Assignment No:08 - Assignment based on Exploratory Data Analysis (EDA) using pandas in Python

Part-A

In [1]:
```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset (CSV file)
# For demonstration, we will use a sample dataset
url = "/Users/umeshsangule/Downloads/tips.csv"
df = pd.read_csv(url)

# Display first 5 rows
print("First 5 rows of the dataset:")
print(df.head(10))
```

```
First 5 rows of the dataset:
   total_bill   tip     sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun  Dinner     2
1       10.34  1.66    Male     No  Sun  Dinner     3
2       21.01  3.50    Male     No  Sun  Dinner     3
3       23.68  3.31    Male     No  Sun  Dinner     2
4       24.59  3.61  Female     No  Sun  Dinner     4
5       25.29  4.71    Male     No  Sun  Dinner     4
6        8.77  2.00    Male     No  Sun  Dinner     2
7       26.88  3.12    Male     No  Sun  Dinner     4
8       15.04  1.96    Male     No  Sun  Dinner     2
9       14.78  3.23    Male     No  Sun  Dinner     2
```

In [2]:
```
# Shape of dataset
print("\nShape of dataset (rows, columns):", df.shape)

# Column names
print("\nColumn names:", df.columns)

# Data types of each column
print("\nData types:")
print(df.dtypes)

# Info summary
print("\nDataset Info:")
print(df.info())
```

```
Shape of dataset (rows, columns): (244, 7)

Column names: Index(['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'siz
e'], dtype='object')

Data types:
total_bill    float64
tip           float64
sex            object
smoker         object
day            object
time           object
size            int64
dtype: object

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    object
 3   smoker      244 non-null    object
 4   day         244 non-null    object
 5   time        244 non-null    object
 6   size        244 non-null    int64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.5+ KB
None
```

In [3]:

```python
# Summary statistics for numerical columns
print("\nDescriptive statistics:")
print(df.describe())

# Summary for categorical columns
print("\nCategorical columns summary:")
print(df.describe(include='object'))
```

```
Descriptive statistics:
       total_bill         tip        size
count  244.000000  244.000000  244.000000
mean    19.785943    2.998279    2.569672
std      8.902412    1.383638    0.951100
min      3.070000    1.000000    1.000000
25%     13.347500    2.000000    2.000000
50%     17.795000    2.900000    2.000000
75%     24.127500    3.562500    3.000000
max     50.810000   10.000000    6.000000

Categorical columns summary:
         sex smoker  day    time
count    244    244  244     244
unique     2      2    4       2
top     Male     No  Sat  Dinner
freq     157    151   87     176
```

In [4]:
```python
# Check missing values
print("\nMissing values in each column:")
print(df.isnull().sum())
```

```
Missing values in each column:
total_bill    0
tip           0
sex           0
smoker        0
day           0
time          0
size          0
dtype: int64
```

In [5]:
```python
# Display last 5 rows
print("\nLast 5 rows of the dataset:")
print(df.tail())

# Random sample of 5 rows
print("\nRandom 5 rows:")
print(df.sample(5))
```

```
Last 5 rows of the dataset:
     total_bill   tip     sex smoker   day    time  size
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2

Random 5 rows:
     total_bill   tip     sex smoker   day    time  size
155       29.85  5.14  Female     No   Sun  Dinner     5
95        40.17  4.73    Male    Yes   Fri  Dinner     4
35        24.06  3.60    Male     No   Sat  Dinner     3
194       16.58  4.00    Male    Yes  Thur   Lunch     2
96        27.28  4.00    Male    Yes   Fri  Dinner     2
```

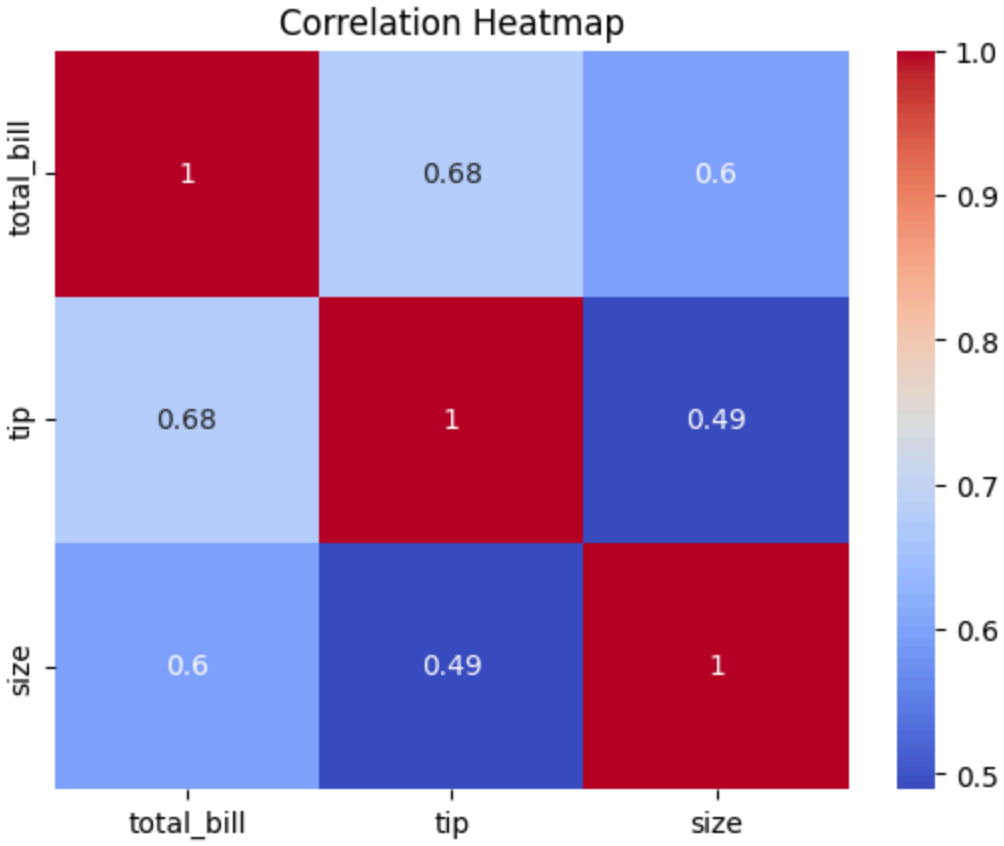
In [6]:
```
# Count of unique values in 'day' column
print("\nValue counts of 'day' column:")
print(df['day'].value_counts())

# Count of unique values in 'sex' column
print("\nValue counts of 'sex' column:")
print(df['sex'].value_counts())
```

```
Value counts of 'day' column:
Sat     87
Sun     76
Thur    62
Fri     19
Name: day, dtype: int64

Value counts of 'sex' column:
Male      157
Female     87
Name: sex, dtype: int64
```

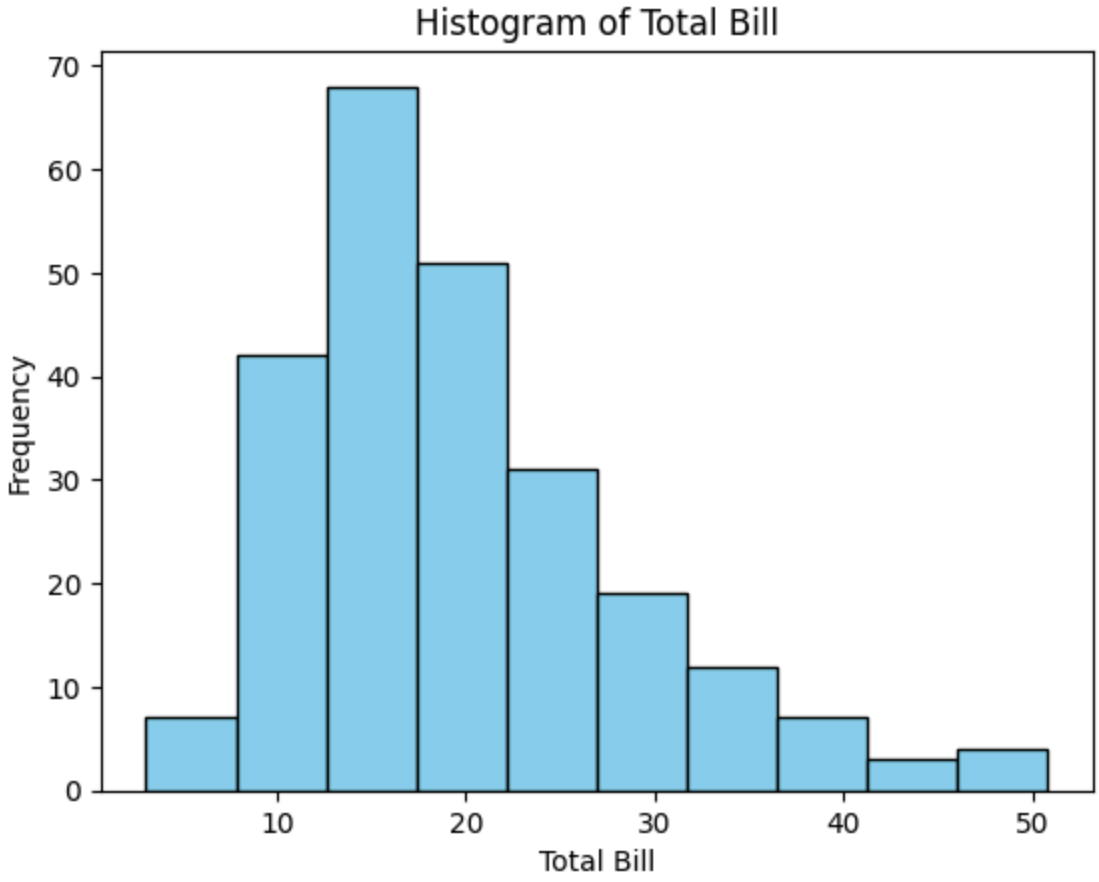
In [7]:
```
# Correlation between numerical columns
print("\nCorrelation matrix:")
print(df.corr())

# Heatmap of correlation
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

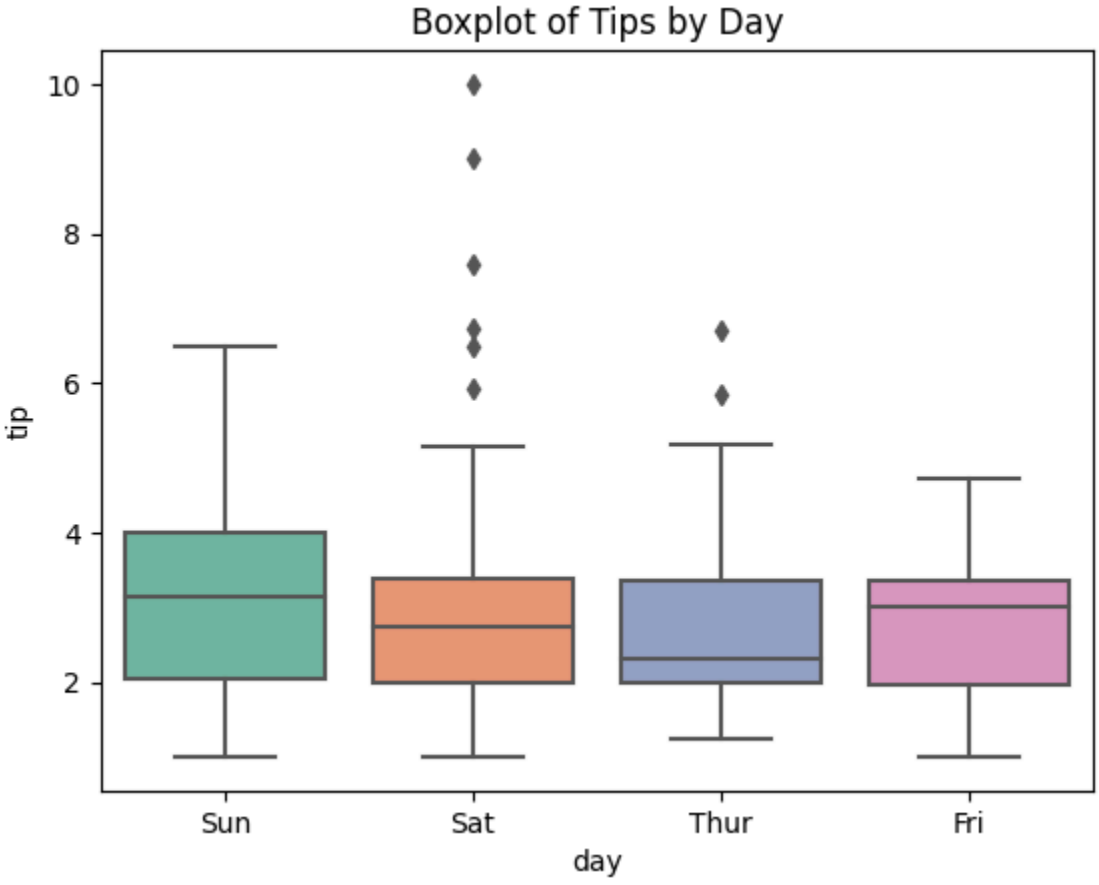
```
Correlation matrix:
            total_bill       tip      size
total_bill    1.000000  0.675734  0.598315
tip           0.675734  1.000000  0.489299
size          0.598315  0.489299  1.000000
```

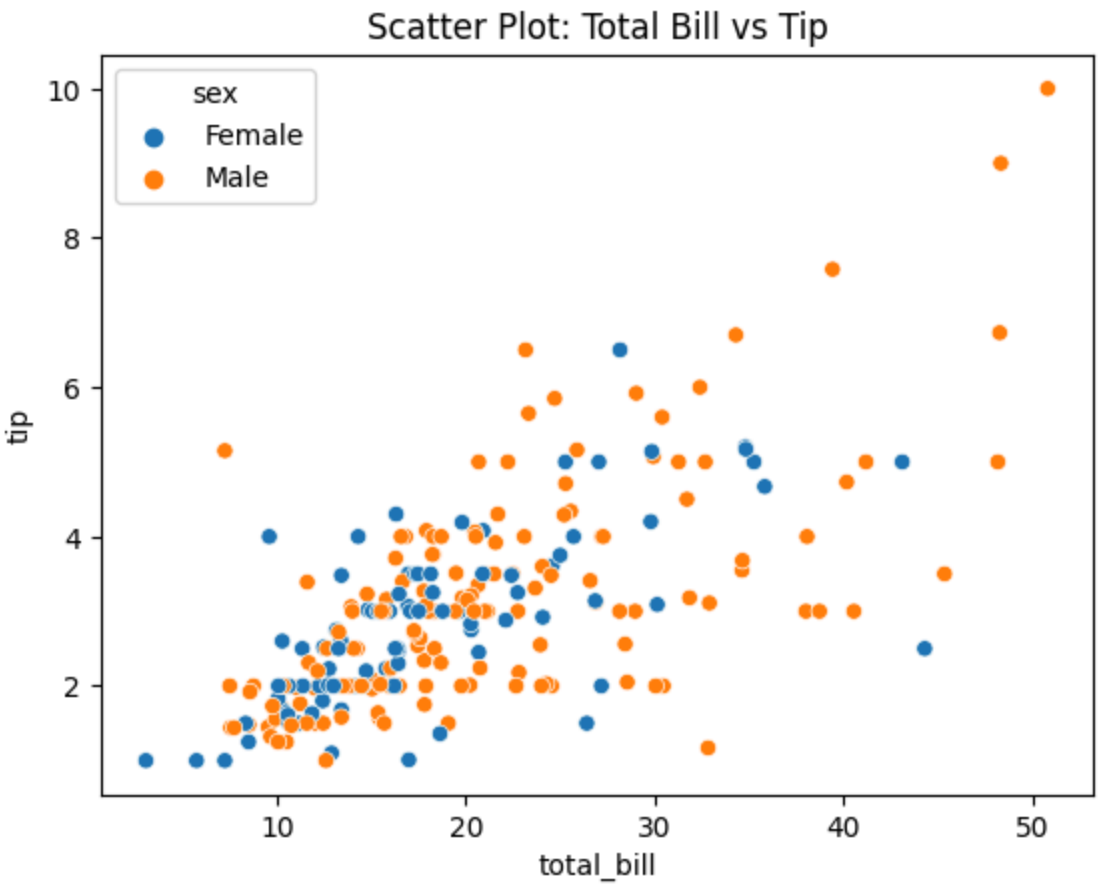
```
/var/folders/5z/nv5hpy0s20n928hd0mgd_swr0000gn/T/ipykernel_9652/1828567939.p
y:3: FutureWarning: The default value of numeric_only in DataFrame.corr is depr
ecated. In a future version, it will default to False. Select only valid column
s or specify the value of numeric_only to silence this warning.
  print(df.corr())
/var/folders/5z/nv5hpy0s20n928hd0mgd_swr0000gn/T/ipykernel_9652/1828567939.p
y:6: FutureWarning: The default value of numeric_only in DataFrame.corr is depr
ecated. In a future version, it will default to False. Select only valid column
s or specify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
```



In [8]:
```
# Histogram of 'total_bill'
plt.hist(df['total_bill'], bins=10, color='skyblue', edgecolor='black')
plt.title("Histogram of Total Bill")
plt.xlabel("Total Bill")
plt.ylabel("Frequency")
plt.show()

# Boxplot of 'tip' by 'day'
sns.boxplot(x='day', y='tip', data=df, palette='Set2')
plt.title("Boxplot of Tips by Day")
plt.show()

# Scatter plot between 'total_bill' and 'tip'
sns.scatterplot(x='total_bill', y='tip', data=df, hue='sex')
plt.title("Scatter Plot: Total Bill vs Tip")
plt.show()
```


Histogram of Total Bill
Frequency
Total Bill
0
10
20
30
40
50
60
70
10
20
30
40
50


Boxplot of Tips by Day
tip
day
2
4
6
8
10
Sun
Sat
Thur
Fri



Part-B

In [9]:
```
pip install ydata-profiling
```

```
Requirement already satisfied: ydata-profiling in /Library/Frameworks/Python.fr
amework/Versions/3.10/lib/python3.10/site-packages (4.0.0)
Requirement already satisfied: scipy<1.10,>=1.4.1 in /Library/Frameworks/Pytho
n.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(1.9.3)
Requirement already satisfied: pandas!=1.4.0,<1.6,>1.1 in /Library/Frameworks/P
ython.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profilin
g) (1.5.2)
Requirement already satisfied: matplotlib<3.7,>=3.2 in /Library/Frameworks/Pyth
on.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(3.6.2)
Requirement already satisfied: pydantic<1.11,>=1.8.1 in /Library/Frameworks/Pyt
hon.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(1.10.4)
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in /Users/umeshsangule/Librar
y/Python/3.10/lib/python/site-packages (from ydata-profiling) (6.0)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in /Users/umeshsangule/Libra
ry/Python/3.10/lib/python/site-packages (from ydata-profiling) (3.1.2)
Requirement already satisfied: visions==0.7.5 in /Library/Frameworks/Python.fra
mework/Versions/3.10/lib/python3.10/site-packages (from visions[type_image_pat
h]==0.7.5->ydata-profiling) (0.7.5)
Requirement already satisfied: numpy<1.24,>=1.16.0 in /Library/Frameworks/Pytho
n.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(1.23.5)
Requirement already satisfied: htmlmin==0.1.12 in /Library/Frameworks/Python.fr
amework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(0.1.12)
Requirement already satisfied: phik<0.13,>=0.11.1 in /Library/Frameworks/Pytho
n.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(0.12.3)
Requirement already satisfied: requests<2.29,>=2.24.0 in /Library/Frameworks/Py
thon.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profilin
g) (2.28.2)
Requirement already satisfied: tqdm<4.65,>=4.48.2 in /Library/Frameworks/Pytho
n.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(4.64.1)
Requirement already satisfied: seaborn<0.13,>=0.10.1 in /Library/Frameworks/Pyt
hon.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profiling)
(0.12.2)
Requirement already satisfied: multimethod<1.10,>=1.4 in /Library/Frameworks/Py
thon.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profilin
g) (1.9.1)
Requirement already satisfied: statsmodels<0.14,>=0.13.2 in /Library/Framework
s/Python.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profi
ling) (0.13.5)
Requirement already satisfied: typeguard<2.14,>=2.13.2 in /Library/Frameworks/P
ython.framework/Versions/3.10/lib/python3.10/site-packages (from ydata-profilin
g) (2.13.3)
Requirement already satisfied: attrs>=19.3.0 in /Library/Frameworks/Python.fram
ework/Versions/3.10/lib/python3.10/site-packages (from visions==0.7.5->vision
s[type_image_path]==0.7.5->ydata-profiling) (25.3.0)
Requirement already satisfied: networkx>=2.4 in /Library/Frameworks/Python.fram
ework/Versions/3.10/lib/python3.10/site-packages (from visions==0.7.5->vision
s[type_image_path]==0.7.5->ydata-profiling) (3.0)
```

```
Requirement already satisfied: tangled-up-in-unicode>=0.0.4 in /Library/Framewo
rks/Python.framework/Versions/3.10/lib/python3.10/site-packages (from vision
s==0.7.5->visions[type_image_path]==0.7.5->ydata-profiling) (0.2.0)
Requirement already satisfied: imagehash in /Library/Frameworks/Python.framewor
k/Versions/3.10/lib/python3.10/site-packages (from visions[type_image_pat
h]==0.7.5->ydata-profiling) (4.3.1)
Requirement already satisfied: Pillow in /Library/Frameworks/Python.framework/V
ersions/3.10/lib/python3.10/site-packages (from visions[type_image_pat
h]==0.7.5->ydata-profiling) (9.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in /Users/umeshsangule/Library/P
ython/3.10/lib/python/site-packages (from jinja2<3.2,>=2.11.1->ydata-profiling)
(2.1.1)
Requirement already satisfied: contourpy>=1.0.1 in /Library/Frameworks/Python.f
ramework/Versions/3.10/lib/python3.10/site-packages (from matplotli
b<3.7,>=3.2->ydata-profiling) (1.0.6)
Requirement already satisfied: cycler>=0.10 in /Library/Frameworks/Python.frame
work/Versions/3.10/lib/python3.10/site-packages (from matplotlib<3.7,>=3.2->yda
ta-profiling) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in /Library/Frameworks/Pytho
n.framework/Versions/3.10/lib/python3.10/site-packages (from matplotli
b<3.7,>=3.2->ydata-profiling) (4.38.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /Library/Frameworks/Pytho
n.framework/Versions/3.10/lib/python3.10/site-packages (from matplotli
b<3.7,>=3.2->ydata-profiling) (1.4.4)
Requirement already satisfied: packaging>=20.0 in /Users/umeshsangule/Library/P
ython/3.10/lib/python/site-packages (from matplotlib<3.7,>=3.2->ydata-profilin
g) (22.0)
Requirement already satisfied: pyparsing>=2.2.1 in /Library/Frameworks/Python.f
ramework/Versions/3.10/lib/python3.10/site-packages (from matplotli
b<3.7,>=3.2->ydata-profiling) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in /Library/Frameworks/Pyth
on.framework/Versions/3.10/lib/python3.10/site-packages (from matplotli
b<3.7,>=3.2->ydata-profiling) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /Library/Frameworks/Python.frame
work/Versions/3.10/lib/python3.10/site-packages (from panda
s!=1.4.0,<1.6,>1.1->ydata-profiling) (2022.6)
Requirement already satisfied: joblib>=0.14.1 in /Library/Frameworks/Python.fra
mework/Versions/3.10/lib/python3.10/site-packages (from phik<0.13,>=0.11.1->yda
ta-profiling) (1.2.0)
Requirement already satisfied: typing-extensions>=4.2.0 in /Library/Frameworks/
Python.framework/Versions/3.10/lib/python3.10/site-packages (from pydantic<1.1
1,>=1.8.1->ydata-profiling) (4.13.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /Library/Frameworks/
Python.framework/Versions/3.10/lib/python3.10/site-packages (from requests<2.2
9,>=2.24.0->ydata-profiling) (3.0.1)
Requirement already satisfied: idna<4,>=2.5 in /Users/umeshsangule/Library/Pyth
on/3.10/lib/python/site-packages (from requests<2.29,>=2.24.0->ydata-profiling)
(3.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /Library/Frameworks/Pyt
hon.framework/Versions/3.10/lib/python3.10/site-packages (from requests<2.2
9,>=2.24.0->ydata-profiling) (1.26.14)
Requirement already satisfied: certifi>=2017.4.17 in /Library/Frameworks/Pytho
n.framework/Versions/3.10/lib/python3.10/site-packages (from requests<2.2
9,>=2.24.0->ydata-profiling) (2022.12.7)
```

```
Requirement already satisfied: patsy>=0.5.2 in /Library/Frameworks/Python.frame
work/Versions/3.10/lib/python3.10/site-packages (from statsmodels<0.14,>=0.1
3.2->ydata-profiling) (0.5.3)
Requirement already satisfied: six in /Library/Frameworks/Python.framework/Vers
ions/3.10/lib/python3.10/site-packages (from patsy>=0.5.2->statsmodels<0.1
4,>=0.13.2->ydata-profiling) (1.16.0)
Requirement already satisfied: PyWavelets in /Library/Frameworks/Python.framewo
rk/Versions/3.10/lib/python3.10/site-packages (from imagehash->visions[type_ima
ge_path]==0.7.5->ydata-profiling) (1.4.1)

[notice] A new release of pip is available: 25.1.1 -> 25.2
[notice] To update, run: pip install --upgrade pip
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```python
import pandas as pd
from ydata_profiling import ProfileReport

# Load dataset
url = "/Users/umeshsangule/Downloads/tips.csv"
df = pd.read_csv(url)

# Create pandas-profiling report
profile = ProfileReport(df, title="Automated EDA Report", explorative=True)

# Specify the folder and file name (change the path as per your system)
output_path = r"/Users/umeshsangule/Downloads/EDA_Report.html"

# Save the report
profile.to_file(output_path)

print(f"EDA report generated and saved at: {output_path}")
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
/Library/Frameworks/Python.framework/Versions/3.10/lib/python3.10/site-package
s/ydata_profiling/model/pandas/discretize_pandas.py:52: FutureWarning: In a fut
ure version, `df.iloc[:, i] = newvals` will attempt to set the values inplace i
nstead of always setting a new array. To retain the old behavior, use either `d
f[df.columns[i]] = newvals` or, if columns are non-unique, `df.isetitem(i, newv
als)`
  discretized_df.loc[:, column] = self._discretize_column(
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
EDA report generated and saved at: /Users/umeshsangule/Downloads/EDA_Report.htm
l
```

In [ ]:

In [ ]: