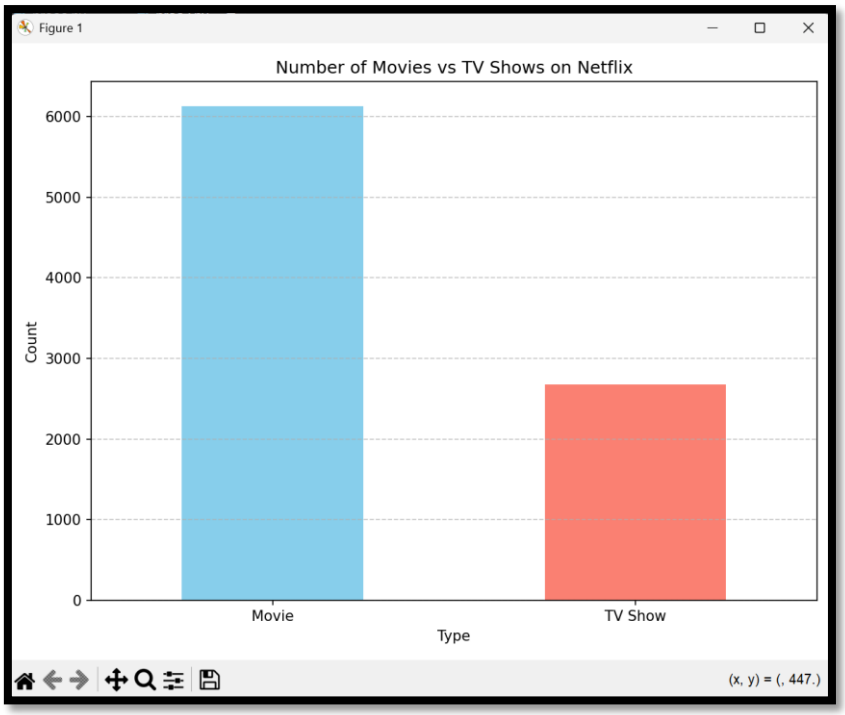
DATA VISULISATION OF THE DATASET

1) How many Movies vs TV Shows?
   CODE-

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

df = pd.read_csv(r"C:\Users\soume\Desktop\DATA ANALYST\SQL
PROJECT_ADV\netflix_titles.csv")

type_counts = df['type'].value_counts()

plt.figure(figsize=(8, 6))
type_counts.plot(kind='bar', color=['skyblue', 'salmon'])
plt.title('Number of Movies vs TV Shows on Netflix')
plt.xlabel('Type')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

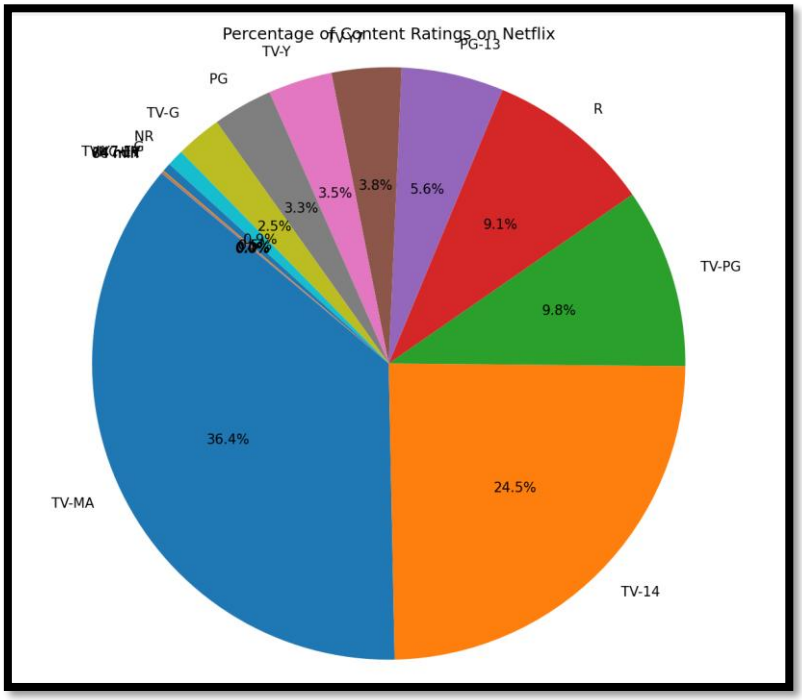
   OUTPUT



2) What is the percentage of each content rating (PG, R, TV-MA)?
CODE-

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


df = pd.read_csv(r"C:\Users\soume\Desktop\DATA ANALYST\SQL PROJECT_ADV\netflix_titles.csv")

rating_counts = df['rating'].value_counts(normalize=True) * 100

plt.figure(figsize=(10, 8))
plt.pie(rating_counts, labels=rating_counts.index, autopct='%1.1f%%', startangle=140)
plt.title('Percentage of Content Ratings on Netflix')
plt.axis('equal')
plt.tight_layout()
plt.show()

rating_percentages = rating_counts.round(2).reset_index()
rating_percentages.columns = ['Rating', 'Percentage']
print(rating_percentages)
```

OUTPUT-



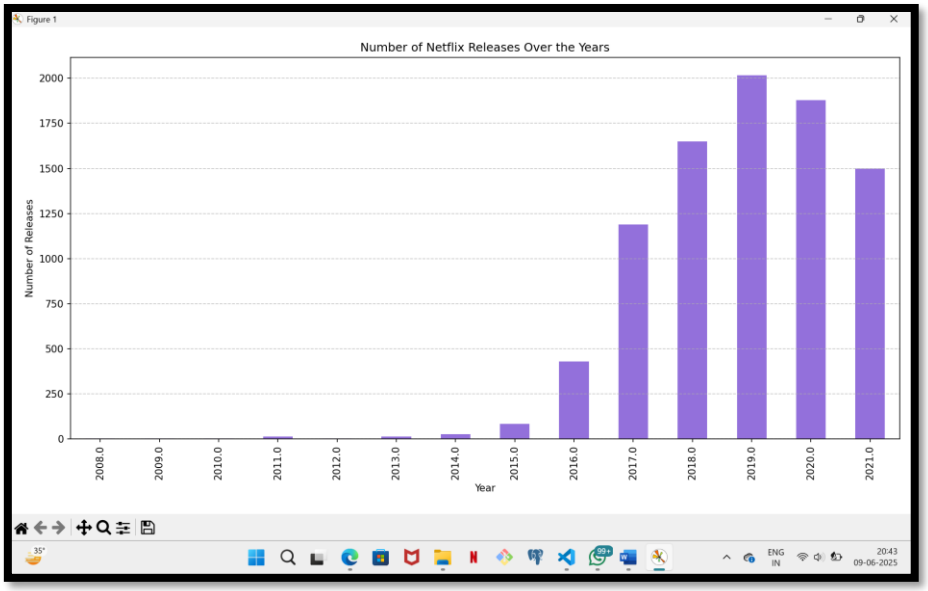
3) What is the distribution of movie durations?

CODE-

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

# Load the dataset

df = pd.read_csv(r"C:\Users\soume\Desktop\DATA ANALYST\SQL PROJECT_ADV\netflix_titles.csv")

# Strip leading/trailing spaces in 'date_added' column
df['date_added'] = df['date_added'].str.strip()

# Convert 'date_added' to datetime format
df['date_added'] = pd.to_datetime(df['date_added'], format="%B %d, %Y", errors='coerce')

# Extract year from 'date_added'
df['year_added'] = df['date_added'].dt.year
```

```
# Drop rows with missing year
df_clean = df.dropna(subset=['year_added'])

# Count number of releases by year
release_trend = df_clean['year_added'].value_counts().sort_index()

# Plot the trend
plt.figure(figsize=(12, 6))
release_trend.plot(kind='bar', color='mediumpurple')
plt.title('Number of Netflix Releases Over the Years')
plt.xlabel('Year')
plt.ylabel('Number of Releases')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

OUTPUT-



4) Top 10 countries with the highest number of shows?
   CODE-

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

# Load the dataset

df = pd.read_csv(r"C:\Users\soume\Desktop\DATA ANALYST\SQL PROJECT_ADV\netflix_titles.csv")

# Drop missing values in 'country'
df_clean = df.dropna(subset=['country']).copy()  # Use .copy() to avoid SettingWithCopyWarning

# Split and explode countries
df_clean.loc[:, 'country'] = df_clean['country'].str.split(',')  # Explicitly use .loc
df_exploded = df_clean.explode('country')

# Strip extra whitespace
df_exploded['country'] = df_exploded['country'].str.strip()

# Count top 10 countries
country_counts = df_exploded['country'].value_counts().head(10)

# Plot
plt.figure(figsize=(10, 6))
country_counts.plot(kind='bar', color='coral')
plt.title('Top 10 Countries with the Highest Number of Netflix Shows')
plt.xlabel('Country')
plt.ylabel('Number of Shows')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

   OUTPUT-

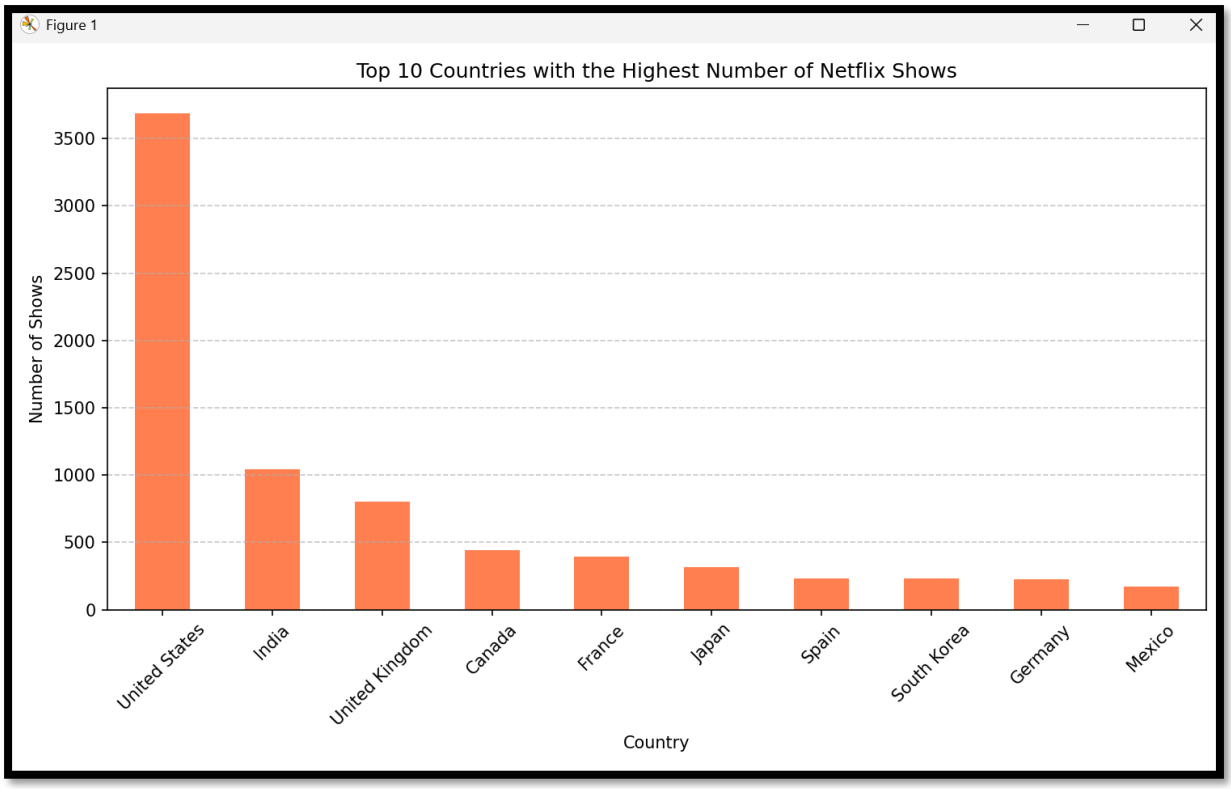


Figure 1
Top 10 Countries with the Highest Number of Netflix Shows
Number of Shows
0
500
1000
1500
2000
2500
3000
3500
United States
India
United Kingdom
Canada
France
Japan
Spain
South Korea
Germany
Mexico
Country