Midterm-2024

- Name: Devi Sri Swetha Tanuku
- Student Number: N01623362

Example1: (2 Mark)

Create a 4x3 integer array using a range between 200 and 300 such that the difference between each element in a row is 12, and the difference between corresponding elements in adjacent rows is 8.

In [ ]:

Example2: (2 Mark)

Here are two datasets representing the hours studied and scores obtained by students in a statistics course. The data is as follows:

- Hours Studied (X): (4, 5, 6, 7, 8, 5, 4, 6, 8, 5)
- Exam Scores (Y): (75, 80, 85, 90, 95, 78, 72, 88, 92, 80)

Tasks:

1. calculate the variance of the exam scores.
2. calculate the Pearson correlation coefficient between the hours studied and the exam scores.

In [13]:
```
import scipy
from scipy.stats.stats import pearsonr
import warnings
warnings.filterwarnings('ignore')

hours_studied = (4, 5, 6, 7, 8, 5, 4, 6, 8, 5)
exam_scores = (75, 80, 85, 90, 95, 78, 72, 88, 92, 80)

pearsonr_coefficient, p_value = pearsonr(hours_studied, exam_scores)
print('PeasonR Correlation Coefficient %0.3f'% (pearsonr_coefficient))
```

```
PeasonR Correlation Coefficient 0.973
```

Example3:(3 Mark)

First, create a NumPy array named data1 with 10 random integers between 10 and 100. Create additional datasets: data2: 10 random float values between 0 and 1.

Create a 1x2 grid of subplots where:

The first subplot shows a line plot of data1. The second subplot displays a scatter plot of data2.

In [111…
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

data1 = np.random.randint(10,100,10)
data1
```

Out[111…
```
array([72, 36, 30, 35, 10, 83, 78, 85, 38, 39])
```

In [113…
```
data2 = np.random.rand(10)
data2
```

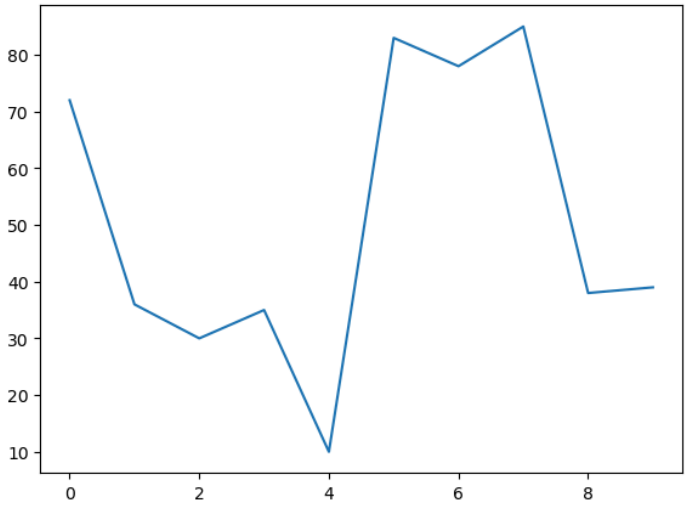
Out[113…
```
array([0.77111277, 0.73478229, 0.66693925, 0.02331512, 0.65190153,
       0.08996992, 0.60874632, 0.34348601, 0.3949714 , 0.87761603])
```

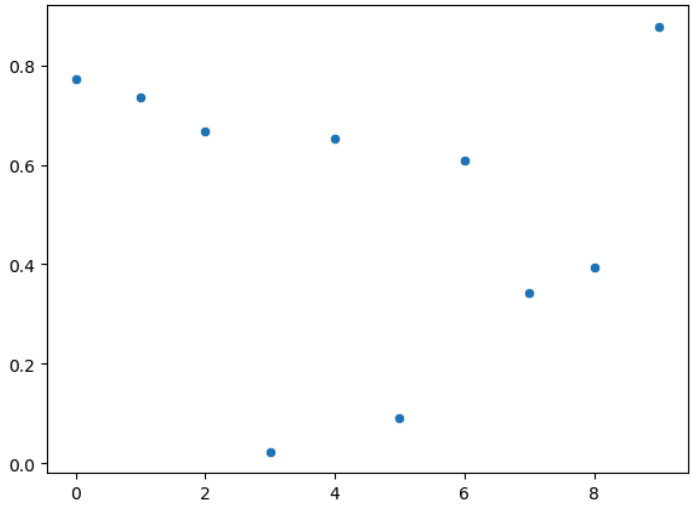
In [119…
```
plt.figure(figsize=(15,5))

plt.subplot(1,2,1)
sns.lineplot(data = data1)

plt.subplot(1,2,2)
sns.scatterplot(data = data2)
```

Out[119…
```
<Axes: >
```





Example4: (8 Mark)

1. Read a CSV dataset and create a DataFrame from it.
2. Check Missing values
3. Check Duplicates
4. Check data type
5. Calculate the average annual income and average spending score for all customers.
6. Analyze the distribution of customers' age and visualize it using a histogram. Are there any patterns or trends?
7. Determine whether there is any correlation between annual income and spending score. Provide statistical evidence.
8. Create a scatter plot to visualize the relationship between age and spending score. Are there any insights to be gained from this plot?
9. Identify the gender distribution of customers in the mall and visualize it using a pie chart.

In [125…
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

df = pd.read_csv("D:\\AIML\\5000 0NB_Data Analytics\\midterm\Mall_Customers.csv")
df
```

Out[125…

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

In [129…
```
df.dtypes
```

Out[129…
```
CustomerID                int64
Gender                   object
Age                       int64
Annual Income (k$)        int64
Spending Score (1-100)    int64
dtype: object
```

In [133…
```
df_missing = pd.read_csv("D:\\AIML\\5000 0NB_Data Analytics\\midterm\Mall_Customers.csv", na_values= ["NaN"])
df_missing
```

Out[133…

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

In [135…
```
df_missing.dtypes
```

Out[135…
```
CustomerID                int64
Gender                   object
Age                       int64
Annual Income (k$)        int64
Spending Score (1-100)    int64
dtype: object
```

In [149…
```
annualIncome= df["Annual Income (k$)"].mean()
annualIncome
```

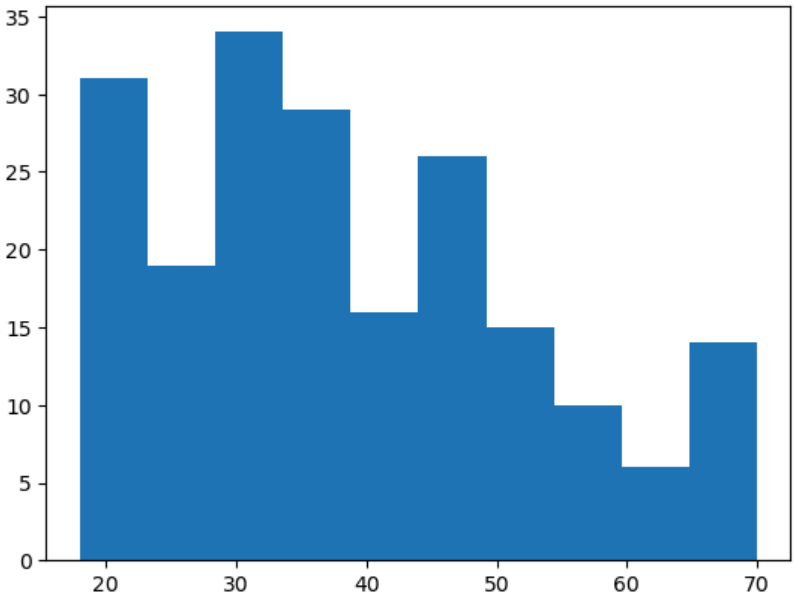
Out[149…
```
60.56
```

In [151…
```
annual_spending= df["Spending Score (1-100)"].mean()
annual_spending
```

Out[151…
```
50.2
```

In [155…
```
age = df["Age"]
plt.hist(age)
plt.plot
```

Out[155…
```
<function matplotlib.pyplot.plot(*args: 'float | ArrayLike | str', scalex: 'bool' = True, scaley: 'bool' = True, data=None, **kwargs)
-> 'list[Line2D]'>
```



There is no pattern or trend

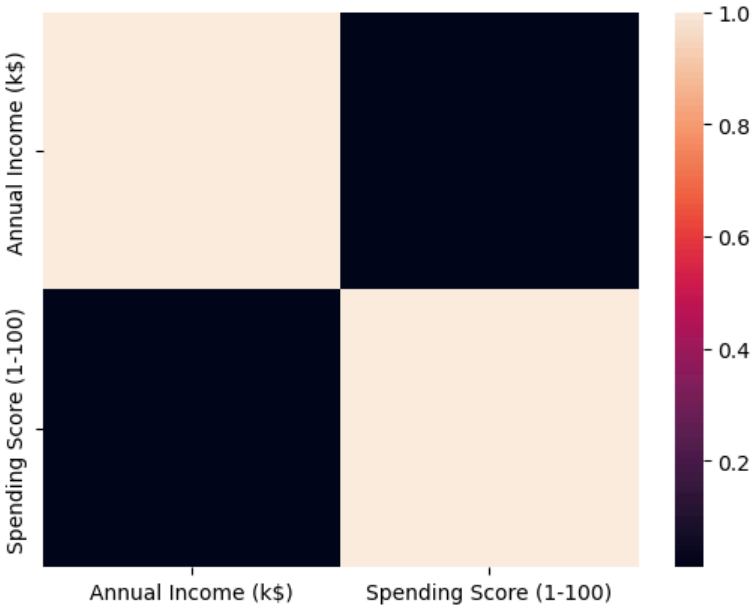
In [160…
```
x = df[["Annual Income (k$)","Spending Score (1-100)"]]
x.corr()
```

Out[160…

	Annual Income (k$)	Spending Score (1-100)
Annual Income (k$)	1.000000	0.009903
Spending Score (1-100)	0.009903	1.000000

In [162…
```
sns.heatmap(x.corr())
```

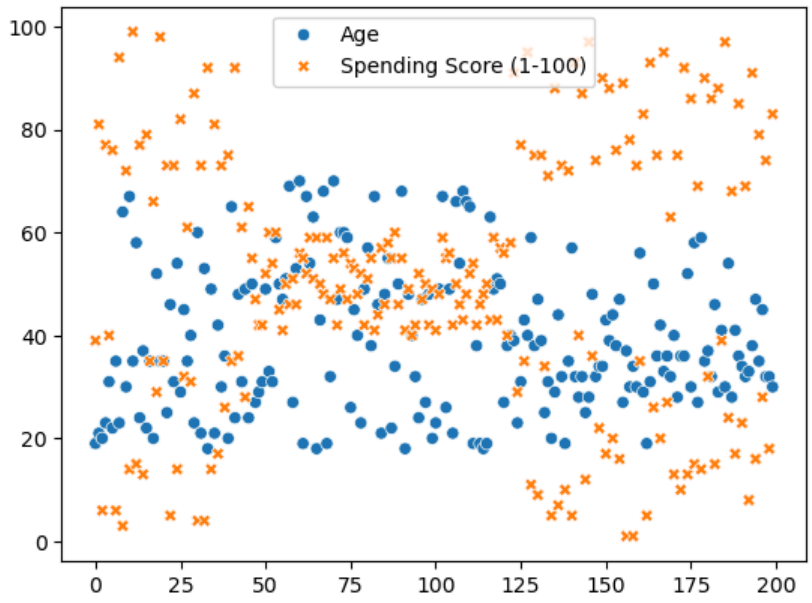
Out[162… <Axes: >



In [ ]:

In [166…
```
y = df[["Age","Spending Score (1-100)"]]
sns.scatterplot(y)
```

Out[166… <Axes: >



In [ ]: