Introduction to Matplotlib

Course Code: CPE 031	**Program**: Computer Engineering
Course Title: Visualization and Data Analysis	**Date Performed**: 10/22/24
Section: CPE21S4	**Date Submitted**: 10/22/24
Name: CALVIN EARL PLANTA	**Instructor**: Prof. Sayo

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 plots, scatter plots, histograms, and box plots, demonstrating an understanding of the library's syntax and functionality.

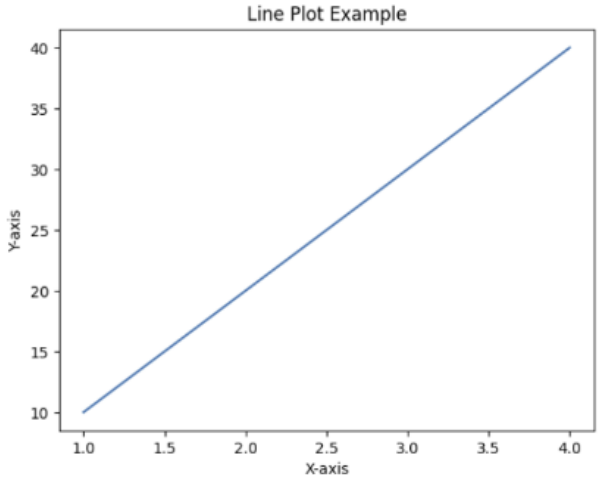
2. Customize visual elements such as titles, labels, and legends to enhance the clarity and aesthetics of their plots, applying best practices in data visualization.

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.)**

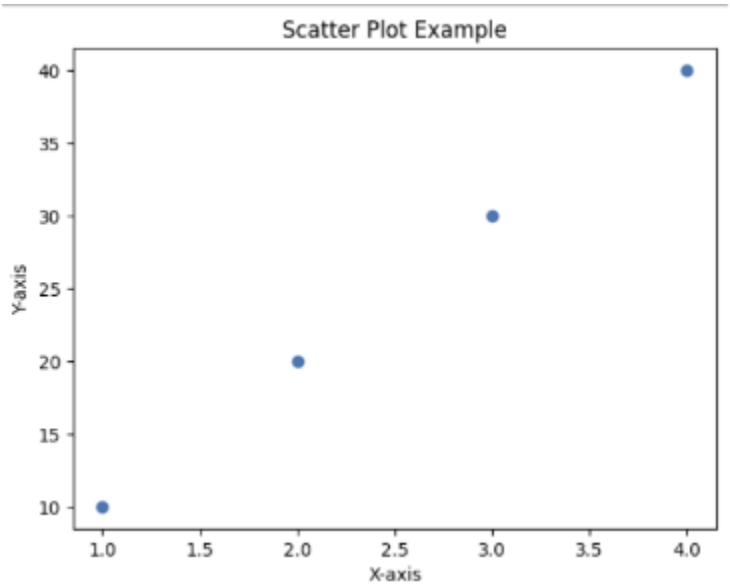
1. Line Plot

```python
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()
```



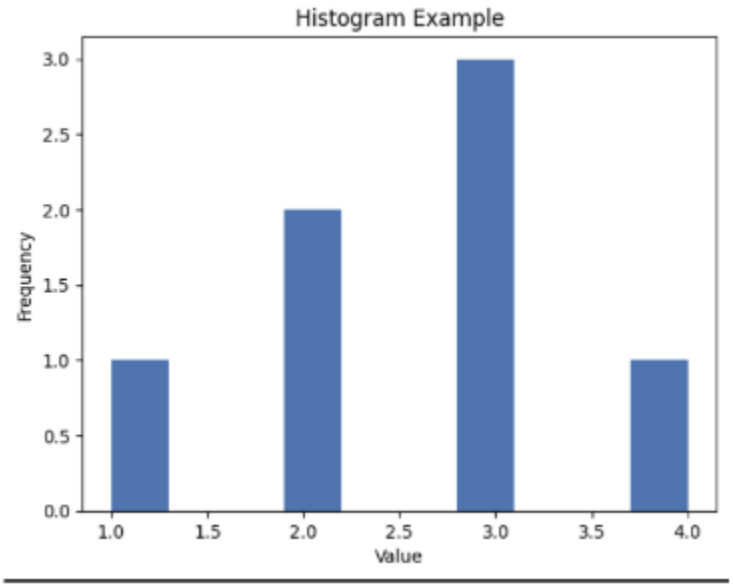
2. Scatter Plot

```
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()
```



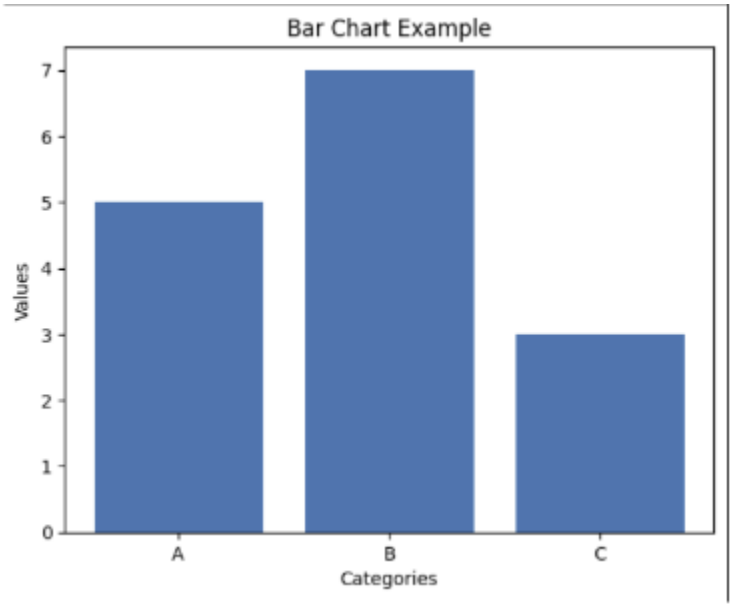
3. Histogram

```
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()
```



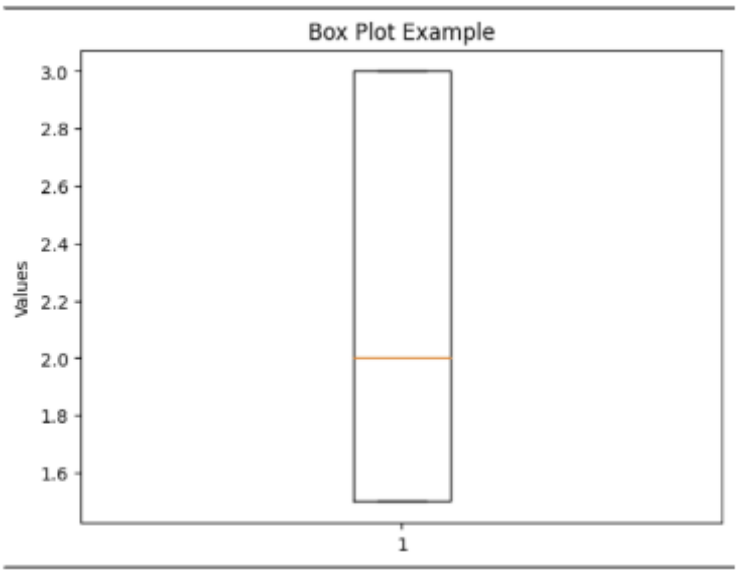
4. Bar Chart

```
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()
```



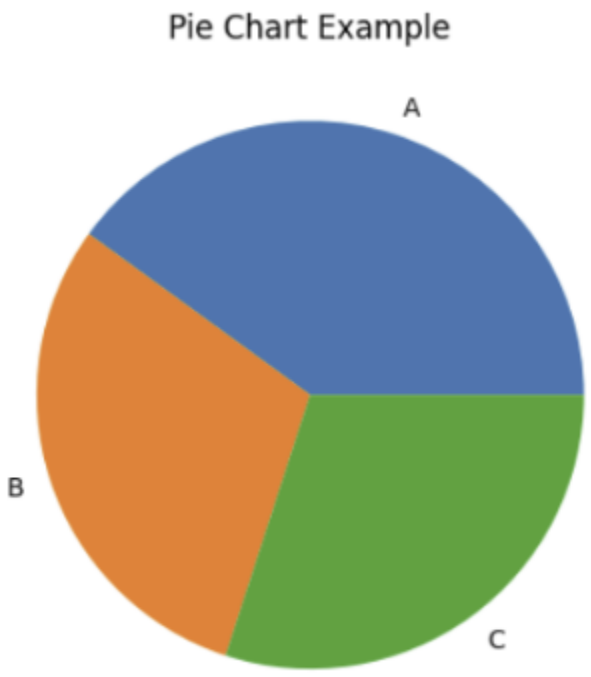
5. Box plot

```
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()
```



6. Pie chart

```python
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()
```



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

```
from google.colab import files
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
uploaded = files.upload()

data = pd.read_csv('most_subscribed_youtube_channels.csv')
data.head(50)
```

Choose Files most_subsc...hannels.csv

- **most_subscribed_youtube_channels.csv**(text/csv) - 70911 bytes, last modified: 10/22/2024 - 100% done

```
Saving most_subscribed_youtube_channels.csv to most_subscribed_youtube_channels (1).csv
```

	Unnamed: 0	rank	Youtuber	subscribers	video views	video count	category	started
0	0	1	T-Series	222,000,000	198,459,090,822	17,317	Music	2006
1	1	2	YouTube Movies	154,000,000	0	0	Film & Animation	2015
2	2	3	Cocomelon - Nursery Rhymes	140,000,000	135,481,339,848	786	Education	2006
3	3	4	SET India	139,000,000	125,764,252,686	91,271	Shows	2006
4	4	5	Music	116,000,000	0	0	NaN	2013
5	5	6	PewDiePie	111,000,000	28,469,458,228	4,497	Gaming	2010

3. **Import the matplotlib.pyplot**

```
import matplotlib.pyplot as plt
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.**

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

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.

OUTPUT

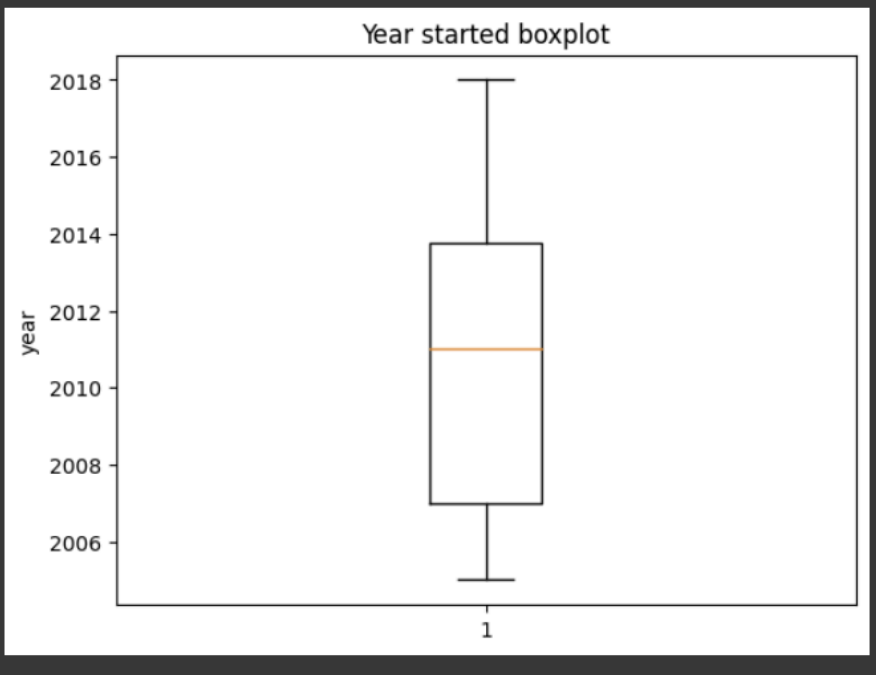
I. Introduction

The dataset that I have chosen from Kaggle is about the top 1000 most subscribed Youtube channels and their subscriber count. The data set also shows additional information such as the Youtube channels' total number of views and videos uploaded, type of video categories that the Youtube channel mainly makes, and the year in which the channel was created.

II. Questions

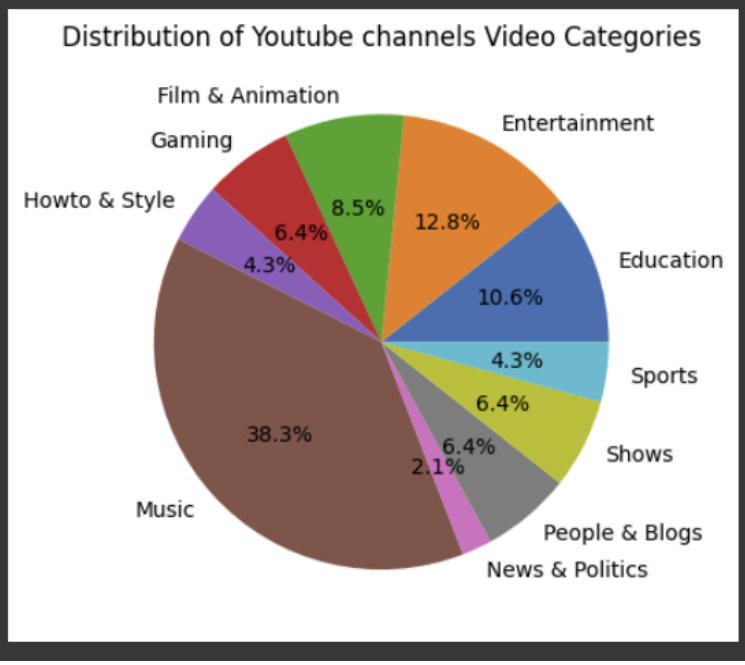
1. What is the mean year in which the Youtube channels were created?

```
#1. Box plot

plt.boxplot(ytchannel['started'])
plt.title("Year started boxplot")
plt.ylabel("year")
plt.show()
```



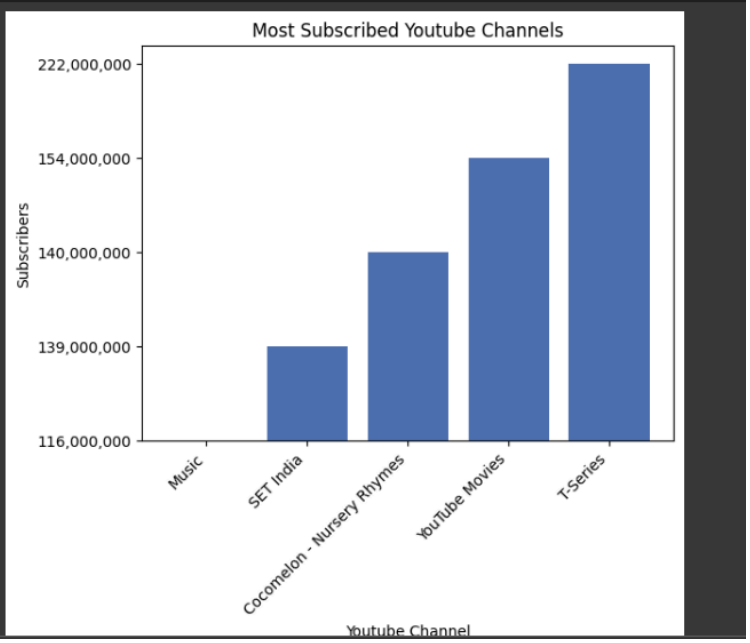
2. What is the percentage for each video category that the Youtube channels mainly create?

```
#2. Pie chart

category_counts = ytchannel.groupby('category')['category'].count()
plt.pie(category_counts, labels=category_counts.index, autopct='%1.1f%%')
plt.title('Distribution of Youtube channels Video Categories')
plt.show()
```



3. What is the most appropriate pyplot visualization for displaying the top 5 Youtube channels and their subscriber count?

```
#3. Bar graph

channels = ytchannel['Youtuber'][:5][::-1]
subcount = ytchannel['subscribers'][:5][::-1]

plt.bar(channels, subcount)
plt.title("Most Subscribed Youtube Channels")
plt.xlabel("Youtube Channel")
plt.ylabel("Subscribers")
plt.xticks(rotation=45, ha='right')

plt.show()
```



III. Visualizations and Observation

1. Box plot of the Youtube channels' year of creation - I used the box plot visualization to represent the year created for each Youtube channel. The years range from 2005 to 2018 and have a mean year of 2013.
2. Pie chart of Youtube video categories - I used the pie chart visualization to represent the most used video category that the top 50 Youtube channels create. I also showed the percentage of each category for more accurate and precise information, which I was able to do by adding autopct='%1.1f%%' to the code. Results of the chart shows that most of the Youtube channels make music content, while the least used category is News & Politics.

3. Bar graph of the most subscribed Youtube channels - I used the bar graph visualization to display the names of the most popular channels and their subscriber count. For the names of the Youtube channels in the x label, I used plt.xticks(rotation=45, ha='right') to keep the names from overlapping each other. The bars in the graph are arranged in a stair-like manner, as in the dataset, it is already arranged by subscriber count in descending order.

IV. Insights

The chosen dataset shows the 1000 (trimmed to 50) most subscribed Youtube channels, along with other information such as total number of views, total videos uploaded and such. We can create various visualizations out of the information that the dataset offers, very much like what I have demonstrated.