```
import pandas as pd
```

```
df = pd.read_csv("Mall.csv")
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
/tmp/ipython-input-1088209665.py in <cell line: 0>()
----> 1 df = pd.read_csv("Mall.csv")

                                   4 frames
/usr/local/lib/python3.12/dist-packages/pandas/io/common.py in get_handle(path_or_buf, mode, encoding, compression, memory_map, is_text, errors, storage_options)
    871         if ioargs.encoding and "b" not in ioargs.mode:
    872             # Encoding
--> 873             handle = open(
    874                 handle,
    875                 ioargs.mode,

FileNotFoundError: [Errno 2] No such file or directory: 'Mall.csv'
```

Next steps: Explain error

```
import pandas as pd

df = pd.read_csv(r"C:\Users\hrith\Downloads\Mall.csv")
print(df.head())
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
/tmp/ipython-input-1407374617.py in <cell line: 0>()
      1 import pandas as pd
      2
----> 3 df = pd.read_csv(r"C:\Users\hrith\Downloads\Mall.csv")
      4 print(df.head())

                                   4 frames
/usr/local/lib/python3.12/dist-packages/pandas/io/common.py in get_handle(path_or_buf, mode, encoding, compression, memory_map, is_text, errors, storage_options)
    871         if ioargs.encoding and "b" not in ioargs.mode:
    872             # Encoding
--> 873             handle = open(
    874                 handle,
    875                 ioargs.mode,

FileNotFoundError: [Errno 2] No such file or directory: 'C:\\Users\\hrith\\Downloads\\Mall.csv'
```

Next steps: Explain error

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

Choose files Mall.csv

- **Mall.csv**(text/csv) - 3981 bytes, last modified: 03/09/2025 - 100% done

```
Saving Mall.csv to Mall.csv
```

```
import pandas as pd

df = pd.read_csv("Mall.csv")
print(df.head())
print(df.info())
```

```
   CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
0           1    Male   19                  15                      39
1           2    Male   21                  15                      81
2           3  Female   20                  16                       6
3           4  Female   23                  16                      77
4           5  Female   31                  17                      40
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              200 non-null    int64 
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64 
 3   Annual Income (k$)      200 non-null    int64 
 4   Spending Score (1-100)  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
None
```

```
import pandas as pd

df = pd.read_csv("Mall.csv")

print("Shape of data:", df.shape)   # rows × columns
print("\nData Info:")
print(df.info())
print("\nFirst 5 Rows:")
print(df.head())
```

```
Shape of data: (200, 5)

Data 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

First 5 Rows:
   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
```

```
print("Missing values per column:")
print(df.isnull().sum())
```

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

```
print("Duplicate rows:", df.duplicated().sum())
```

```
Duplicate rows: 0
```

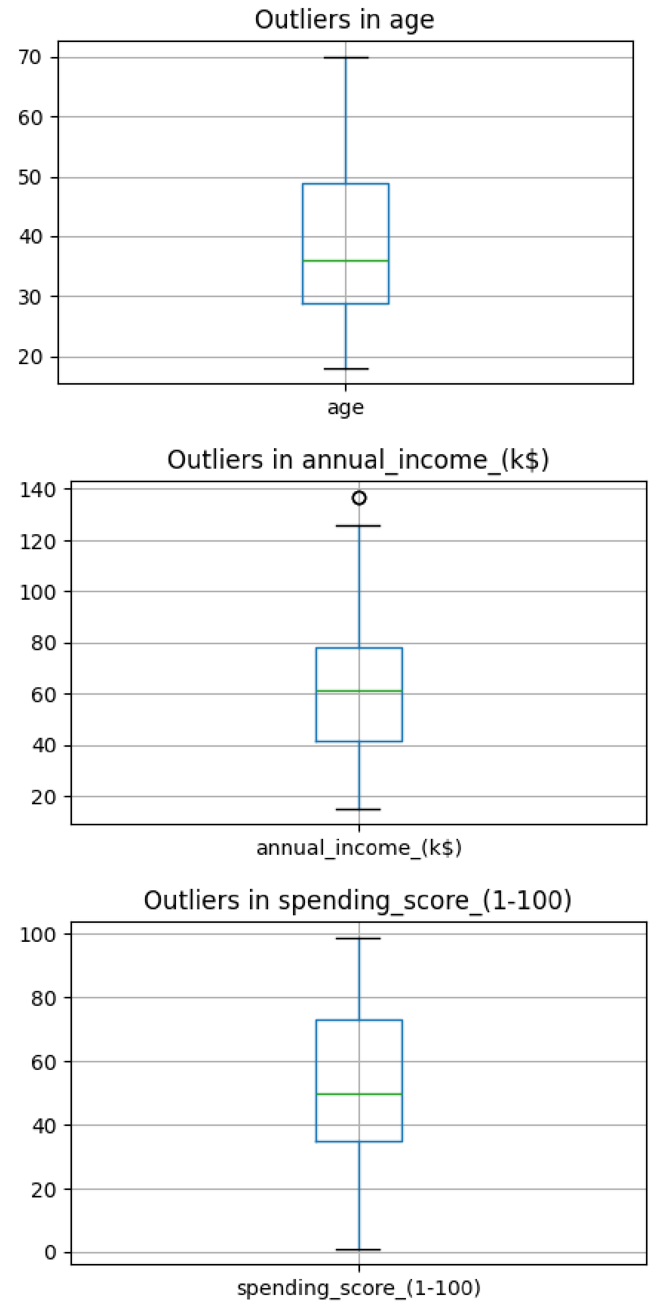
```
df.columns = df.columns.str.strip().str.lower().str.replace(" ", "_")

print("Cleaned column names:", df.columns)
```

```
Cleaned column names: Index(['customerid', 'gender', 'age', 'annual_income_(k$)',
       'spending_score_(1-100)'],
      dtype='object')
```

```
df['gender'] = df['gender'].str.strip().str.capitalize()

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

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

```
import matplotlib.pyplot as plt

num_cols = ['age', 'annual_income_(k$)', 'spending_score_(1-100)']

for col in num_cols:
    plt.figure(figsize=(5,3))
    df.boxplot(column=col)
    plt.title(f"Outliers in {col}")
    plt.show()
```



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

```
<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$)      200 non-null    int64 
 4   spending_score_(1-100)  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
None
   customerid  gender  age  annual_income_(k$)  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
```

```
df['gender'] = df['gender'].map({'Male': 0, 'Female': 1})
print(df['gender'].value_counts())
```

```
Series([], Name: count, dtype: int64)
```

```
print(df.columns)
```

```
Index(['gender', 'age', 'annual_income_(k$)', 'spending_score_(1-100)'], dtype='object')
```

```
df = pd.get_dummies(df, columns=['gender'])
```

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
numeric_cols = ['age', 'annual_income_(k$)', 'spending_score_(1-100)']

df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
```