


Nama: Albihan NIM: 064002200031		MODUL 8 Nama Dosen: Dedy Sugiarto
Hari/Tanggal: Senin, 7 Agustus 2023	Praktikum Statistika	Nama Asisten Laboratorium 1. Elen Fadilla Estri 064002000008 2. Rukhy Zaifa Aduhalim 064002000041

Eksplorasi Data Menggunakan Python

1. Teori Singkat

histogram berguna untuk memberikan gambaran ukuran tendensi sentral dan kesimetrisan data pengamatan. Penyajian grafis lainnya yang bisa merangkum informasi lebih detail mengenai distribusi nilai-nilai data pengamatan adalah Box and Whisker Plots atau lebih sering disebut dengan BoxPlot atau Box-Plot (kotak-plot) saja. Seperti namanya, Box and Whisker, bentuknya terdiri dari Box (kotak) dan whisker.

Box-plot atau boxplot (juga dikenal sebagai diagram box-and-whisker) merupakan suatu box (kotak berbentuk bujur sangkar). Boxplot adalah salah satu cara dalam statistik deskriptif untuk menggambarkan secara grafik dari data numeris melalui lima ukuran sebagai berikut:

- Nilai observasi terkecil,
- Kuartil terendah atau kuartil pertama (Q1), yang memotong 25% dari data terendah
- Median (Q2) atau nilai pertengahan,
- Kuartil tertinggi atau kuartil ketiga (Q3), yang memotong 25% dari data terbesar
- Nilai observasi terbesar.

Dalam boxplot juga ditunjukkan, jika ada, nilai outlier dari observasi. Boxplot dapat digunakan untuk menunjukkan perbedaan antara populasi tanpa menggunakan asumsi distribusi statistik yang mendasarinya. Karenanya, boxplot tergolong dalam statistik non-parametrik. Jarak antara bagian-bagian dari box menunjukkan derajat dispersi (penyebaran) dan skewness (kecondongan) dalam data. Dalam penggambarannya, boxplot dapat digambarkan secara horizontal maupun vertikal.

2. Alat dan Bahan

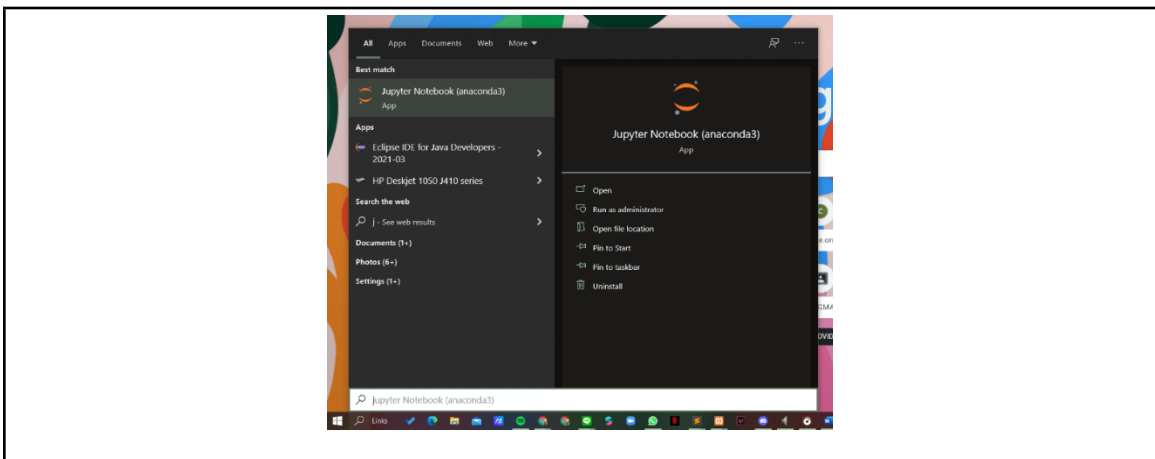
Hardware : Laptop/PC

Software : R Studio

3. Elemen Kompetensi

a. Latihan pertama – Praktikum

1. Buka Jupyter Notebook atau Google Colab di Browser



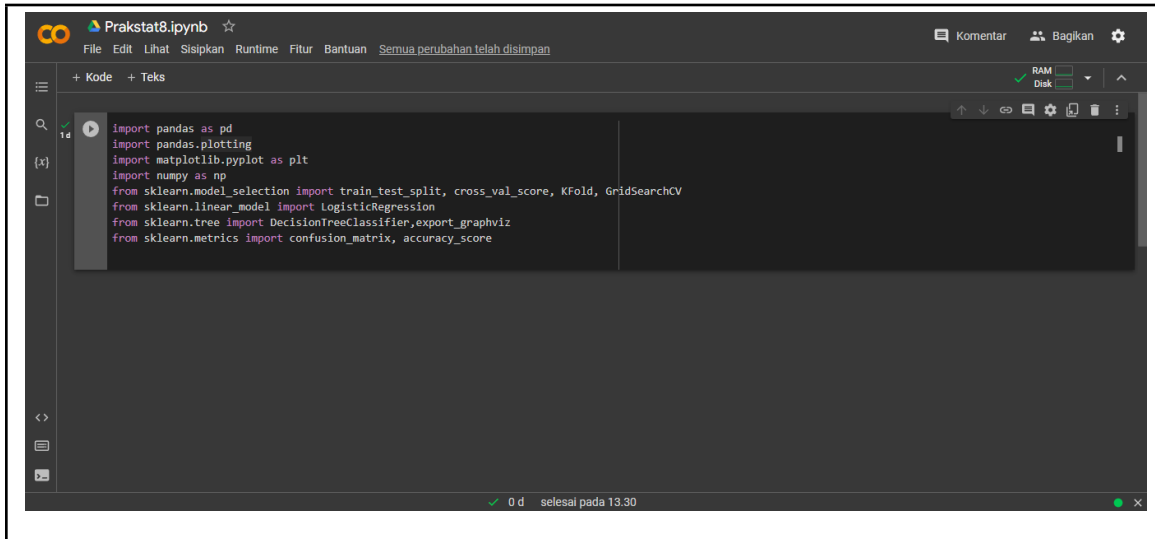
2. Lalu jalankan script berikut dan berikan output (gunakan nama variable data dengan nama masing-masing)

```
import pandas as pd
from pandas.tools import plotting
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score, KFold,
GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.metrics import confusion_matrix, accuracy_score
```



```
from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifier
```

Output:

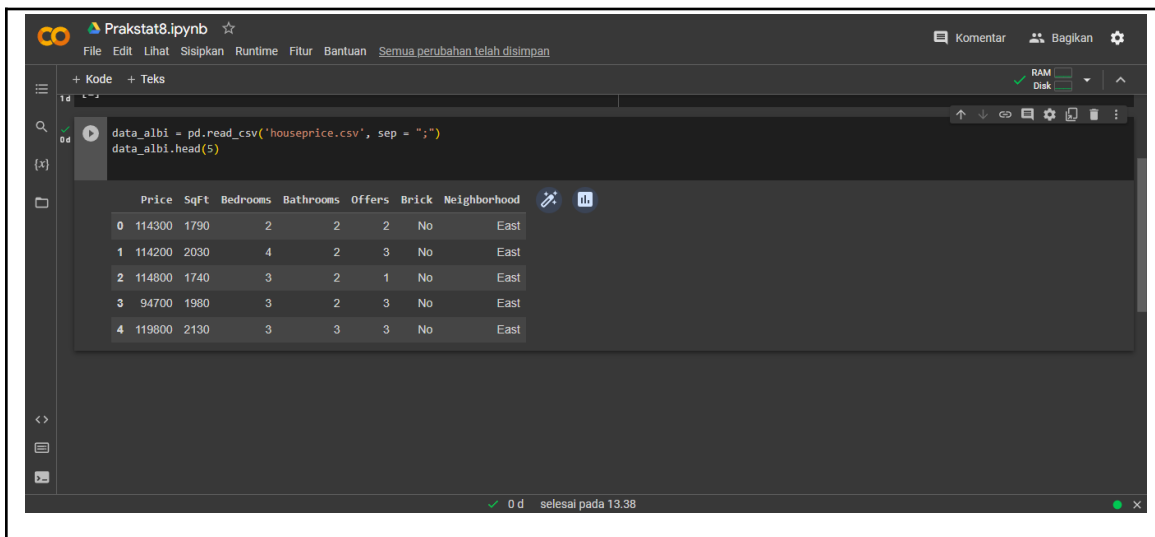


```
import pandas as pd
import pandas.plotting
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score, KFold, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.metrics import confusion_matrix, accuracy_score
```

3. Script

```
data_nama = pd.read_csv('C:/prakstatik/houseprice.csv')
data_nama.head(5)
```

Output:



```
data_albi = pd.read_csv('houseprice.csv', sep = ";")
data_albi.head(5)
```

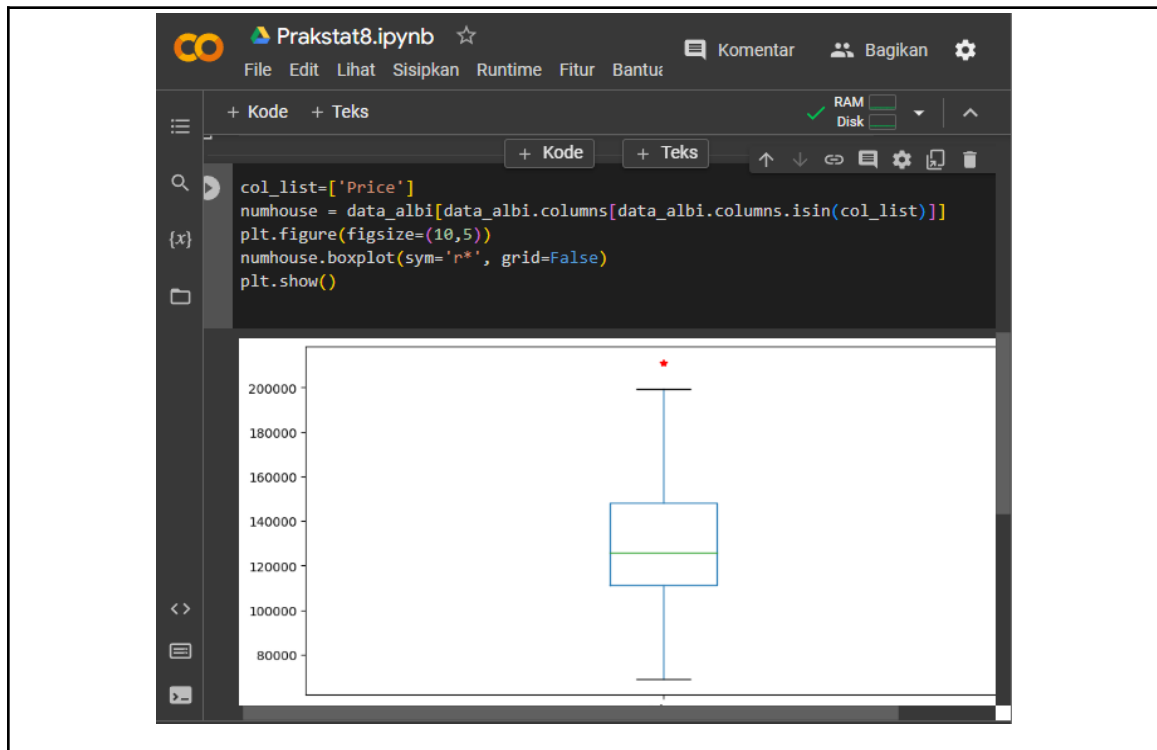
	Price	SqFt	Bedrooms	Bathrooms	Offers	Brick	Neighborhood
0	114300	1790	2	2	2	No	East
1	114200	2030	4	2	3	No	East
2	114800	1740	3	2	1	No	East
3	94700	1980	3	2	3	No	East
4	119800	2130	3	3	3	No	East



4. Boxplot dari harga rumah/Price

```
col_list=['Price']  
numhouse = data_nama[data_nama.columns[data_nama.columns.isin(col_list)]]  
plt.figure(figsize=(10,5))  
numhouse.boxplot(sym='r*', grid=False)  
plt.show()
```

Output:

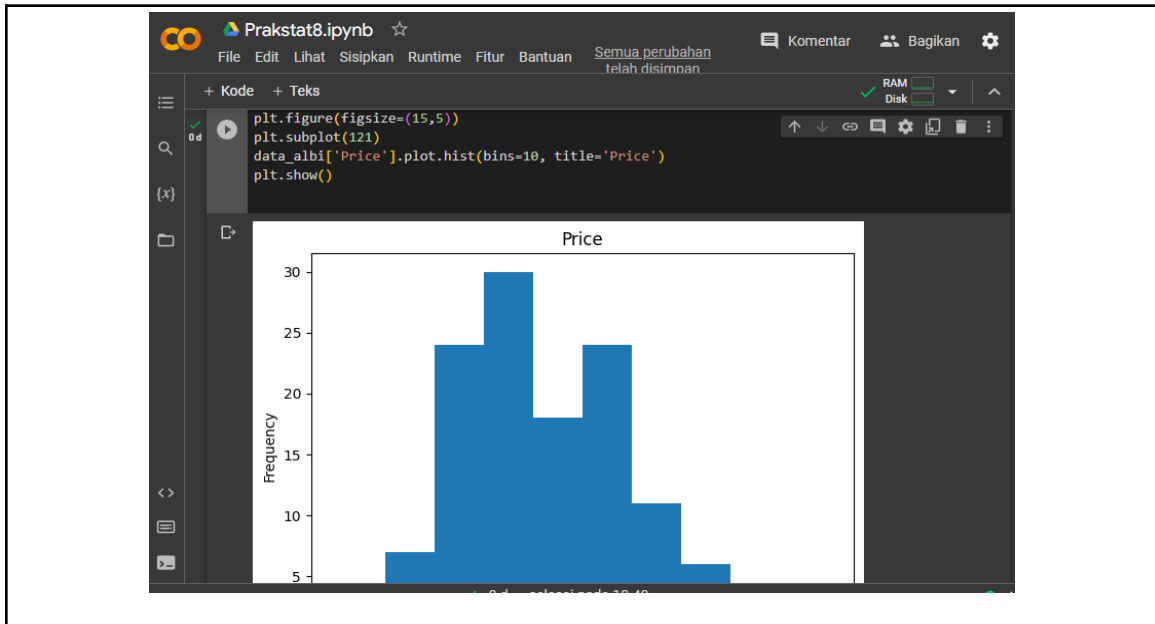


5. Histogram dari Price

```
plt.figure(figsize=(15,5))  
plt.subplot(121)  
data_nama['Price'].plot.hist(bins=10, title='Price')  
plt.show()
```



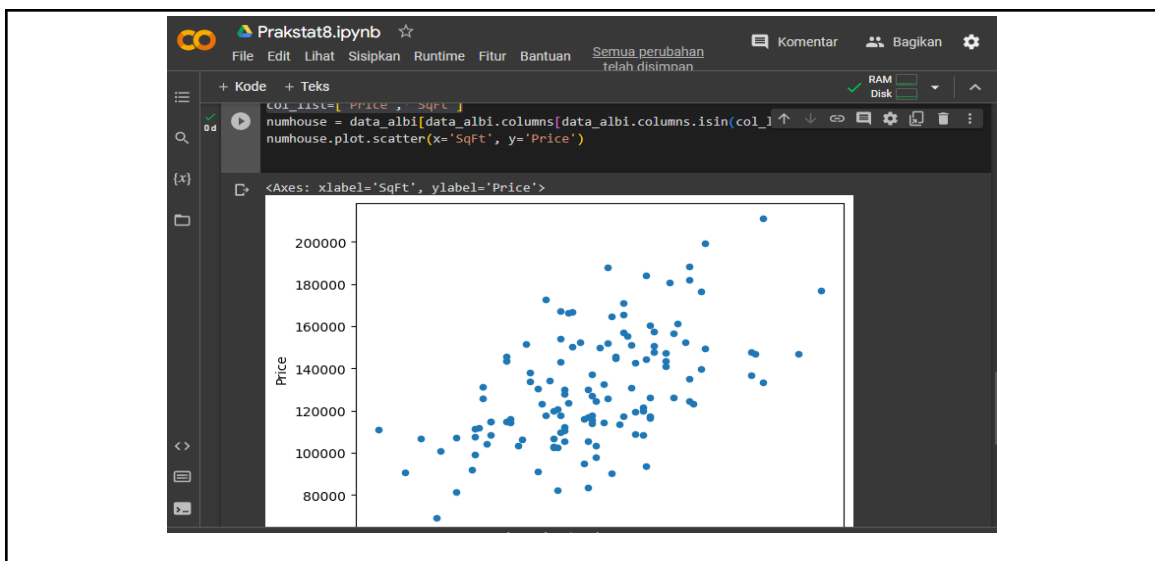
Output:



6. Scatter dari Price

```
col_list=['Price', 'SqFt']
numhouse = data_nama[data_nama.columns[data_nama.columns.isin(col_list)]]
numhouse.plot.scatter(x='SqFt', y='Price')
```

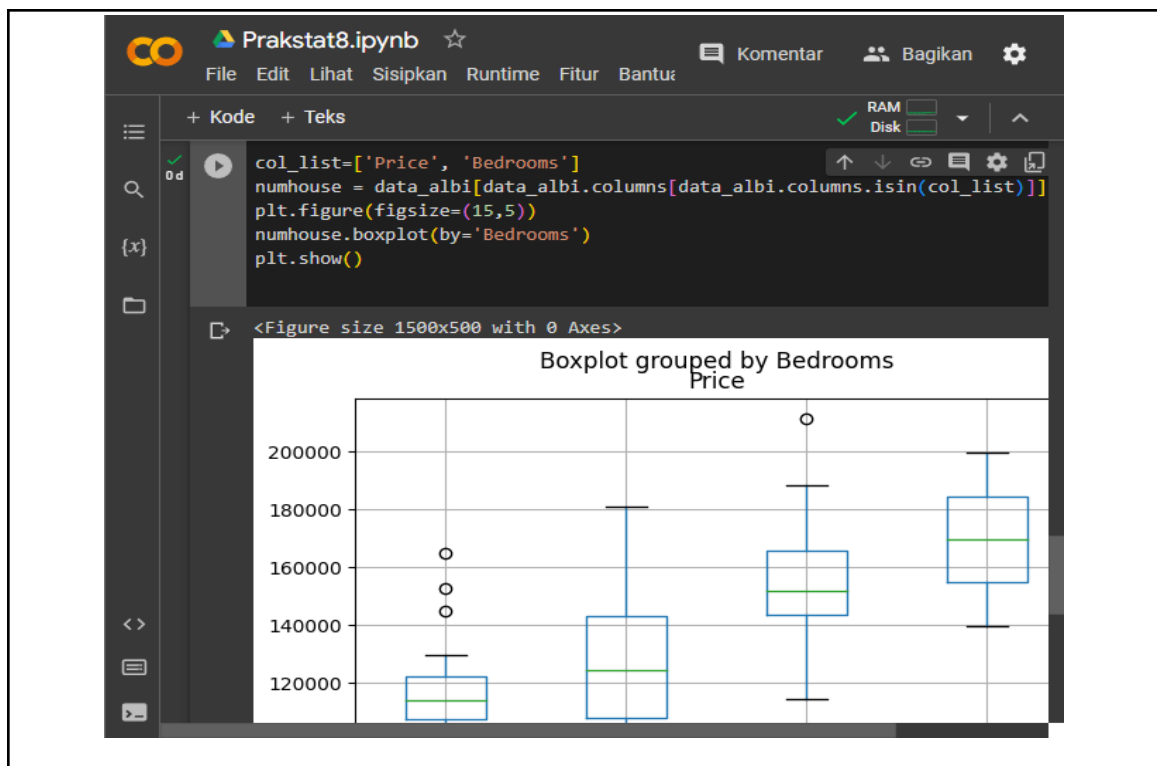
Output:



7. Group Boxplot Berdasarkan Bedrooms

```
col_list=['Price', 'Bedrooms']  
numhouse = data_nama[data_nama.columns[data_nama.columns.isin(col_list)]]  
plt.figure(figsize=(15,5))  
numhouse.boxplot(by='Bedrooms')  
plt.show()
```

Output:



b. Latihan Kedua – Tugas

1. Buatlah Scatter Plot Harga Rumah Berdasarkan Bedrooms!

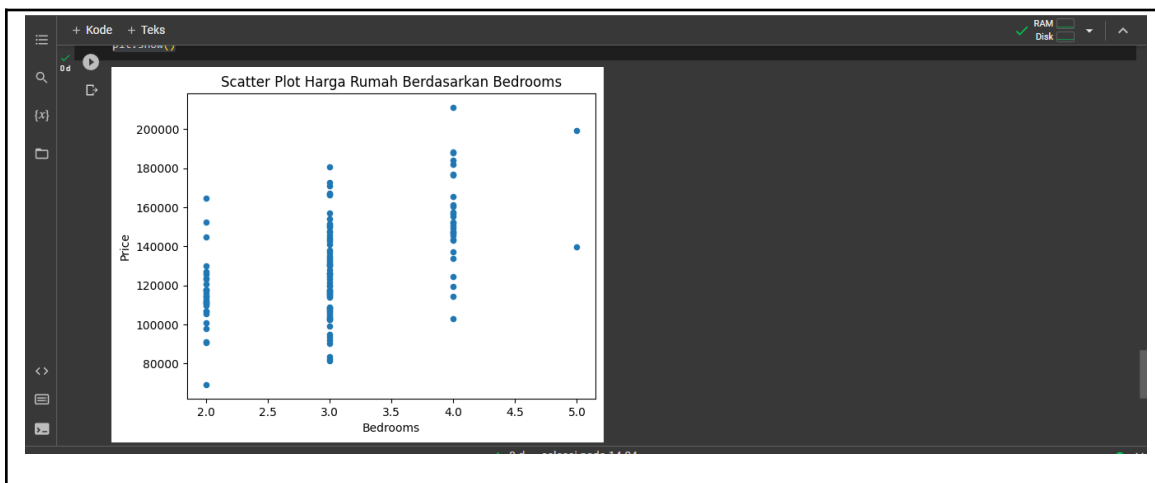
Script:

```
col_list = ['Price', 'Bedrooms']  
  
numhouse = data_albi[col_list]  
  
numhouse.plot.scatter(x='Bedrooms', y='Price')
```



```
plt.title('Scatter Plot of Price vs. Bedrooms')  
plt.xlabel('Bedrooms')  
plt.ylabel('Price')  
plt.show()
```

Output:



Penjelasan: bertujuan untuk membaca data dari kolom 'Price' dan 'Bedrooms' dalam DataFrame `data_albi`, dan kemudian membuat scatter plot untuk memvisualisasikan hubungan antara harga (Price) dan jumlah kamar tidur (Bedrooms). Scatter plot tersebut menunjukkan titik-titik data di mana setiap titik mewakili sebuah observasi dari properti dengan harga tertentu dan jumlah kamar tidur tertentu.

2. Buatlah Scatter Plot Harga Rumah berdasarkan Bathrooms!

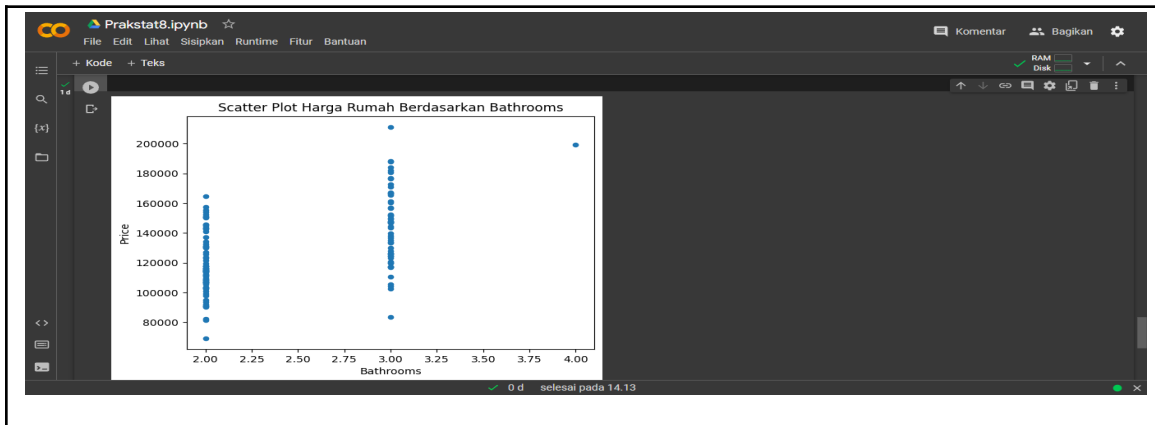
Script:

```
col_list = ['Price', 'Bathrooms']  
numhouse = data_albi[col_list]  
numhouse.plot.scatter(x='Bathrooms', y='Price')  
plt.title('Scatter Plot Harga Rumah Berdasarkan Bathrooms')
```



```
plt.xlabel('Bathrooms')  
plt.ylabel('Price')  
plt.show()
```

Output:



Penjelasan: bertujuan untuk membaca data dari kolom '**Price**' dan '**Bathrooms**' dalam DataFrame `data_albi`, dan kemudian membuat scatter plot untuk memvisualisasikan hubungan antara harga (**Price**) dan jumlah kamar mandi (**Bathrooms**). Scatter plot tersebut menunjukkan titik-titik data di mana setiap titik mewakili sebuah observasi dari properti dengan harga tertentu dan jumlah kamar mandi tertentu. Dengan visualisasi ini, kita dapat melihat bagaimana harga rumah berubah berdasarkan jumlah kamar mandi yang dimiliki.

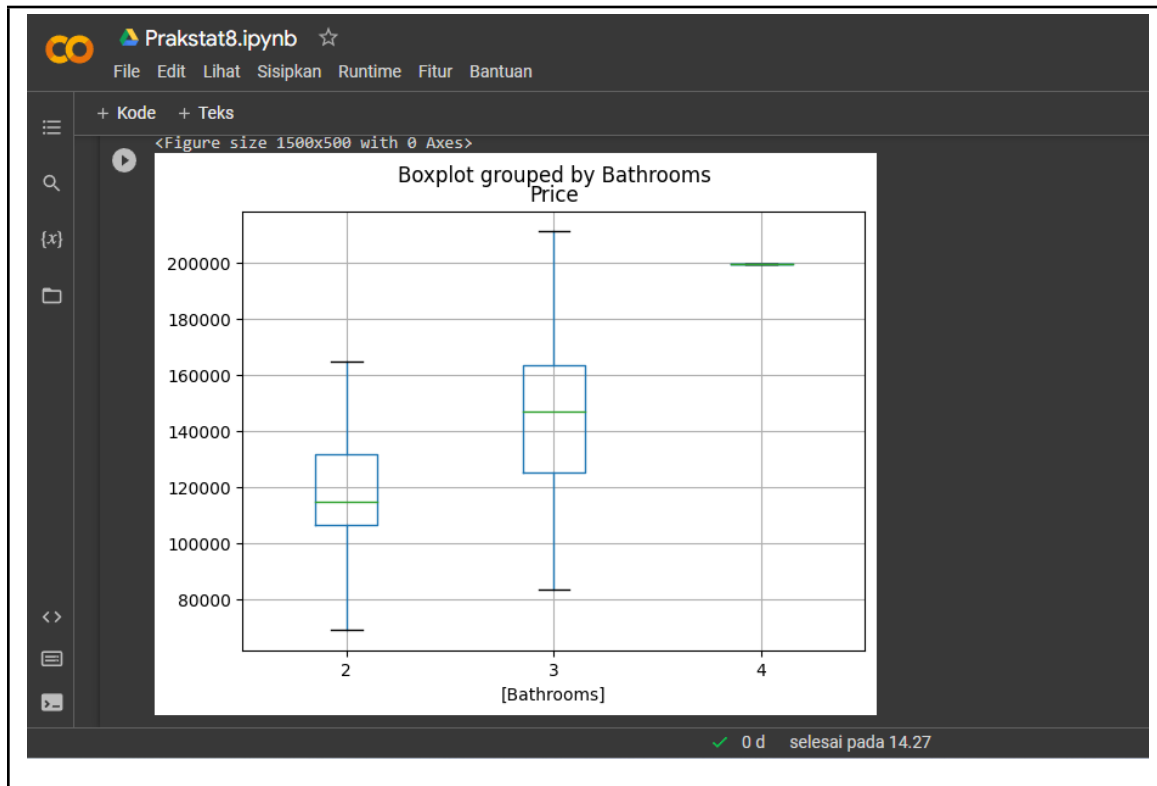
3. Buatlah Grouped Boxplot berdasarkan Bathrooms dengan Price sebagai perbandingannya!

Script:

```
col_list=['Price', 'Bathrooms']  
numhouse = data_albi[data_albi.columns[data_albi.columns.isin(col_list)]]  
plt.figure(figsize=(15,5))  
numhouse.boxplot(by='Bathrooms')  
plt.show()
```



Output:



Penjelasan: bertujuan untuk membaca data dari kolom '**Price**' dan '**Bathrooms**' dalam DataFrame `data_albi`, kemudian membuat **box plot** untuk memvisualisasikan distribusi harga (**Price**) berdasarkan jumlah kamar mandi (**Bathrooms**). Box plot ini memungkinkan kita untuk melihat bagaimana harga rumah tersebar atau berdistribusi untuk setiap nilai unik dalam kolom '**Bathrooms**'. **Box plot** juga membantu mengidentifikasi adanya pencilan (**outliers**) dan melihat perbedaan harga antara kategori jumlah kamar mandi.

4. File Praktikum

Github Repository:



5. Soal Latihan

Soal:

1. Apa yang dimaksud Exploratory Data Analysis?
2. Mengapa EDA diperlukan melakukan dalam melakukan analisis data?

Jawaban:

1. **EDA** adalah suatu pendekatan untuk menganalisis dan memahami data secara visual dan deskriptif sebelum melakukan analisis lebih lanjut atau membangun model. Tujuan utama dari EDA adalah untuk mendapatkan wawasan awal tentang data, mengidentifikasi pola, mengenali anomali, dan mengungkap hubungan antara variabel. EDA melibatkan penggunaan metode statistik sederhana dan teknik visualisasi data untuk merangkum, menggambarkan, dan menyajikan data agar mudah dipahami.
2. Dengan melakukan **EDA** secara cermat, analisis data dapat mengidentifikasi potensi masalah, membuat keputusan yang lebih baik, dan menghasilkan pemahaman yang lebih dalam tentang data yang digunakan. Hal ini merupakan langkah awal yang penting sebelum melakukan analisis lebih lanjut atau pemodelan yang lebih kompleks.

6. Kesimpulan

- a. Dalam pengerjaan praktikum Statistika, histogram berguna untuk memberikan gambaran ukuran tendensi sentral dan kesimetrisan data pengamatan. Penyajian grafis lainnya yang bisa merangkum informasi lebih detail mengenai distribusi nilai-nilai data pengamatan adalah Box and Whisker Plots atau lebih sering disebut dengan BoxPlot atau Box-Plot (kotak-plot) saja. Seperti namanya, Box and Whisker, bentuknya terdiri dari Box (kotak) dan whisker.
- b. Kita juga dapat mengetahui boxplot juga ditunjukkan, jika ada, nilai outlier dari observasi. Boxplot dapat digunakan untuk menunjukkan perbedaan antara populasi tanpa menggunakan asumsi distribusi statistik yang mendasarinya. Karenanya, boxplot tergolong dalam statistik non-parametrik. Jarak antara bagian-bagian dari box menunjukkan derajat dispersi (penyebaran) dan skewness (kecondongan) dalam data. Dalam penggambarannya, boxplot dapat digambarkan secara horizontal maupun vertikal.

7. Cek List (✓)

No	Elemen Kompetensi	Penyelesaian	
		Selesai	Tidak Selesai
1.	Latihan Pertama	<input checked="" type="checkbox"/>	



2.	Latihan Kedua	<input checked="" type="checkbox"/>	
----	---------------	-------------------------------------	--

8. Formulir Umpan Balik

No	Elemen Kompetensi	Waktu Pengerjaan	Kriteria
1.	Latihan Pertama	30 Menit	1
2.	Latihan Kedua	40 Menit	1

Keterangan:

1. Menarik
2. Baik
3. Cukup
4. Kurang

