


Nama: Mendari pertiwi NIM: 064002200037	 Praktikum Statistika	MODUL 8 Nama Dosen: Dedy Sugiarto
Hari/Tanggal: senin, 7 agustus 2023		Nama Asisten Labratorium 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

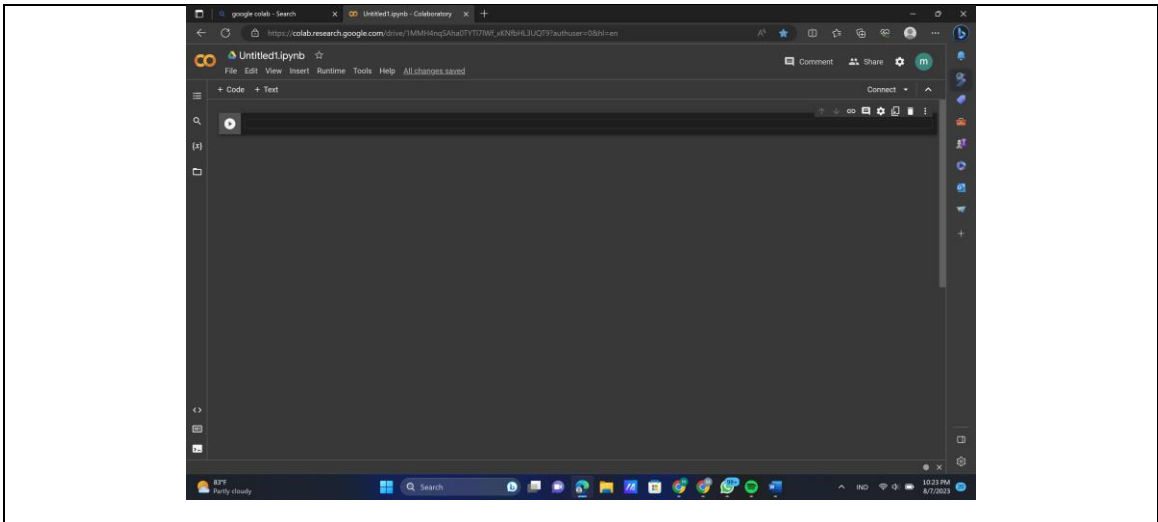
Hardware : Laptop/PC

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
```

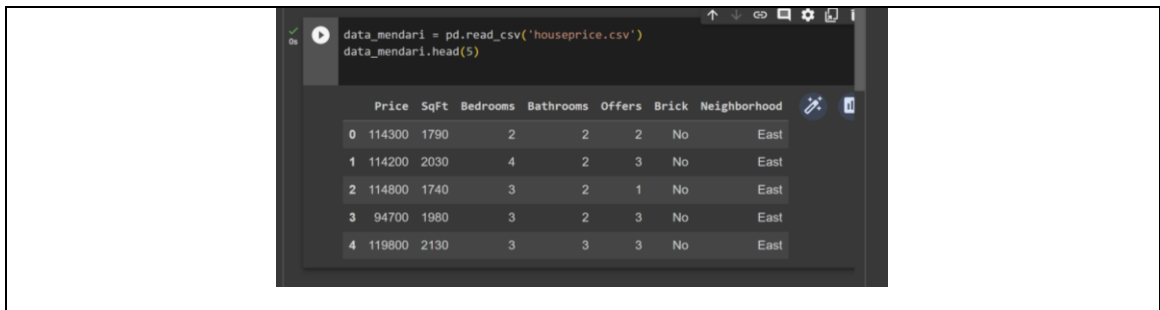
Output:

```
import pandas as pd
#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
```

3. Script

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

Output:

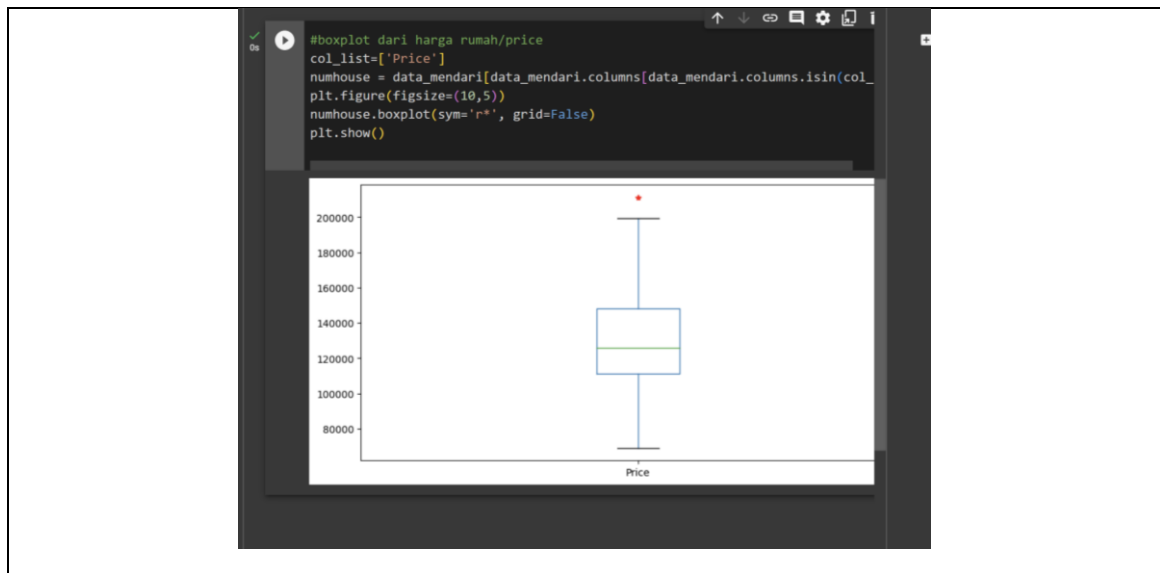


4. Boxplot dari harga rumah/Price

```
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:

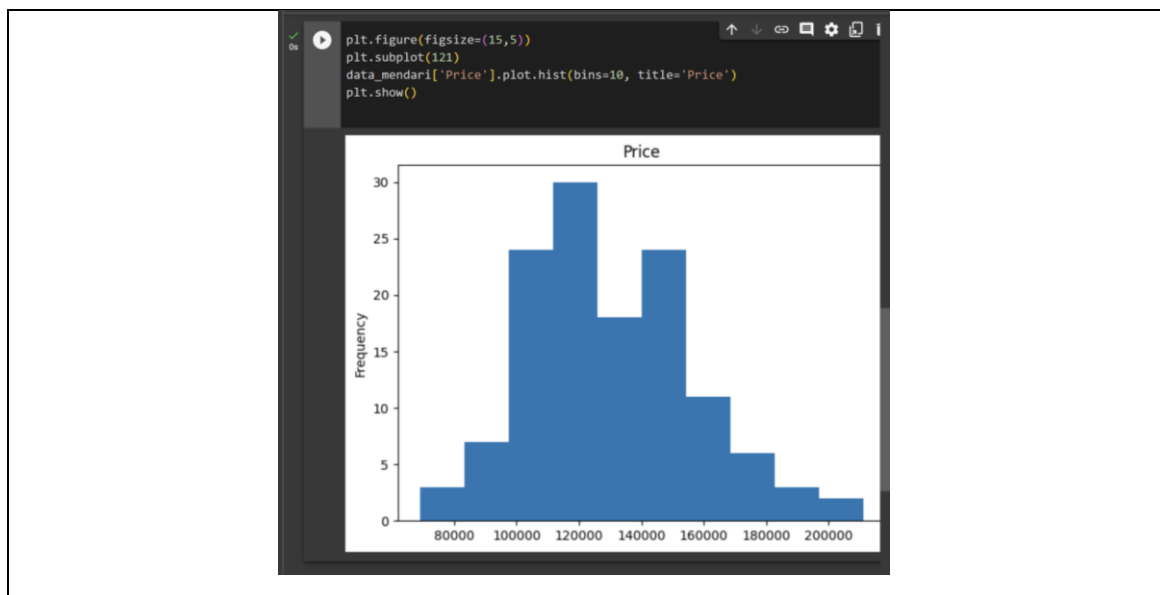




5. Histogram dari Price

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

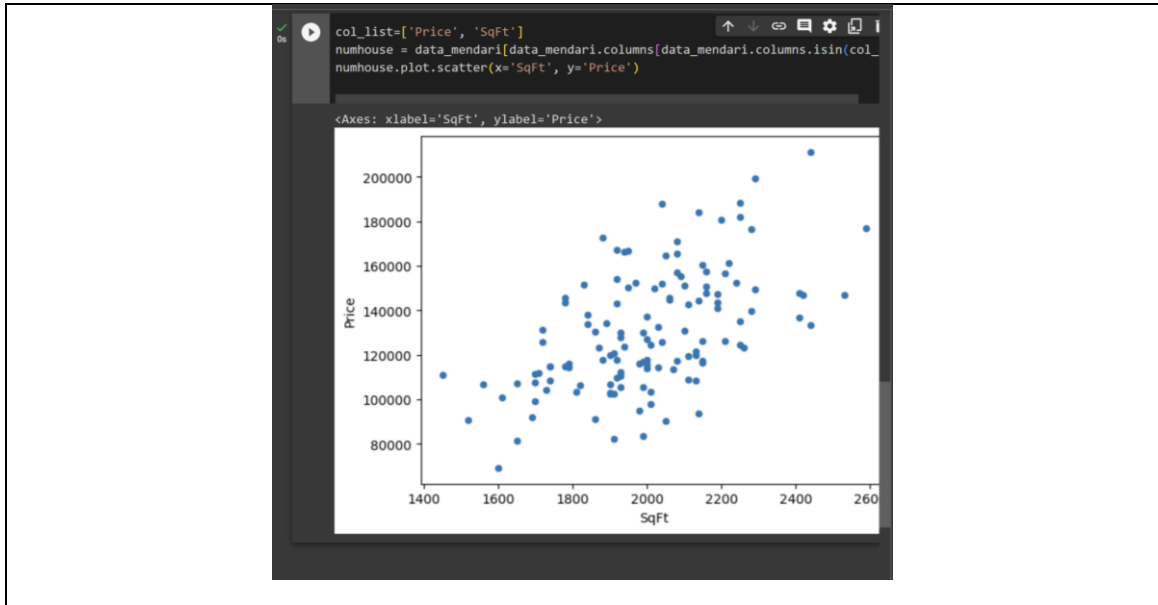
Output:



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:

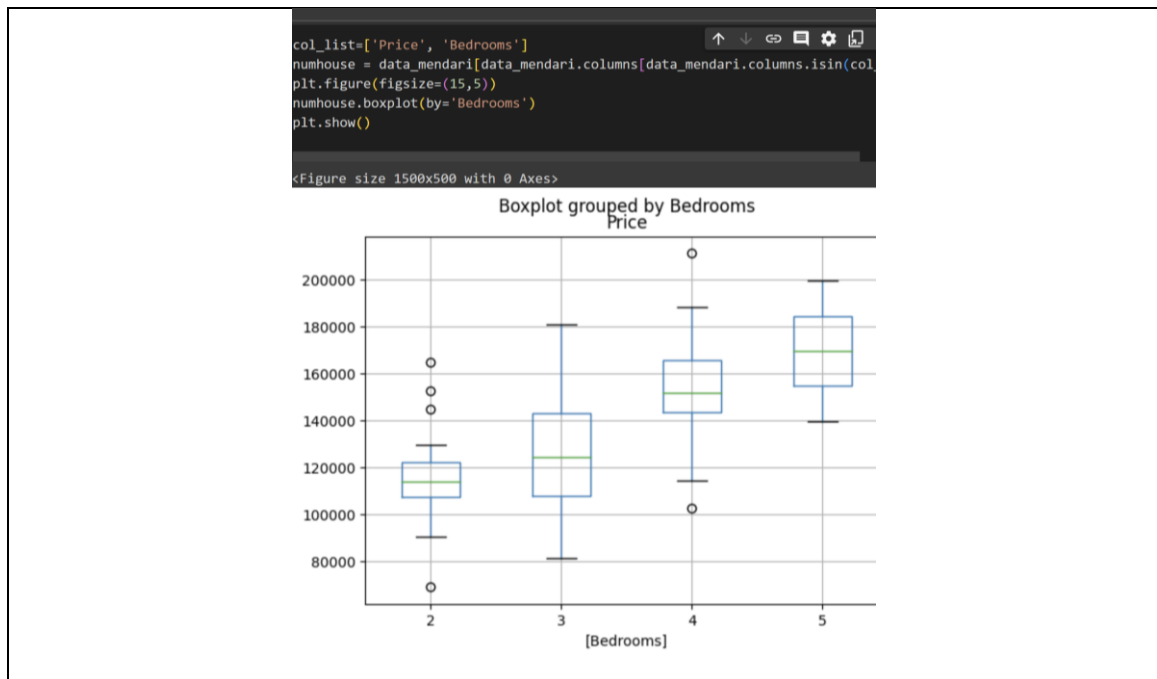


7. Group Boxplot Berdasarkan Bedrooms

```
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:





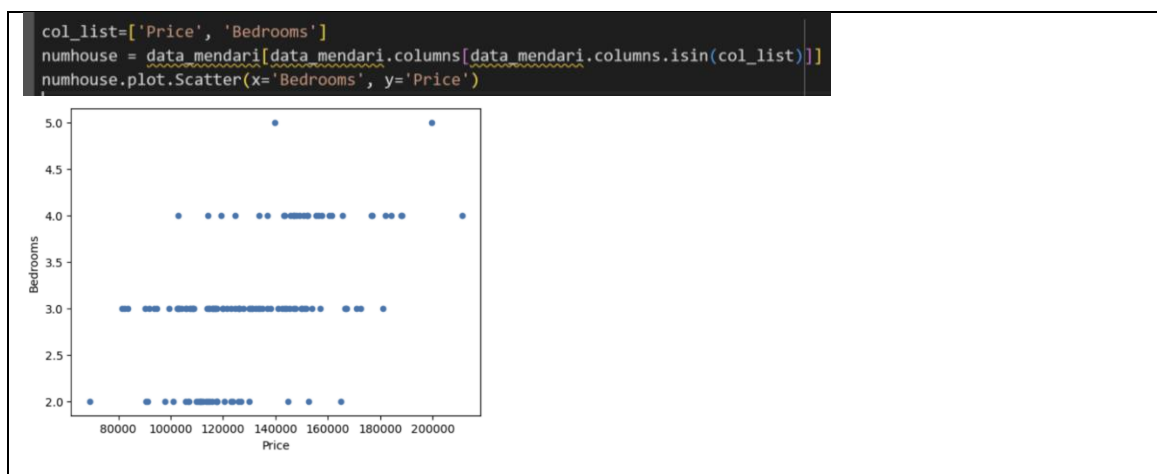
b. Latihan Kedua – Tugas

1. Buatlah Scatter Plot Harga Rumah Berdasarkan Bedrooms!

Script:

```
col_list=['Price', 'Bedrooms']
numhouse = data_mendari[data_mendari.columns[data_mendari.columns.isin(col_list)]]
numhouse.plot.Scatter(x='Bedrooms', y='Price')
```

Output:



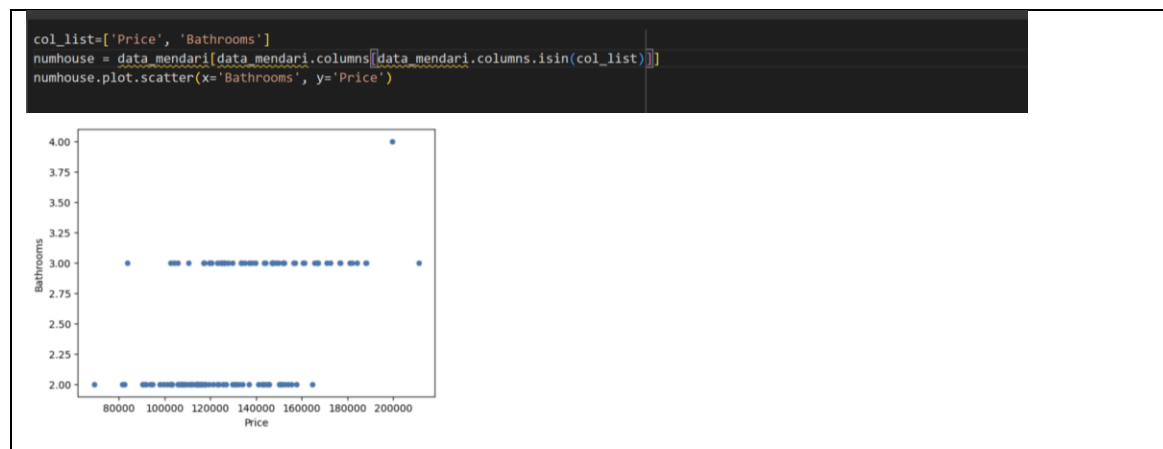
Penjelasan: Diagram di atas yaitu Scatter Plot harga rumah/Price berdasarkan jumlah kamar tidurnya (Bedrooms). Area price berada di sumbu Y dan Bedrooms berada di sumbu X.

2. Buatlah Scatter Plot Harga Rumah berdasarkan Bathrooms!

Script:

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

Output:



Penjelasan: ?

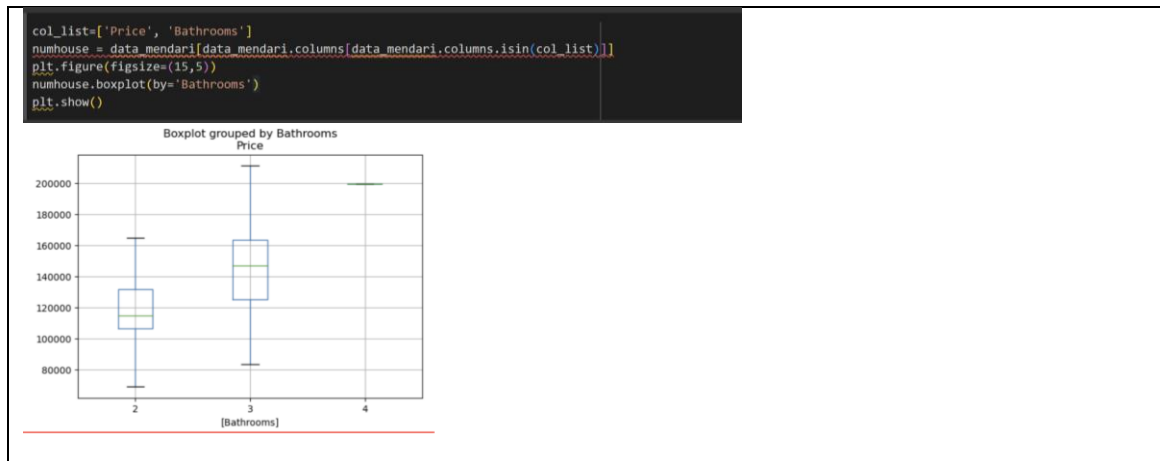
3. Buatlah Grouped Boxplot berdasarkan Bathrooms dengan Price sebagai perbandingannya!

Script:

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



Output:



Penjelasan: gambar diatas adalah Grouped Boxplot dengan membandingkan harga rumah (price) dengan jumlah kamar mandi (bathrooms).

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. Exploratory Data Analysis (EDA) adalah suatu pendekatan analisis dalam ilmu data yang digunakan untuk mengeksplorasi, menganalisis, dan memahami dataset secara mendalam. EDA bertujuan untuk mengidentifikasi pola, tren, anomali, serta hubungan antar variabel dalam dataset.
2. Exploratory Data Analysis diperlukan karena memiliki beberapa manfaat penting dalam analisis data:

- Memahami Data: EDA membantu dalam memahami karakteristik, struktur, dan kompleksitas data sebelum dilakukan analisis lebih lanjut. Ini memungkinkan analisis data untuk memiliki wawasan awal tentang data yang sedang ditangani.
- Mengidentifikasi Anomali dan Kesalahan: EDA membantu mengidentifikasi anomali atau kesalahan dalam data, seperti nilai-nilai yang hilang atau tidak masuk akal. Dengan



mengidentifikasi masalah ini, langkah-langkah perbaikan dapat diambil sebelum analisis lebih lanjut.

- Menemukan Pola dan Hubungan: Melalui visualisasi dan analisis statistik, EDA dapat membantu mengungkapkan pola, tren, dan hubungan antar variabel dalam data. Hal ini dapat membantu dalam mengambil keputusan yang lebih baik dan mengidentifikasi faktor-faktor penting.
- Pemilihan Fitur: EDA dapat membantu dalam pemilihan fitur yang paling relevan untuk analisis atau pemodelan lebih lanjut. Ini dapat mengurangi dimensi data dan memfokuskan pada fitur-fitur yang memiliki dampak signifikan.
- Penyusunan Hipotesis: Melalui EDA, Anda dapat mengembangkan hipotesis awal tentang bagaimana variabel-variabel dalam dataset dapat saling mempengaruhi. Hipotesis-hipotesis ini kemudian dapat diuji secara lebih rinci.
- Komunikasi Hasil: EDA menghasilkan visualisasi dan temuan yang mudah dipahami, sehingga memudahkan dalam berkomunikasi dengan orang lain, termasuk rekan kerja atau pemangku kepentingan lainnya.

6. Kesimpulan

- a. Dalam pengerjaan praktikum Statistika, dapat melihat bagaimana menerapkan berbagai teknik Eksplorasi Data menggunakan bahasa pemrograman Python. belajar tentang visualisasi data, analisis deskriptif, distribusi data, korelasi antar variabel, serta teknik-teknik lainnya yang membantu memahami dataset secara lebih mendalam. Praktikum ini memberi alat yang kuat untuk mengidentifikasi pola, tren, dan anomali dalam data, serta membantu dalam pengambilan keputusan yang berdasarkan fakta.
- b. Kita juga dapat mengetahui pentingnya langkah-langkah Eksplorasi Data dalam tahap awal analisis data. Dengan melakukan EDA, dapat mengidentifikasi nilai-nilai yang hilang, mengatasi outlier, memahami distribusi data, dan menemukan korelasi yang dapat membantu kami dalam mengembangkan hipotesis atau model yang lebih kompleks. Melalui visualisasi dan analisis statistik, kami dapat membuka wawasan baru tentang karakteristik dataset dan menemukan informasi berharga yang mungkin tidak terlihat pada pandangan pertama. Kesimpulannya, EDA adalah langkah penting dalam mengurai kompleksitas data dan memberikan fondasi yang kokoh untuk analisis lebih lanjut.



7. Cek List (✓)

No	Elemen Kompetensi	Penyelesaian	
		Selesai	Tidak Selesai
1.	Latihan Pertama	✓	
2.	Latihan Kedua	✓	

8. Formulir Umpan Balik

No	Elemen Kompetensi	Waktu Pengerjaan	Kriteria
1.	Latihan Pertama	20 Menit	menarik
2.	Latihan Kedua	30 Menit	menarik

Keterangan:

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