

Introduction to Matplotlib

Course Code: CPE 031

Date Performed: Oct 22, 2024

Section: CPE21S4

Date Submitted:

Name: Guariño, Danica T.

Instructor: Engr. Edcel B. Artificio

Intended Learning Outcomes (ILO):

By the end of this laboratory session, learners will be able to:

1. Utilize Matplotlib's pyplot interface to create a variety of visualizations, including line

1. Utilize Matplotlib's pyplot interface to create a variety of visualizations, including line plots, scatter plots, histograms, and box plots, demonstrating an understanding of the library's syntax and functionality.

2. Customize visual elements such as titles, labels, and legends to enhance the clarity and aesthetics of their plots, applying best practices in data visualization.
3. Analyze and interpret visual data representations to extract meaningful insights, effectively communicating findings through well-structured graphical presentations.

Part 1: Perform the following codes, and understand the difference between line plot, scatter plot, histogram, bar chart, box plot, and pie chart using matplotlib's pyplot sub-module. **(Provide a screenshot of your output.)**

1. Line Plot

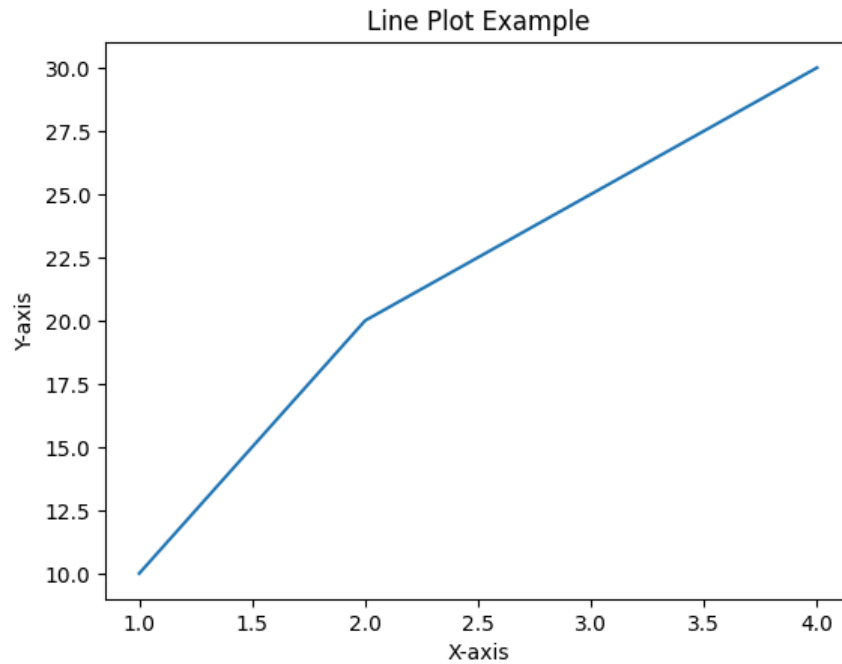
```
import matplotlib.pyplot as plt

x = [1, 2, 3, 4]
y = [10, 20, 25, 30]

plt.plot(x, y)

plt.title("Line Plot Example")
plt.xlabel("X-axis")
plt.ylabel("Y-axis")
plt.show()
```

Output:

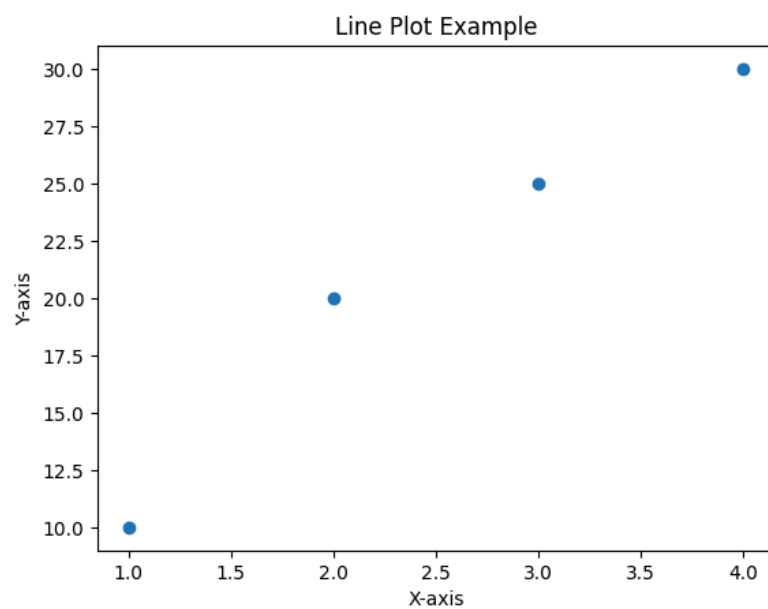


2. Scatter Plot

```
import matplotlib.pyplot as plt

x = [1, 2, 3, 4]
y = [10, 20, 25, 30]
plt.scatter(x, y)
plt.title("Scatter Plot Example")
plt.xlabel("X-axis")
plt.ylabel("Y-axis")
plt.show()
```

Output:

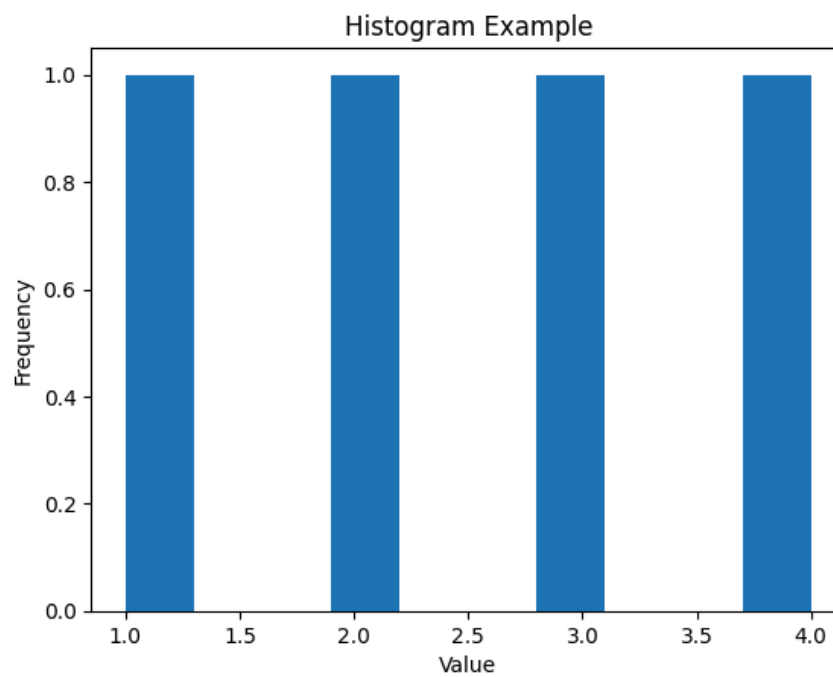


3. Histogram

```
import matplotlib.pyplot as plt

data = [1, 2, 2, 3, 3, 3, 4]
plt.hist(data)
plt.title("Histogram Example")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```

Output:

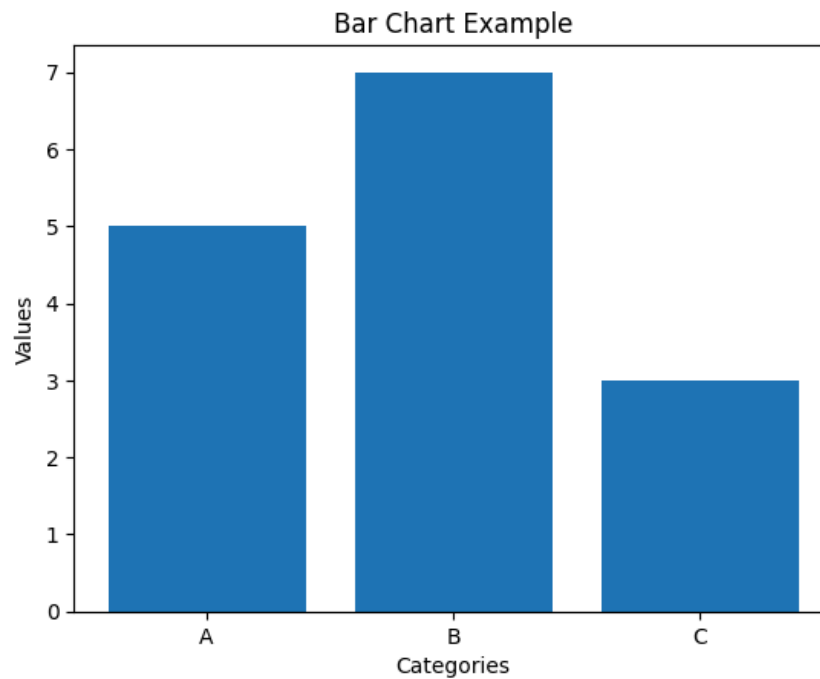


4. Bar Chart

```
import matplotlib.pyplot as plt

categories = ['A', 'B', 'C']
values = [5, 7, 3]
plt.bar(categories, values)
plt.title("Bar Chart Example")
plt.xlabel("Categories")
plt.ylabel("Values")
plt.show()
```

Output:



5. Box plot

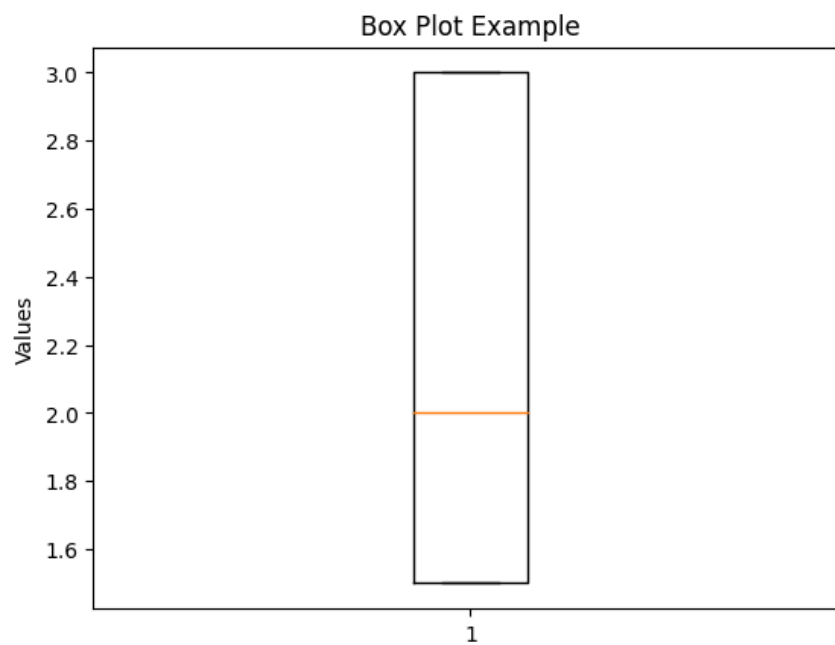
```
import matplotlib.pyplot as plt

data = [[1.5]*10 + [2]*10 + [3]*10]

plt.boxplot(data)

plt.title("Box Plot Example")
plt.ylabel("Values")
plt.show()
```

Output:



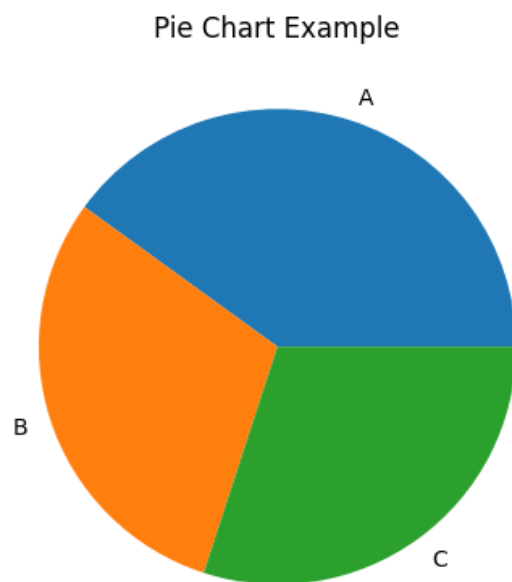
6. Pie chart

```
import matplotlib.pyplot as plt

labels = ['A', 'B', 'C']
sizes = [40, 30, 30]

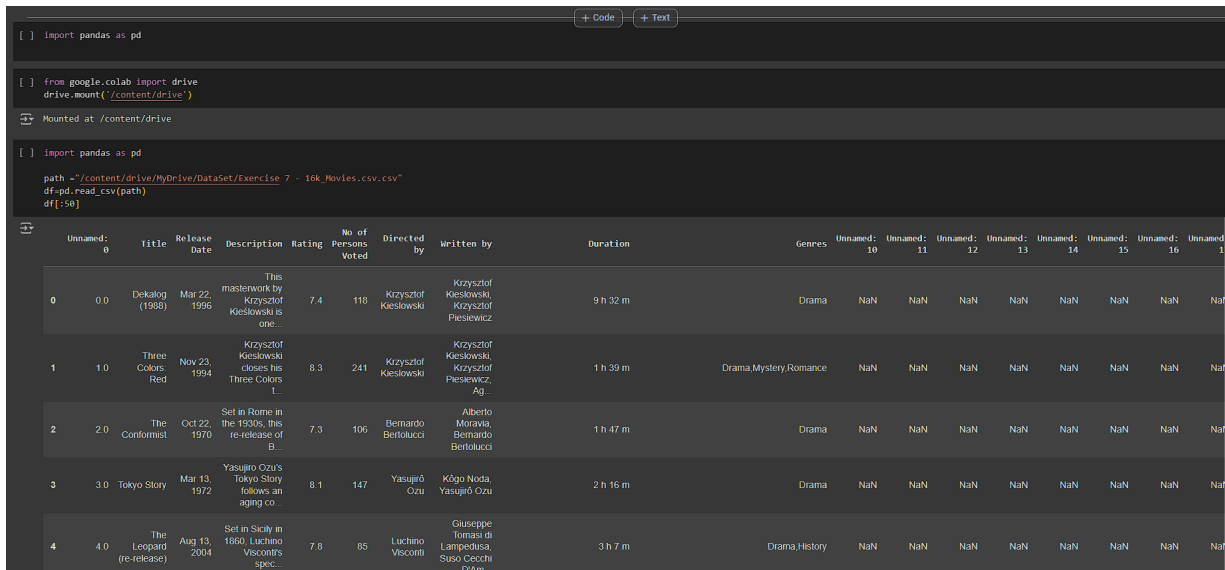
plt.pie(sizes, labels=labels)
plt.title("Pie Chart Example")
plt.show()
```

Output:



Part 2: Refer to the instructions below.

1. **Find a dataset for this activity:** Please visit Kaggle and look for a new dataset that would allow you to perform visualization and analysis using matplotlib.
2. **Creating a dataframe from your CSV file:** Once you have successfully loaded your dataset, you need to create a dataframe from your uploaded CSV file.



```
[ ] import pandas as pd

[ ] from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive

[ ] import pandas as pd

path = "/content/drive/MyDrive/Dataset/Exercise 7 - 16k_Movies.csv"
df=pd.read_csv(path)
df[:50]
```

	Unnamed: 0	Title	Release Date	Description	Rating	No of Persons Voted	Directed by	Written by	Duration	Genres	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13	Unnamed: 14	Unnamed: 15	Unnamed: 16	Unnamed: 17
0	0.0	Dekalog (1988)	Mar 22, 1996	This masterpiece by Krzysztof Kieslowski is one...	7.4	110	Krzysztof Kieslowski	Krzysztof Kieslowski, Krzysztof Piesiewicz	9 h 32 m	Drama	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1.0	Three Colors: Red	Nov 23, 1994	Krzysztof Kieslowski closes his Three Colors t...	8.3	241	Krzysztof Kieslowski	Krzysztof Kieslowski, Krzysztof Piesiewicz, Ag...	1 h 39 m	Drama,Mystery,Romance	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2.0	The Conformist	Oct 22, 1970	Set in Rome in the 1930s, this re-release of B...	7.3	106	Bernardo Bertolucci	Alberto Moravia, Bernardo Bertolucci	1 h 47 m	Drama	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	3.0	Tokyo Story	Mar 19, 1972	Yasujiro Ozu's Tokyo Story follows an aging co...	8.1	147	Yasujiro Ozu	Kôga Noda, Yasujiro Ozu	2 h 16 m	Drama	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	4.0	The Leopard (re-release)	Aug 13, 2004	Set in Sicily in 1860, Luciano Visconti's spec...	7.6	65	Luciano Visconti	Giuseppe Tomasi di Lampedusa, Suso Cecchi d'Am...	3 h 7 m	Drama,History	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

3. Import the matplotlib.pyplot

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

4. **Based on your chosen dataset, you will develop three questions that you will answer using pyplot visualizations. This means that you will need to produce at least three pyplot visualizations. You are also required to make certain customizations on your data vizes.**

- Question 1. How many movies were released each year?

Question 1. How many movies were released each year?

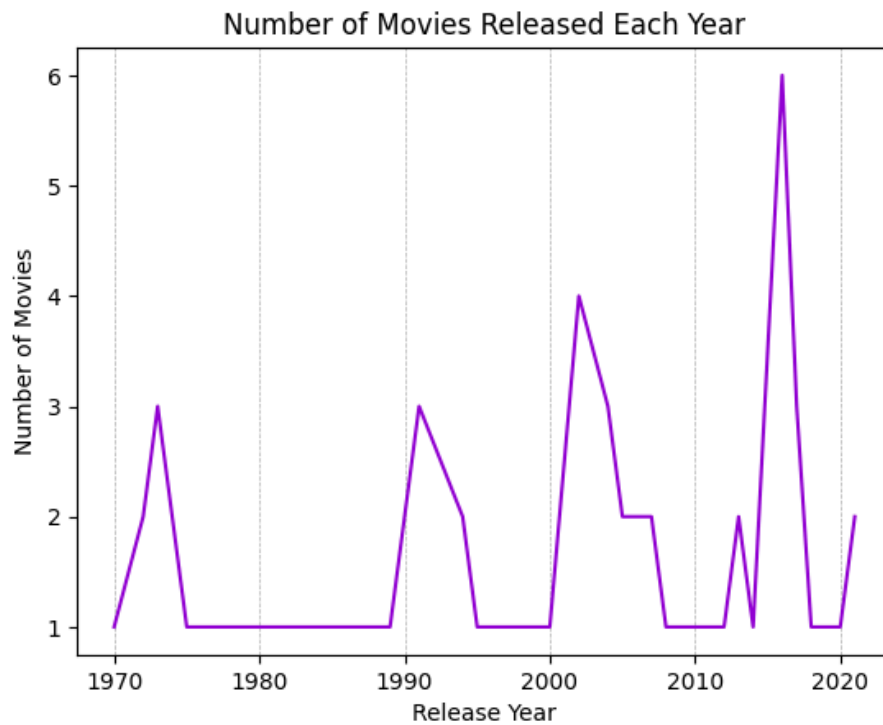
```
import matplotlib.pyplot as plt
import pandas as pd

df['Release Date'] = pd.to_datetime(df['Release Date'], errors='coerce')
df['release_year'] = df['Release Date'].dt.year

movie_counts = df.groupby('release_year')['Title'].count().reset_index()
movie_counts.columns = ['release_year', 'count']
movie_counts = movie_counts.sort_values(by='release_year')
print(movie_counts)

# Create the line plot
plt.plot(movie_counts['release_year'], movie_counts['count'], c= 'darkviolet')
plt.title('Number of Movies Released Each Year')
plt.xlabel('Release Year')
plt.ylabel('Number of Movies')
plt.grid(axis='x',linestyle='--',linewidth=0.50)
plt.show()
```

Output:



Observation: Over the years, the pace of movie releases changed. Peaks occur in 2020, 1995, 2005, and 1975. 2020 is the year with the highest. There were times when the making of films was relatively slow. This was particularly from 1980–1990 to 2010–2015.

- Question 2. What is the relationship between the number of persons voted and the rating of the movies?

Question 2. What is the relationship between the number of persons voted and the rating of the movies?

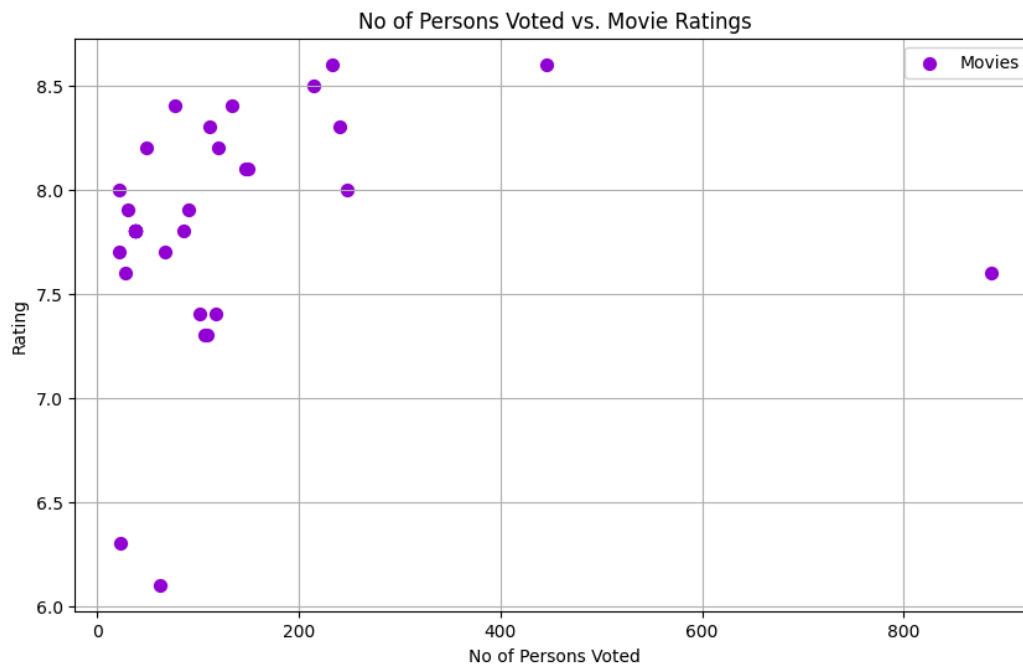
```
[44] import matplotlib.pyplot as plt
import pandas as pd

path = "/content/drive/MyDrive/DataSet/Exercise 7 - 16k_Movies.csv.csv"
df = pd.read_csv(path)
df['No of Persons Voted'] = pd.to_numeric(df['No of Persons Voted'], errors='coerce')
print(df.columns)

plt.figure(figsize=(10, 6))

plt.scatter(df['No of Persons Voted'], df['Rating'], s=50, c='darkviolet', label='Movies')
plt.legend()
plt.xlabel('No of Persons Voted')
plt.ylabel('Rating')
plt.title('No of Persons Voted vs. Movie Ratings')
plt.grid()
plt.show()
```

Output:



Observation: The scatter plot shows the relationship between the quantity of votes and the corresponding movie ratings. Regardless of their ratings, the majority of the films often receive a moderate amount of votes. Nonetheless, the majority of the extremely high number of votes for a select few films show a range of ratings. This implies that, while there are some really popular films with a wide range of ratings, the majority do not necessarily get more popular as their ratings rise.

- Question 3. What is the average rating for movies directed by each director?

Question 3. What is the average rating for movies directed by each director?

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

try:
    df = pd.read_csv("/content/drive/MyDrive/DataSet/Exercise 7 - 16k_Movies.csv.csv")
except FileNotFoundError:
    print("Error: The specified file was not found.")
    exit()

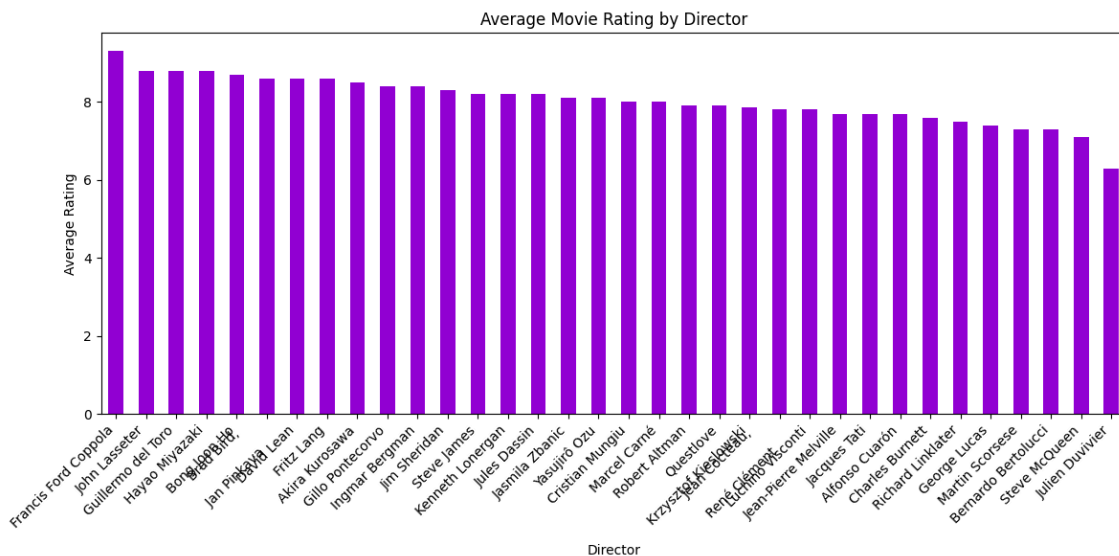
print("Columns in the DataFrame:", df.columns)
print("First few rows of the DataFrame:\n", df.head())

director_col = 'Directed by'
df_cleaned = df.dropna(subset=[director_col, 'Rating'])

if df_cleaned.empty:
    print("Warning: The cleaned DataFrame is empty. No data to analyze.")
else:
    avg_ratings = df_cleaned.groupby(director_col)['Rating'].mean().sort_values(ascending=False)

    plt.figure(figsize=(12, 6))
    avg_ratings.plot(kind="bar", color="darkviolet", rot=90)
    plt.title("Average Movie Rating by Director")
    plt.xlabel("Director")
    plt.ylabel("Average Rating")
    plt.xticks(rotation=45, ha="right")
    plt.tight_layout()
    plt.show()
```

Output:



Observation: We can see from this bar chart that different directors receive different average ratings for their films. For example, Francis Ford Coppola receives the highest average rating, while Julien Duvivier receives the lowest. Based on the average ratings of their films, this type of representation highlights the differences in popularity and success among different directors. It makes it simple to compare the average rating of each film and determine which director consistently produces these kinds of films.

5. Provide observations for each of your data viz, then **produce one insight not longer**

than five sentences given your three observations. Your output shall follow this outline:

- a. Introduction (Describe your dataset)
- b. Questions
- c. Visualization and Observation
- d. Insight

Insights:

My dataset shows 50 different movies and for each movies, my data set shows information about its film like the who directed, how many ratings, the hour duration and etc. In this activity, I came up with 3 questions that I'll be answered by my pyplot visualizations. The first question is about how many movies were released each year, the second talks about the relationship between the number of persons voted and the rating of the movies, and the third and last question talks about the average rating for movies directed by each director. These questions helps me to dive deeper intomy dataset.. By examining the number of movies released each year, I identified production trends over time. Exploring the relationship between the number of votes and movie ratings shows audience engagement and preferences. Assessing the average ratings for movies directed by each director highlights the success and popularity of different directors.

6. Your grade will depend on the quality of the question, difficulty/complexity of the visualization, and value-add of the insight that you will generate.