```
from google.colab import files
uploaded = files.upload()
```

Choose Files Mall_Customers.csv

- **Mall_Customers.csv**(text/csv) - 3981 bytes, last modified: 4/21/2025 - 100% done

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid")

df = pd.read_csv('Mall_Customers.csv')
df.head()
```

	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

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
None
```

```
print(df.describe())
```

```
       CustomerID         Age  Annual Income (k$)  Spending Score (1-100)
count  200.000000  200.000000          200.000000              200.000000
mean   100.500000   38.850000           60.560000               50.200000
std     57.879185   13.969007           26.264721               25.823522
min      1.000000   18.000000           15.000000                1.000000
25%     50.750000   28.750000           41.500000               34.750000
50%    100.500000   36.000000           61.500000               50.000000
75%    150.250000   49.000000           78.000000               73.000000
max    200.000000   70.000000          137.000000               99.000000
```

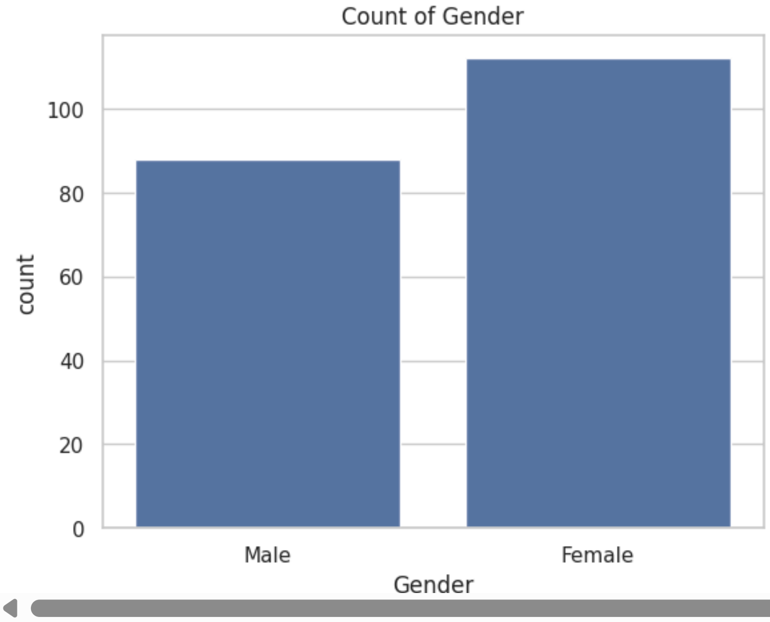
```
print(df.isnull().sum())
```

```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

```
print(df['Gender'].value_counts())
```

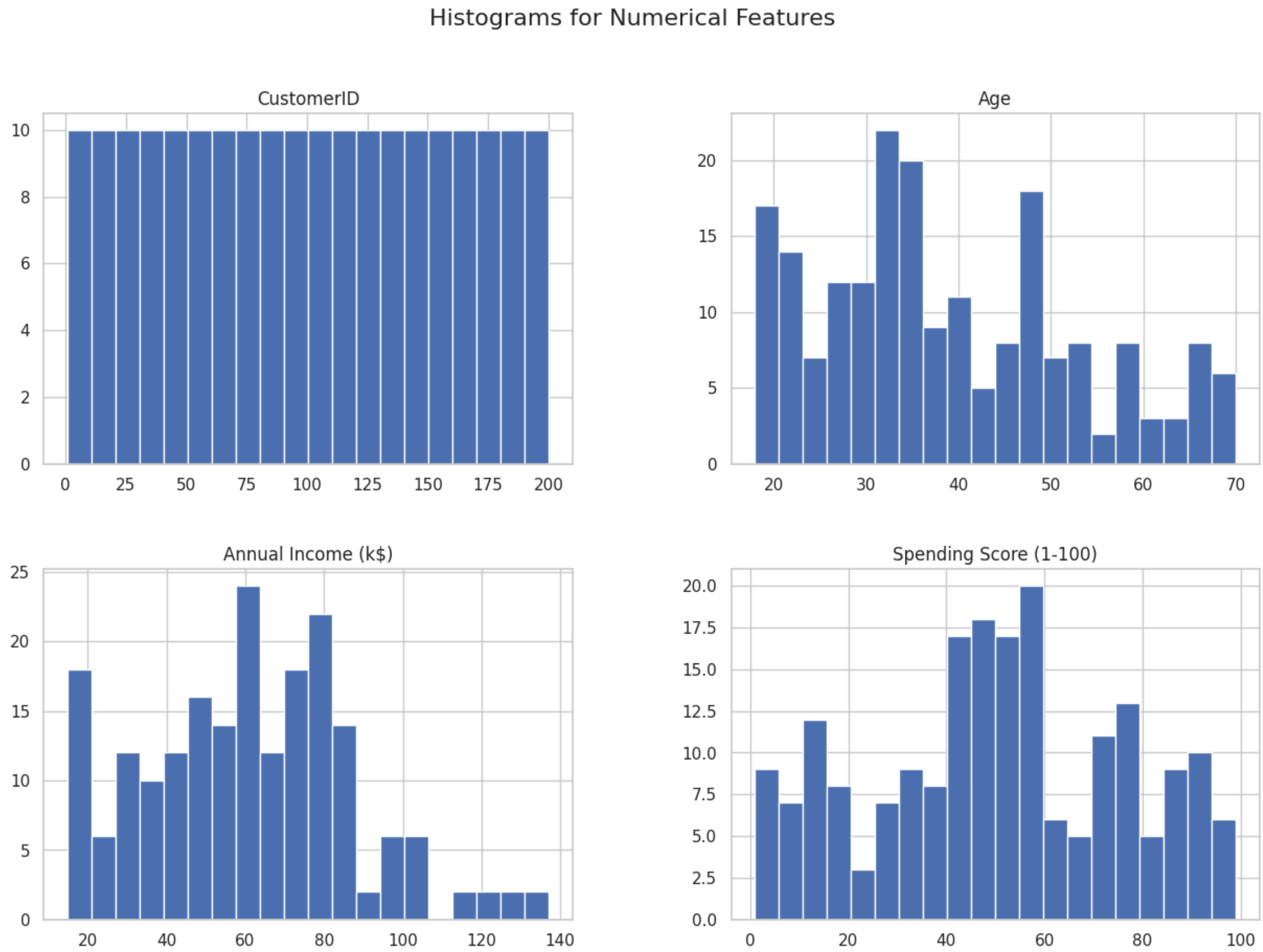
```
Gender
Female    112
Male       88
Name: count, dtype: int64
```

```
sns.countplot(x='Gender', data=df)
plt.title('Count of Gender')
plt.show()
```



Observation: The dataset has a nearly balanced gender distribution, but there are slightly more Female customers than Male customers.

```
# Histograms
df.hist(figsize=(15,10), bins=20)
plt.suptitle('Histograms for Numerical Features', fontsize=16)
plt.show()
```



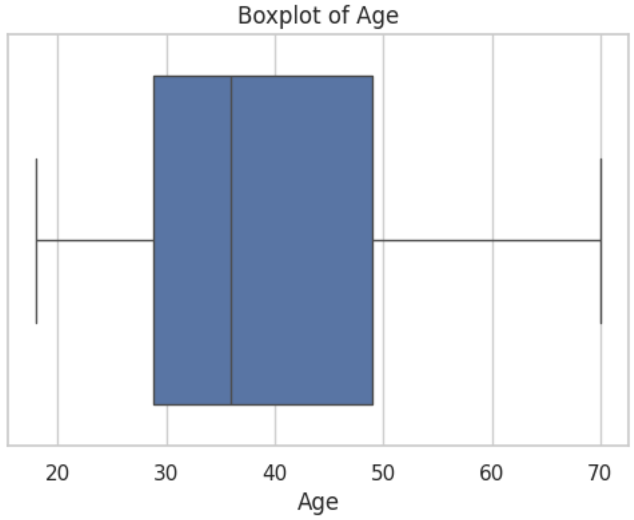
Observation:

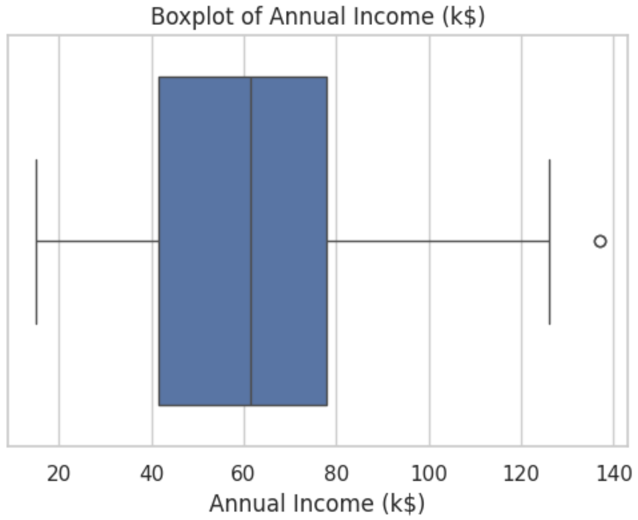
Age is somewhat right-skewed, with most customers between 20–40 years old.

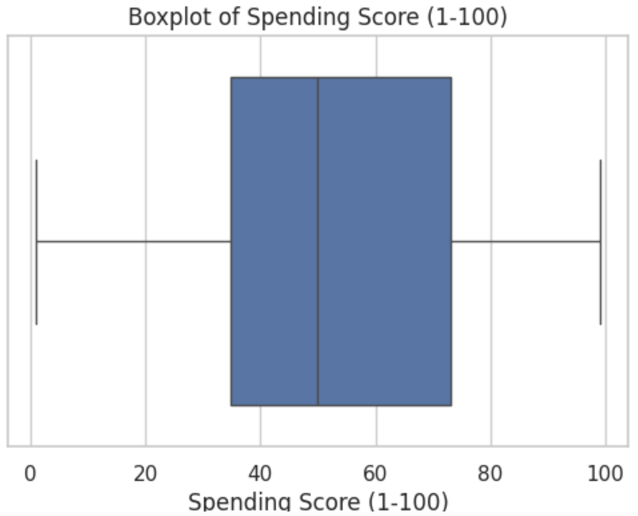
Annual Income distribution appears fairly uniform with some concentration around 40k–80k.

Spending Score shows two peaks: one at the lower end and one at the higher end, suggesting two major customer groups.

```
# Boxplots
for col in ['Age', 'Annual Income (k$)', 'Spending Score (1-100)']:
    plt.figure(figsize=(6,4))
    sns.boxplot(x=df[col])
    plt.title(f'Boxplot of {col}')
    plt.show()
```







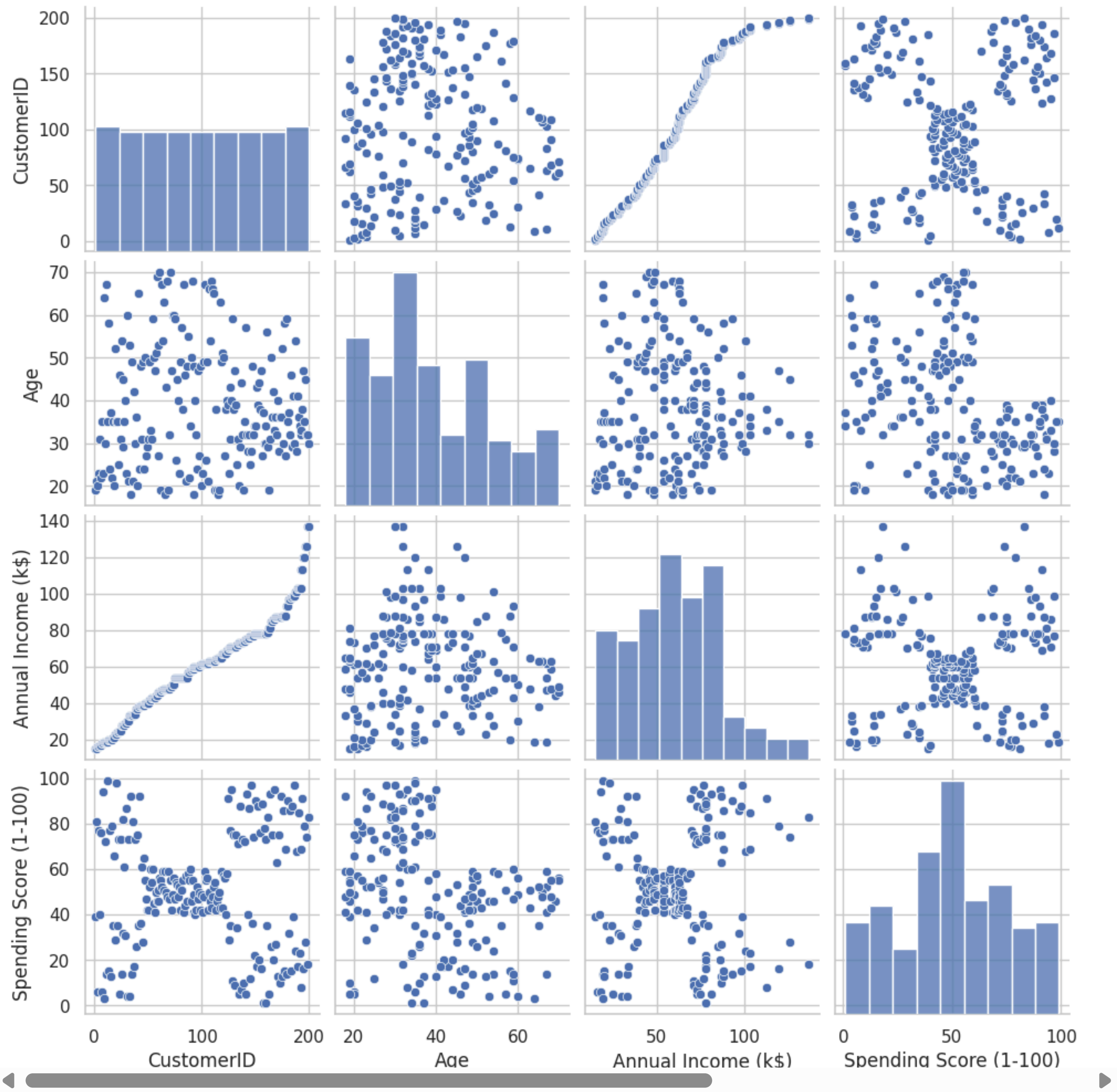
Observation:

Age boxplot shows no significant outliers.

Annual Income boxplot shows a wider spread but no extreme outliers.

Spending Score shows some variation but no serious outliers. Overall, the data appears clean without extreme anomalies.

```
sns.pairplot(df)
plt.show()
```



Observation:

We can visually confirm the earlier findings: clusters exist in the Spending Score vs. Annual Income relationship.

Gender does not show a strong pattern in Age, Income, or Spending when viewed individually.

```
# Scatter plot between Annual Income and Spending Score
plt.figure(figsize=(8,6))
sns.scatterplot(x='Annual Income (k$)', y='Spending Score (1-100)', data=df, hue='Gender')
plt.title('Annual Income vs Spending Score by Gender')
plt.show()
```

Observation:

Customers are visibly divided into distinct groups:

High income, low spending.

Low income, high spending.

Middle range clusters.

Some high-income customers do not spend much, and vice versa — suggesting different customer behaviors.

Final Summary: The Mall Customers dataset offers key insights into customer demographics and spending behavior:

The gender distribution is almost even, with a slight female dominance.

Most customers are aged between 20 and 40 years.

Spending behavior is not strongly related to income or age.

There are clear clusters of customer groups based on their Annual Income and Spending Score, which may be helpful for customer segmentation and targeted marketing.

No significant outliers were detected, and the dataset is relatively clean.

Further clustering techniques (like K-Means) could be applied to segment the customers more precisely based on their characteristics.