**EDA Documentation**

EDA stands for Exploratory Data Analysis and it is the process of **analyzing and summarizing datasets** to understand their main characteristics, patterns, relationships, and anomalies often using **visual methods** like Bar graphs, pie charts and many different charts and plots.

In here we have given a dataset which is called netflix\_tites.csv in this dataset we have completely Cleaned the data, completed all the normalization so we can easily perform machine learning in that dataset.

Now we move to the EDA part where we have made multiple visuals to understand the trend and distribution of the data.

***Content Growth Over Time***

plt.bar(years - bar\_width/2, movies, width=bar\_width, label='Movies')

plt.bar(years + bar\_width/2, tv\_shows, width=bar\_width, label='TV Shows')

plt.title("Netflix Content Growth by Year", fontsize=16, fontweight='bold')

plt.xlabel("Release Year", fontsize=12)

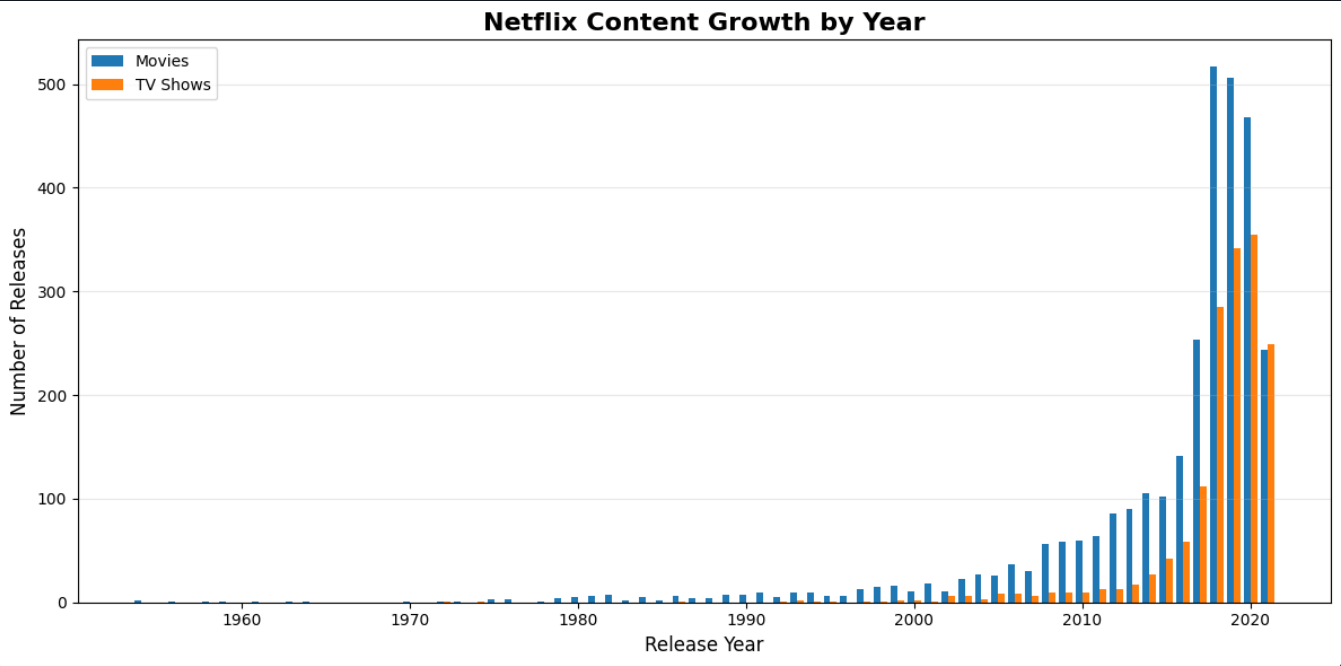
plt.ylabel("Number of Releases", fontsize=12)

plt.legend()

plt.grid(alpha=0.3, axis='y')

plt.tight\_layout()

plt.show()



In this visual we can see the growth of the content over time based on release\_year, Total number of contents released on that year and types of the content.

*Distributions Analysis*

* Based on Content Type

plt.figure(figsize=(6, 5))

plt.bar(pandas\_df['type'], pandas\_df['count'], color=['#FF5C5C', '#5C85FF'])

plt.title('Movies vs TV Shows Count')

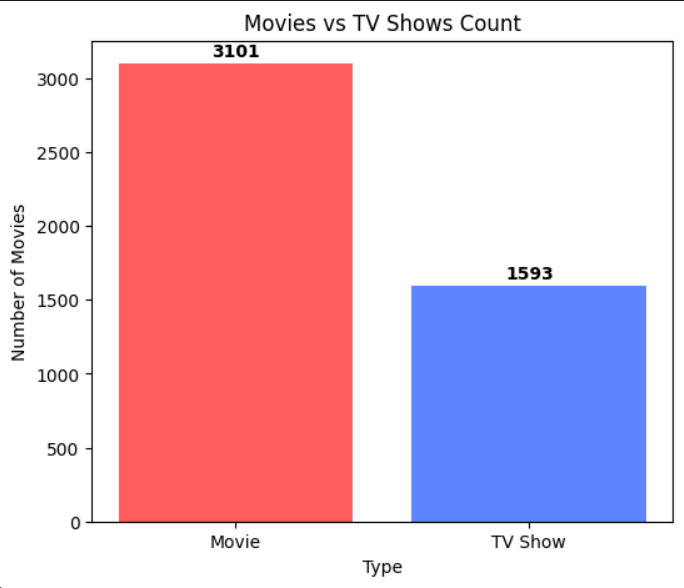
plt.xlabel('Type')

plt.ylabel('Number of Movies')

for i, val in enumerate(pandas\_df['count']):

plt.text(i, val + 50, str(val), ha='center', fontweight='bold')

plt.show()



In this visual it shown the distributions of total count of content on the based on it’s content type, As it is given there is only two type of contents here Movie and TV show.

* Based On Genres

top\_genres = genre\_df.groupBy("genre").count().orderBy("count", ascending=False).limit(10)

display(top\_genres)

genre\_pd = top\_genres.toPandas()

plt.figure(figsize=(10, 5))

plt.bar(genre\_pd['genre'], genre\_pd['count'], color='#F7B267', edgecolor='black')

plt.title('Top 10 Genres', fontsize=14, fontweight='bold')

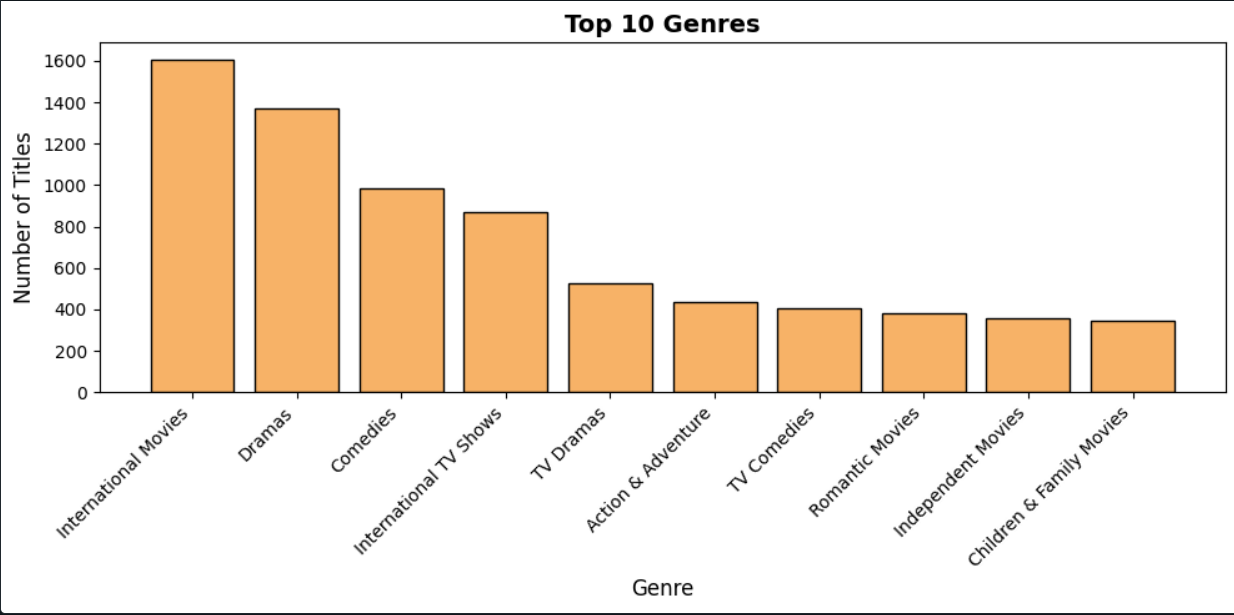
plt.xlabel('Genre', fontsize=12)

plt.ylabel('Number of Titles', fontsize=12)

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()



In this visual it shows the distribution of content based on the content genres there are more than 50+ type of genres.

* Based on Content Ratings

rating\_df = df.groupBy("rating").count().orderBy("count", ascending=False).limit(10)

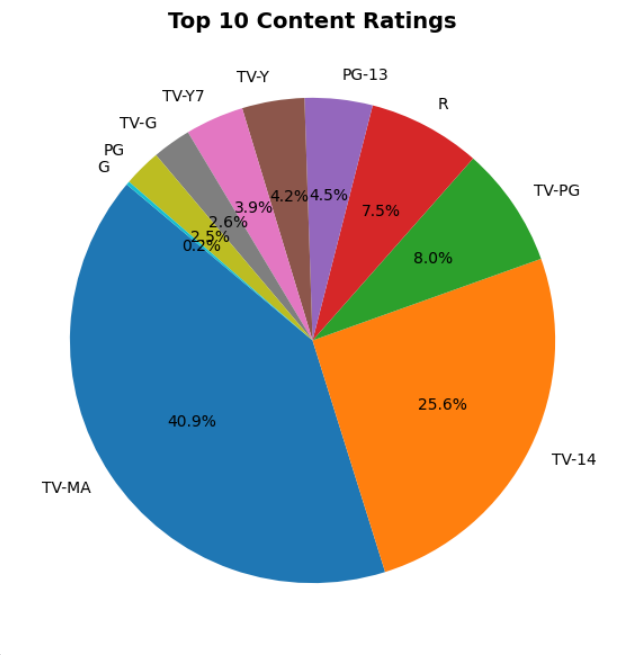
rating\_pd = rating\_df.toPandas()

plt.figure(figsize=(7,7))

plt.pie(rating\_pd['count'], labels=rating\_pd['rating'], autopct='%1.1f%%', startangle=140)

plt.title('Top 10 Content Ratings', fontsize=14, fontweight='bold')

plt.show()



* Country Level Analysis

plt.figure(figsize=(12, 6))

plt.bar(top\_countries\_pd['country'], top\_countries\_pd['count'], color='#4ECDC4', edgecolor='black')

plt.title('Top 15 Countries by Netflix Content', fontsize=14, fontweight='bold')

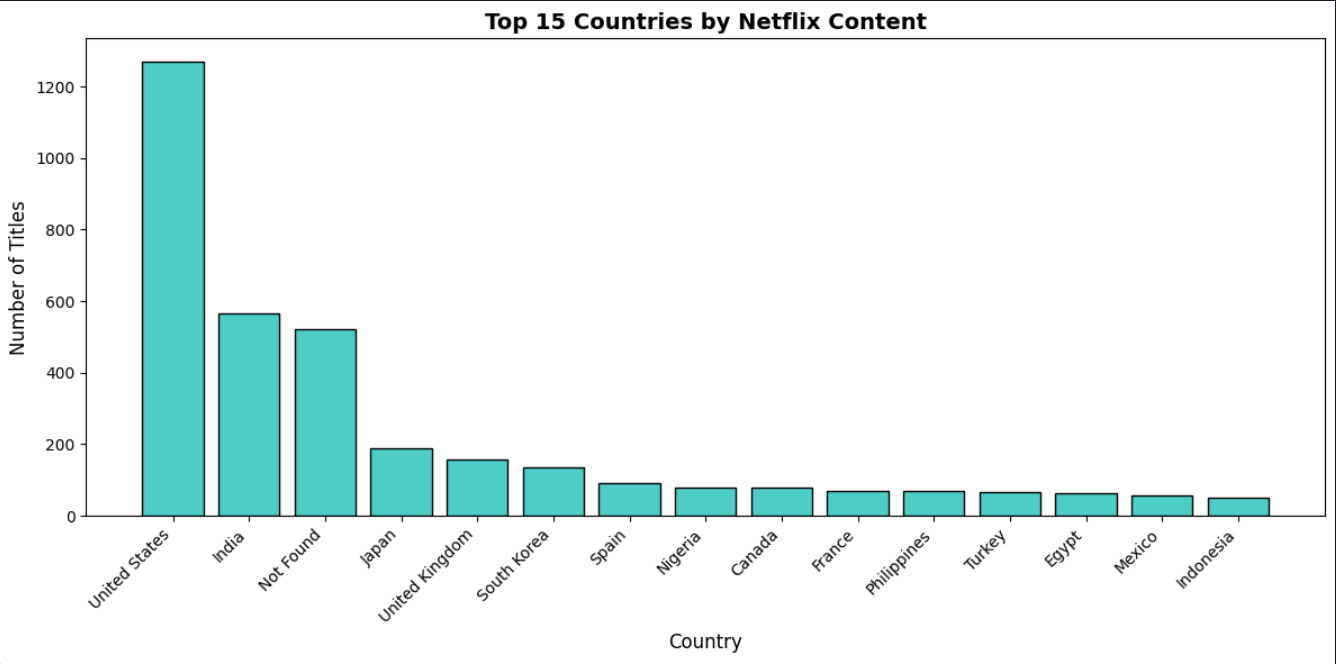
plt.xlabel('Country', fontsize=12)

plt.ylabel('Number of Titles', fontsize=12)

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()



In this visual it is showing the top 15 countries in Non-Increasing Order which is based on the Counting of the content released on there country.