**BAHRIA UNIVERSITY, Karachi Campus)**

# Department of Software Engineering

# ASSIGNMENT # 02 – Fall 2024

|  |  |  |
| --- | --- | --- |
| Course Title: **IDS** |  | Course Code: |
| Class: **BSE – 5(A)** |  | Shift: **Morning** |
| Course Instructor: **Aamana** |  | Date: **31stOct 2024** |
| Due Date: **14 Nov 2024**  Student Name: **Muhammad Umer** |  | Max. Marks: **5.0 Marks**  Reg No: **81731** |

**Instructions:**

1. Deadline for the submission of Assignment on is **24** Nov 2024.
2. Upload word file containing screenshots under each task and also executable notebook file
3. If you submit your assignment after the given deadline then 2 Marks will be deducted for the late submissions.

**Question No. 1: [CLO#02, 5.0 marks]**

**After performing all the data cleaning steps on the given dataset of Netflix analyze below given questions.**

1: Using Python, calculate and display the percentage of movies versus TV shows in the dataset. Create a pie chart or bar graph to represent your findings visually.

**Solution:**

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv('Netflix\_Raw\_Dataset.csv')

type\_counts = df['type'].value\_counts()

total = len(df)

movie\_percentage = (type\_counts.get('Movie', 0) / total) \* 100

tv\_show\_percentage = (type\_counts.get('TV Show', 0) / total) \* 100

print(f"Movie Percentage: {movie\_percentage}%")

print(f"TV Show Percentage: {tv\_show\_percentage}%")

labels = ['Movies', 'TV Shows']

sizes = [movie\_percentage, tv\_show\_percentage]

colors = ['#ff9999', '#66b3ff']

explode = (0.1, 0)

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

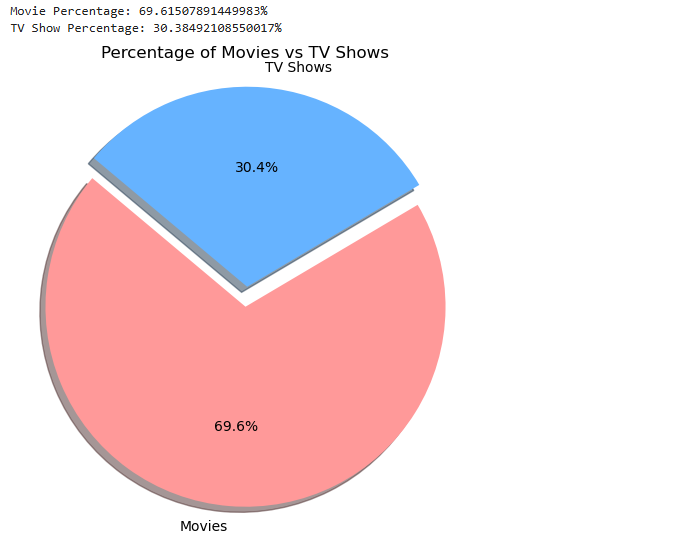
plt.pie(sizes, explode=explode, labels=labels, colors=colors, autopct='%1.1f%%', shadow=True, startangle=140)

plt.title('Percentage of Movies vs TV Shows')

plt.axis('equal')

plt.show()

**Output:**



2: Determine the distribution of ratings for both movies and TV shows. Present the result in a tabular format, and visualize it using a bar chart.

Code:

rating\_distribution = df.groupby(['type', 'rating']).size().unstack(fill\_value=0)

print("Rating Distribution:")

print(rating\_distribution)

rating\_distribution.plot(kind='bar', stacked=True, figsize=(10, 6), colormap='viridis')

plt.title('Rating Distribution for Movies vs TV Shows')

plt.xlabel('Rating')

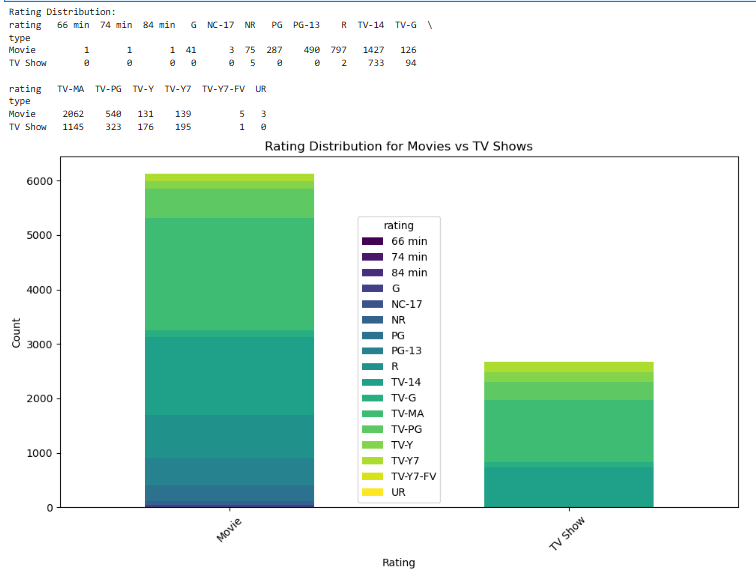
plt.ylabel('Count')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

Output:



3: Extract the names of actors from the dataset and calculate the top 5 actors based on their frequency of appearances. Describe how you would identify and list them using Python.

Code:

actors\_list = df['cast'].dropna().str.split(',').explode().str.strip()

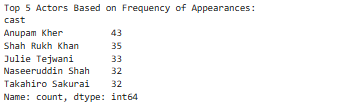
actor\_counts = actors\_list.value\_counts()

top\_5\_actors = actor\_counts.head(5)

print("Top 5 Actors Based on Frequency of Appearances:")

print(top\_5\_actors)

Output:



4: List the top 3 directors who have directed the most movies in the Action genre. Describe how you would filter the data and rank the directors to answer this.

Code:

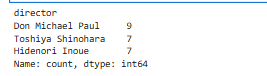
action\_movies = df[df['listed\_in'].str.contains('Action', na=False)]

director\_counts = action\_movies['director'].value\_counts()

top\_3\_directors = director\_counts.head(3)

print(top\_3\_directors)

Output:



5: Determine the number of movies and shows produced by Pakistan on Netflix, and identify the actor who appears most frequently in them. Outline your approach for analyzing this subset of the dataset.

Code:

pakistan\_shows = df[df['country'] == 'Pakistan']

type\_counts\_pakistan = pakistan\_shows['type'].value\_counts()

actors\_list\_pakistan = pakistan\_shows['cast'].dropna().str.split(',').explode().str.strip()

actor\_counts\_pakistan = actors\_list\_pakistan.value\_counts()

most\_frequent\_actor = actor\_counts\_pakistan.head(1)

print(type\_counts\_pakistan)

print(most\_frequent\_actor)

Output:

