**LAPORAN**

**MEMPREDIKSI HARGA RUMAH DENGAN MULTIPLE LINEAR**

**REGRESSION DI PYTHON BESERTA VISUALISASINYA**

**MENGGUNAKAN GOOGLE STUDIO**



**Oleh :**

**Khusnul Khotimah (192400024)**

**PROGRAM STUDI STATISTIKA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS PGRI ADI BUANA SURABAYA**

**2021**

**DESKRIPSI DATA**

Data yang saya ambil adalah data sekunder. Dari data tersebut akan saya gunakan untuk memprediksi harga rumah. Sampel yang dipakai adalah 551, dengan 7 variabel. Variabel independennya adalah variabel bedrooms, bathrooms, floors, grade, yr\_built, dan yr\_renovated. Sedangkan variabel dependennnya adalah variabel price.

|  |  |  |
| --- | --- | --- |
| **No.** | **Variabel** | **Keterangan** |
| 1. | Bedrooms (X1) | Banyak kamar tidur |
| 2. | Bathrooms (X2) | Banyak kamar mandi |
| 3. | Floors (X3) | Banyak lantai |
| 4. | Grade (X4) | Peringkat |
| 5. | Yr\_built (X5) | Tahun pembuatan rumah |
| 6. | Yr\_renovated (X6) | Tahun renovasi rumah |
| 7. | Price (Y) | Harga rumah |

**PREPROCESSING DATA**

1. Langkah pertama dimulai dengan menyalakan library yang diperlukan yaitu pandas, numpy, matplotlib, dan lainnya seperti berikut ini.

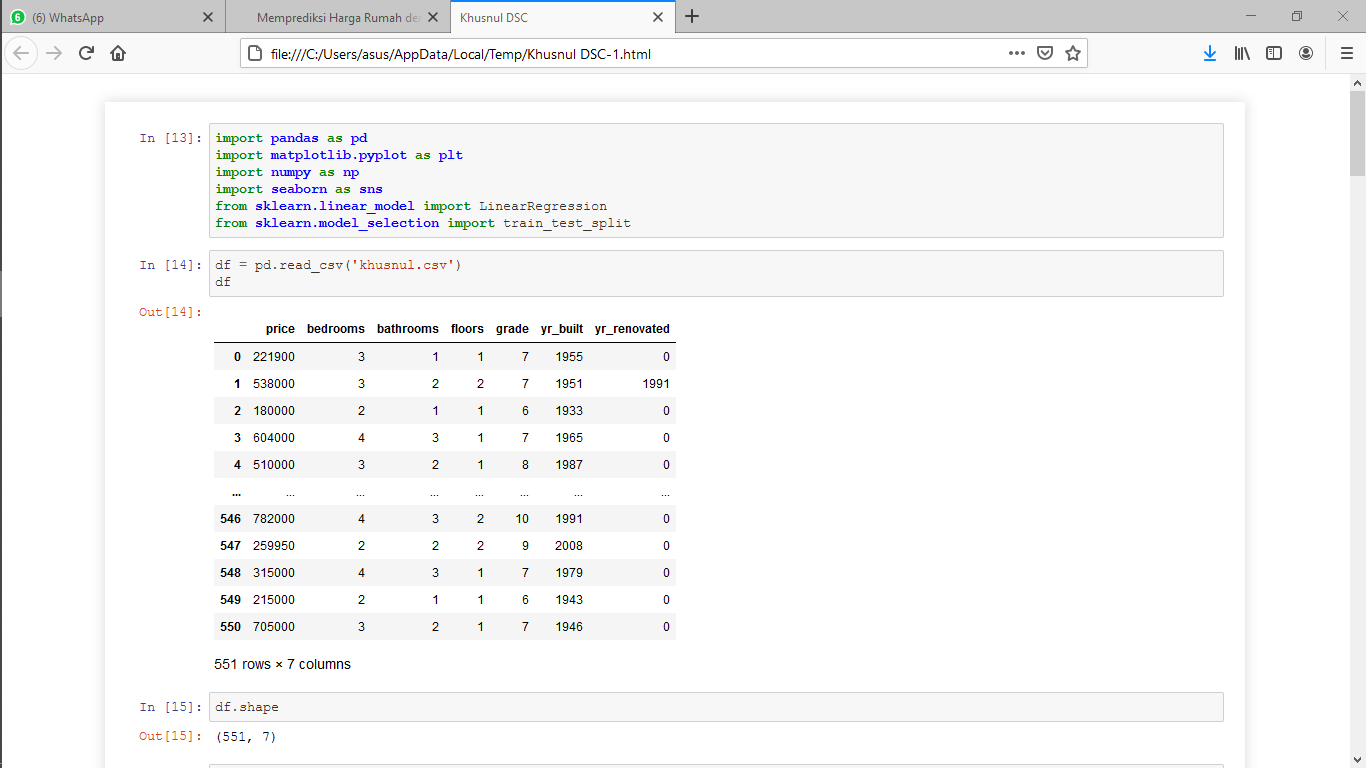
import pandas as pd  
import matplotlib.pyplot as plt  
import numpy as np  
import seaborn as sns  
from sklearn.linear\_model import LinearRegression  
from sklearn.model\_selection import train\_test\_split

1. Selanjutnya, mengimpor dataset yang akan digunakan. Dengan syntax :

df = pd.read\_csv('khusnul.csv')

df

Dengan hasil output sebagai berikut:



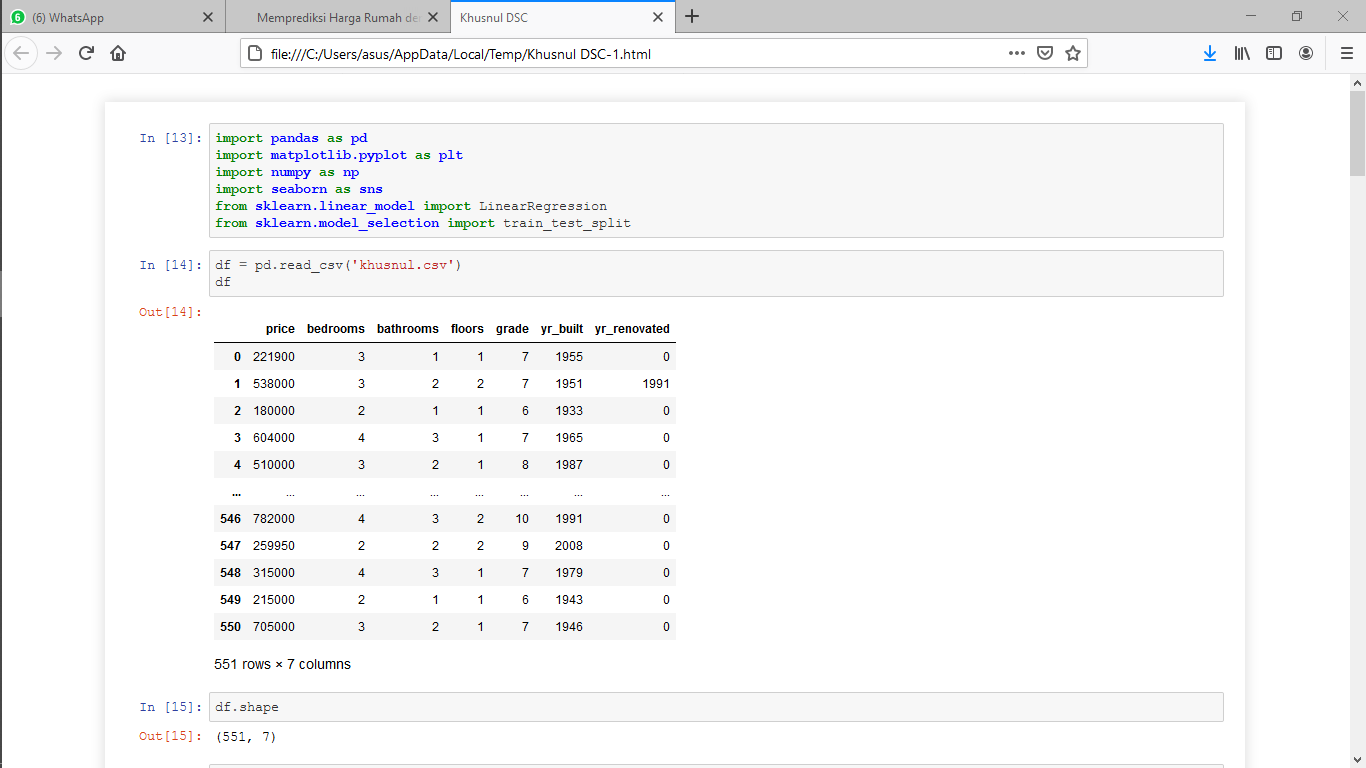
Data yang saya pakai saya namakan “**df**” dengan data asli yang diambil berbentuk file csv dengan nama “khusnul.csv”, lalu run. Kemudian untuk melihat datanya bisa memanggilnya dengan nama yang sudah dibuat tadi yaitu **df**.

Variabel yang digunakan yaitu price, bedrooms, bathrooms, floors, grade, yr\_built, dan yr\_renovated, dengan data sebanyak 551.

1. Untuk melihat jumlah kolom dan baris datanya, bisa menggunakan :

df.shape

Output :

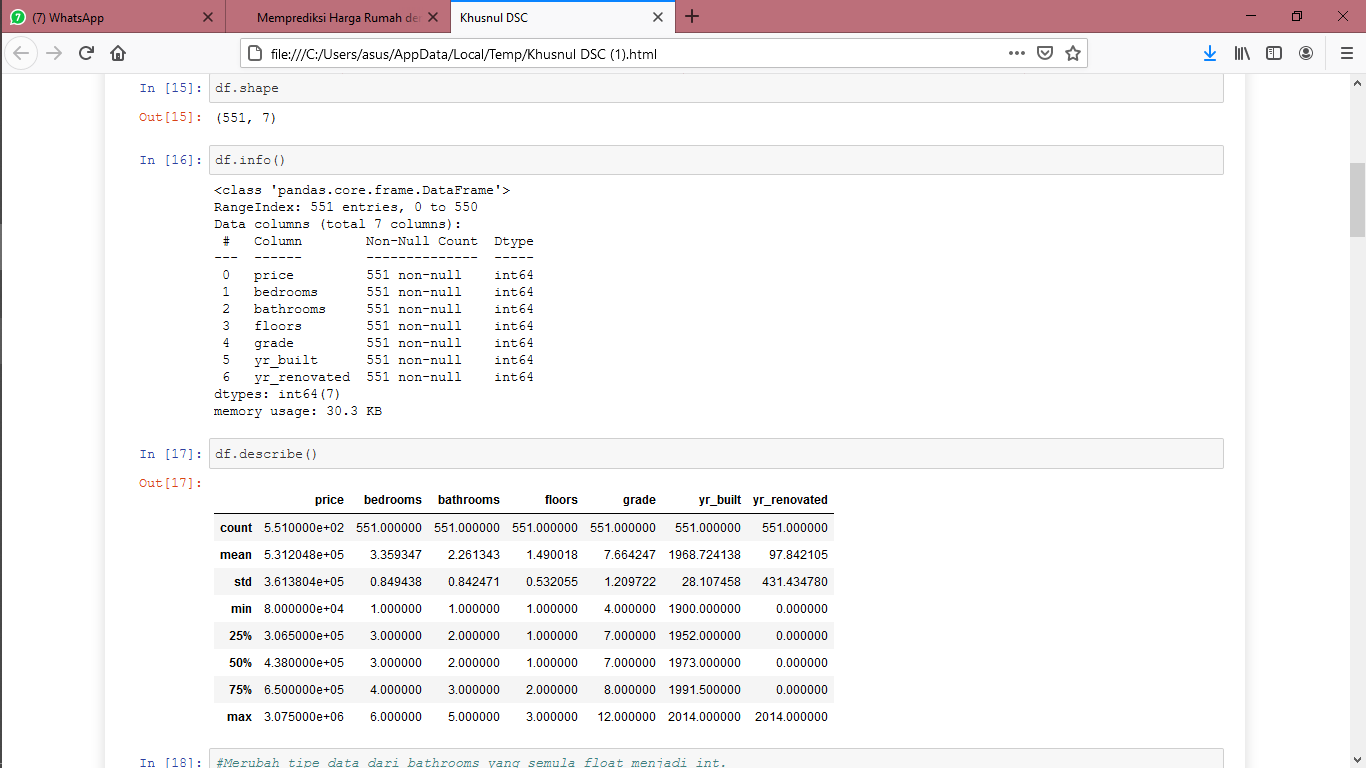


Jadi, kolomnya sebanyak 7 dan barisnya sebanyak 551.

1. Kemudian, melihat informasi dari data yang akan digunakan yaitu dengan memakai syntax :

df.info()

Output :

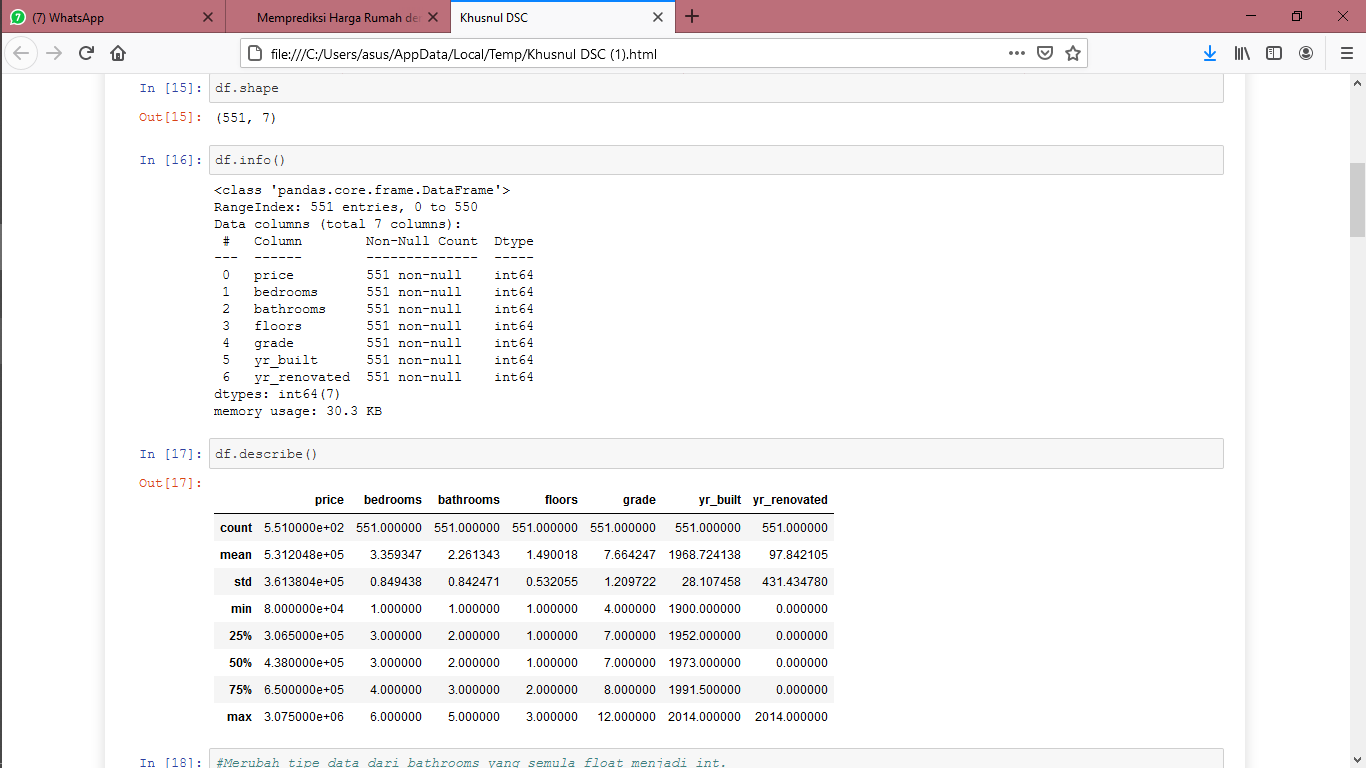


Dari hasil output tersebut diketahui jumlah masing-masing variabel, type data nya berupa numerik dan memory yang digunakan sebesar 30,3 KB.

1. Untuk melihat statistika deskriptif dari data mulai dari mean, kuartil, st.deviasi, dan lainnya bisa menggunakan syntax sebagai berikut :

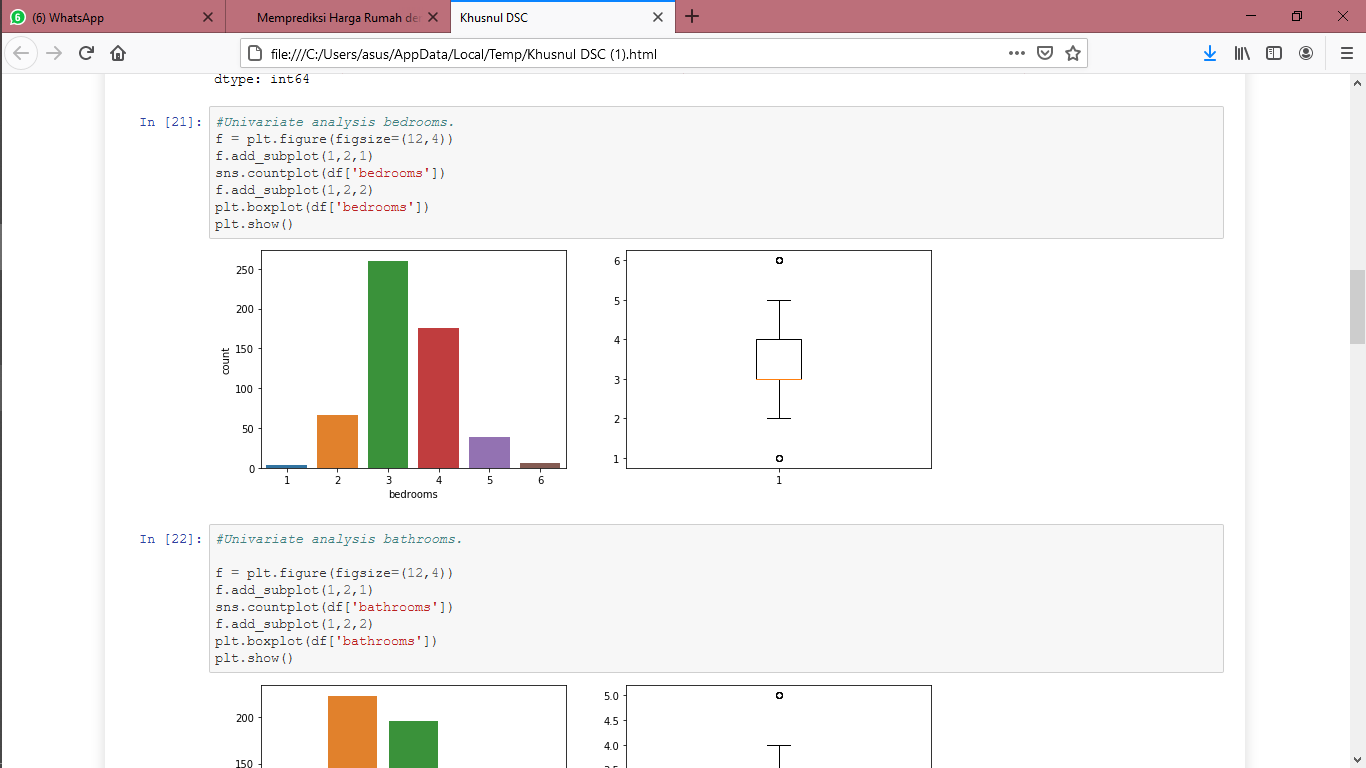
df.info()

Output :

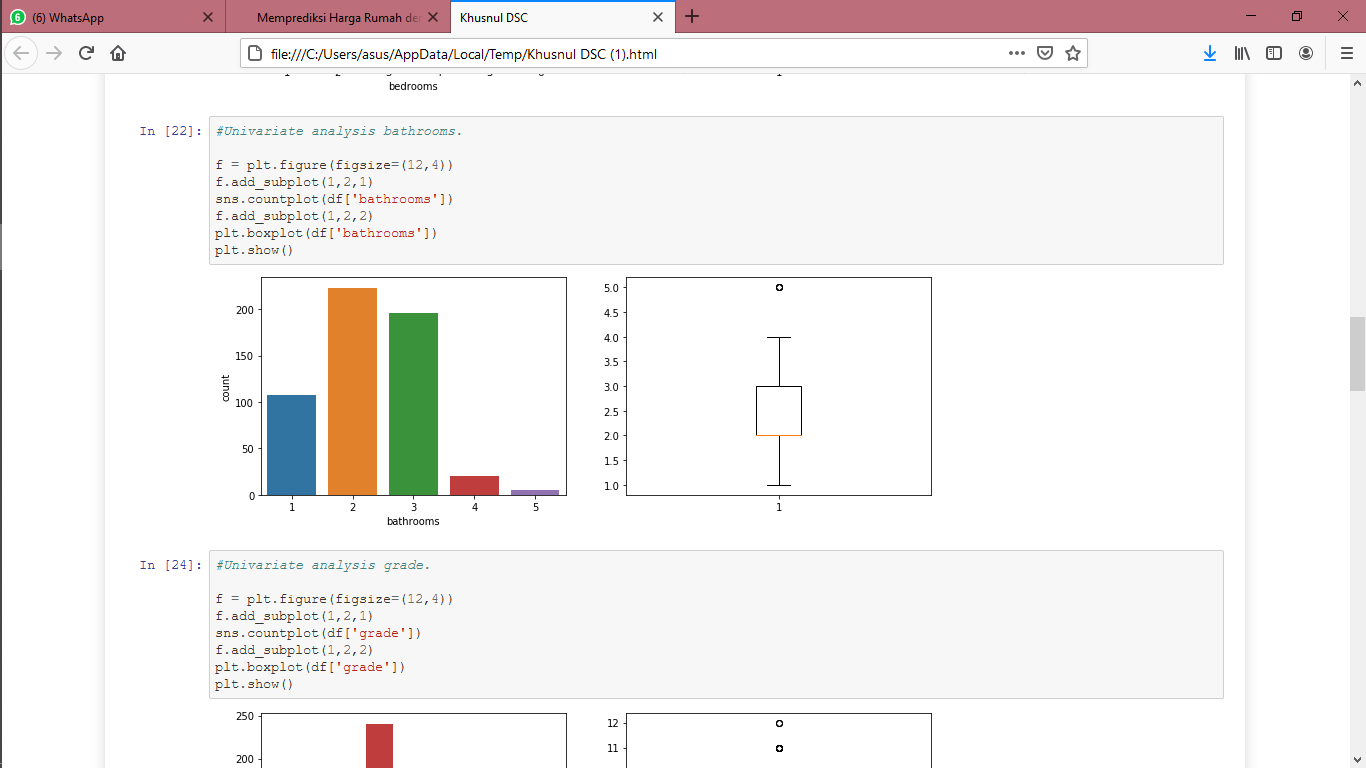


1. Kemudian melakukan Eksplorasi Data untuk mengenal data lebih jauh. Dengan menganalisis secara univariate dan bivariatenya.

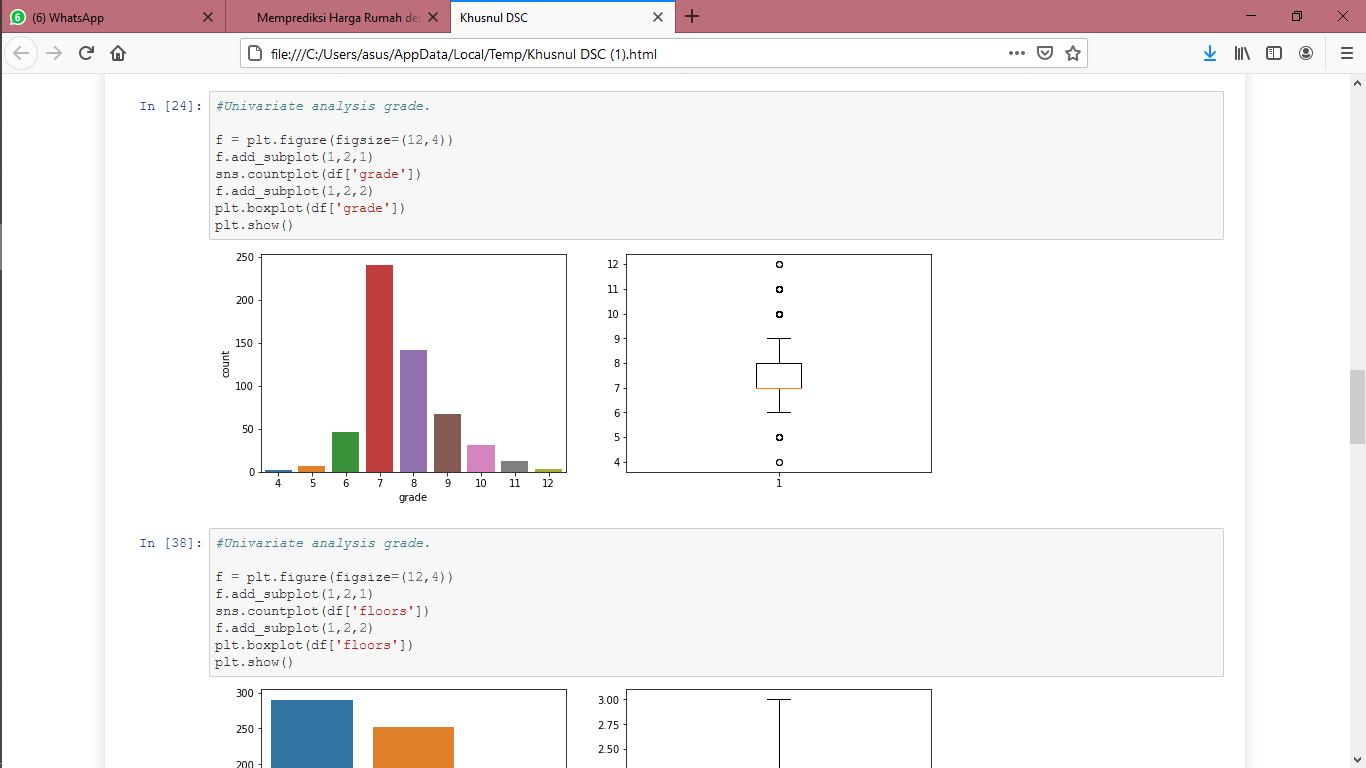
**Univariate Analysis**

****

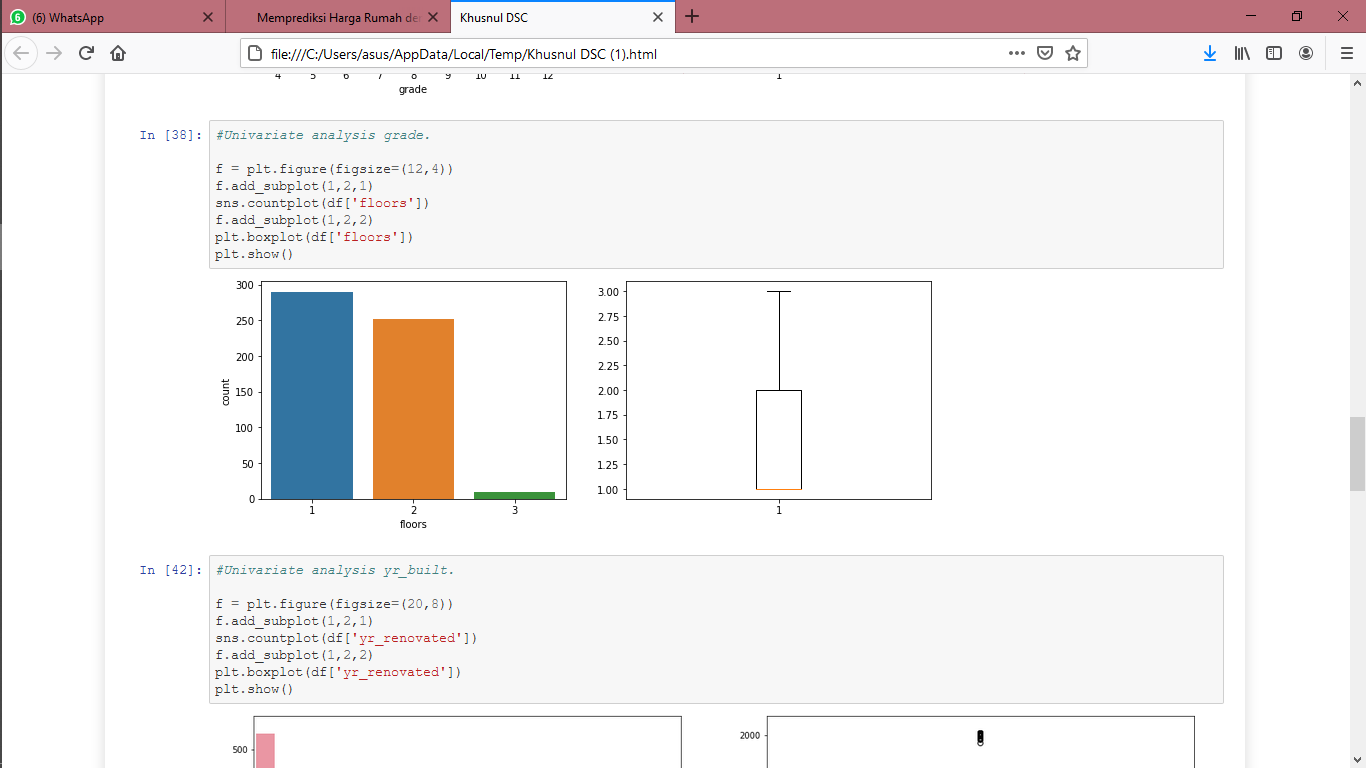
Diketahui bahwa jumlah kamar tidur kebanyakan ada 3.

****

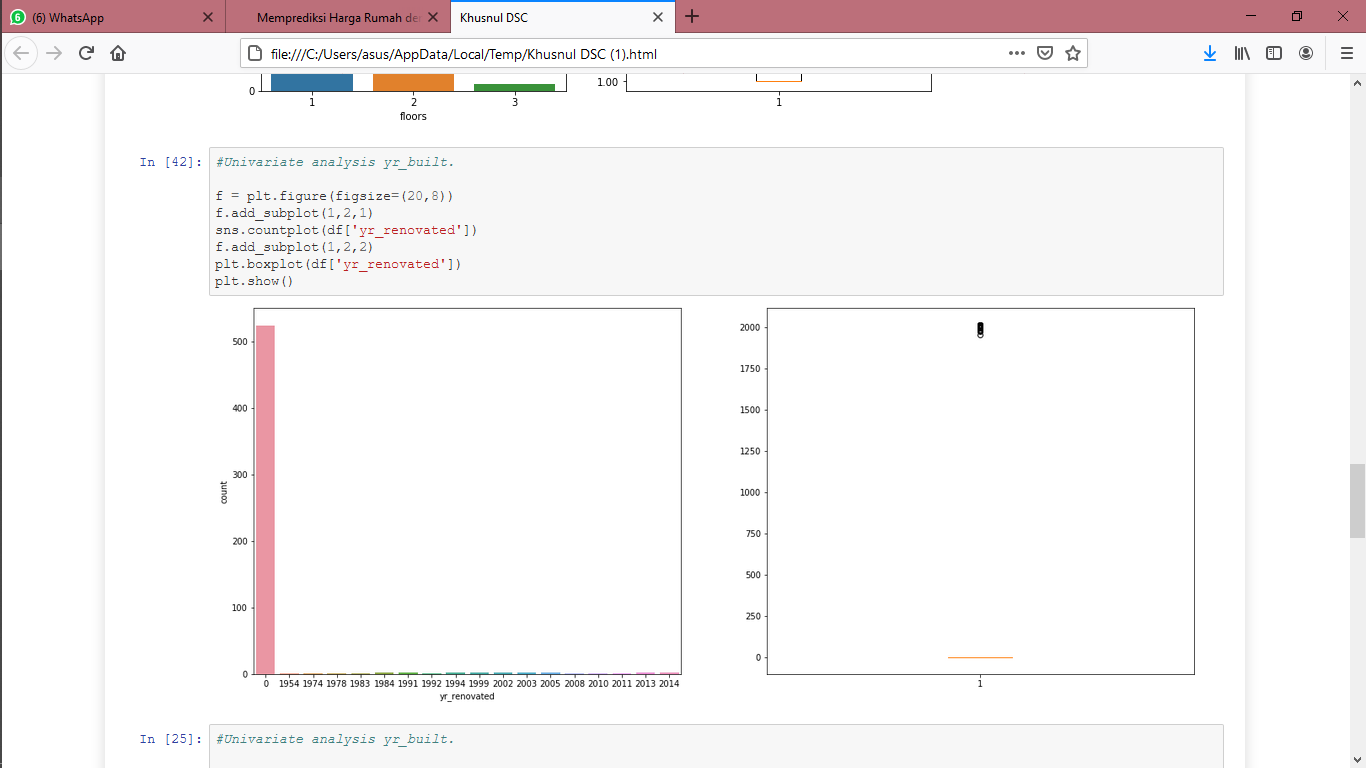
Diketahui bahwa jumlah kamar mandi sebagian besar di angka 2 sampai dengan 3.



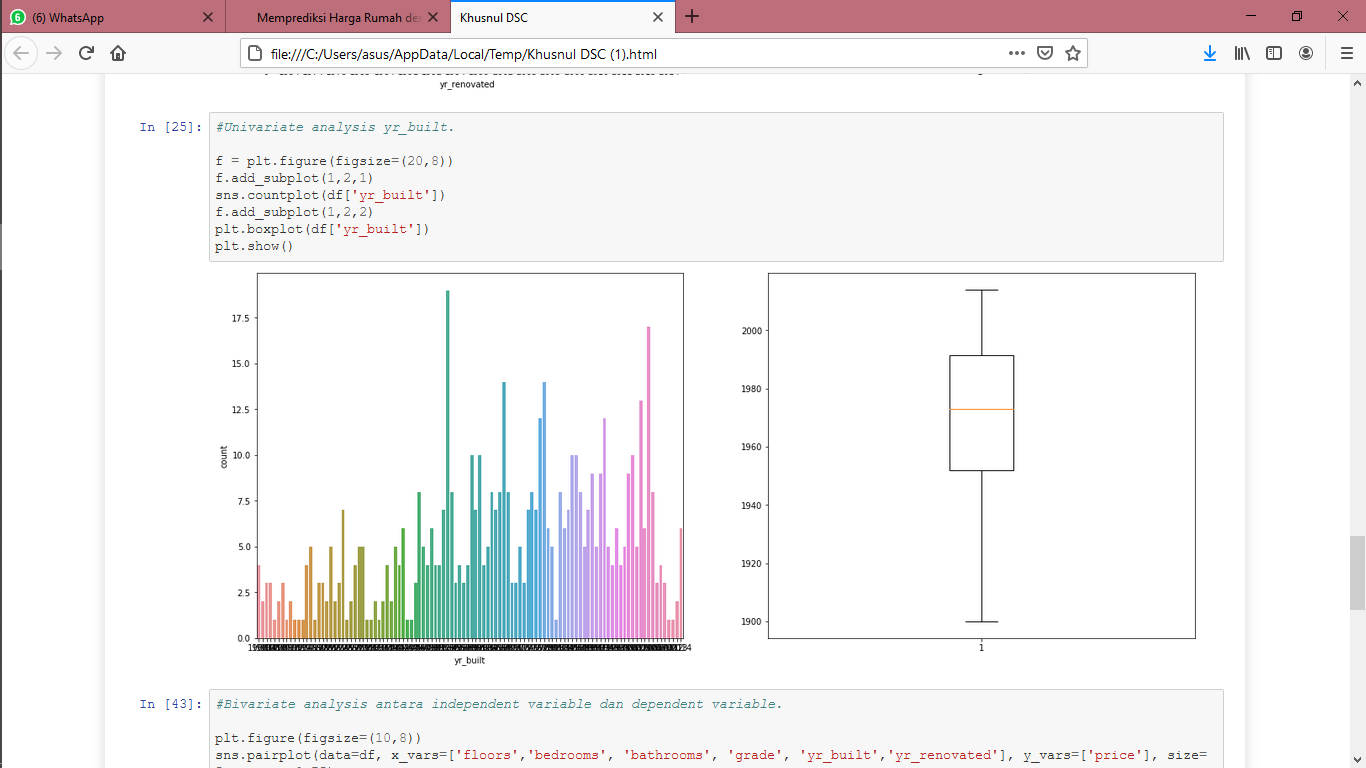
Diketahui grade terbanyak didominasi di angka 7.



Diketahui bahwa kebanyakan hanya ada 1 lantai.

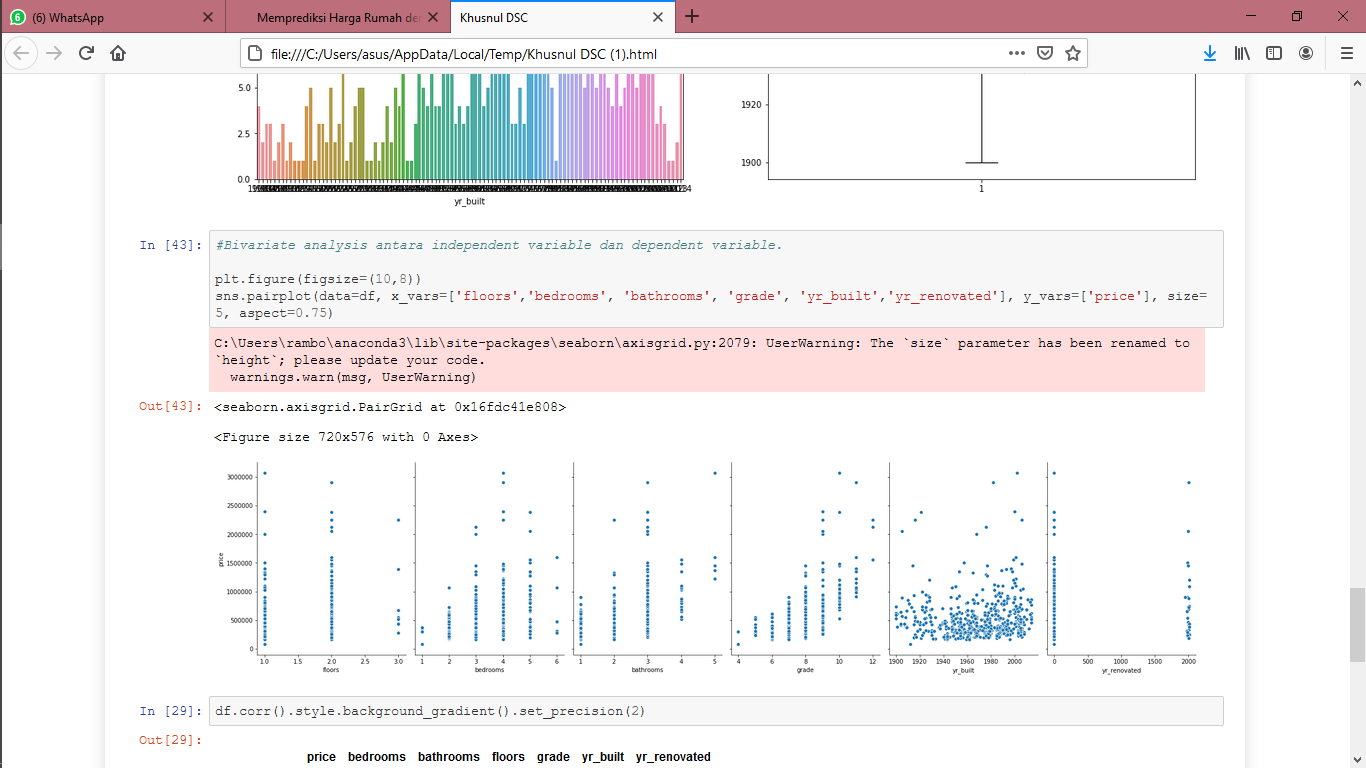


Diketahui bahwa kebanyakan tidak melakukan renovasi rumah jadi tahun renovasian bernilai 0 (tidak ada).



Diketahui bahwa density terdapat disekitar tahun 1960-1980 an dan data tidak memiliki outlier.

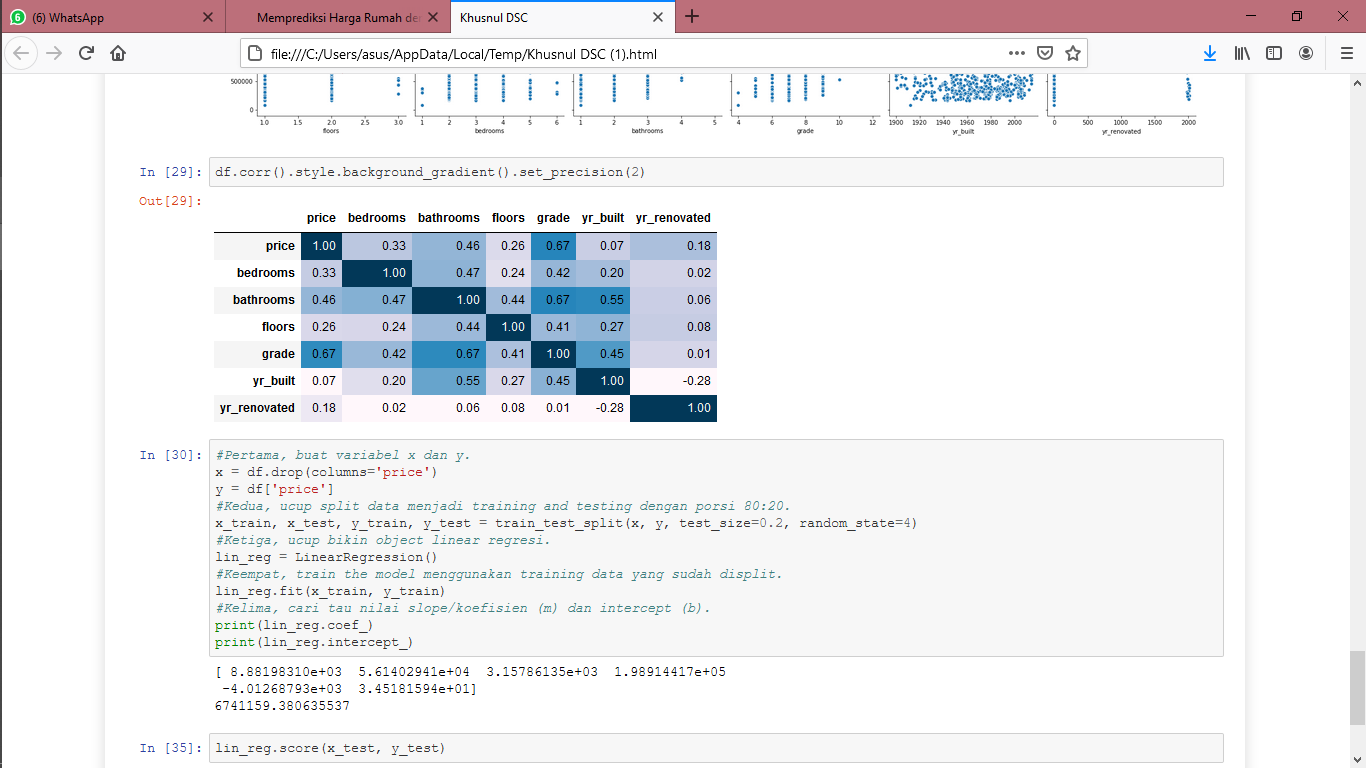
**Bivariate Analysis**



1. Selanjutnya, untuk mengetahui korelasi atau hubungan dari variabel dependen dan independen maka dapat dilakukan dengan :

df.corr().style.background\_gradient().set\_precision(2)

Hasil output :



Diketahui bahwa nilai korelasi yr\_renovated sangat mendekati nol bahkan bernilai negatif artinya bahwa renovasian rumah tidak mempengaruhi harga rumah.

**MODELLING DAN EVALUATING**

1. Syntax yang digunakan :

#Pertama, buat variabel x dan y.

x = df.drop(columns='price')

y = df['price']

#Kedua, ucup split data menjadi training and testing dengan porsi 80:20.

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=4)

#Ketiga, ucup bikin object linear regresi.

lin\_reg = LinearRegression()

#Keempat, train the model menggunakan training data yang sudah displit.

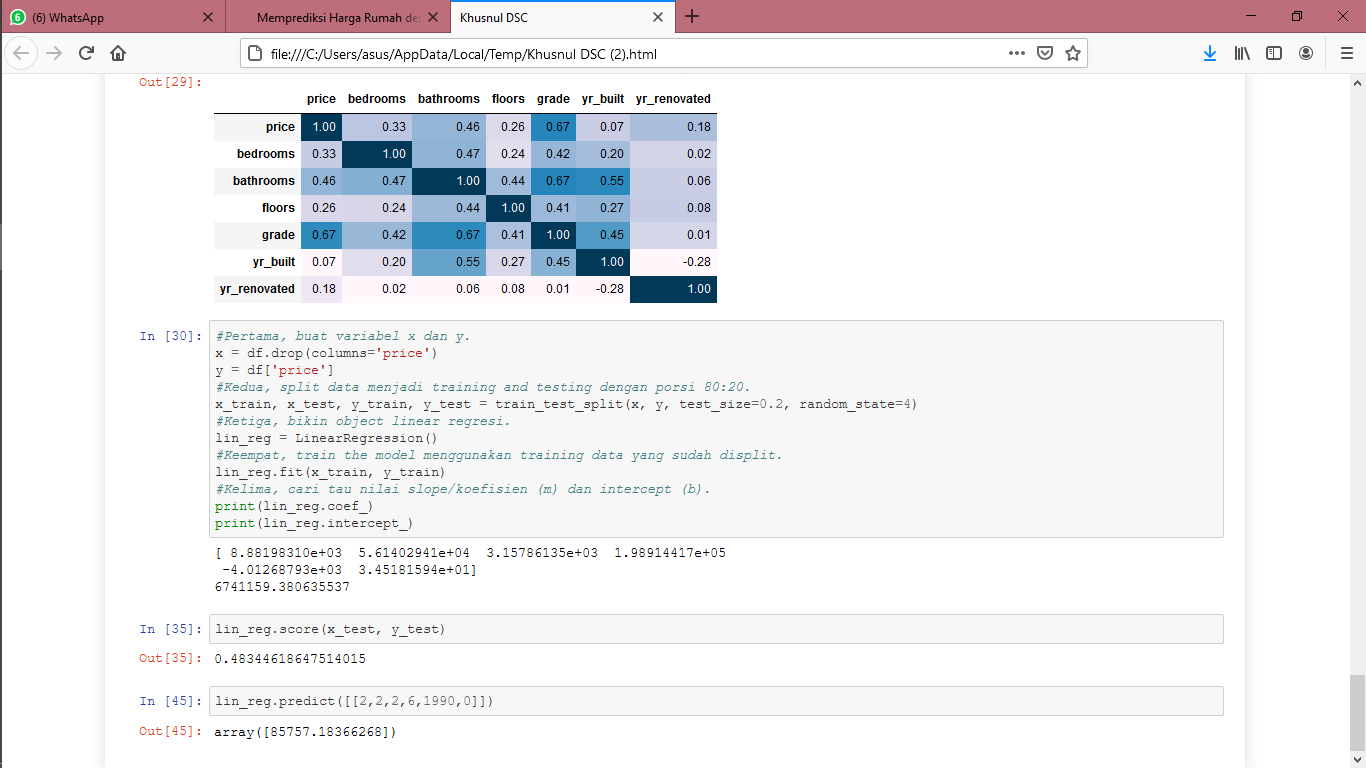
lin\_reg.fit(x\_train, y\_train)

#Kelima, cari tau nilai slope/koefisien (m) dan intercept (b).

print(lin\_reg.coef\_)

print(lin\_reg.intercept\_)

Output:



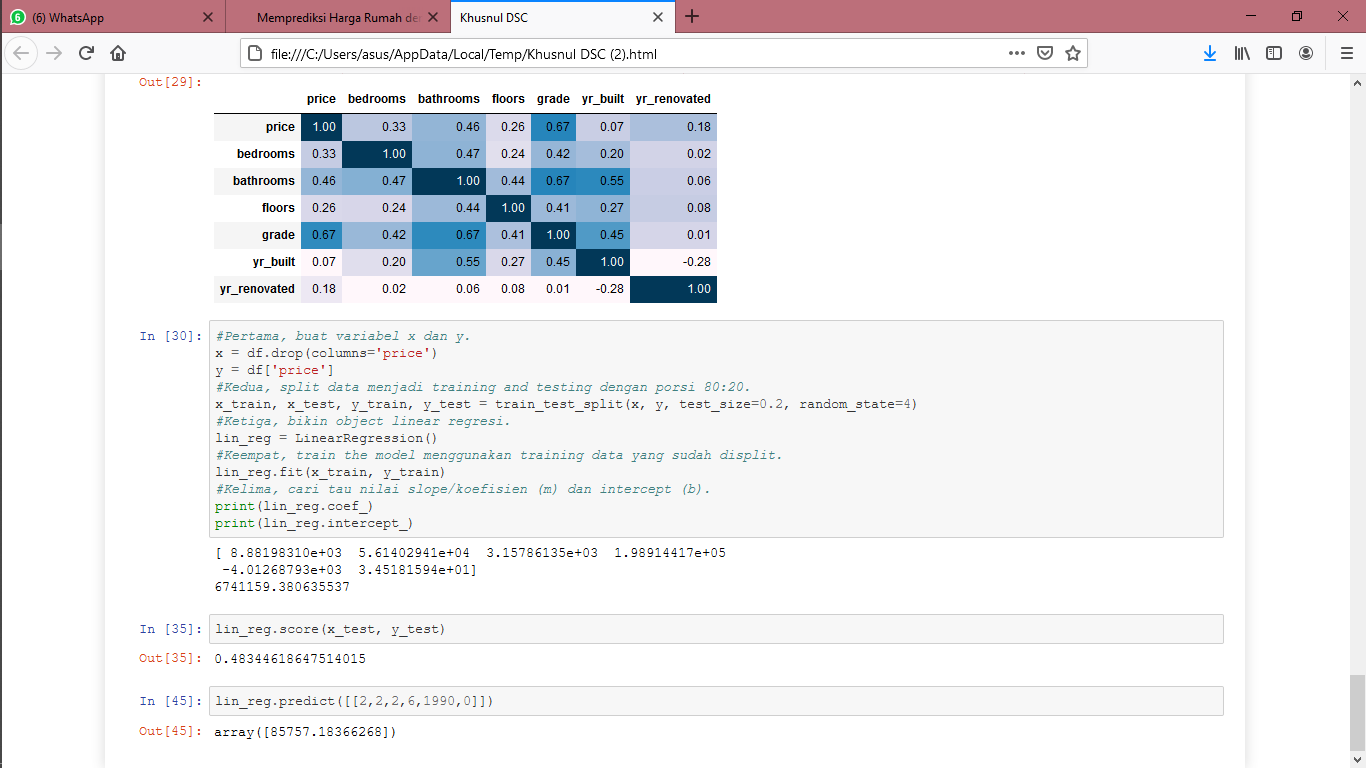
Dari nilai m dan b diatas, kalau dimasukan ke dalam rumus menjadi:

**Y = 6741159.38 + 8.88X1 + 5.61X2 + 3.16X3 + 1.99X4 – 4.01X5 + 3.45X6**

1. Kemudian, mencari tahu score akurasi dari model menggunakan testing data yang sudah displit.

lin\_reg.score(x\_test, y\_test)

Hasil :



**Diperoleh hasil score accuracy dari Model Machine Learning sebesar 48,34 %.**

1. Langkah terakhir melakukan prediksi terhadap kriteria rumah yang diinginkan.

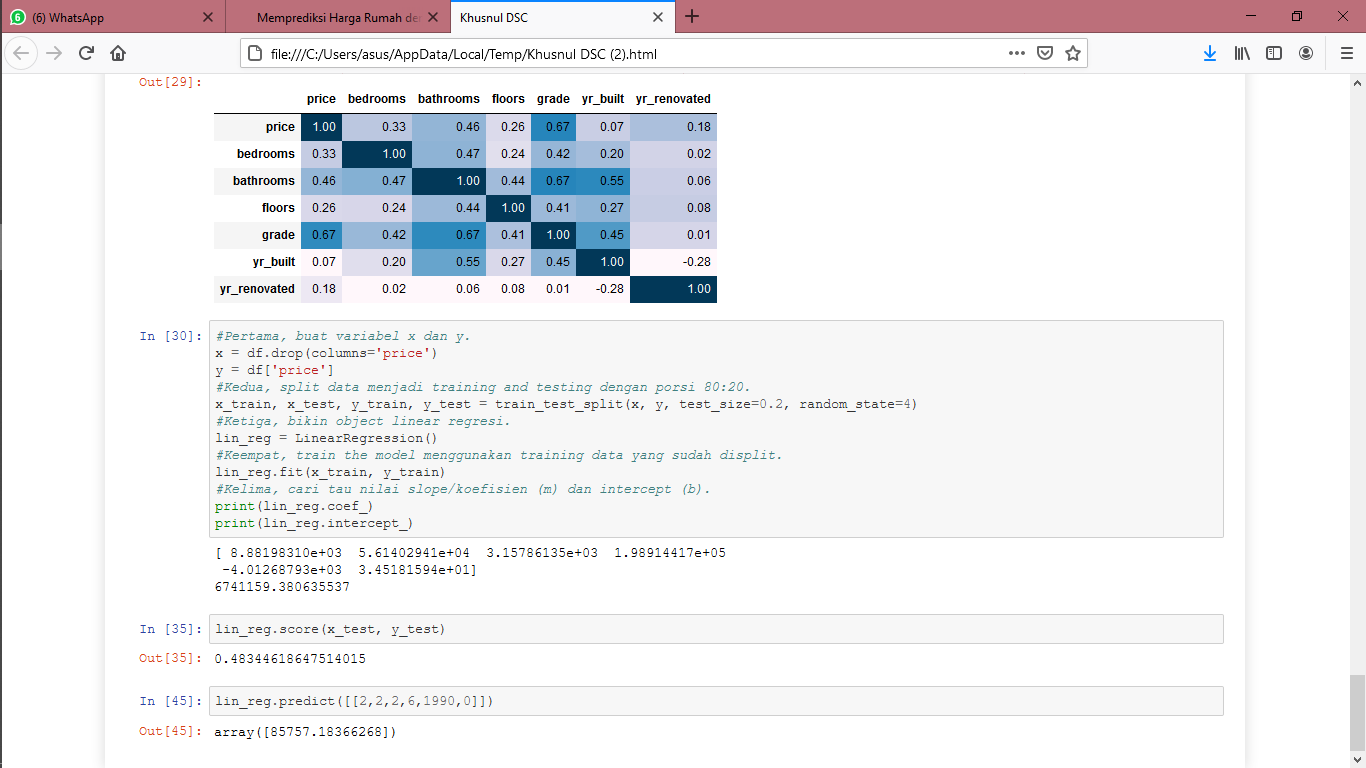
Dengan kriteria yang diinginkan yaitu :

1. Jumlah bedrooms = 2
2. Jumlah bathrooms = 2
3. Terdiri dari 2 lantai
4. Dengan gradenya 6
5. Tahun pembuatan rumahnya tahun 1990
6. Tahun renovasi rumah = 0, artinya sama sekali belum pernah melakukan renovasi rumah.

Maka prediksi yang digunakan untuk melihat harga rumahnya yaitu :

lin\_reg.predict([[2,2,2,6,1990,0]])

Output :



**Jadi, prediksi yang diperoleh yaitu 85757.18 US$.**

**KESIMPULAN**

Diperoleh nilai keakuratan dari Model Machine Learning: Multiple Linear Regression sebesar 48,34 %. Artinya keakuratan dari model ini tidak cukup akurat untuk memprediksi harga rumah.

Dari kriteria rumah yang diinginkan, kita dapat memprediksi harga rumah yaitu dengan harga 85757,18 US$.