# Netflix Data Analysis

Exploring Trends and Insights

# Objective

01) Import and Clean the Netflix database

02) Handle Missing values, convert data formats, and prepare the data

03) Analyze trends and distributions using Pandas and Seaborn

04) Explore the data through various visualizations

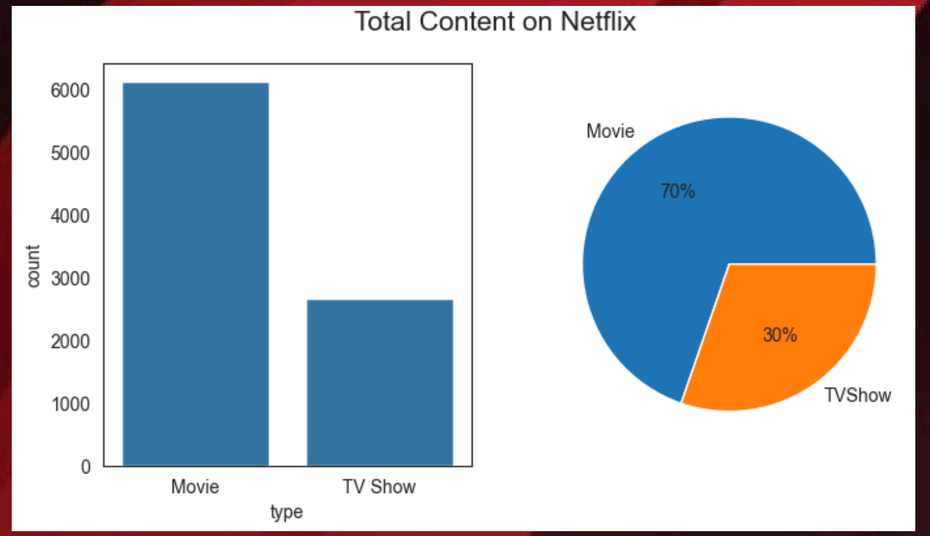
05) Extract key patterns based on content type, genres, ratings , and countries

# Importing Required Libraries and cleaning the dataset

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

data = pd.read_csv(r"C:\Users\91931\Downloads\Intern\Dataset\ netflix1.csv")

# Checking for missing values.
data.isnull().sum()

# Drop duplicates if present.
data.drop_duplicates(inplace=True)

# Convert date_added to datetime
data['date_added'] = pd.to_datetime(data['date_added'])
```

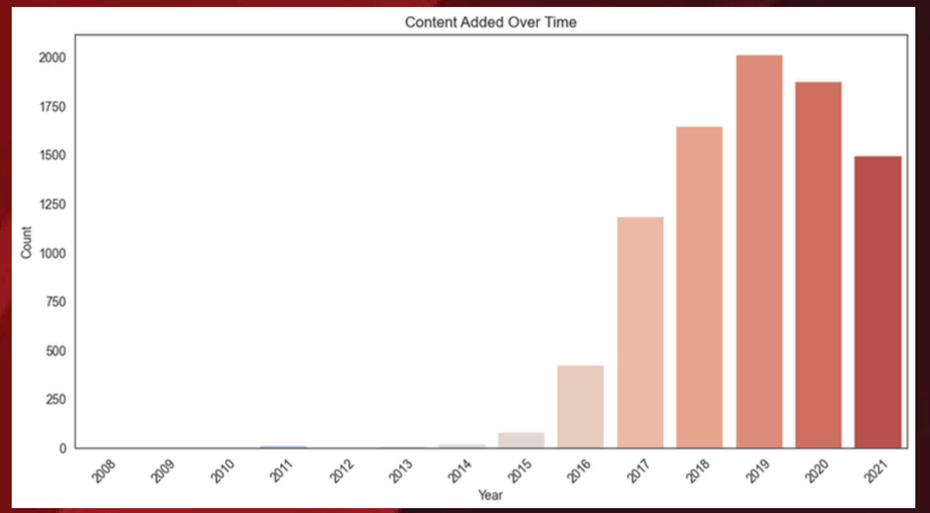
Exploratory Data Analysis

```
# Content Type Distribution(Movies vs TV Shows)
type_counts = data['type'].value_counts()
# Plot the distribution
freq = data['type'].value_counts()
fig,axes = plt.subplots(1,2,figsize=(8,4))
sns.countplot(data,x=data['type'],ax=axes[0])
plt.pie(freq,labels=['Movie','TVShow'],autopct='%.0f%%')
plt.suptitle('Total Content on Netflix',fontsize=15)
```



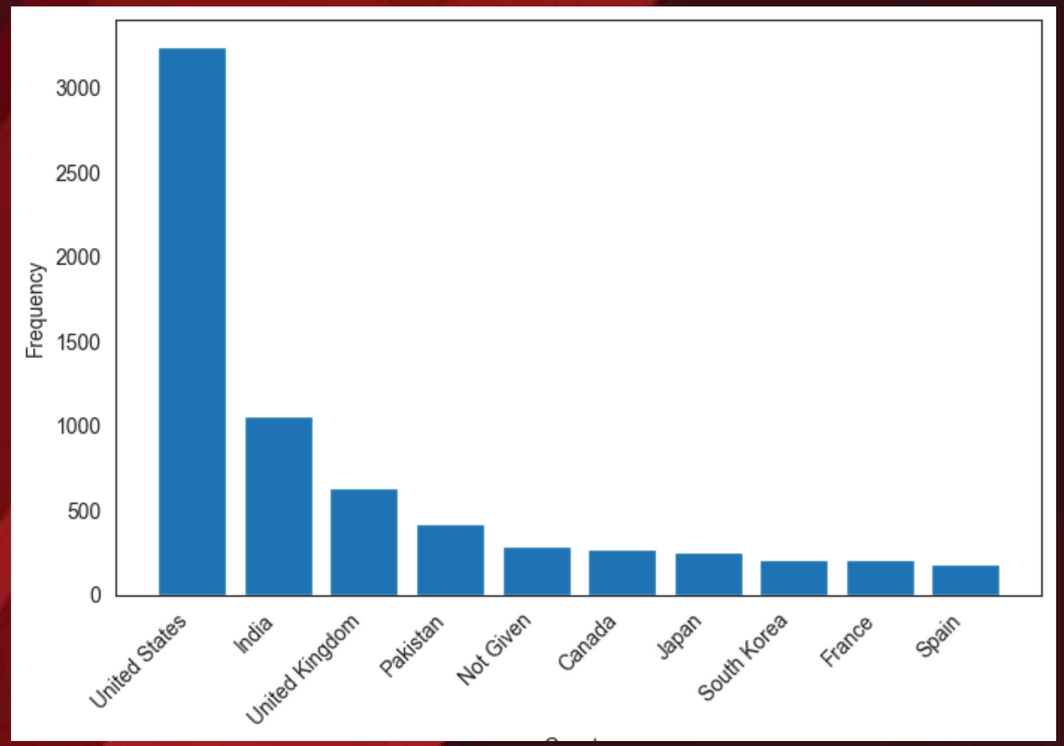

Exploratory Data Analysis

```
### Content Added Over Time

# Extract year and month from 'date_added'

data['year_added'] = data['date_added'].dt.year
data['month_added'] = data['date_added'].dt.month

# Plot content added over the years

plt.figure(figsize=(12, 6))
sns.countplot(x='year_added', data=data, palette='coolwarm')
plt.title('Content Added Over Time')
plt.xlabel('Year')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```



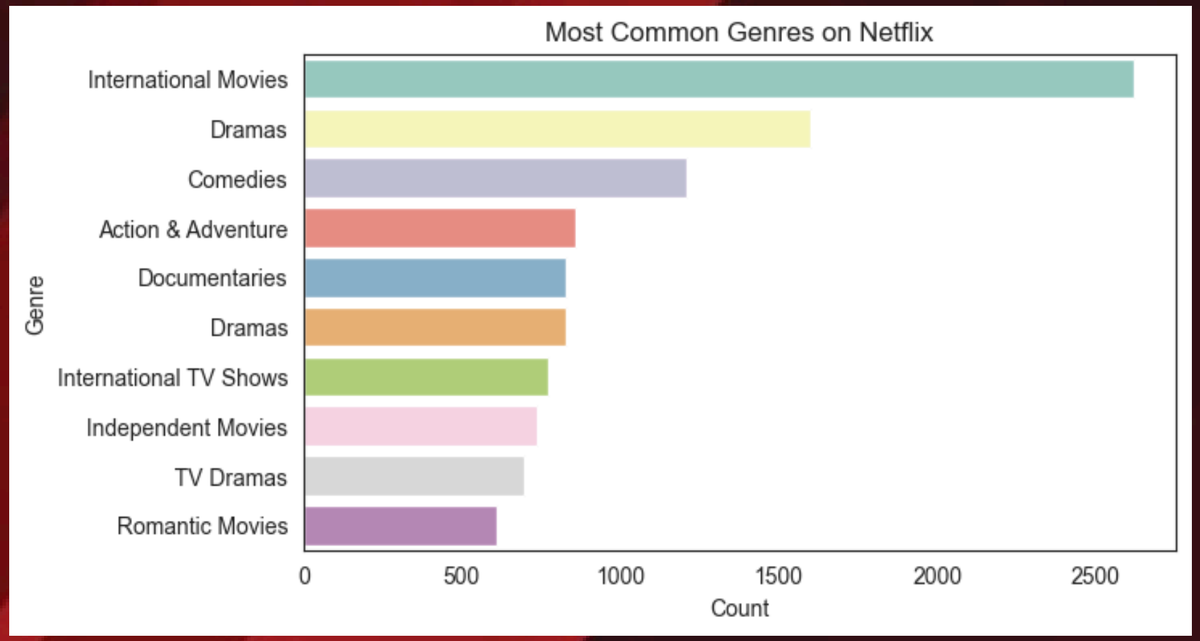
Top 10 Countries with most Content

```
top_ten_countries=data['country'].value_counts().reset_index()
 .sort_values(by='count',ascending=False)[:10]
plt.figure(figsize=(8,5))
plt.bar(top_ten_countries['country'],
top_ten_countries['count'])
plt.xticks(rotation=45,ha='right')
plt.xlabel("Country")
plt.ylabel("Frequency")
plt.suptitle("Top10 countries with most Content On Netflix")
plt.show()
```



Most Common Genres

```
# Split the 'listed_in' column and count genres

data['genres'] = data['listed_in'].apply(lambda x: x.split(','))
all_genres = sum(data['genres'], [])
genre_counts = pd.Series(all_genres).value_counts().head(10)

# Plot the most common genres
plt.figure(figsize=(7, 4))

sns.barplot(x=genre_counts.values,y=genre_counts.index,
palette='Set3')
sns.set_style("white")
# plt.grid(False)
plt.title('Most Common Genres on Netflix')
plt.xlabel('Count')
plt.ylabel('Genre')
plt.show()
```



# Predictive Analysis

Predictive Analysis

```
#Convert duration into numeric:
data['duration_num'] = data['duration'].str.extract('(\d+)').astype(float)

# Encode categorical variables[Label Encoding or One-Hot encoding]

from sklearn.preprocessing import LabelEncoder

label_cols = ['rating','director','country']
for col in label_cols:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col].astype(str))
```

Random Forest

```
# Train a model(Random Forest)

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train,Y_train)

Y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(Y_test,Y_pred))
print("Classification Report:\n", classification_report(Y_test,Y_pred))
```

Preparing Features and Target

```
# Prepare Features and target

features = ['release_year','duration_num','rating','country','year','month']
X = data[features].dropna()
Y = data.loc[X.index, 'type'].apply(lambda x:1 if x == 'Movie' else 0)#BinaryTarget

# Split data
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2,
random_state=42)
```

Random Forest

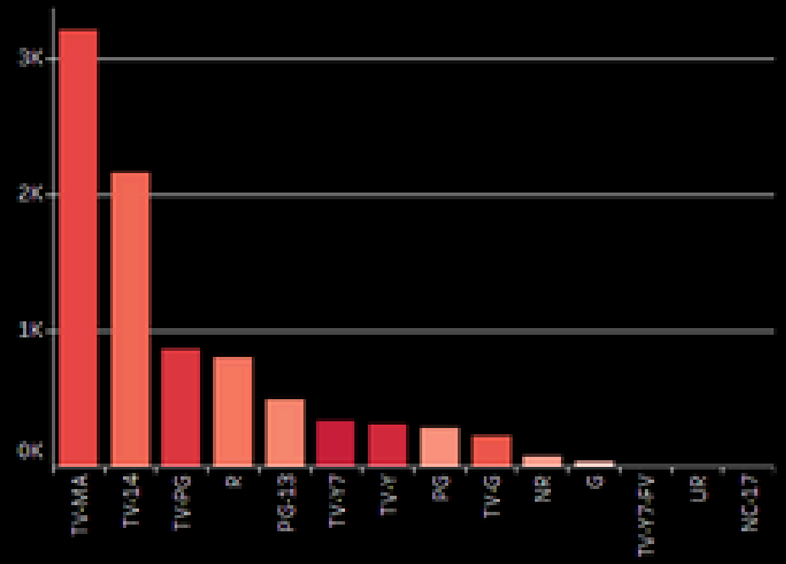
```
Accuracy: 0.997155858930603
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       520
           1       1.00      1.00      1.00      1238

    accuracy                           1.00      1758
   macro avg       1.00      1.00      1.00      1758
weighted avg       1.00      1.00      1.00      1758
```

# Key Insights

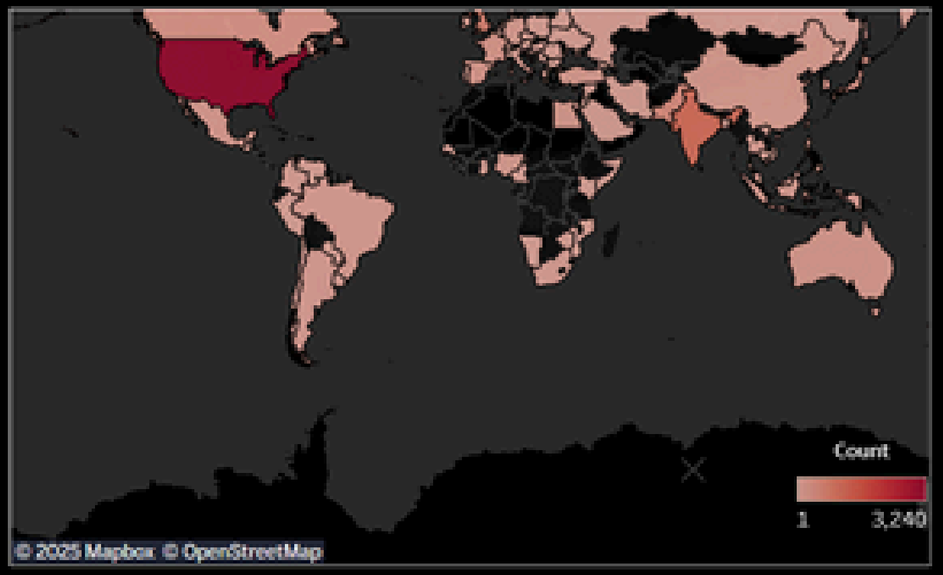
- Netflix has more movies than the TV Shows.
- TV -MA and TV-14 are the most common Ratings.
- Most Content added during 2019, 2020 and 2021.
- United States and India are the top content producers.
- Most common genres on Netflix are International Movies, Dramas, and Comedies.

# Netflix Dasboard

## Ratings

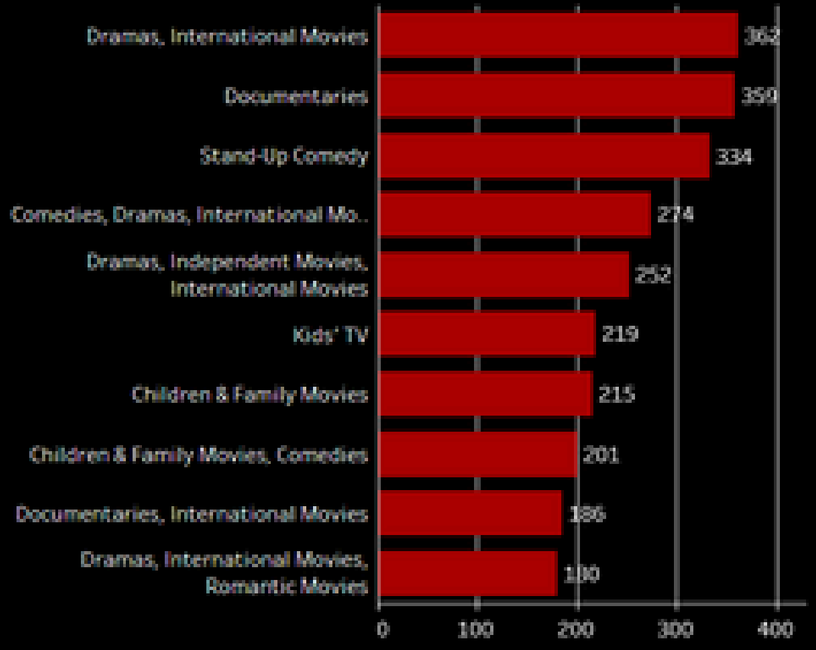



## Total Movies and TV Shows by country




## Movies and TV Shows Released over the Years


## Top 10 Genre




## Movies and TV Shows Distribution