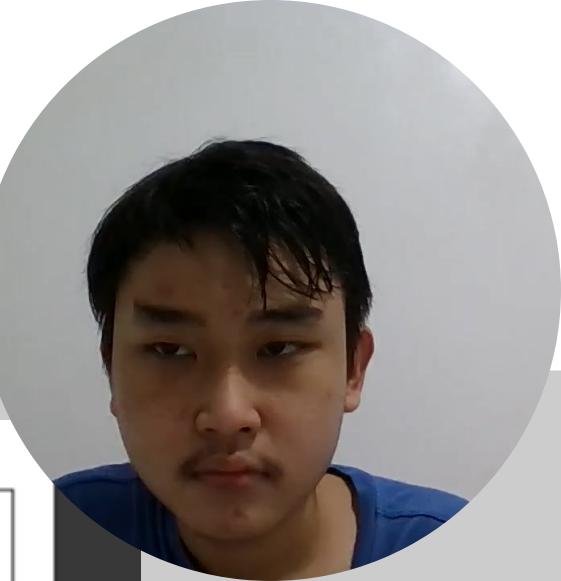
```
Mengecek atribut variabel kategorik

[4] df.describe(include=[object])

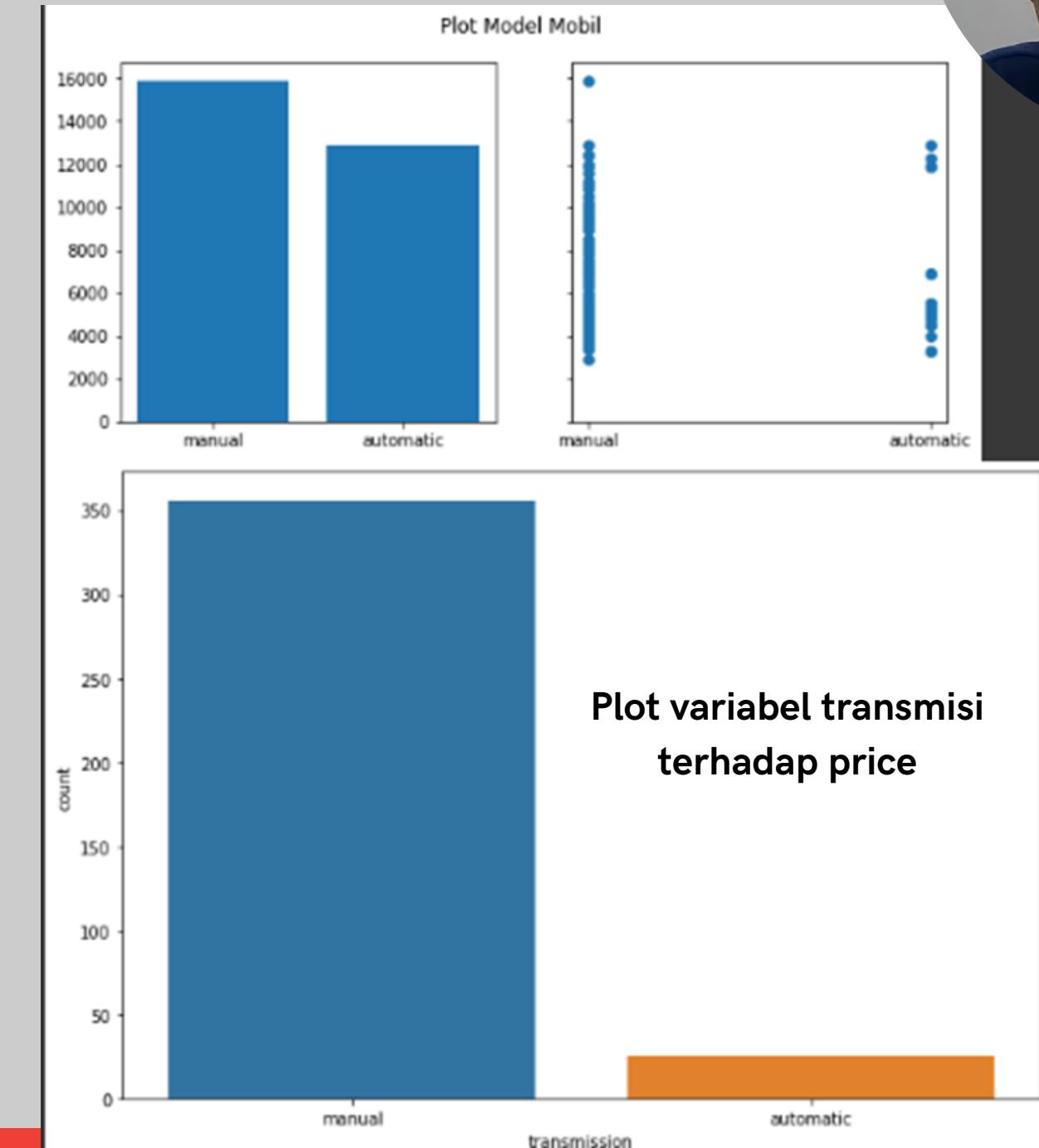
      model  transmission
count    380          380
unique     4              2
top    lounge        manual
freq    212          355
```



Melihat persebaran data dengan beberapa jenis plot untuk variabel kategorik



Diperoleh dari visualisasi grafik tersebut bahwa model mobil yang paling banyak ditemukan dalam data adalah model 'lounge' dan yang paling sedikit adalah 'star' dengan persebaran harga dibanding model yang cukup mirip, namun di model mobil 'lounge' cukup banyak juga berkisar di harga 10000 euro.



Diperoleh dari visualisasi grafik tersebut bahwa model mobil yang paling banyak ditemukan dalam data adalah jenis transmisi mobil yang paling banyak ditemukan di dalam data adalah 'manual', dengan rata-rata berkisar di harga 5000 euro.

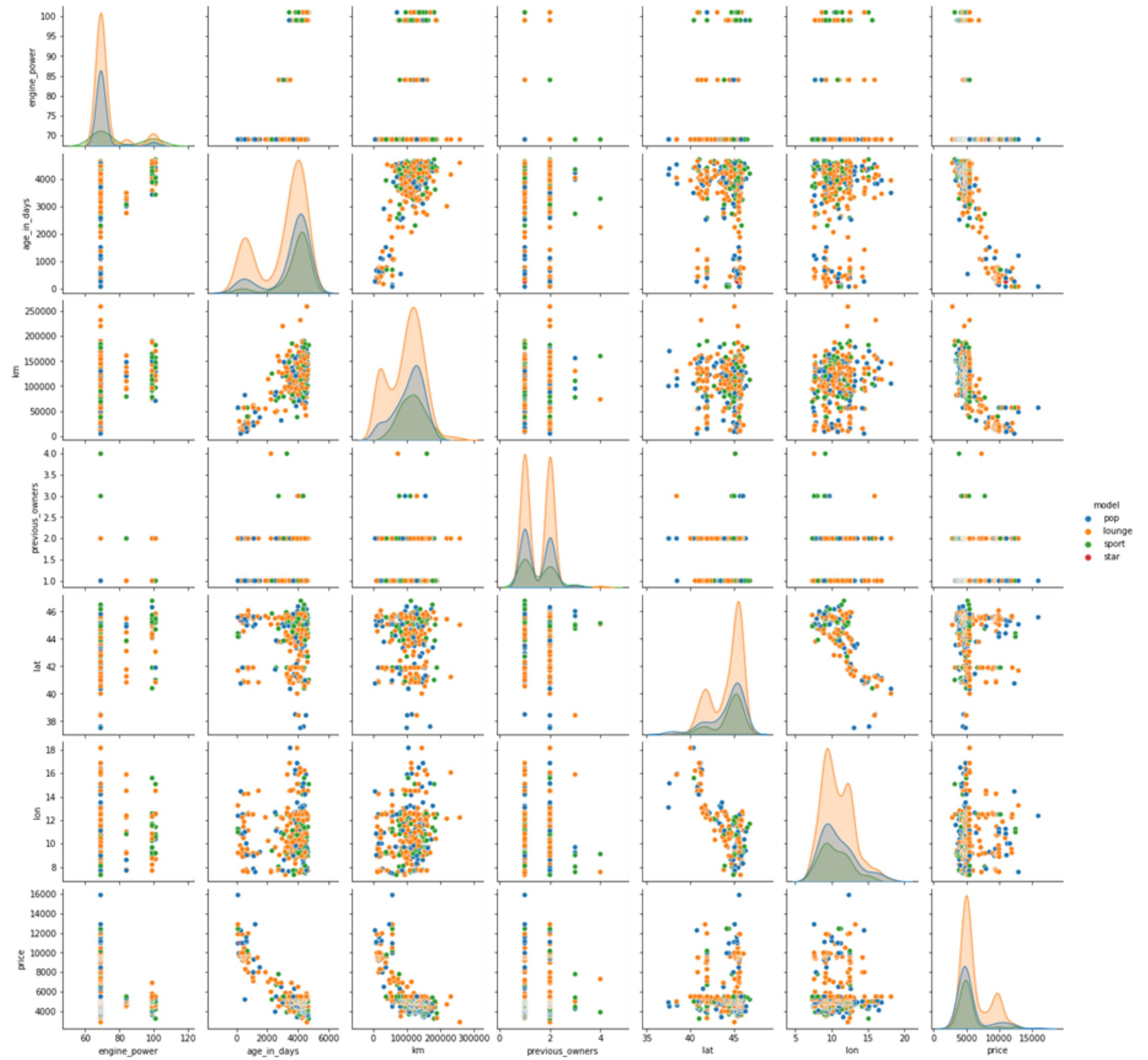


2.3.2 Variabel Numerik

Melakukan cek atribut variabel numerik dan melakukan plotting dengan pair plot untuk melihat scatterplot dari atribut variabel numerik.

[8] df.describe().T

	count	mean	std	min	25%	50%	75%	max
engine_power	380.0	73.015789	9.984672	69.000000	69.000000	69.000000	69.000000	101.000000
age_in_days	380.0	3310.265789	1437.555063	91.000000	3097.250000	3927.000000	4322.000000	4719.000000
km	380.0	102196.250000	47115.355234	4981.000000	76611.750000	112000.000000	135000.000000	259000.000000
previous_owners	380.0	1.510526	0.560244	1.000000	1.000000	1.000000	2.000000	4.000000
lat	380.0	44.257712	1.767518	37.510872	43.514196	45.069679	45.556942	46.781651
lon	380.0	10.742644	2.167753	7.320720	9.159140	10.301505	12.346783	18.168011
price	380.0	5881.655263	2170.617946	2890.000000	4600.000000	5000.000000	5500.000000	15900.000000



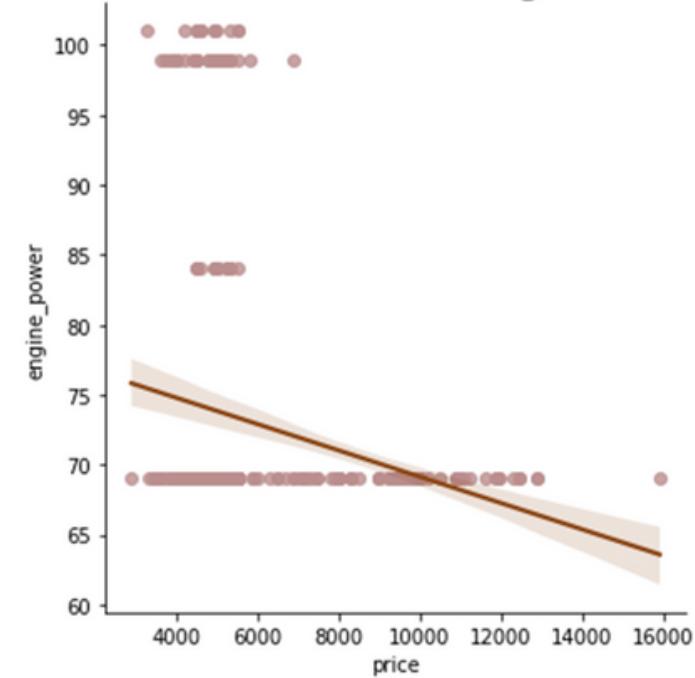
Diperoleh statistik dari variabel numerik tersebut. Serta dari pairplot dapat dilihat bahwa 'price' atau harga yang merupakan variabel yang akan kita prediksi mempunyai grafik yang right skewed (memiliki lebih banyak data di kiri).



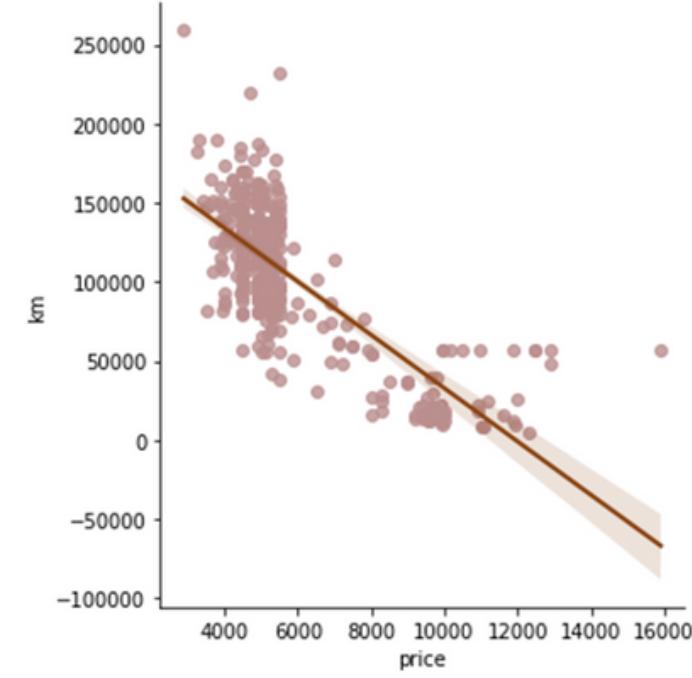
2.4 Mengecek Outlier pada Data

Kemudian, kita juga dapat mengecek outlier yang mungkin ada dari data tersebut dengan membuat scatterplot antara setiap variabel numerik dengan variabel price sebagai variabel target.

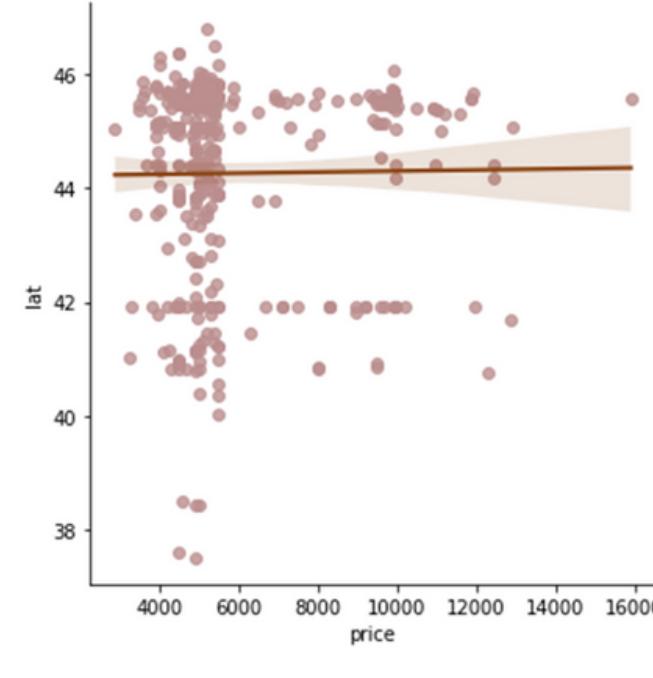
Line Plot on Price vs Engine Power



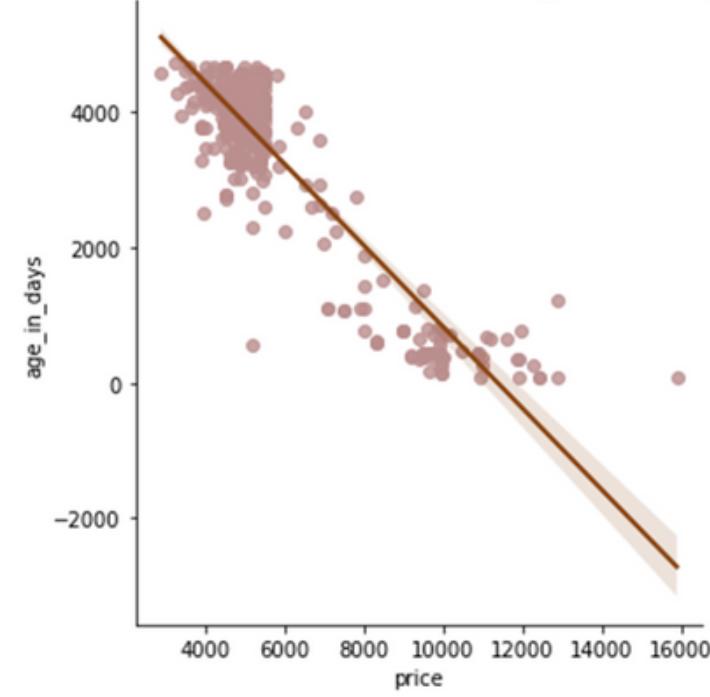
Line Plot on Price vs Kilometers



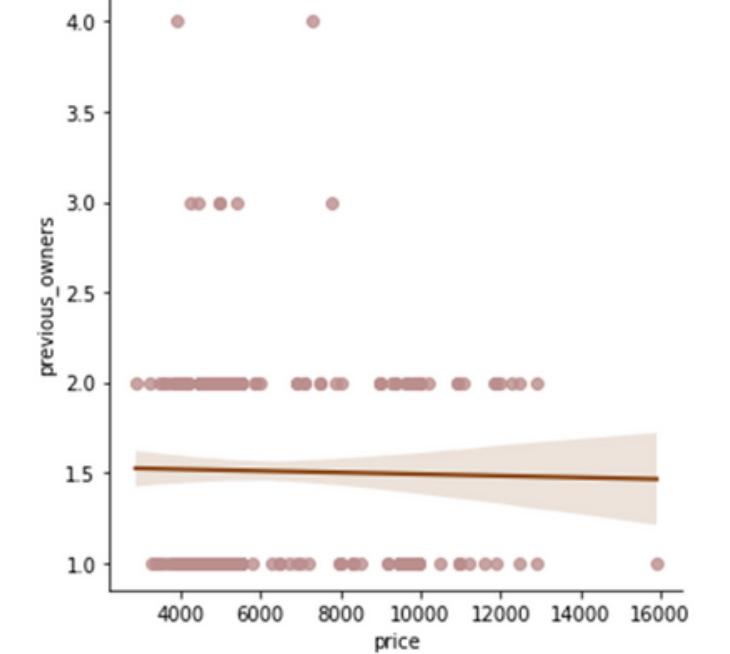
Line Plot on Price vs Latitude



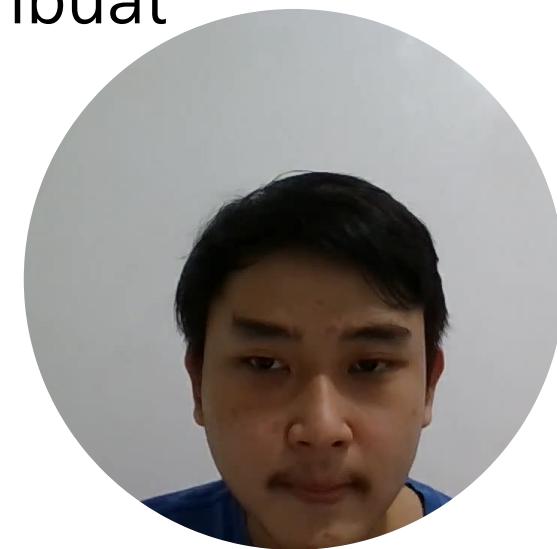
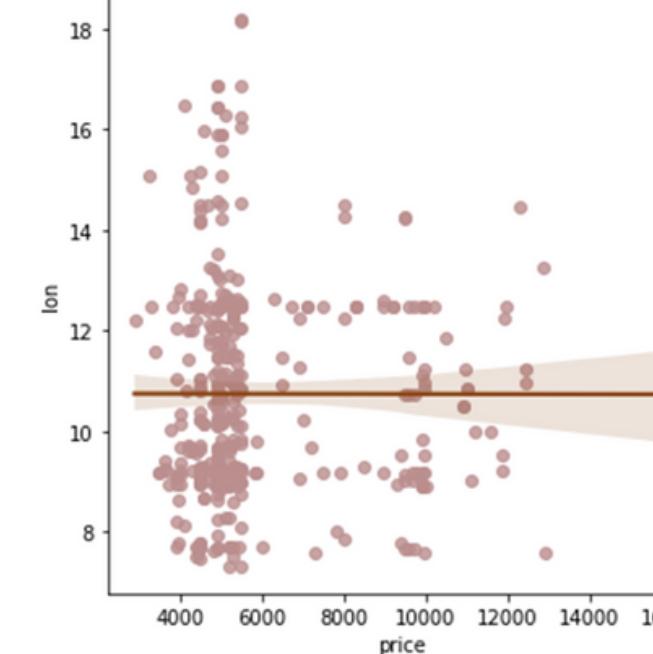
Line Plot on Price vs Age in Days



Line Plot on Price vs Previous Owners



Line Plot on Price vs Longitude

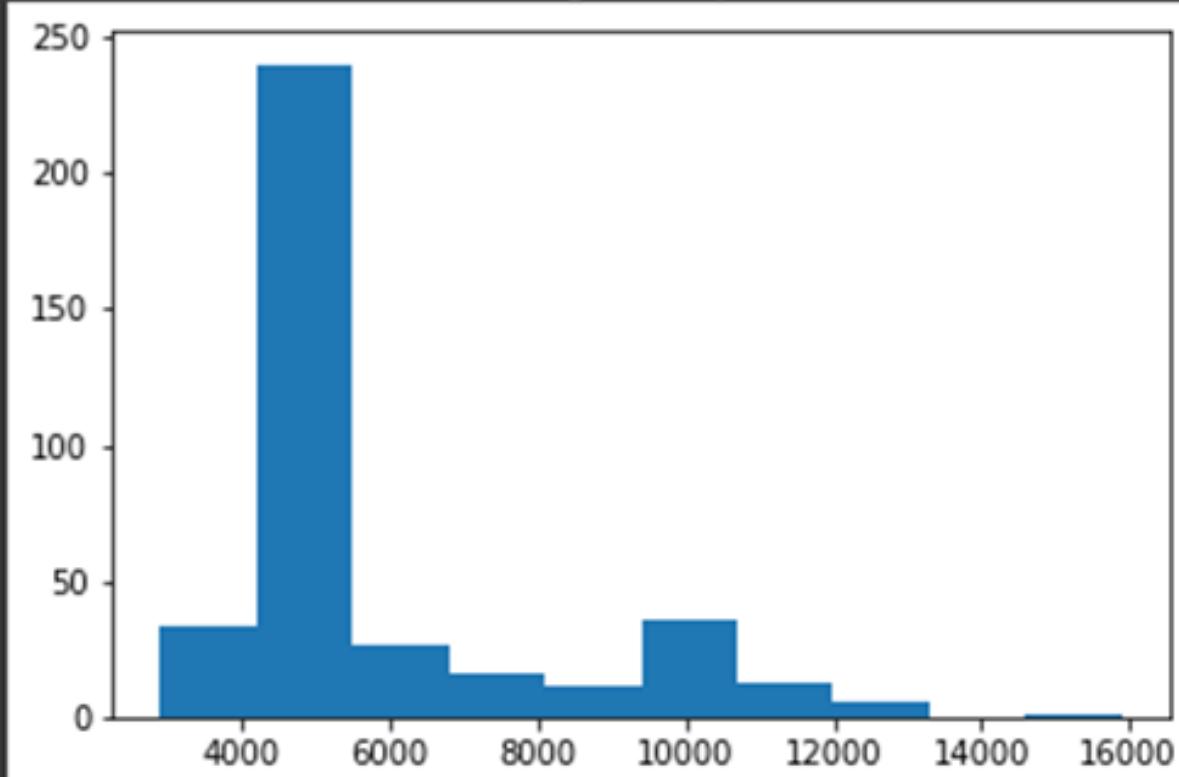


Dari grafik di samping, terlihat beberapa outlier dari masing-masing variabel. Data yang menjadi outlier sebetulnya tidak merusak persebaran data secara krusial dan juga data tersebut masih valid dalam observasi.

Karena kita akan memprediksi harga ('price'), maka akan dicoba untuk melihat persebaran data dari price dan boxplotnya.

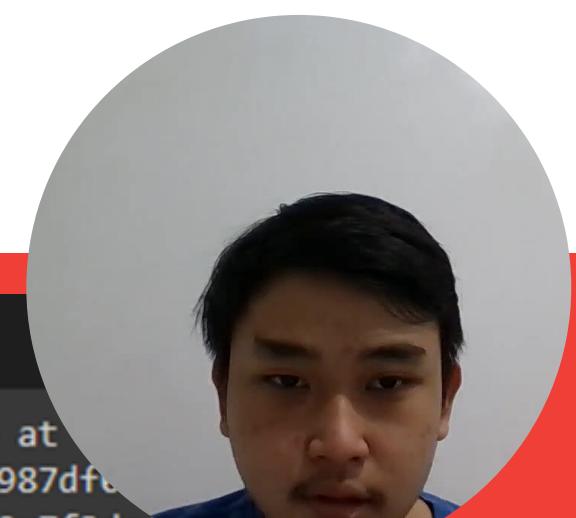
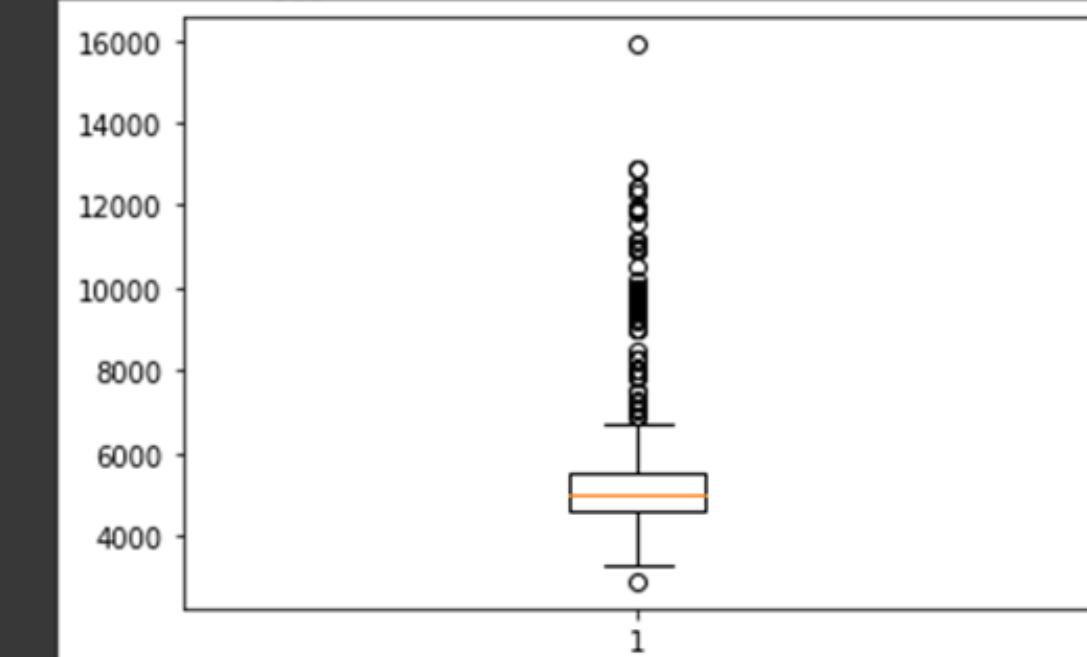
```
[11] plt.hist(df['price'])
```

```
(array([ 33., 240., 26., 16., 11., 36., 12., 5., 0., 1.]),  
 array([ 2890., 4191., 5492., 6793., 8094., 9395., 10696., 11997.,  
        13298., 14599., 15900.]),  
<a list of 10 Patch objects>)
```



```
[12] plt.boxplot(df['price'])
```

```
{'whiskers': [matplotlib.lines.Line2D at  
              <matplotlib.lines.Line2D at 0x7f3da4987df0>],  
 'caps': [matplotlib.lines.Line2D at 0x7f3da4987e00>,  
          <matplotlib.lines.Line2D at 0x7f3da49954f0>],  
 'boxes': [matplotlib.lines.Line2D at 0x7f3da4987730>],  
 'medians': [matplotlib.lines.Line2D at 0x7f3da4995850>],  
 'fliers': [matplotlib.lines.Line2D at 0x7f3da4995b50>],  
 'means': []}
```



Dapat dilihat dari boxplot tersebut bahwa ada outlier dimana harga jualnya sangat tinggi sendiri mencapai 16000 euro dan sebenarnya ada juga outlier yang dibawah 3000 euro (namun tidak kita masukkan sebagai outlier karena masih dekat dengan boxplot)

Outlier tersebut dapat kita lihat sebagai berikut.

```
[13]: outliers = df[df['price'] > 14000]
outliers.head()

model engine_power transmission age_in_days km previous_owners lat lon price
97 pop 69 manual 91 56779 1 45.580879 12.36937 15900
```



Karena outlier tersebut tersebar cukup jauh dan hanya ada 1 observasi, maka Kita dapat menghilangkan/mendrop nilai tersebut.

```
▶ #Karena outlier tersebut tersebar cukup jauh dan hanya ada 1 observasi, maka Kita dapat menghilangkan/mendrop nilai tersebut
df_preprop = df.drop(outliers.index)
```

Mengecek bahwa outlier tersebut sudah dihapus dalam data yang baru (berkurang 1 dari yang sebelumnya 380).

```
▶ df_preprop.shape
#mengecek bahwa outlier tersebut sudah dihapus dalam data yang baru (berkurang 1 dari yang sebelumnya 380)

(379, 9)
```

2.5 Membuat *Dummy Variable*

PRE-PROCESSING

Alasan membuat dummy variable adalah karena model regresi linier hanya dapat menerima int atau float. Jadi, dengan menggunakan dummy variabel, variabel kategorik tersebut dapat berubah menjadi int supaya dapat digunakan dalam membuat model.

```
df_prep[‘model’] = df_prep[‘model’].astype(‘category’)
df_prep[‘transmission’] = df_prep[‘transmission’].astype(‘category’)
```

```
<class ‘pandas.core.frame.DataFrame’>
Int64Index: 379 entries, 0 to 379
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   model            379 non-null    category
 1   engine_power     379 non-null    int64   
 2   transmission     379 non-null    category
 3   age_in_days      379 non-null    int64   
 4   km                379 non-null    int64   
 5   previous_owners   379 non-null    int64   
 6   price             379 non-null    int64  
dtypes: category(2), int64(5)
memory usage: 18.8 KB
```

Info dataset setelah dilakukan perubahan tipe data kategori dan penghapusan variabel

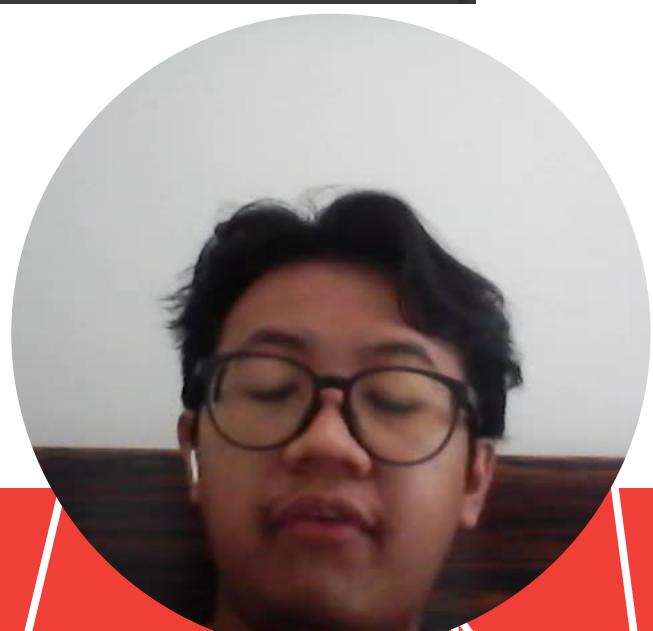
Alasan penghapusan variabel koordinat bujur dan lintang adalah karena kedua variabel variabel tersebut tidak bisa berdiri sendiri untuk memiliki pengaruh (yang berarti) dalam model.



Dengan Melakukan dummy encoding akan dibentuk df_preprop.

```
# karena model ML hanya menerima int atau float. Jadi, akan menggunakan dummy variabel untuk mengubah variabel kategorik tersebut menjadi int supaya dapat  
# Dengan Melakukan dummy encoding akan dibentuk df_preprop  
df_preprop_dummy = pd.get_dummies(df_preprop)  
df_preprop_dummy.tail()
```

	engine_power	age_in_days	km	previous_owners	price	model_lounge	model_pop	model_sport	model_star	transmission_automatic	transmission_manual
375	69	4474	55976	2	5500	1	0	0	0	0	1
376	69	4200	134717	1	5500	1	0	0	0	0	1
377	69	3470	113344	1	5500	1	0	0	0	0	1
378	69	3712	130000	1	5500	0	1	0	0	1	0
379	99	4566	96000	1	5500	1	0	0	0	0	1



Untuk mengakomodir pembuatan model dalam regresi, dibentuk juga dummy variabel untuk n-1 kategori.

PRE-PROCESSING

```
[46] # Untuk mengakomodir pembuatan model dalam regresi, dibentuk juga dummy variabel untuk n-1 kategori

df_prepop_dummies = df_prepop.copy()

df_prepop_dummies['model_lounge'] = np.where(df_prepop_dummies['model']=='lounge', 1, 0)
df_prepop_dummies['model_pop'] = np.where(df_prepop_dummies['model']=='pop', 1, 0)
df_prepop_dummies['model_sport'] = np.where(df_prepop_dummies['model']=='sport', 1, 0)
df_prepop_dummies['transmission_manual'] = np.where(df_prepop_dummies['transmission']=='manual', 1, 0)

df_prepop_dummies = df_prepop_dummies.drop(labels = ['model', 'transmission'], axis =1)

df_prepop_dummies.tail()
```



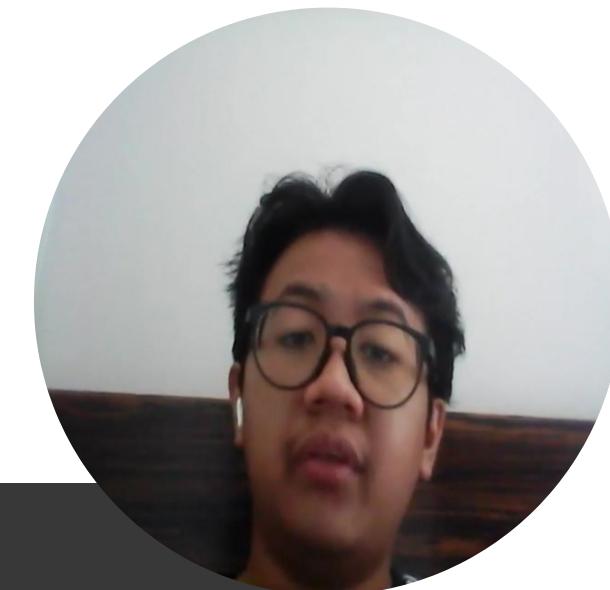
	engine_power	age_in_days	km	previous_owners	price	model_lounge	model_pop	model_sport	transmission_manual
375	69	4474	55976	2	5500	1	0	0	1
376	69	4200	134717	1	5500	1	0	0	1
377	69	3470	113344	1	5500	1	0	0	1
378	69	3712	130000	1	5500	0	1	0	0
379	99	4566	96000	1	5500	1	0	0	1

Didapatkan setelah melakukan cek untuk data dengan dummy variable menggunakan .info() adalah sebagai berikut:

```
df_prepred_dummy.info()  
df_prepred_dummies.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 379 entries, 0 to 379  
Data columns (total 11 columns):  
 #   Column           Non-Null Count  Dtype     
 ---  --  
 0   engine_power     379 non-null    int64    
 1   age_in_days     379 non-null    int64    
 2   km               379 non-null    int64    
 3   previous_owners 379 non-null    int64    
 4   price            379 non-null    int64    
 5   model_lounge    379 non-null    uint8    
 6   model_pop        379 non-null    uint8    
 7   model_sport      379 non-null    uint8    
 8   model_star       379 non-null    uint8    
 9   transmission_automatic 379 non-null    uint8    
 10  transmission_manual 379 non-null    uint8    
dtypes: int64(5), uint8(6)  
memory usage: 20.0 KB
```

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 379 entries, 0 to 379  
Data columns (total 9 columns):  
 #   Column           Non-Null Count  Dtype     
 ---  --  
 0   engine_power     379 non-null    int64    
 1   age_in_days     379 non-null    int64    
 2   km               379 non-null    int64    
 3   previous_owners 379 non-null    int64    
 4   price            379 non-null    int64    
 5   model_lounge    379 non-null    int64    
 6   model_pop        379 non-null    int64    
 7   model_sport      379 non-null    int64    
 8   transmission_manual 379 non-null    int64    
dtypes: int64(9)  
memory usage: 29.6 KB
```



2.6

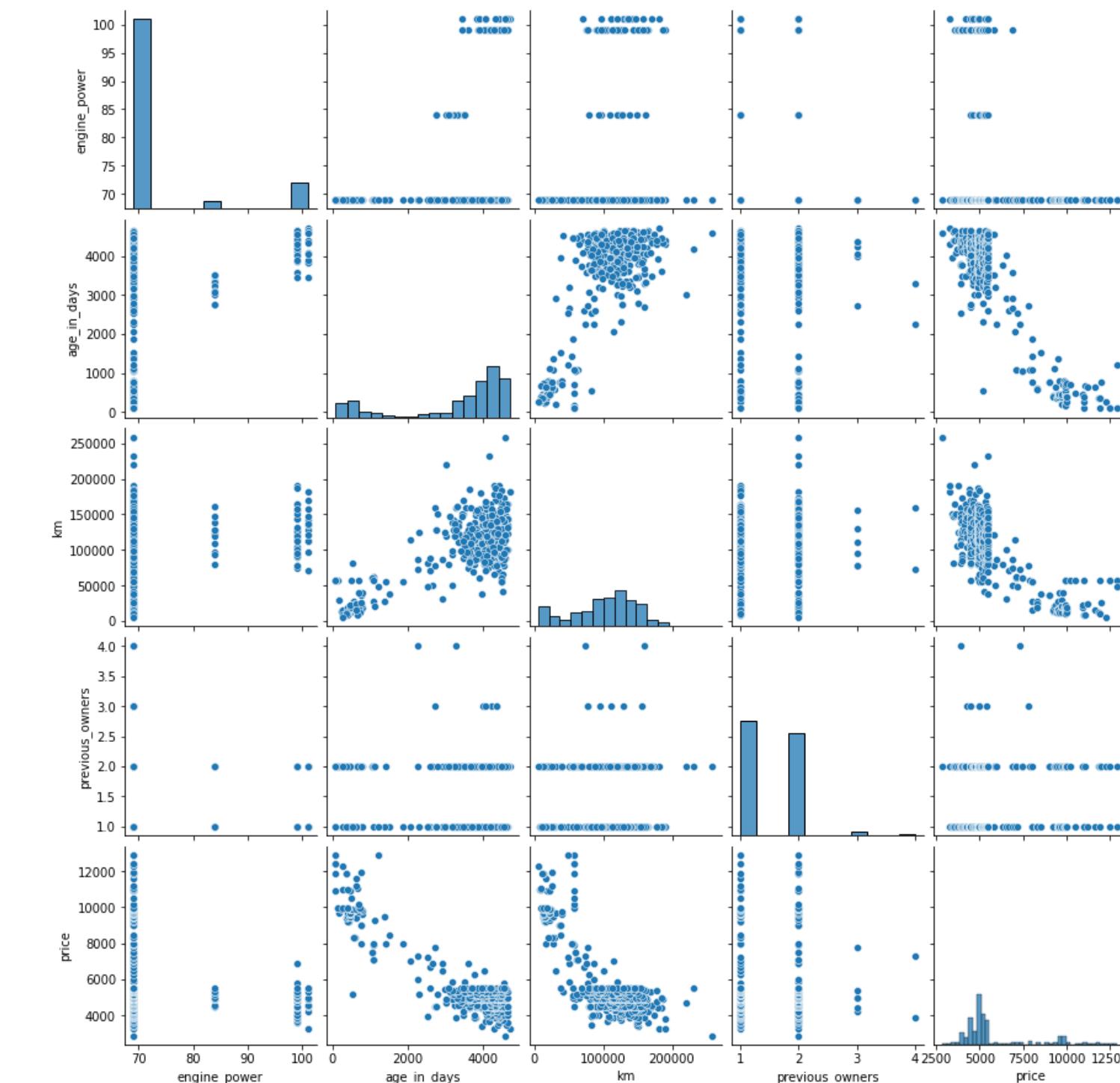
Exploratory Data Analysis

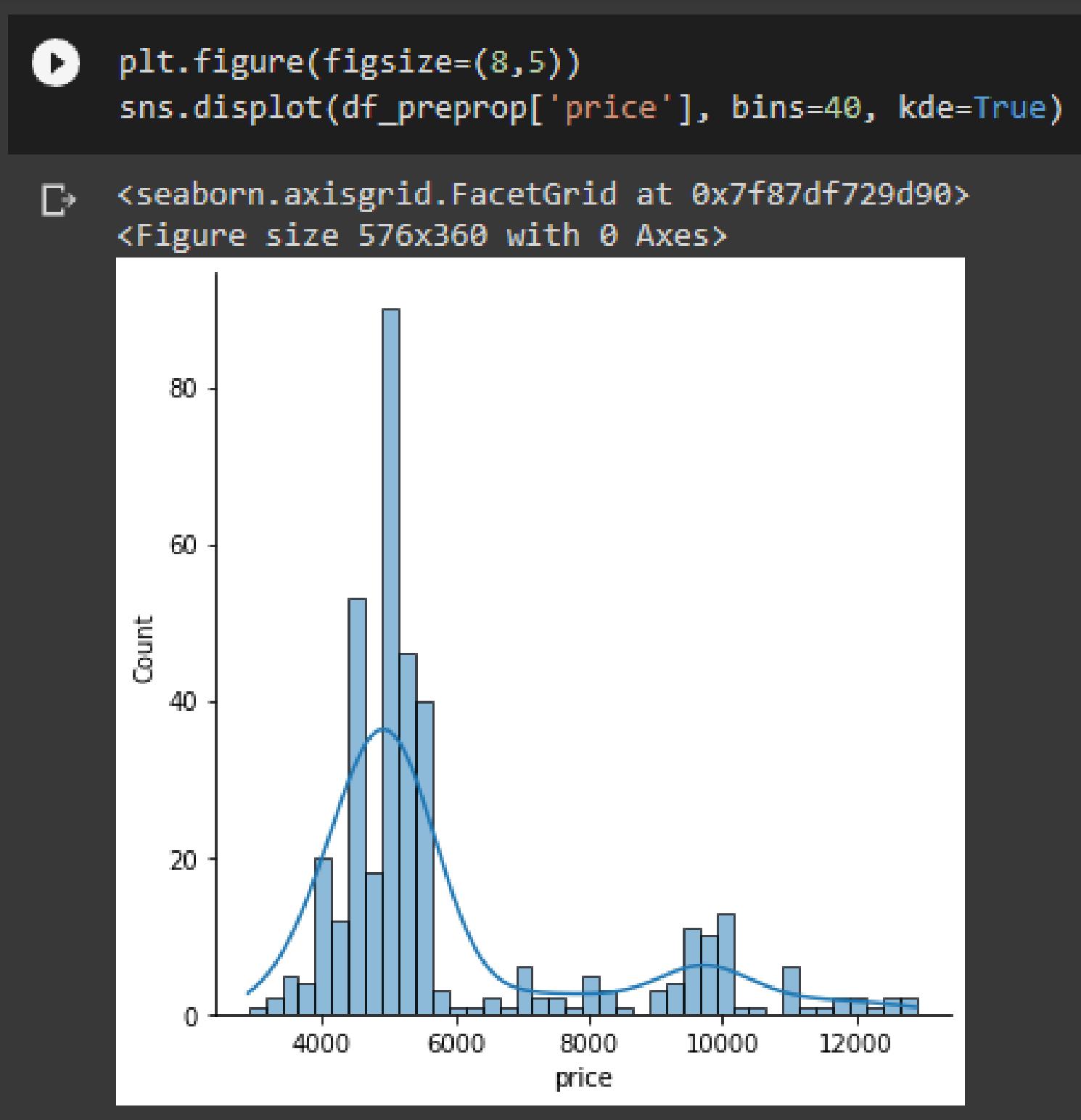
Setelah dilakukan pre-processing akan dilakukan visualisasi data untuk membantu memahami hasil data yang akan digunakan untuk pemodelan.

Dari pairplot di samping, dapat dilihat distribusi atribut tunggal dan hubungan dua atribut data.

```
# Visualisasi pairplot  
sns.pairplot(df_prep)
```

2.6.1 *Pair Plot*





Dari visualisasi tersebut, dapat dilihat distribusi atau sebaran harga jual mobil setelah dilakukan preprocessing. Distribusinya masih memiliki sifat yang sama seperti sebelum di preprocessing yakni right skewed (memiliki lebih banyak data di kiri).

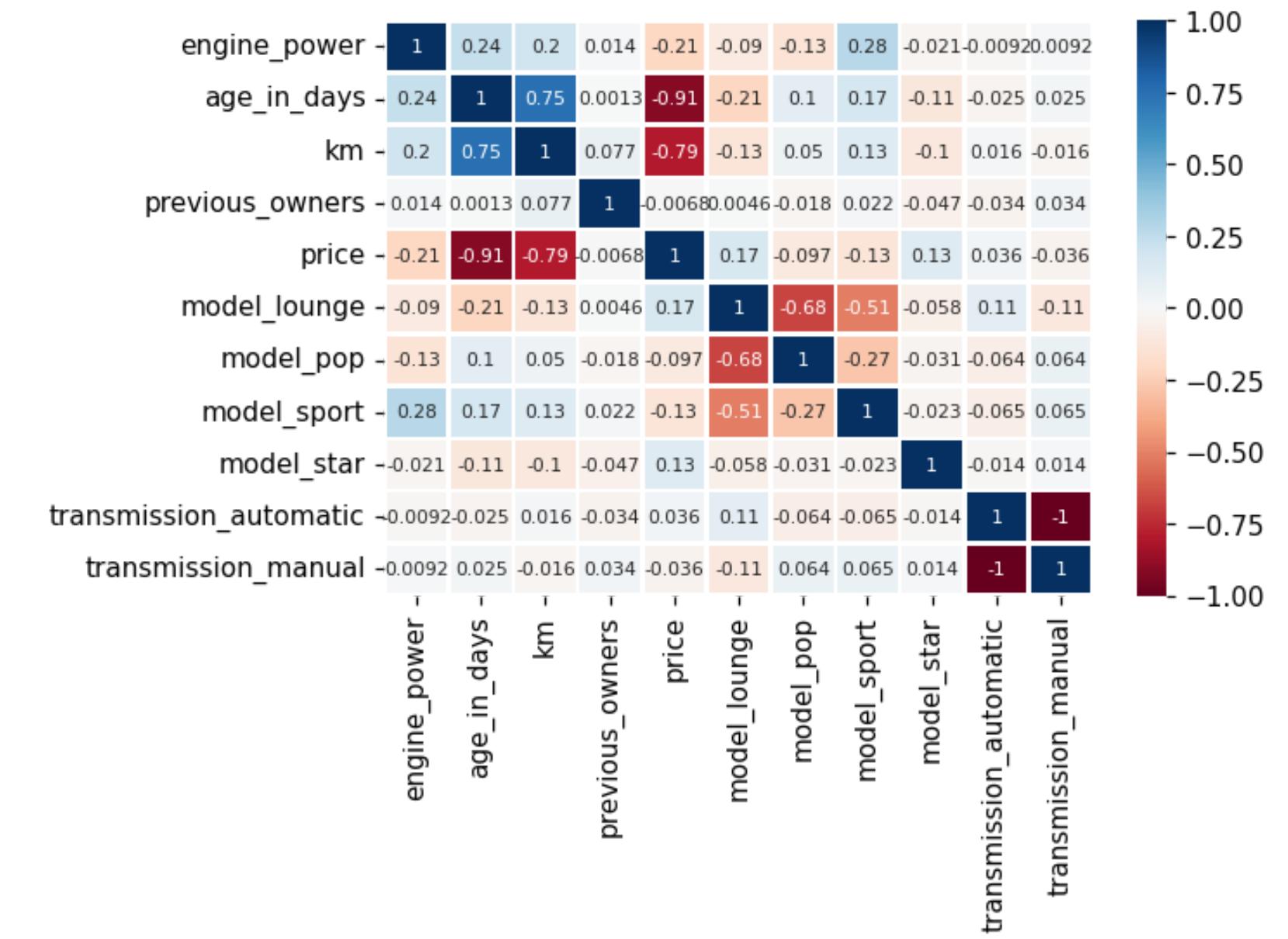


2.6.2 Heatmap Correlation Matrix

```
[ ] abs(df_preprop_dummy.corr()['price']).sort_values()[:-1][:11]
```

	price	age_in_days	km	engine_power	model_lounge	model_sport	model_star	model_pop	transmission_manual	transmission_automatic	previous_owners
price	1.000000										
age_in_days	0.912773	1.000000									
km	0.789244		1.000000								
engine_power	0.205478			1.000000							
model_lounge	0.171268				1.000000						
model_sport	0.129281					1.000000					
model_star	0.125253						1.000000				
model_pop	0.096627							1.000000			
transmission_manual	0.036481								1.000000		
transmission_automatic	0.036481									1.000000	
previous_owners	0.006827										1.000000
Name: price, dtype: float64											

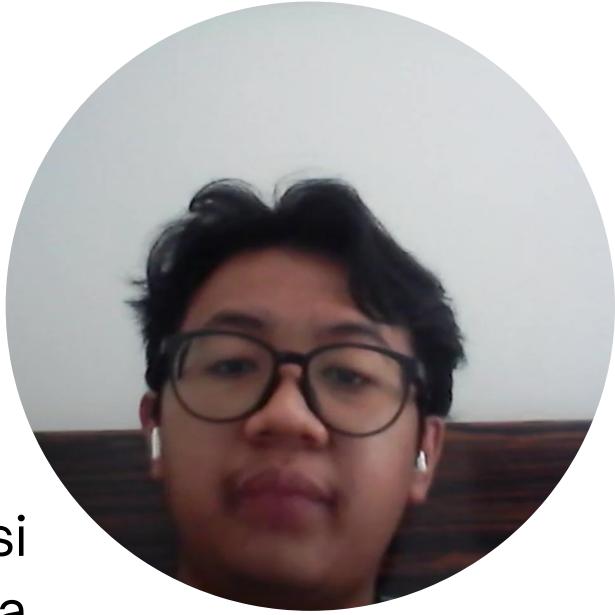
```
#Heatmap setelah di preprocessing (melihat korelasi antar variabel)
plt.figure(dpi=110)
sns.heatmap(df_preprop_dummy.corr(), linewidths=1,cmap="RdBu", annot=True, annot_kws={'size':7})
```





KESIMPULAN

- Terdapat korelasi negatif sebesar (0.91) antara umur mobil bekas dengan harga jual mobil bekas. Semakin muda umur dari mobil bekas, semakin mahal harga jual mobil bekas, begitu juga sebaliknya.
- Terdapat korelasi negatif sebesar (0.79) antara jumlah kilometer yang telah ditempuh dengan harga jual mobil bekas. Semakin rendah jumlah kilometer, semakin mahal harga jual mobil bekas, begitu juga sebaliknya.
- Sedangkan untuk kekuatan mesin, model, transmisi mobil, dan jumlah kepemilikan mobil korelasinya terhadap harga jual mobil kecil, sehingga tidak terlalu berpengaruh pada harga jual mobil.
- Persebaran data akan lebih terpusat untuk harga mobil bekas yang cenderung murah, selain dari faktor kelas ekonomi suatu negara, dikarenakan dalam rentang harga mobil tersebut mayoritas orang-orang akan lebih tertarik melakukan pembelian mobil bekas.
- Kedua variabel kategori (model & transmisi) yang dibentuk dummy variabelnya tidak terlalu mempengaruhi harga jual mobil bekas.
- Banyak variabel yang memiliki nilai korelasi yang kecil, sehingga ketika akan melakukan pemodelan, variabel tersebut berkemungkinan besar tidak signifikan secara statistik (*not statistically significant*)



HASIL

Dari hasil pre-processing, kami akan mengajukan 4 model linier yang dapat digunakan untuk melakukan prediksi terhadap ‘price’ (keterangan tentang model ada di bagian 3). Berikut adalah hipotesis untuk teknik regresi pada proses selanjutnya:

- Hipotesis Model 1

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0 \text{ (model tidak berguna)}$$

$$H_1: \text{Minimal salah satu dari } \beta_j \neq 0; j = 1, 2, 3, 4, 5, 6, 7, 8 \text{ (model berguna)}$$

- Hipotesis Model 2

$$H_0: \beta_1 = \beta_2 = 0 \text{ (model tidak berguna)}$$

$$H_1: \text{Minimal salah satu dari } \beta_j \neq 0; j = 1, 2 \text{ (model berguna)}$$

- Hipotesis Model 3

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0 \text{ (model tidak berguna)}$$

$$H_1: \text{Minimal salah satu dari } \beta_j \neq 0; j = 1, 2, 3, 4, 5, 6, 7, 8 \text{ (model berguna)}$$

- Hipotesis Model 4

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \text{ (model tidak berguna)}$$

$$H_1: \text{Minimal salah satu dari } \beta_j \neq 0; j = 1, 2, 3, 4 \text{ (model berguna)}$$



Modeling



SUMMARY DATA

```
# mengintip isi data  
summary(reg_data)  
head(reg_data)
```

```
engine_power    age_in_days      km      previous_owners  
Min.   : 69.00  Min.   : 91   Min.   : 4981  Min.   :1.000  
1st Qu.: 69.00  1st Qu.:3136  1st Qu.: 77303  1st Qu.:1.000  
Median : 69.00  Median :3927  Median :112000  Median :1.000  
Mean   : 73.03  Mean   :3319  Mean   :102316  Mean   :1.512  
3rd Qu.: 69.00  3rd Qu.:4322  3rd Qu.:135000  3rd Qu.:2.000  
Max.   :101.00  Max.   :4719  Max.   :259000  Max.   :4.000  
      price      model_lounge      model_pop      model_sport  
Min.   :2890  Min.   :0.0000  Min.   :0.0000  Min.   :0.0000  
1st Qu.:4600  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  
Median :5000  Median :1.0000  Median :0.0000  Median :0.0000  
Mean   :5855  Mean   :0.5594  Mean   :0.2665  Mean   :0.1715  
3rd Qu.:5500  3rd Qu.:1.0000  3rd Qu.:1.0000  3rd Qu.:0.0000  
Max.   :12900  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  
transmission_manual  
Min.   :0.000  
1st Qu.:1.000  
Median :1.000  
Mean   :0.934  
3rd Qu.:1.000  
Max.   :1.000  
A data.frame: 6 x 9
```

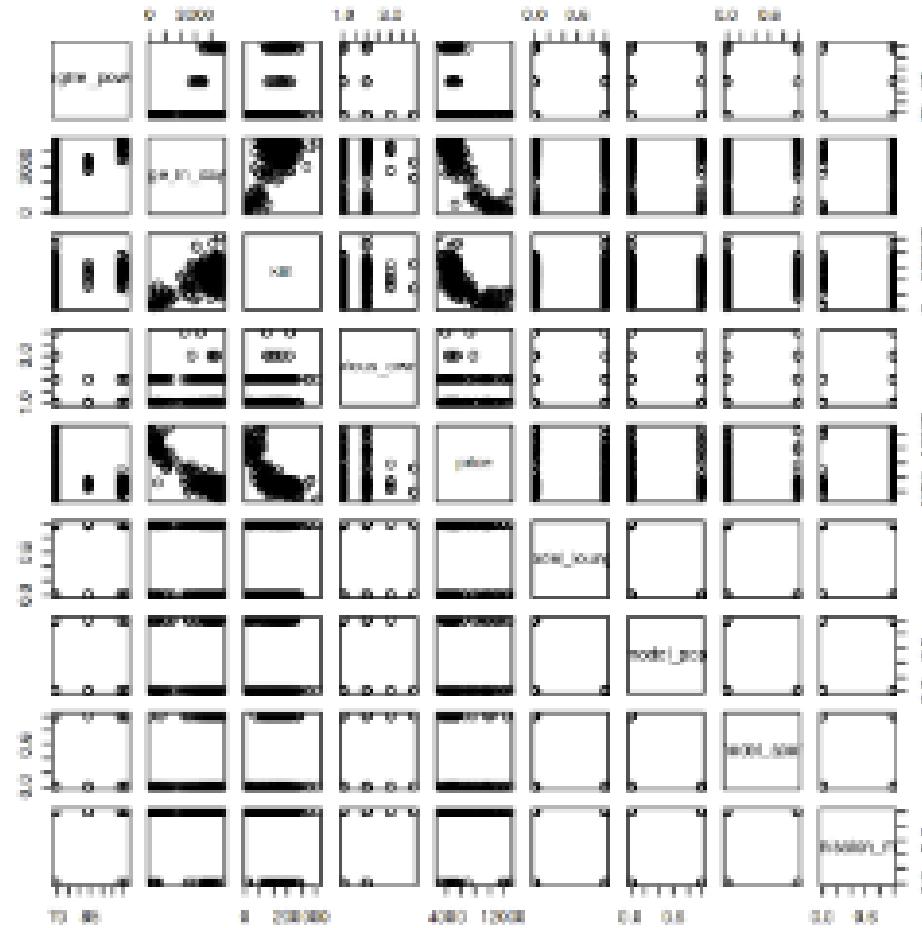
	engine_power	age_in_days	km	previous_owners	price	model_lounge	model_pop	model_sport	transmission_manual
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	69	4474	56779	2	4490	0	1	0	1
2	69	2708	160000	1	4500	1	0	0	1
3	69	3470	170000	2	4500	1	0	0	0
4	69	3288	132000	2	4700	0	0	1	1
5	69	3712	124490	2	4790	0	0	1	1
6	69	3684	91000	1	4900	1	0	0	1

KORELASI VARIABEL

```
# melihat korelasi  
cor(reg_data)  
plot(reg_data)
```

A matrix: 9 x 9 of type dbl

	engine_power	age_in_days	km	previous_owners	price	model_lounge	model_pop	model_sport	transmission_manual
engine_power	1.000000000	0.240853589	0.20116751	0.014113003	-0.205477816	-0.089758218	-0.13311525	0.27717108	0.009221188
age_in_days	0.240853589	1.000000000	0.75274126	0.001349008	-0.912773267	-0.212085412	0.10424687	0.17197716	0.025354108
km	0.201167511	0.752741263	1.00000000	0.077022105	-0.789243558	-0.131994176	0.04951047	0.12981717	-0.015629650
previous_owners	0.014113003	0.001349008	0.07702211	1.000000000	-0.006826529	0.004585495	-0.01812025	0.02161629	0.034130039
price	-0.205477816	-0.912773267	-0.78924356	-0.006826529	1.000000000	0.171267755	-0.09662679	-0.12928133	-0.036480997
model_lounge	-0.089758218	-0.212085412	-0.13199418	0.004585495	0.171267755	1.000000000	-0.679122288	1.000000000	-0.107395063
model_pop	-0.133115253	0.104246874	0.04951047	-0.018120246	-0.096626791	-0.679122288	1.000000000	-0.27423959	0.064008323
model_sport	0.277171082	0.171977162	0.12981717	0.021616288	-0.129281330	-0.512627086	-0.27423959	1.000000000	0.064509839
transmission_manual	0.009221188	0.025354108	-0.01562965	0.034130039	-0.036480997	-0.107395063	0.06400832	0.06450984	1.000000000



MODEL 1

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_{5A} + \beta_6 x_{5B} + \beta_7 x_{5C} + \beta_8 x_{6A}$$

```
Call:  
lm(formula = price ~ engine_power + age_in_days + km + previous_owners +  
model_lounge + model_pop + model_sport + transmission_manual,  
data = reg_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-4834.2	-483.4	-38.7	391.3	4237.7

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	1.133e+04	8.741e+02	12.962	< 2e-16 ***		
engine_power	2.910e+00	4.401e+00	0.661	0.509		
age_in_days	-1.090e+00	4.508e-02	-24.178	< 2e-16 ***		
km	-1.073e-02	1.340e-03	-8.006	1.55e-14 ***		
previous_owners	5.179e+01	7.410e+01	0.699	0.485		
model_lounge	-8.907e+02	8.078e+02	-1.103	0.271		
model_pop	-8.710e+02	8.114e+02	-1.074	0.284		
model_sport	-7.350e+02	8.149e+02	-0.902	0.368		
transmission_manual	-2.078e+02	1.669e+02	-1.245	0.214		

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

Residual standard error: 800.3 on 370 degrees of freedom

Multiple R-squared: 0.8594, Adjusted R-squared: 0.8563

F-statistic: 282.6 on 8 and 370 DF, p-value: < 2.2e-16

```
vif(model1)
```

engine_power: 1.14215012141284 age_in_days: 2.45189045221362 km: 2.35343895409535 previous_owners: 1.01772913143801 model_lounge: 95.1727191411132 model_pop: 76.1516414950934 model_sport: 55.8436357607877 transmission_manual: 1.01597019074067

Dapat dilihat bahwa model_lounge, model_pop, dan model_sport memiliki nilai VIF > 10, sehingga kami simpulkan terdapat multikolinearitas pada model ini. Dengan demikian, model ini tidak akan digunakan lebih lanjut.

MODEL 2

Kami mencoba melakukan metode *Stepwise Regression* dari variabel-variabel pada model 1. Prosesnya sebagai berikut.

```
stepwise<-lm(price~, reg_data)  
step(stepwise)
```

```
Start: AIC=5076.08  
price ~ engine_power + age_in_days + km + previous_owners + model_lounge + model_pop + transmission_manual  
                                         Df Sum of Sq   RSS   AIC  
- transmission_manual  1    945170 238969379 5071.3  
- model_pop            1   1179377 239203585 5071.7  
<none>                           238024208 5071.8  
- model_lounge          1   1670766 239694974 5072.4  
- km                   1   40915638 278939846 5129.9  
- age_in_days           1   386310393 624334681 5435.3  
  
Step: AIC=5071.78  
price ~ age_in_days + km + model_lounge + model_pop + transmission_manual  
                                         Df Sum of Sq   RSS   AIC  
- engine_power          1    280038 237240869 5074.5  
- previous_owners        1    312795 237273626 5074.6  
- model_sport            1    520942 237481773 5074.9  
- model_pop              1    738096 237698926 5075.3  
- model_lounge           1    778708 237739539 5075.3  
- transmission_manual    1    992491 237953322 5075.7  
<none>                           236960831 5076.1  
- km                   1    41051634 278012465 5134.6  
- age_in_days            1    374397449 611358280 5433.3  
  
Step: AIC=5071.28  
price ~ age_in_days + km + model_lounge + model_pop  
                                         Df Sum of Sq   RSS   AIC  
- model_pop              1    1152029 240121408 5071.1  
<none>                           238969379 5071.3  
- model_lounge           1    1475720 240445099 5071.6  
- km                   1    40406877 279376256 5128.5  
- age_in_days            1    388313248 627282619 5435.0  
  
Step: AIC=5074.53  
price ~ age_in_days + km + previous_owners + model_lounge + model_pop + model_sport + transmission_manual  
                                         Df Sum of Sq   RSS   AIC  
- previous_owners        1    315894 237556763 5073.0  
- model_sport             1    502740 237743608 5073.3  
- model_pop               1    754699 237995568 5073.7  
- model_lounge            1    784610 238025478 5073.8  
- transmission_manual    1    995715 238236583 5074.1  
<none>                           237240869 5074.5  
- km                   1    40899394 278140263 5132.8  
- age_in_days            1    377873353 615114222 5433.6  
  
Step: AIC=5071.11  
price ~ age_in_days + km + model_lounge  
                                         Df Sum of Sq   RSS   AIC  
- model_lounge           1    441278 240562686 5069.8  
<none>                           240121408 5071.1  
- km                   1    40143113 280264520 5127.7  
- age_in_days            1    387585184 627706592 5433.3  
  
Step: AIC=5069.8  
price ~ age_in_days + km  
                                         Df Sum of Sq   RSS   AIC  
<none>                           240562686 5069.8  
- km                   1    40580692 281143378 5126.9  
- age_in_days            1    394863457 635426142 5435.9  
  
Call:  
lm(formula = price ~ age_in_days + km, data = reg_data)  
  
Coefficients:  
(Intercept)  age_in_days      km  
1.054e+04 -1.086e+00 -1.056e-02
```



MODEL 2 (Lanjutan)

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Call:

```
lm(formula = price ~ age_in_days + km, data = reg_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-3878.3	-414.3	-65.7	401.8	4188.8

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	1.054e+04	1.073e+02	98.241	<2e-16 ***		
age_in_days	-1.086e+00	4.371e-02	-24.843	<2e-16 ***		
km	-1.056e-02	1.326e-03	-7.964	2e-14 ***		

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

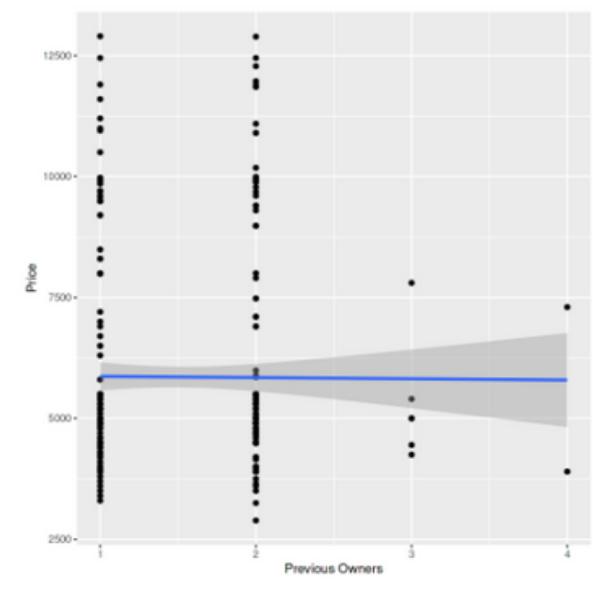
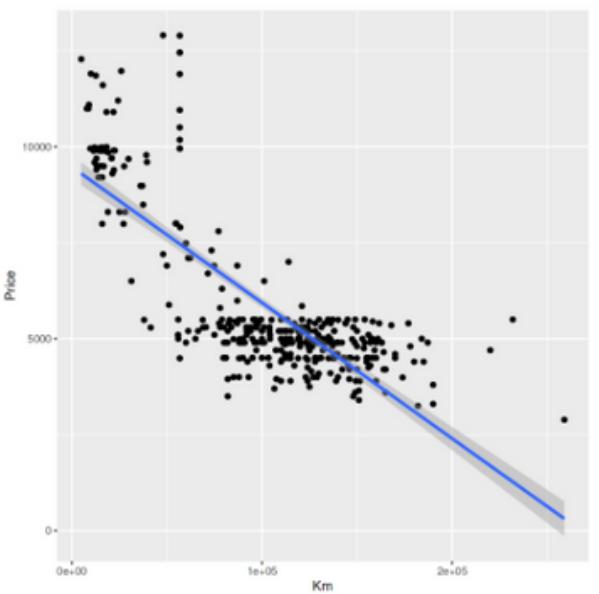
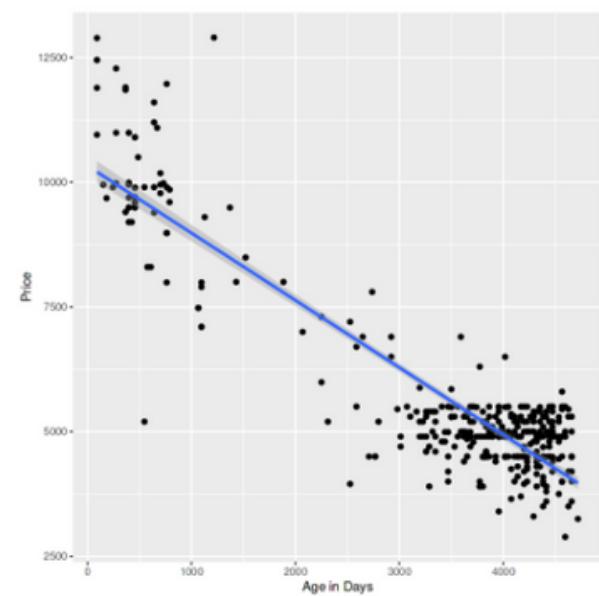
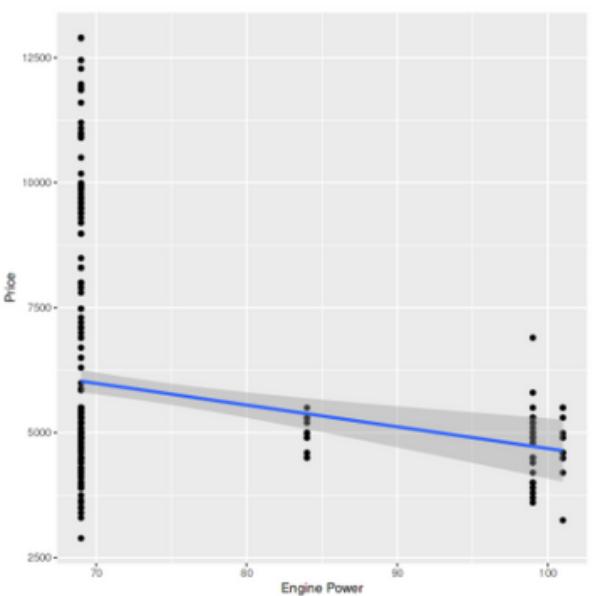
Residual standard error: 799.9 on 376 degrees of freedom

Multiple R-squared: 0.8572, Adjusted R-squared: 0.8565

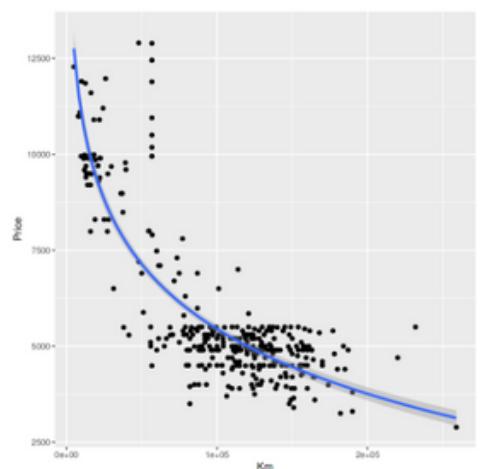
F-statistic: 1129 on 2 and 376 DF, p-value: < 2.2e-16

Model ini memiliki R² dan Ra² yang bernilai tinggi, sehingga kami menganggap model cukup bagus.

MODEL 3



Pada plot awal untuk regresi standar, plot (3) terlihat masih kurang fitted terhadap sebaran data. Selanjutnya plot (3) akan diubah menjadi regresi dengan logaritma, terlihat data telah cukup fitted terhadap model



MODEL 3 (Lanjutan)

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \log(x_3) + \beta_4 x_4 + \beta_5 x_{5A} + \beta_6 x_{5B} + \beta_7 x_{5C} + \beta_8 x_{6A}$$

Call:

```
lm(formula = price ~ engine_power + age_in_days + km + previous_owners +
  model_lounge + model_pop + model_sport + transmission_manual,
  data = reg_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-3459.3	-446.7	-56.0	375.9	4574.2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.936e+04	1.228e+03	15.765	<2e-16 ***
engine_power	1.572e+00	4.311e+00	0.365	0.716
age_in_days	-9.694e-01	5.204e-02	-18.628	<2e-16 ***
km	-9.841e+02	9.973e+01	-9.866	<2e-16 ***
previous_owners	7.212e+01	7.278e+01	0.991	0.322
model_lounge	-1.289e+02	7.978e+02	-0.162	0.872
model_pop	-1.281e+02	8.009e+02	-0.150	0.881
model_sport	5.655e+01	8.051e+02	0.070	0.944
transmission_manual	-1.695e+02	1.634e+02	-1.038	0.300

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 784.1 on 370 degrees of freedom

Multiple R-squared: 0.865, Adjusted R-squared: 0.8621

F-statistic: 296.3 on 8 and 370 DF, p-value: < 2.2e-16

```
vif(model3)
```

engine_power: 1.14144449787587 age_in_days: 3.40408718718144 km: 3.34977612383972 previous_owners:

1.0226992192097 model_lounge: 96.7161340456702 model_pop: 77.2827439649471 model_sport: 56.7682656936969

transmission_manual: 1.01361229356681

Dapat dilihat bahwa model_lounge, model_pop, dan model_sport juga memiliki nilai VIF>10, sehingga kami simpulkan terdapat multikolinearitas. Dengan demikian, model ini tidak akan digunakan lebih lanjut.

MODEL 4

Kami mencoba melakukan metode *Stepwise Regression* dari variabel-variabel pada model 3. Prosesnya sebagai berikut.

```
stepwise<-lm(price~., reg_data)
step(stepwise)
```

Start: AIC=5060.61
price ~ engine_power + age_in_days + km + previous_owners + model_lounge +
model_pop + model_sport + transmission_manual

	Df	Sum of Sq	RSS	AIC
- model_sport	1	3033	227486047	5058.6
- model_pop	1	13819	227496833	5058.6
- model_lounge	1	16060	227499075	5058.6
- engine_power	1	81723	227564737	5058.8
- previous_owners	1	603723	228086737	5059.6
- transmission_manual	1	661984	228144998	5059.7
<none>			227483014	5060.6
- km	1	50529451	278012465	5134.6
- age_in_days	1	213344402	448827416	5309.4

Step: AIC=5058.62
price ~ engine_power + age_in_days + km + previous_owners + model_lounge +
model_pop + transmission_manual

	Df	Sum of Sq	RSS	AIC
- engine_power	1	82606	227568654	5058.6
- previous_owners	1	606966	228093013	5057.6
- transmission_manual	1	662484	228148531	5057.7
- model_pop	1	1142114	228628161	5058.5
<none>			227486047	5058.6
- model_lounge	1	1564391	229050438	5059.2
- km	1	51292077	278778125	5133.7
- age_in_days	1	213687958	441174005	5307.6

Step: AIC=5056.76
price ~ age_in_days + km + previous_owners + model_lounge + model_pop +
transmission_manual

	Df	Sum of Sq	RSS	AIC
- previous_owners	1	611727	228180381	5055.8
- transmission_manual	1	664763	228233417	5055.9
<none>			227568654	5056.8
- model_pop	1	1482145	228970799	5057.1
- model_lounge	1	1795254	229363907	5057.7
- km	1	51322121	278890775	5131.8
- age_in_days	1	216338021	443906675	5308.0



Step: AIC=5055.77
price ~ age_in_days + km + model_lounge + model_pop + transmission_manual

	Df	Sum of Sq	RSS	AIC
- transmission_manual	1	618770	228799150	5054.8
<none>			228180381	5055.8
- model_pop	1	1430852	229611232	5056.1
- model_lounge	1	1811816	229992197	5056.8
- km	1	50759466	278939846	5129.9
- age_in_days	1	221934198	450114571	5311.3

Step: AIC=5054.8
price ~ age_in_days + km + model_lounge + model_pop

	Df	Sum of Sq	RSS	AIC
<none>			228799150	5054.8

Step: AIC=5054.8
price ~ age_in_days + km + model_lounge + model_pop

	Df	Sum of Sq	RSS	AIC
- model_pop	1	1487268	238206419	5055.1
- model_lounge	1	1647926	238447876	5055.5
- km	1	50577105	279376256	5128.5
- age_in_days	1	222357936	451157086	5310.1

Call:
lm(formula = price ~ age_in_days + km + model_lounge + model_pop,
data = reg_data)

Coefficients:

(Intercept)	age_in_days	km	model_lounge	model_pop
19314.7484	-0.9737	-887.9285	-184.4377	-188.1653

MODEL 4 (Lanjutan)

$$\tilde{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3A} + \beta_4 x_{3B}$$

```
model4 <- lm(formula = price ~ age_in_days + km + model_lounge + model_pop, data = reg_data)
summary(model4)

Call:
lm(formula = price ~ age_in_days + km + model_lounge + model_pop,
    data = reg_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3346.5 -434.4   -51.6   374.0  4529.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.931e+04 9.798e+02 19.713 <2e-16 ***
age_in_days -9.737e-01 5.107e-02 -19.065 <2e-16 ***
km          -8.879e+02 9.765e+01 -9.093 <2e-16 ***
model_lounge -1.844e+02 1.124e+02 -1.641 0.102    
model_pop    -1.882e+02 1.241e+02 -1.517 0.130    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 782.2 on 374 degrees of freedom
Multiple R-squared:  0.8642,    Adjusted R-squared:  0.8628 
F-statistic: 595.1 on 4 and 374 DF,  p-value: < 2.2e-16
```

Lalu perhatikan bahwa uji hipotesis 2,yaitu uji F dengan tingkat signifikansi 0.05, diperoleh p-value<2.2e-16<0.05, maka H₀ ditolak , cukup untuk $\beta_j = 0, j=1,\dots,8$ (model berguna).

Model ini kurang baik dikarenakan Pr(>|t|) dari beberapa variabel cukup besar lebih dari 0.05 yang menyatakan bahwa variabel-variabel tersebut kurang signifikan. Jika dibandingkan dengan model 2, dari hasil summary pada kedua model, diperoleh bahwa model 2 memiliki R²=0.8572 dan Ra²=0.8565, sedangkan model 4 memiliki R²=0.8642 dan Ra²=0.8628





PENGOLAHAN DATA DAN ANALISIS DATA



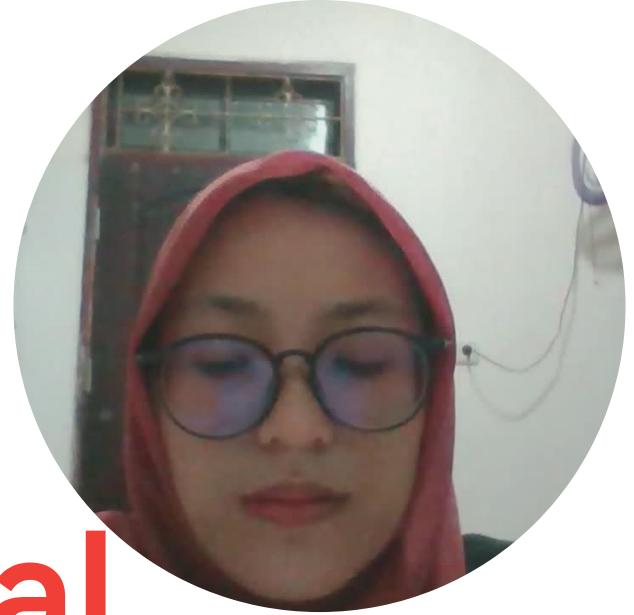
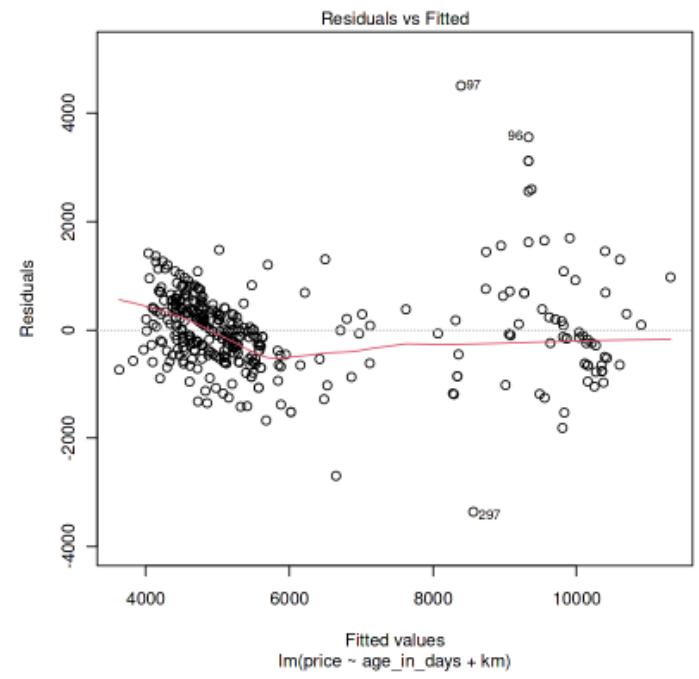
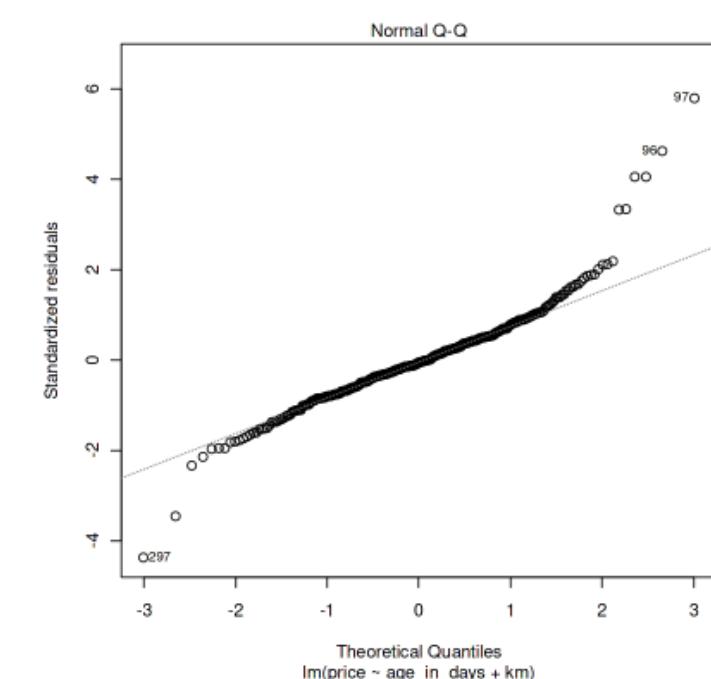
Final Model

Dari keempat model dapat disimpulkan bahwa model 2 adalah model terbaik dan akan digunakan. Model matematikanya dapat ditulis sebagai berikut

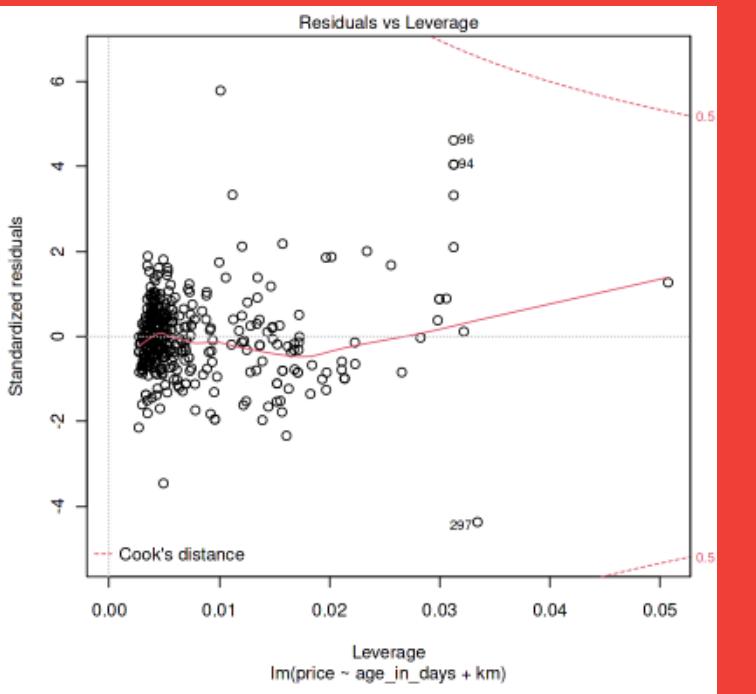
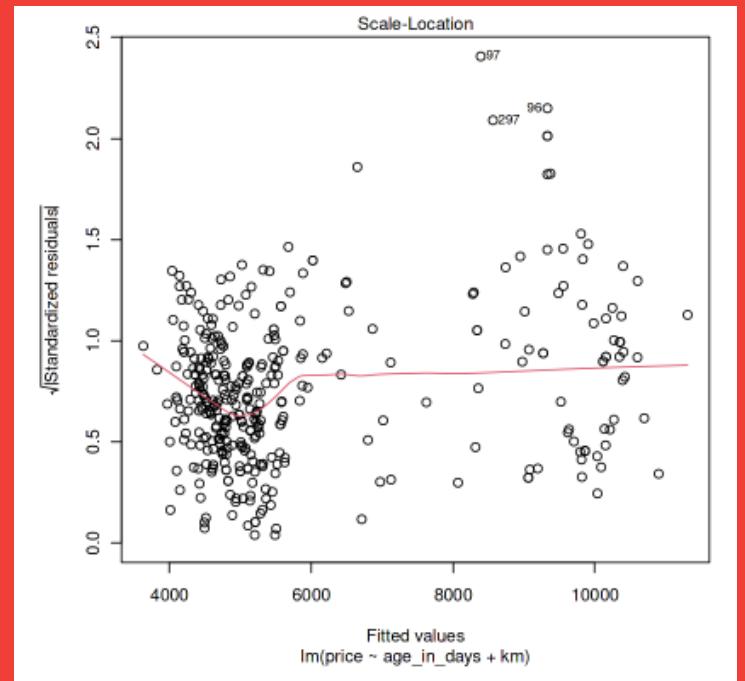
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Analisis Residual

Pada bagian ini, akan dilihat apakah asumsi normalitas pada residual, linearitas, dan heteroskedastisitas terpenuhi pada model atau tidak



Analisis Residual (Lanjutan)



Untuk menguji asumsi pada model, kami melakukan visualisasi dalam bentuk plot untuk melihat apakah asumsi terpenuhi atau tidak. Dari plot Normal Q-Q, dapat dilihat bahwa data berada disekitar garis lurus, maka residualnya terdistribusi secara normal, sehingga dapat dikatakan model regresi memenuhi asumsi Normalitas. Dari plot Scale Location, dapat dilihat bahwa variasinya tersebar luas .

Multikolinearitas

multikolinearitas pada model ini.

```
vif(model12)
```

```
age_in_days: 2.30744066903809 km: 2.30744066903809
```

Dapat dilihat bahwa nilai VIF pada variabel age_in_days = $2.30744066903809 < 10$ dan nilai VIF pada variabel km = $2.30744066903809 < 10$. Dengan demikian, dapat dikatakan bahwa tidak terdapat multikolinearitas pada model ini.



Analisis Model

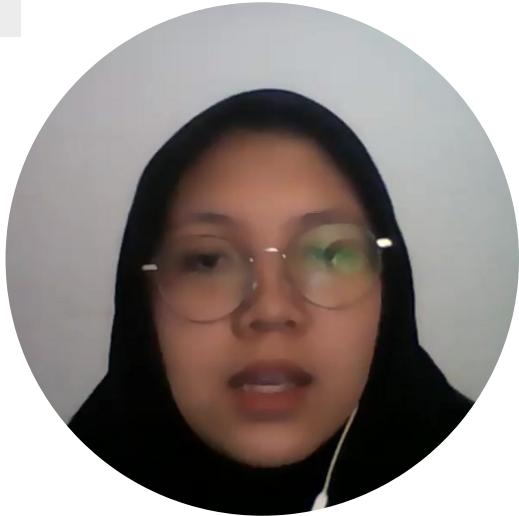
Diperoleh summary dari model terbaik ini, yaitu

```
Call:  
lm(formula = price ~ age_in_days + km, data = reg_data)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-3878.3 -414.3   -65.7   401.8  4188.8  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 1.054e+04 1.073e+02  98.241 <2e-16 ***  
age_in_days -1.086e+00 4.371e-02 -24.843 <2e-16 ***  
km          -1.056e-02 1.326e-03 -7.964 2e-14 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 799.9 on 376 degrees of freedom  
Multiple R-squared:  0.8572,    Adjusted R-squared:  0.8565  
F-statistic: 1129 on 2 and 376 DF,  p-value: < 2.2e-16
```

$$Y = 1.054e+04 + (-1.086e+00)X_1 + (-1.056e-02)X_2 + \varepsilon; \varepsilon \sim NIID(0, \sigma^2)$$

Model yang dibuat memiliki nilai adjusted R-squared di sekitar 0.8565 dan R-squared sekitar 0.8572. Nilai-nilai tersebut sudah berada di atas 0.75 yang menjadi batas bahwa adanya korelasi yang kuat, sehingga model yang dibuat dapat diterima dan dapat digunakan dengan cukup baik.

Berdasarkan hipotesis bagian 2, bahwa dengan menggunakan uji hipotesis pada bagian 2, yaitu uji F, dengan tingkat signifikansi 0. 05, diperoleh p-value<2.2e-16<0.05, maka H0 ditolak. Dengan demikian, dengan 0.05, terdapat cukup bukti untuk mengatakan bahwa minimal salah satu dari $\beta_j \neq 0 ; j = 1, 2$. Maka, dapat dikatakan bahwa model ini berguna. Untuk pengujian hipotesis pada model lainnya yang ditulis pada bagian dua menyatakan bahwa keempat H0 ditolak.



Analisis Model (Lanjutan)

Kami telah menguji seluruh asumsi yang dibutuhkan, model ini telah memenuhi asumsi normalitas, heteroskedastisitas, linearitas, dan multikolinearitas. Kami juga telah menguji model dengan menggunakan uji-F disimpulkan model berguna, uji-t disimpulkan kedua variabel prediktor tersebut signifikan, menghitung R² diperoleh nilai yang baik dan Ra² diperoleh nilai yang baik.

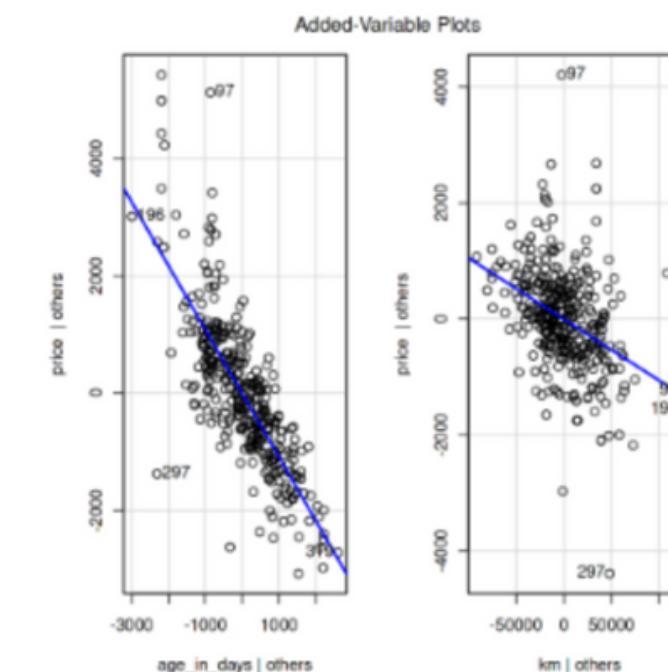
Interval kepercayaan dari model ini adalah sebagai berikut.

```
confint(mode12)
```

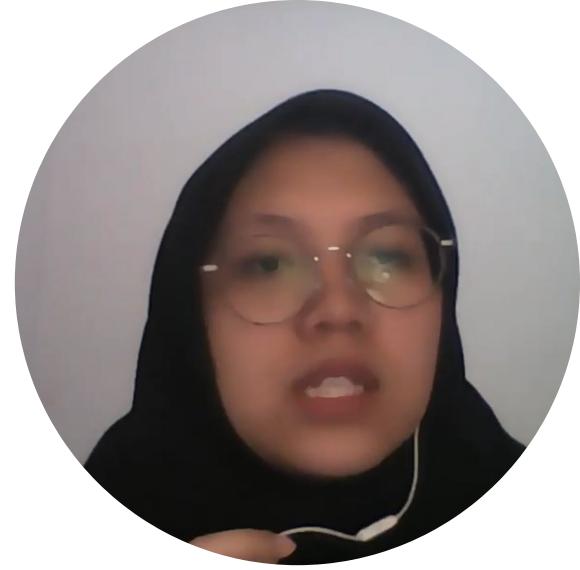
	2.5 %	97.5 %
(Intercept)	1.032848e+04	1.075038e+04
age_in_days	-1.171726e+00	-9.998485e-01
km	-1.317063e-02	-7.954888e-03

selanjutnya kami akan melakukan analisis terhadap korelasi dari masing-masing variabel prediktor terhadap variabel respon. Dari model yang dibuat, kita dapat mengetahui korelasi dari nilai estimated parameter setiap variabel. Jika *estimated parameter* bernilai positif, maka hubungannya adalah korelasi positif dan sebaliknya.

- age_in_days dengan price berkorelasi negatif. Hal ini berarti makin muda umur dari mobil bekas, makin mahal harga jual mobil bekas, begitu juga sebaliknya.
- km dengan price berkorelasi negatif. Hal ini berarti makin rendah jumlah kilometer, makin mahal harga jual mobil bekas, begitu juga sebaliknya.



KESIMPULAN



Hasil analisis yang dilakukan dapat menjawab pertanyaan variabel yang paling mempengaruhi harga jual mobil bekas, yaitu umur dari mobil dalam hari dan jarak kilometer yang telah ditempuh mobil. Makin muda umur dari mobil bekas, makin mahal harga jual mobil bekas, begitu juga sebaliknya. Makin rendah jumlah kilometer, makin mahal harga jual mobil bekas, begitu juga sebaliknya. Oleh karena itu, disarankan apabila calon pembeli ingin membeli mobil bekas dengan harga termurah, maka hal yang dapat dijadikan pertimbangan pertama kali adalah umur mobil bekas dan jarak kilometer yang telah ditempuh mobil bekas. Dari hasil analisis, variabel-variabel lainnya memang tidak terlalu berpengaruh terhadap harga mobil bekas, namun sebenarnya variabel-variabel tersebut tetap ada pengaruhnya, walaupun sedikit, sehingga setelah mempertimbangkan umur mobil dan jarak kilometer yang telah ditempuh mobil bekas, calon pembeli dapat mempertimbangkan variabel-variabel lain ketika ingin membeli sebuah mobil bekas dengan harga termurah.



TERIMA KASIH