

NAMA : DINDA AULIA
NIM : 41822010062
MATKUL : Visualisasi Data Analitik I

Tugas Besar 1 Visualisasi Data

Pada tugas besar 1 ini menggunakan data iris dari UCI yang terdapat pada google drive:

Dengan nama folder data/iris.

Data iris ini adalah data dengan banyaknya data 150 terdiri dari 3 kelas dengan masing-masing kelas berisi 50 data.

Data tersebut terdiri dari 5 atribut/kolom dengan keterangan atribut seperti keterangan di bawah:

1. Panjang sepal dalam cm
2. Lebar sepal dalam cm
3. Panjang Petal dalam cm
4. Lebar Petal dalam cm
5. Kelas Atribut:

- a. Iris Setosa
- b. Iris Versicolor
- c. Iris Virginica

Berikut ini Kode untuk menentukan nama-nama kelas target dan kemungkinan sedang melakukan klasifikasi data.

```
[5]: iris.keys()

[5]: dict_keys(['data', 'target', 'frame', 'target_names', 'DESCR', 'feature_names', 'filename', 'data_module'])

[6]: dataset = pd.DataFrame(data = np.c_[iris['data']],
                           columns = iris['feature_names'])
dataset['class'] = np.c_[list(map((lambda x: iris.target_names[x]), iris.target))]

[7]: dataset
```

Ini hasil dari klasifikasi data

[7]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	class
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

150 rows x 5 columns

Keterangan : PadaTabel diatas memiliki lima kolom yang bertitel 'sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)', dan 'class'. Terdapat baris data dengan nilai numerik di bawah setiap kolom. Kelas dikategorikan menjadi 'setosa' atau 'virginica'. Tabel tersebut dilabeli memiliki '150 baris x 5 kolom'.

No 1

kode ini dioperasikan untuk menganalisis data secara statistik dengan menghitung nilai Min, Max dan korelasi antar sifat-sifat class

```
[13]: import pandas as pd

data = {
    "Min": [4.3, 2.0, 1.0, 0.1],
    "Max": [7.9, 4.4, 6.9, 2.5],
    "Mean": [5.84, 3.05, 3.76, 1.20],
    "Std Deviasi": [0.83, 0.43, 1.76, 0.76],
    "Class Correlation": [0.7826, -0.4194, 0.9490, 0.9565]}

# Daftar indeks untuk kolom atribut
index = ["Sepal Length", "Sepal Width", "Petal Length", "Petal Width"]

# Membuat DataFrame dari data
df = pd.DataFrame(data, index=index)

# Menampilkan DataFrame
print(df)

# Menyimpan DataFrame ke file CSV
df.to_csv("summary_statistics.csv")

# Jika ingin menyimpan ke file Excel, uncomment baris berikut:
# df.to_excel("summary_statistics.xlsx")
```

Ini hasil Min,Max,Mean

	Min	Max	Mean	Std Deviasi	Class Correlation
Sepal Length	4.3	7.9	5.84	0.83	0.7826
Sepal Width	2.0	4.4	3.05	0.43	-0.4194
Petal Length	1.0	6.9	3.76	1.76	0.9490
Petal Width	0.1	2.5	1.20	0.76	0.9565

Keterangan : Gambar pada tabel di atas data statistik sepatu dan petai dengan lima kolom yaitu "Min"(Menghitung nilai minimum dalam data.), "Max"(Menghitung nilai maksimum dalam data), "Rata-rata"(Menghitung rata-rata mean), "Deviasi Standar(Menghitung standar deviasi)", dan "Korelasi"(Mengambil nilai korelasi dari data class).

No 2

```
# Import necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt

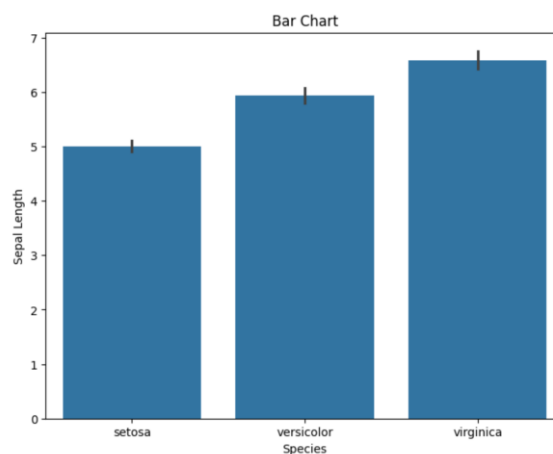
# Load the Iris dataset
iris = sns.load_dataset('iris')

# Create a bar plot using Seaborn
plt.figure(figsize=(8, 6))
sns.barplot(x='species', y='sepal_length', data=iris)

# Add title and Labels
plt.title('Bar Chart')
plt.xlabel('Species')
plt.ylabel('Sepal Length')

# Show the plot
plt.show()
```

Code ini menggunakan library seaborn dan matplotlib. Sns digunakan untuk mengambil dan mengelompokkan data, sedangkan plt untuk visualisasi. dan Bar plot ditampilkan menggunakan plt.show.



Keterangan: Gambar grafik batang yang membandingkan panjang tiga jenis iris: setosa, versicolor, dan virginica. setosa mencapai nilai 5 pada sumbu Y, versicolor 6, dan virginica tertinggi 7. Sumbu Y bernilai "Panjang " skala 0-7 selang 1. Semua batang biru gelap dengan garis kesalahan menunjukkan variasi. Grafik ini dapat membandingkan panjang kelopak tiga iris tanpa melihat gambar.

No 3

```
[2]: # Import necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt

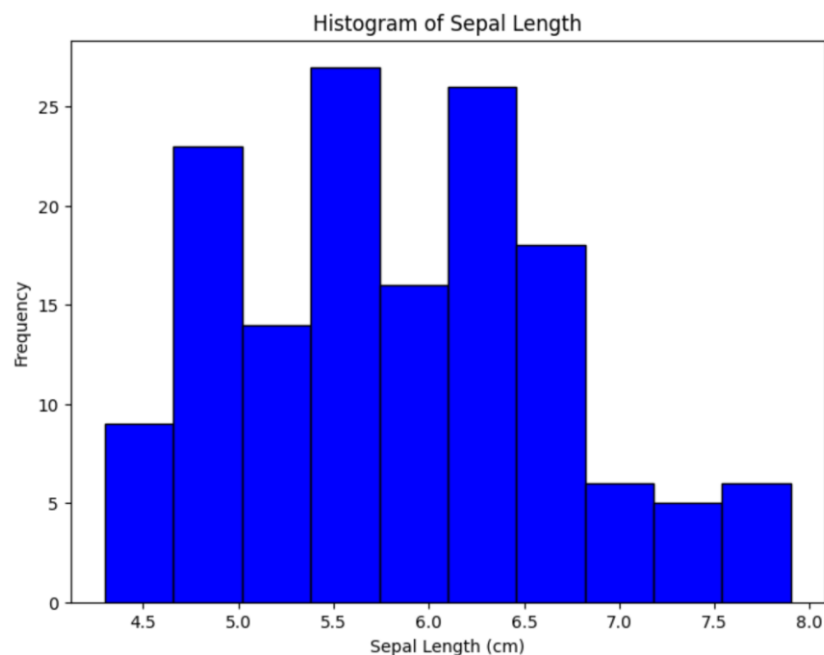
# Load the Iris dataset
iris = sns.load_dataset('iris')

# Create a histogram for the 'sepal_length' column
plt.figure(figsize=(8, 6))
plt.hist(iris['sepal_length'], bins=10, color='blue', edgecolor='black')

# Add title and Labels
plt.title('Histogram of Sepal Length')
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Frequency')

# Show the plot
plt.show()
```

Code python untuk bar plot data iris dengan seaborn dan matplotlib. Pertama import library, lalu data iris dimuat dan disimpan. Digunakan fungsi plt.figure, plt.barplot untuk membuat plot dengan sumbu x spesies dan y panjang kelopak. Lalu tambah label dan judul dengan fungsi plt. Plot ditampilkan dengan plt.show sebagai perbandingan rata-rata panjang kelopak tiga iris.



Keterangan: Gambar diatas menunjukkan distribusi frekuensi panjang sepal tumbuhan dalam kisaran 4,5 hingga 8 cm. Panjang sepal yang paling umum diukur adalah sekitar 6 cm, yang ditandai dengan frekuensi tertinggi. Semakin jauh dari 6 cm, baik ke atas

maupun ke bawah, frekuensinya semakin rendah. Kisaran panjang sepal yang paling sering diukur adalah antara 5 hingga 7 cm. Panjang sepal yang jarang tercatat adalah 4,5-5 cm dan 7,5-8 cm.

No 4

```
[2]: # Import necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Iris dataset
iris = sns.load_dataset('iris')

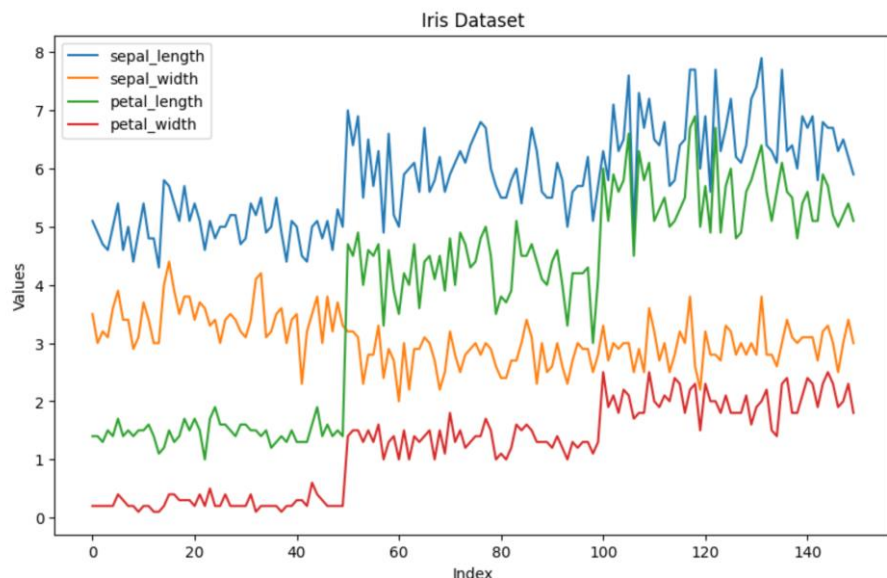
# Plot multiple lines for sepal_length, sepal_width, petal_length, and petal_width
plt.figure(figsize=(10, 6))
plt.plot(iris.index, iris['sepal_length'], label='sepal_length')
plt.plot(iris.index, iris['sepal_width'], label='sepal_width')
plt.plot(iris.index, iris['petal_length'], label='petal_length')
plt.plot(iris.index, iris['petal_width'], label='petal_width')

# Add title and Labels
plt.title('Iris Dataset')
plt.xlabel('Index')
plt.ylabel('Values')

# Show the Legend
plt.legend()

# Display the plot
plt.show()
```

Plot ditampilkan dengan perintah `plt.show()`. Kode untuk visualisasi data dengan plot multilines dari dataset Iris.



Keterangan : Gambar menunjukkan grafik hubungan dua variabel. Sumbu x skala 0-10, sumbu y skala 0-5. Terdapat titik data dua variabel. Secara umum hubungan antar variabel semakin besar x maka y juga besar. Ada variasi beberapa titik. Grafik bentuk linier meski tidak sepenuhnya.

No 5

```
[3]: import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

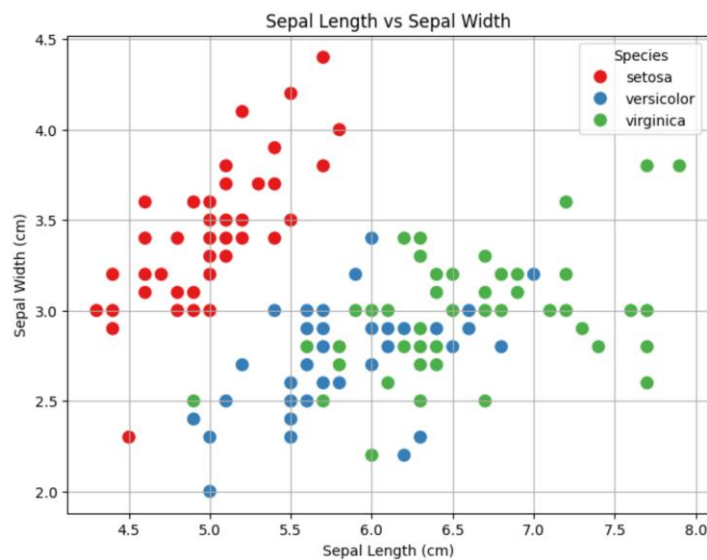
# Load the dataset (Iris dataset)
url = "https://raw.githubusercontent.com/mwaskom/seaborn-data/master/iris.csv"
data = pd.read_csv(url)

# Create a scatter plot using seaborn
plt.figure(figsize=(8, 6))
sns.scatterplot(x='sepal_length', y='sepal_width', hue='species', data=data, palette='Set1', s=100)

# Customize the plot
plt.title('Sepal Length vs Sepal Width')
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Sepal Width (cm)')
plt.legend(title='Species')
plt.grid(True)

# Show the plot
plt.show()
```

Code ini untuk Memuat data dari URL menggunakan `pandas.read_csv()`, Membuat plot scatter menggunakan `matplotlib` dan `seaborn`, Mencantumkan judul, label sumbu, dan legenda pada plot, Menampilkan hasil plot dengan perintah `plt.show()`.



Keterangan : Ada tiga jenis tanaman diwakili titik beda warna: merah, biru, hijau. Titik merah mewakili jenis setosa, biru untuk versicolor, hijau virginica dan Plot berjudul "Panjang vs Lebar Kelopak" dengan sumbu-x panjang kelopak dan y lebar kelopak, Plot berguna untuk melihat hubungan variabel pada tanaman berjenis berbeda.