Exploratory Data Analysis – Netflix Movies & TV Shows Dataset

Step 1: Load the dataset and preview the top 5 rows.

```
import pandas as pd
df = pd.read_csv('netflix_titles.csv')
df.head()
```

Step 2: Check Missing Values in each column

```
df.isnull().sum()
```

Step 3: Fill missing text data with "Unknown" and remove rows missing critical date data.

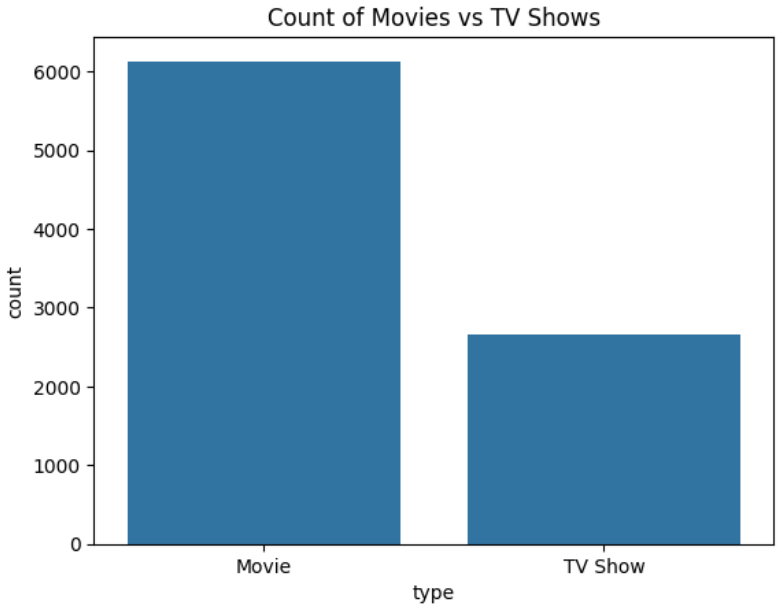
```
df['director'] = df['director'].fillna('Unknown')
df['cast'] = df['cast'].fillna('Unknown')
df['country'] = df['country'].fillna('Unknown')
df = df.dropna(subset=['date_added'])
```

Step 4: Converts string date to date time object and extracts year/month.

```
df['date_added'] = df['date_added'].str.strip()
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
df['year_added'] = df['date_added'].dt.year
df['month_added'] = df['date_added'].dt.month
```

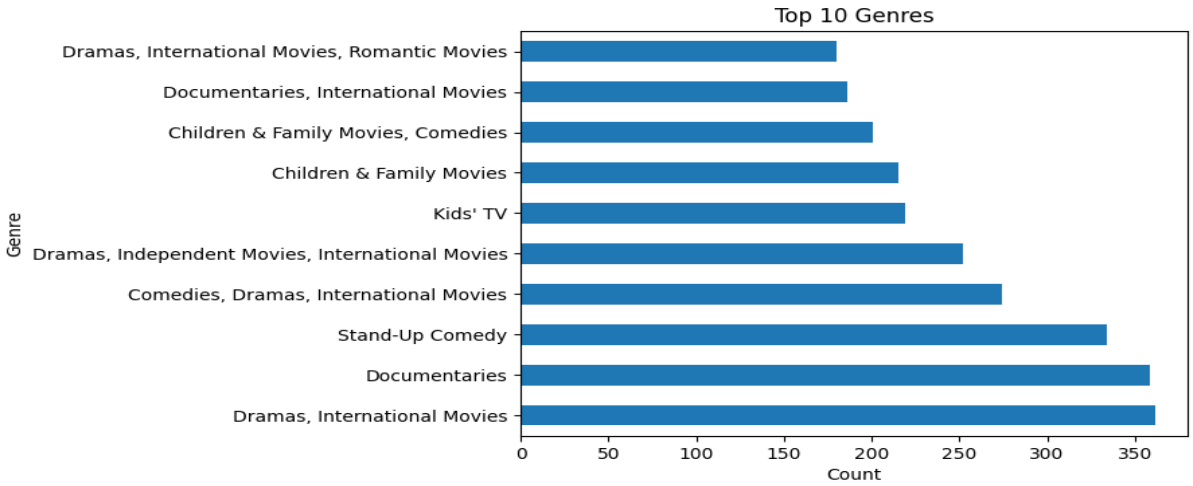
Step 5: Univariate Analysis

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.countplot(data=df, x='type')
plt.title('Count of Movies vs TV Shows')
plt.show()
```



Step 6: Display top 10 genres using horizontal bar plot.

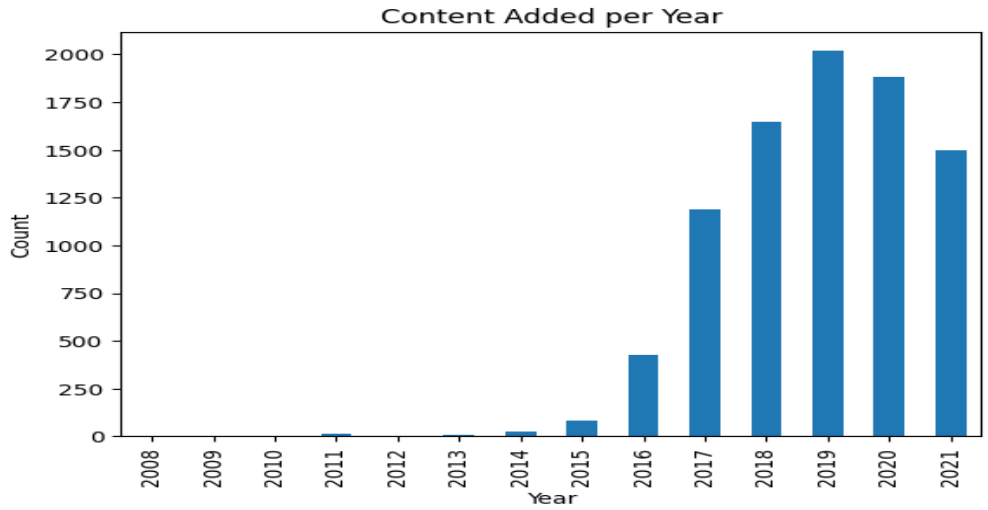
```
df['listed_in'].value_counts().head(10).plot(kind='barh')
plt.title('Top 10 Genres')
plt.xlabel('Count')
plt.ylabel('Genre')
plt.show()
```



Step 7: Shows trend of content addition over years.

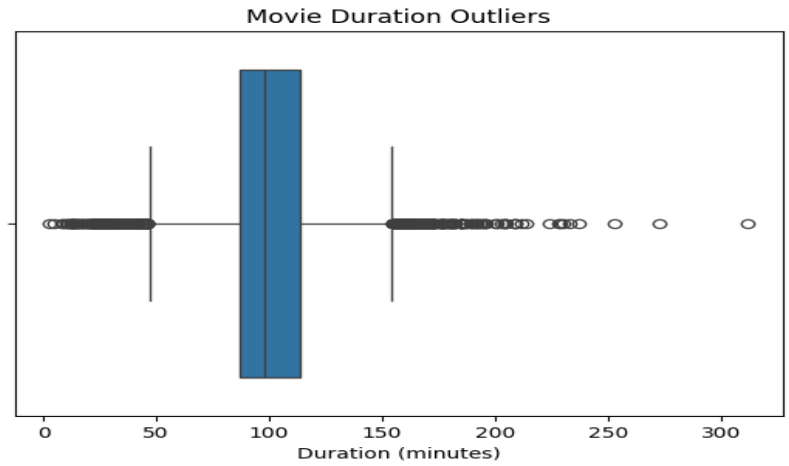
```
df['year_added'].value_counts().sort_index().plot(kind='bar')
plt.title('Content Added per Year')
```

```
plt.xlabel('Year')
plt.ylabel('Count')
plt.show()
```



Step 8: Outlier Detection- Extracts and visualizes duration outliers for movies.

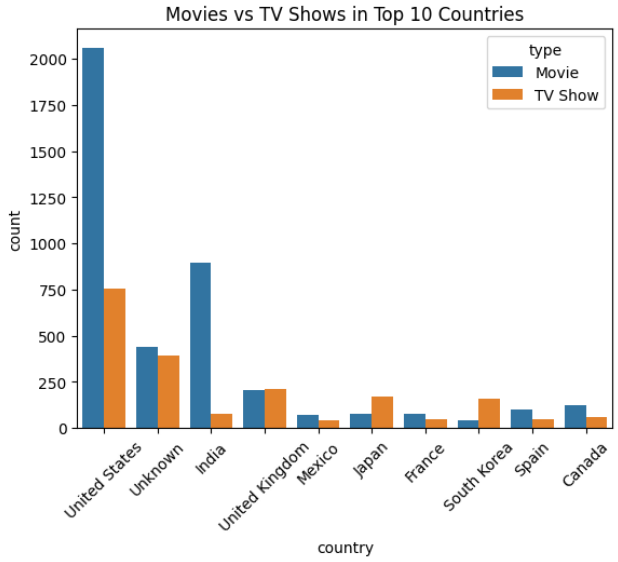
```
df[['duration_int', 'duration_type']] = df['duration'].str.extract(r'(\d+)\s*(\w+)')
df['duration_int'] = pd.to_numeric(df['duration_int'], errors='coerce')
sns.boxplot(data=df[df['type'] == 'Movie'], x='duration_int')
plt.title('Movie Duration Outliers')
plt.xlabel('Duration (minutes)')
plt.show()
```



Step 9(a): Bivariate Analysis- Compares movie vs show count across top countries.

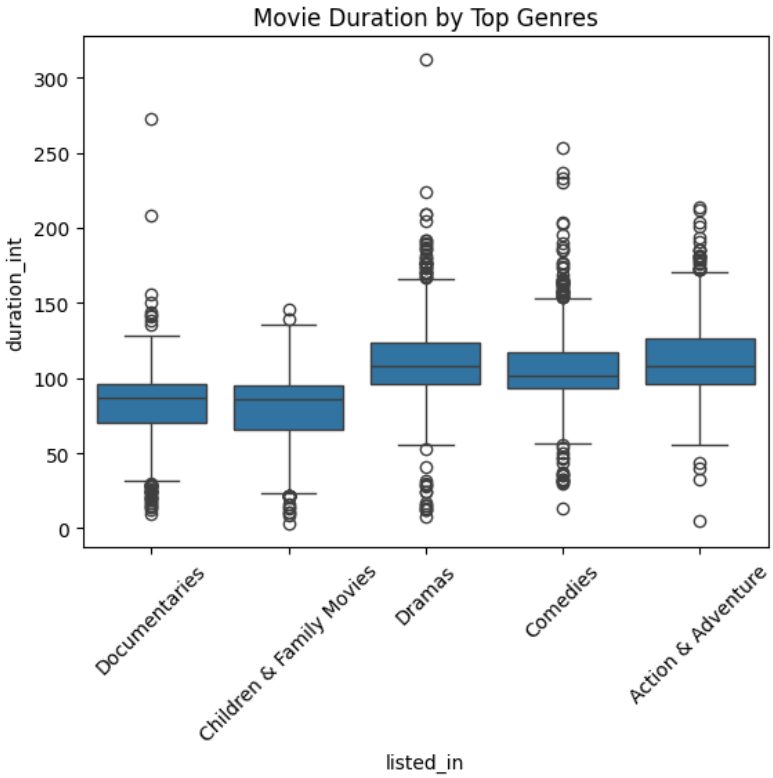
```
top_countries = df['country'].value_counts().head(10).index
sns.countplot(data=df[df['country'].isin(top_countries)], x='country', hue='type')
plt.title('Movies vs TV Shows in Top 10 Countries')
```

```
plt.xticks(rotation=45)
plt.show()
```



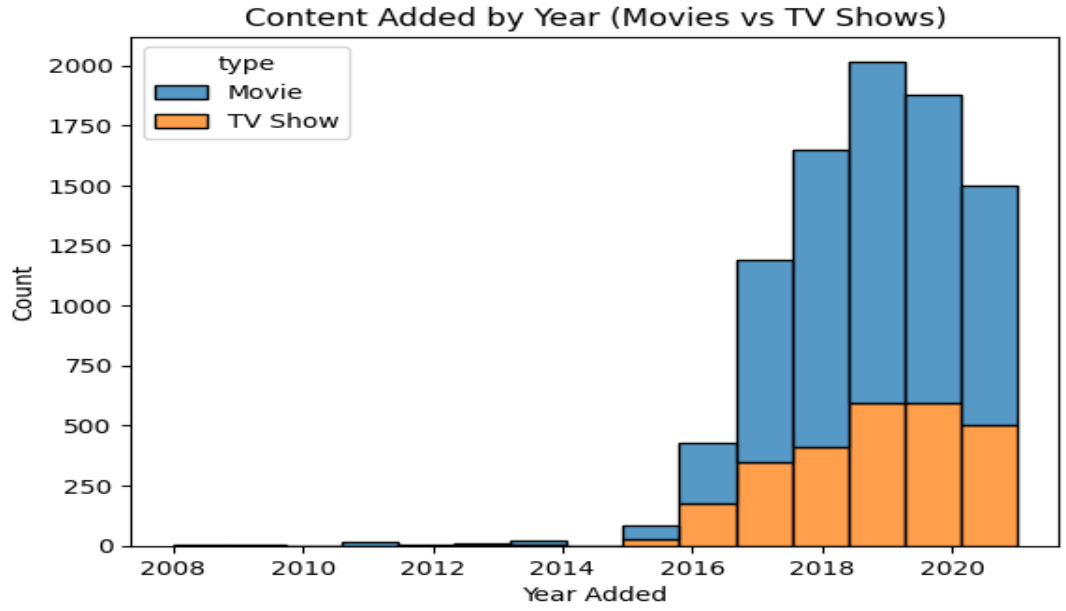
Step 9(b): Compares average movie duration across top genres.

```
movie_df = df[(df['type'] == 'Movie') & (df['duration_int'].notna())].copy()
movie_df['listed_in'] = movie_df['listed_in'].str.split(',').str[0]
top_genres = movie_df['listed_in'].value_counts().head(5).index
sns.boxplot(data=movie_df[movie_df['listed_in'].isin(top_genres)], x='listed_in', y='duration_int')
plt.title('Movie Duration by Top Genres')
plt.xticks(rotation=45)
plt.show()
```



Step 9(c): Histogram of content additions by type and year.

```
sns.histplot(data=df, x='year_added', hue='type', multiple='stack', bins=15)
plt.title('Content Added by Year (Movies vs TV Shows)')
plt.xlabel('Year Added')
plt.ylabel('Count')
plt.show()
```



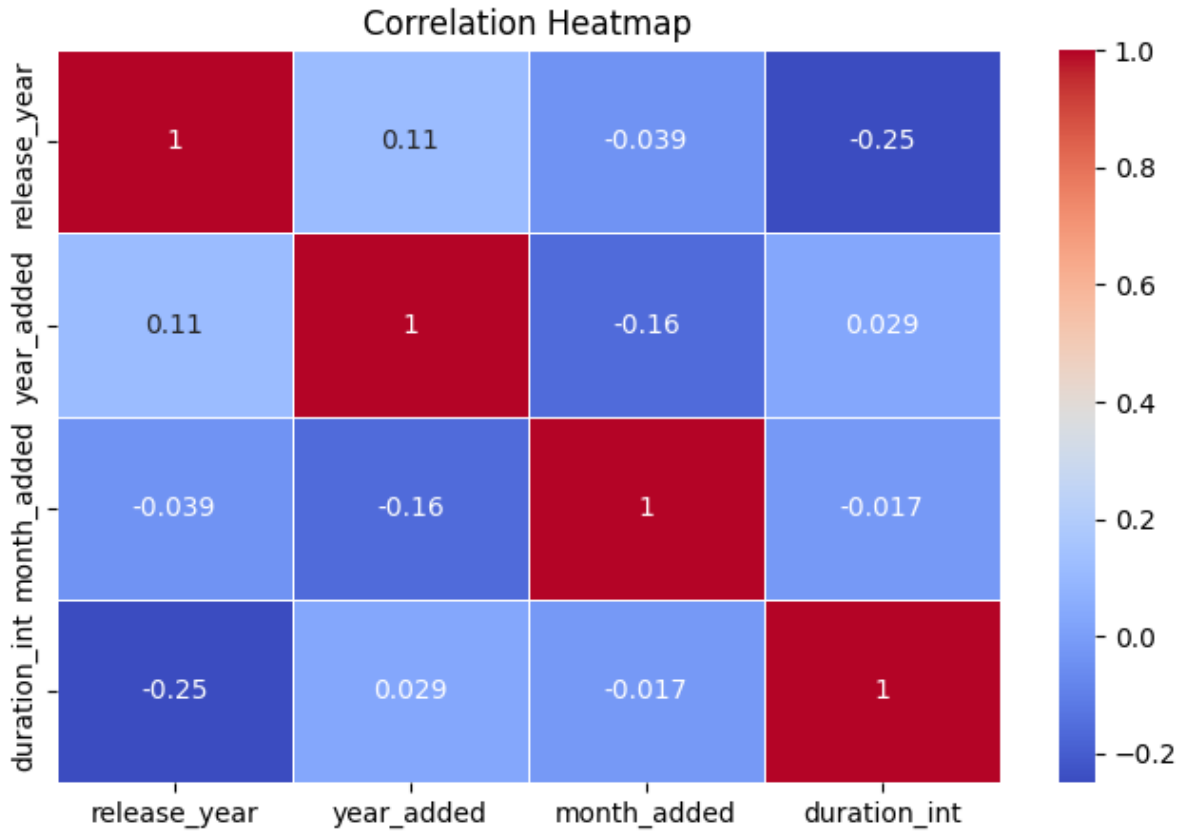
Step 10: Correlation Heatmap- Displays correlation between numeric features like duration and year/month added.

```
plt.figure(figsize=(8, 5))

sns.heatmap(df.select_dtypes(include='number').corr(), annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Heatmap')

plt.show()
```



Summary of Findings

- Movies dominate the platform.

- Top countries include US, India, UK.

- Most additions happened from 2017–2020.

- Genres like Drama and International Movies are common.

- Weak correlation exists between numerical features.