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

import pandas as pd
df = pd.read_csv('movies.csv')
df.head()
```

Choose files movies.csv

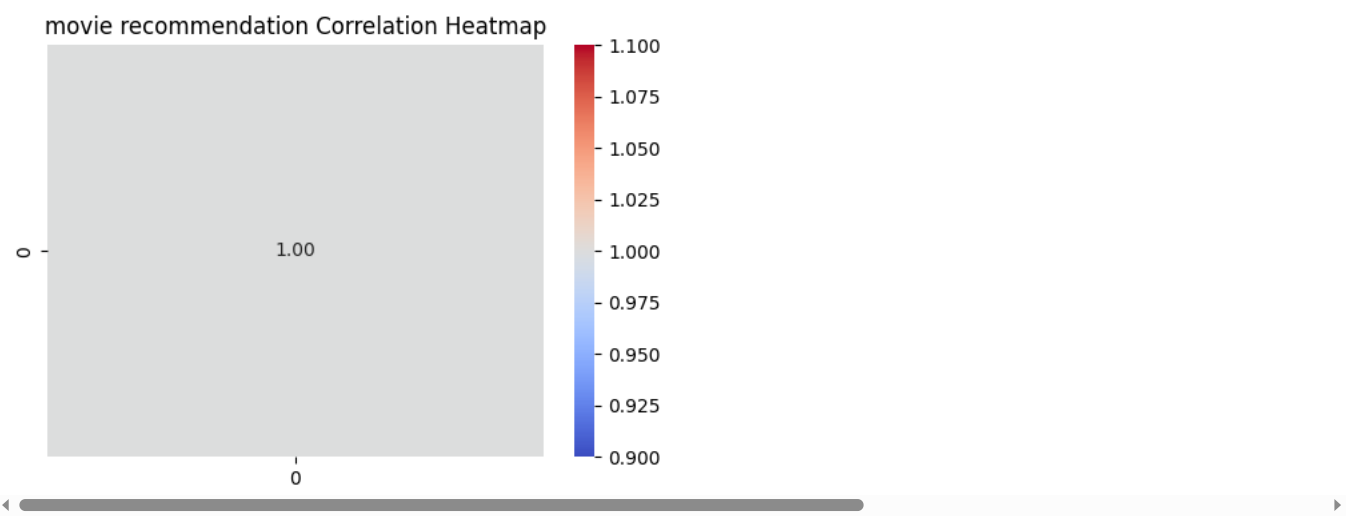
- **movies.csv**(text/csv) - 608 bytes, last modified: 08/05/2025 - 100% done

```
Saving movies.csv to movies.csv
```

	Customer	Age	Watched movie	Related movie	Start time	End time	websites	paid
0	Logeshkannan	19	kaththi	thupakki	03:00	06:00	JioHotstar	150
1	Dhamothiran	19	good bad ugly	vidamuyarchi	09:00	12:00	Amazon prime	100
2	Sakthi	19	kingston	gangers	12:00	03:00	JioHotstar	150
3	Sanjeev	18	beast	gurkha	03:00	06:00	Amazon prime	100
4	Naveen	18	Wrong turn	Thanksgiving	09:00	12:00	JioHotstar	150

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

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



data = {
    'Age': [19, 19, 19, 18, 18, 18, 18, 20, 20],
    'Paid': [150, 100, 150, 100, 150, 100, 150, 100, 150]
}


# Create a DataFrame
df = pd.DataFrame(data)
# Generate Correlation Matrix
correlation_matrix = df.corr()
# Plot Heatmap
plt.figure(figsize=(6, 4))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("movie recommendation Correlation Heatmap")
plt.show()
```




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

# Sample data
data = [10, 12, 14, 15, 17, 20, 30, 100]  # 100 is an outlier

# Convert to DataFrame
df = pd.DataFrame(data, columns=['values'])

# Calculate Q1, Q3, and IQR
Q1 = df['values'].quantile(0.25)
```

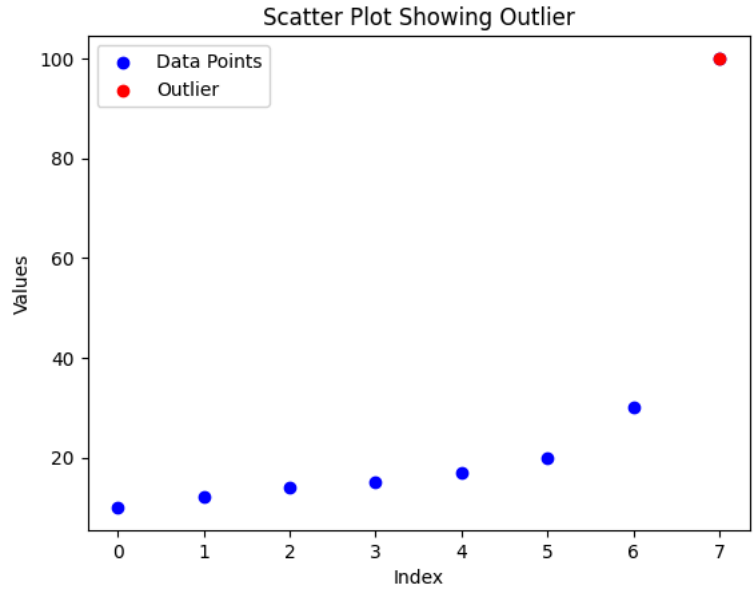
```
Q3 = df['values'].quantile(0.75)
IQR = Q3 - Q1

# Define outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['values'] < lower_bound) | (df['values'] > upper_bound)]

# Plot Boxplot
plt.boxplot(df['values'])
plt.title("Box Plot to Detect Outliers")
plt.show()

# Print outliers
print("Outliers:\n", outliers)
```


```
Outliers:
    values
7      100
```

```
from scipy import stats

# Convert to NumPy array
data_array = np.array(data)

# Calculate Z-scores
z_scores = np.abs(stats.zscore(data_array))

# Find outliers (Z-score > 3)
outliers = data_array[z_scores > 3]

print("Outliers using Z-score method:", outliers)
```

```
Outliers using Z-score method: []
```

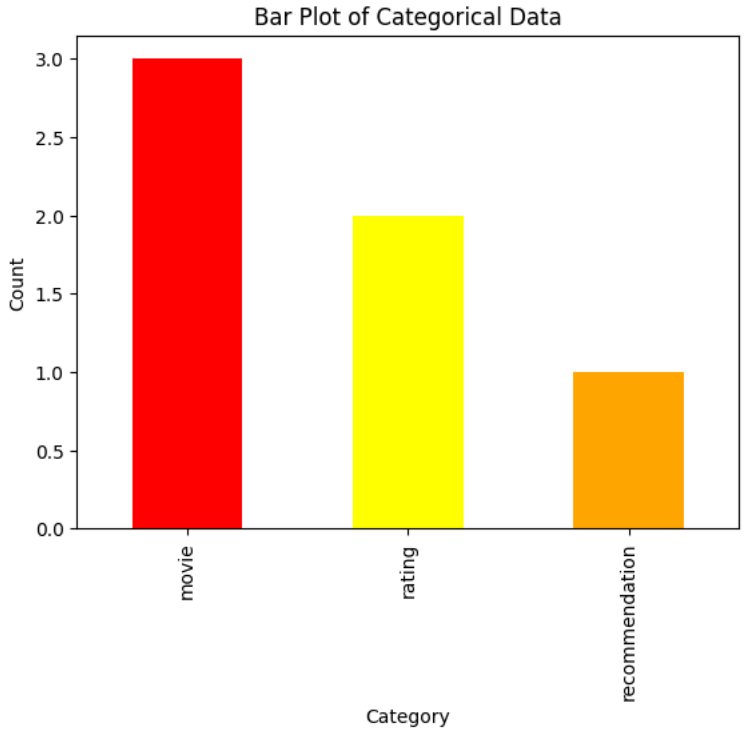
```
plt.scatter(range(len(data)), data, color='blue', label="Data Points")
plt.scatter([data.index(100)], [100], color='red', label="Outlier")  # Highlight outlier
plt.xlabel("Index")
plt.ylabel("Values")
plt.title("Scatter Plot Showing Outlier")
plt.legend()
plt.show()
```



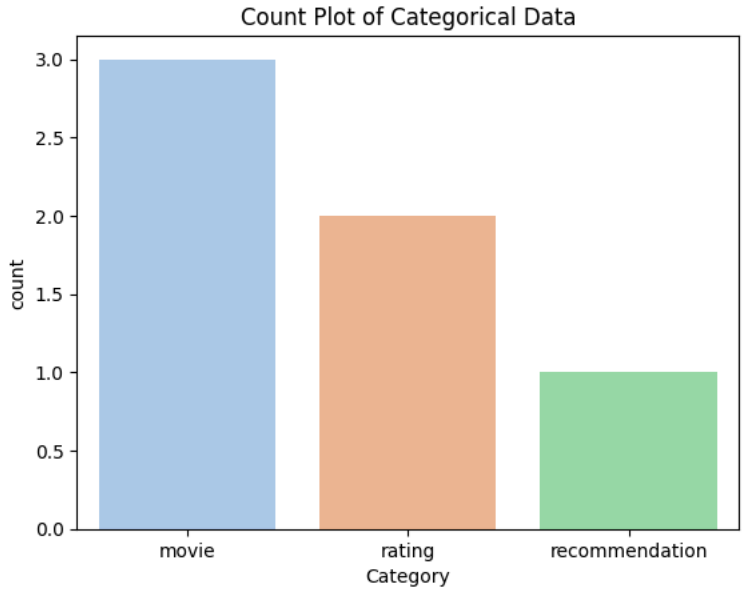

```
import pandas as pd

# Sample categorical data
data = pd.DataFrame({'Category': ['movie', 'rating', 'movie', 'recommendation', 'rating', 'movie']})

# Count occurrences
print(data['Category'].value_counts())

# Frequency table (percentage)
print(data['Category'].value_counts(normalize=True) * 100)
```

```
Category
movie             3
rating            2
recommendation    1
Name: count, dtype: int64
Category
movie             50.000000
rating            33.333333
recommendation    16.666667
Name: proportion, dtype: float64
```

```
import matplotlib.pyplot as plt
import seaborn as sns

# Sample data
categories = ['movie', 'rating', 'movie', 'recommendation', 'rating', 'movie']
df = pd.DataFrame({'Category': categories})

# Bar Plot
df['Category'].value_counts().plot(kind='bar', color=['red', 'yellow', 'orange'])
plt.xlabel("Category")
plt.ylabel("Count")
plt.title("Bar Plot of Categorical Data")
plt.show()

# Count Plot (Using Seaborn)
sns.countplot(x=df['Category'], palette="pastel")
plt.title("Count Plot of Categorical Data")
plt.show()
```




```
<ipython-input-23-d201b68d195a>:16: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `le

  sns.countplot(x=df['Category'], palette="pastel")
```




```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder

# Sample data
df = pd.DataFrame({'entertainment': ['Apple', 'Banana', 'Orange', 'Apple', 'Banana']})

# Label Encoding
label_encoder = LabelEncoder()
df['_Label'] = label_encoder.fit_transform(df['Fruit'])
print(df)

# One-Hot Encoding
df_one_hot = pd.get_dummies(df['Fruit'])
print(df_one_hot)



!pip install ydata-profiling
import pandas as pd
from ydata_profiling import ProfileReport  # Correct import
# Load sample dataset
df = pd.read_csv("https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv")
```

```
# Generate profile report
profile = ProfileReport(df, explorative=True)
# Display report in Colab
profile.to_notebook_iframe()
import pandas as pd
from ydata_profiling import ProfileReport  # Correct import

# Load sample dataset
df = pd.read_csv("https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv")

# Generate profile report
profile = ProfileReport(df, explorative=True)

# Display report in Colab
profile.to_notebook_iframe()
profile.to_file("titanic_report.html")
```

```
Requirement already satisfied: ydata-profiling in /usr/local/lib/python3.11/dist-packages (4.16.1)
Requirement already satisfied: scipy<1.16,>=1.4.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.15.2)
Requirement already satisfied: pandas!=1.4.0,<3.0,>1.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.2.2)
Requirement already satisfied: matplotlib<=3.10,>=3.5 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (3.10.0)
Requirement already satisfied: pydantic>=2 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.11.4)
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (6.0.2)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (3.1.6)
Requirement already satisfied: visions<0.8.2,>=0.7.5 in /usr/local/lib/python3.11/dist-packages (from visions[type_image_path]<0.8.2
Requirement already satisfied: numpy<2.2,>=1.16.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.0.2)
Requirement already satisfied: htmlmin==0.1.12 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.1.12)
Requirement already satisfied: phik<0.13,>=0.11.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.12.4)
Requirement already satisfied: requests<3,>=2.24.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (2.32.3)
Requirement already satisfied: tqdm<5,>=4.48.2 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (4.67.1)
Requirement already satisfied: seaborn<0.14,>=0.10.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.13.2)
Requirement already satisfied: multimethod<2,>=1.4 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.12)
Requirement already satisfied: statsmodels<1,>=0.13.2 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.14.4)
Requirement already satisfied: typeguard<5,>=3 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (4.4.2)
Requirement already satisfied: imagehash==4.3.1 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (4.3.1)
Requirement already satisfied: wordcloud>=1.9.3 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.9.4)
Requirement already satisfied: dacite>=1.8 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (1.9.2)
Requirement already satisfied: numba<=0.61,>=0.56.0 in /usr/local/lib/python3.11/dist-packages (from ydata-profiling) (0.60.0)
Requirement already satisfied: PyWavelets in /usr/local/lib/python3.11/dist-packages (from imagehash==4.3.1->ydata-profiling) (1.8.0
Requirement already satisfied: pillow in /usr/local/lib/python3.11/dist-packages (from imagehash==4.3.1->ydata-profiling) (11.2.1)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2<3.2,>=2.11.1->ydata-profiling
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profi
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profiling
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-prof
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-prof
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profil
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-profi
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib<=3.10,>=3.5->ydata-p
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba<=0.61,>=0.56.0->yda
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas!=1.4.0,<3.0,>1.1->ydata-profilin
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas!=1.4.0,<3.0,>1.1->ydata-profil
Requirement already satisfied: joblib>=0.14.1 in /usr/local/lib/python3.11/dist-packages (from phik<0.13,>=0.11.1->ydata-profiling)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profiling
Requirement already satisfied: pydantic-core==2.33.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profiling)
Requirement already satisfied: typing-extensions>=4.12.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profili
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2->ydata-profilin
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata-profiling)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata-profil
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests<3,>=2.24.0->ydata-profil
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.11/dist-packages (from statsmodels<1,>=0.13.2->ydata-profiling
Requirement already satisfied: attrs>=19.3.0 in /usr/local/lib/python3.11/dist-packages (from visions<0.8.2,>=0.7.5->visions[type_im
Requirement already satisfied: networkx>=2.4 in /usr/local/lib/python3.11/dist-packages (from visions<0.8.2,>=0.7.5->visions[type_im
Requirement already satisfied: puremagic in /usr/local/lib/python3.11/dist-packages (from visions<0.8.2,>=0.7.5->visions[type_image_
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib<=3.10,>=3
Summarize dataset: 100%                                                    41/41 [00:04<00:00,  3.24it/s, Completed]

  0%|          | 0/15 [00:00<?, ?it/s]
 27%|██▋       | 4/15 [00:00<00:00, 37.54it/s]
 53%|█████▎    | 8/15 [00:00<00:00, 33.80it/s]
100%|██████████| 15/15 [00:00<00:00, 38.87it/s]
Generate report structure: 100%                                            1/1 [00:03<00:00,  3.44s/it]

Render HTML: 100%                                                          1/1 [00:00<00:00,  1.02it/s]
```

YData Profiling Report

Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Overview

Brought to you by YData

Overview Alerts 20 Reproduction

Dataset statistics

Number of variables	15
Number of observations	891
Missing cells	869
Missing cells (%)	6.5%
Duplicate rows	53
Duplicate rows (%)	5.9%

Variable types

Categorical	8
Numeric	4
Boolean	3

Total size in memory	398.1 KiB
Average record size in memory	457.6 B

Variables

Select Columns

Summarize dataset: 100% 41/41 [00:04<00:00, 5.74it/s, Completed]

```
  0%|          | 0/15 [00:00<?, ?it/s]
100%|██████████| 15/15 [00:00<00:00, 68.50it/s]
```

Generate report structure: 100% 1/1 [00:05<00:00, 5.11s/it]

Render HTML: 100% 1/1 [00:03<00:00, 3.78s/it]

YData Profiling Report Overview Variables Interactions Correlations Missing values Sample Duplicate rows

deck

Categorical

High correlation **Missing**

Distinct	7
Distinct (%)	3.4%
Missing	688
Missing (%)	77.2%
Memory size	54.6 KiB

C 59
B 47
D 33
E 32
A 15
Other valu... 17

More details

embark_town

Categorical

High correlation

Distinct	3
Distinct (%)	0.3%
Missing	2
Missing (%)	0.2%

Southampt... 644
Cherbourg 168
Queenstown 77

Export report to file: 100% 1/1 [00:00<00:00, 20.57it/s]

```
!pip install dtale
import dtale
# Launch D-Tale dashboard
dtale.show(df)
dtale.show(df)
```

```
Collecting dtale
  Downloading dtale-3.17.0-py2.py3-none-any.whl.metadata (16 kB)
Collecting dash-daq<=0.5.0 (from dtale)
  Downloading dash_daq-0.5.0.tar.gz (642 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 642.7/642.7 kB 15.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: future>=0.14.0 in /usr/local/lib/python3.11/dist-packages (from dtale) (1.0.0)
Requirement already satisfied: missingno in /usr/local/lib/python3.11/dist-packages (from dtale) (0.5.2)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from dtale) (2.2.2)
Collecting squarify (from dtale)
  Downloading squarify-0.4.4-py3-none-any.whl.metadata (600 bytes)
Collecting strsimpy (from dtale)
  Downloading strsimpy-0.2.1-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: six in /usr/local/lib/python3.11/dist-packages (from dtale) (1.17.0)
Requirement already satisfied: xlrd in /usr/local/lib/python3.11/dist-packages (from dtale) (2.0.1)
Requirement already satisfied: beautifulsoup4!=4.13.0b2 in /usr/local/lib/python3.11/dist-packages (from dtale) (4.13.4)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from dtale) (2025.4.26)
Collecting dash-bootstrap-components<=1.7.1 (from dtale)
  Downloading dash_bootstrap_components-1.7.1-py3-none-any.whl.metadata (17 kB)
Collecting lz4 (from dtale)
  Downloading lz4-4.4.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.8 kB)
Requirement already satisfied: cycler in /usr/local/lib/python3.11/dist-packages (from dtale) (0.12.1)
Collecting dash<=2.18.2 (from dtale)
  Downloading dash-2.18.2-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-packages (from dtale) (0.13.2)
Requirement already satisfied: werkzeug in /usr/local/lib/python3.11/dist-packages (from dtale) (3.1.3)
Collecting Flask-Compress (from dtale)
  Downloading Flask_Compress-1.17-py3-none-any.whl.metadata (8.8 kB)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (from dtale) (1.6.1)
Requirement already satisfied: statsmodels in /usr/local/lib/python3.11/dist-packages (from dtale) (0.14.4)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from dtale) (3.4.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from dtale) (2.2.5)
```