Defining Problem Statement and Analysing basic metrics

```
import pandas as pd
csv_path = '/content/netflix.csv'

# Load the dataset
data = pd.read_csv(csv_path)

# 1. Defining Problem Statement
problem_statement = "Netflix aims to make data-driven decisions to
determine the types of shows/movies to produce and develop strategies
for business growth in various countries."
print(problem_statement)
```

```
Netflix aims to make data-driven decisions to determine the types of
shows/movies to produce and develop strategies for business growth in
various countries.
```

```
from google.colab import drive
drive.mount('/content/drive')
```

Observations on the shape of data, data types of all the attributes, (If required), missing value detection, statistical summary

```
# 2. Observations on Basic Metrics
# 2.1. Shape of Data
data_shape = data.shape

# 2.2. Data Types of Attributes
data_types = data.dtypes

# 2.4. Missing Value Detection
missing_values = data.isnull().sum()

# 2.5. Statistical Summary (for numerical attributes)
# Summary statistics for numerical columns
summary_statistics = data.describe()

print("\n2. Observations on Basic Metrics:")
print(f"2.1. Shape of Data: {data_shape}")
print("\n2.2. Data Types of Attributes:")
```

```
print(data_types)
print("\n2.3. Missing Value Detection:")
print(missing_values)
print("\n2.4. Statistical Summary (for numerical attributes):")
print(summary_statistics)
```

```
2. Observations on Basic Metrics:
2.1. Shape of Data: (8807, 12)

2.2. Data Types of Attributes:
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
dtype: object

2.3. Missing Value Detection:
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64

2.4. Statistical Summary (for numerical attributes):
       release_year
count   8807.000000
mean    2014.180198
std        8.819312
min     1925.000000
25%     2013.000000
50%     2017.000000
75%     2019.000000
max     2021.000000
```

```
data.head(5)
```

```
   show_id     type                  title          director
cast  \
2       s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown
cast
3       s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown
cast
10      s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
11      s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
12      s2  TV Show         Blood & Water  Unknown director    Ama
Qamata

          country date_added  release_year rating   duration  \
2   United States 2021-01-01          2020  PG-13     90 min
3   United States 2021-01-01          2020  PG-13     90 min
10   South Africa 2021-01-01          2021  TV-MA  2 Seasons
11   South Africa 2021-01-01          2021  TV-MA  2 Seasons
12   South Africa 2021-01-01          2021  TV-MA  2 Seasons

                 listed_in
description  \
2            Documentaries            As her father nears the end of
his life
3            Documentaries   filmmaker Kirsten Johnson stages his
death in...
10  International TV Shows                    After crossing paths at
a party
11  International TV Shows   a Cape Town teen sets out to prove
whether a ...
12               TV Dramas                    After crossing paths at
a party

    duration_mins  week_of_year  month release_date  days_to_netflix
2            90.0            53      1   2020-01-01              366
3            90.0            53      1   2020-01-01              366
10            NaN            53      1   2021-01-01                0
11            NaN            53      1   2021-01-01                0
12            NaN            53      1   2021-01-01                0
```

Question: Handling null values

a. For categorical variables with null values, update those rows as unknown_column_name. Example : Replace missing value with Unknown Actor for missing value in Actors column.

b. Replace with 0 for continuous variables having null values.

Answer:

The column ***director, cast,country,date_added,rating and description*** has Null values which is replaced with 'Unkown columnName' respectively

```
for col in data.select_dtypes(include=['object']).columns:
    data[col] = data[col].fillna(f'Unknown {col}')
data.head(5)
```

```
  show_id     type                  title          director  \
0      s1    Movie   Dick Johnson Is Dead   Kirsten Johnson
1      s2  TV Show          Blood & Water  Unknown director
2      s3  TV Show              Ganglands   Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans  Unknown director
4      s5  TV Show           Kota Factory  Unknown director

                                                  cast          country
\
0                                        Unknown cast    United States

1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...     South Africa

2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...  Unknown country

3                                        Unknown cast  Unknown country

4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...             India


           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                          listed_in  \
0                                    Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                            Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...

                                        description
0  As her father nears the end of his life, filmm...
1  After crossing paths at a party, a Cape Town t...
2  To protect his family from a powerful drug lor...
3  Feuds, flirtations and toilet talk go down amo...
4  In a city of coaching centers known to train I...
```

Question: Un-nesting the columns

a. Un-nest the columns those have cells with multiple comma separated values by creating multiple row

Answer

The column which contains multiple values are splited into rows using explode

```
def has_comma_separated_values(series):
    return series.apply(lambda x: isinstance(x, str) and ',' in
x).any()

# Identify columns to explode
columns_to_explode = [col for col in data.columns if
has_comma_separated_values(data[col])]

# Explode each identified column
for col in columns_to_explode:
    data[col] = data[col].str.split(',')
    data = data.explode(col).reset_index(drop=True)

# Display the DataFrame after applying explode
data.head(15)
```

```
   show_id     type                  title          director
cast  \
0       s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown
cast
1       s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown
cast
2       s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown
cast
3       s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown
cast
4       s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
5       s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
6       s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
7       s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
8       s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
9       s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
10      s2  TV Show         Blood & Water  Unknown director    Ama
```

```
Qamata
11      s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
12      s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
13      s2  TV Show         Blood & Water  Unknown director    Ama
Qamata
14      s2  TV Show         Blood & Water  Unknown director    Ama
Qamata

          country    date_added  release_year rating   duration  \
0   United States  September 25          2020  PG-13     90 min
1   United States  September 25          2020  PG-13     90 min
2   United States          2021          2020  PG-13     90 min
3   United States          2021          2020  PG-13     90 min
4    South Africa  September 24          2021  TV-MA  2 Seasons
5    South Africa  September 24          2021  TV-MA  2 Seasons
6    South Africa  September 24          2021  TV-MA  2 Seasons
7    South Africa  September 24          2021  TV-MA  2 Seasons
8    South Africa  September 24          2021  TV-MA  2 Seasons
9    South Africa  September 24          2021  TV-MA  2 Seasons
10   South Africa          2021          2021  TV-MA  2 Seasons
11   South Africa          2021          2021  TV-MA  2 Seasons
12   South Africa          2021          2021  TV-MA  2 Seasons
13   South Africa          2021          2021  TV-MA  2 Seasons
14   South Africa          2021          2021  TV-MA  2 Seasons

                  listed_in
description
0           Documentaries            As her father nears the end of
his life
1           Documentaries   filmmaker Kirsten Johnson stages his
death in...
2           Documentaries            As her father nears the end of
his life
3           Documentaries   filmmaker Kirsten Johnson stages his
death in...
4   International TV Shows                    After crossing paths at
a party
5   International TV Shows   a Cape Town teen sets out to prove
whether a ...
6                TV Dramas                    After crossing paths at
a party
7                TV Dramas   a Cape Town teen sets out to prove
whether a ...
8             TV Mysteries                    After crossing paths at
a party
9             TV Mysteries   a Cape Town teen sets out to prove
whether a ...
```

```
10  International TV Shows                    After crossing paths at
a party
11  International TV Shows   a Cape Town teen sets out to prove
whether a ...
12               TV Dramas                    After crossing paths at
a party
13               TV Dramas   a Cape Town teen sets out to prove
whether a ...
14            TV Mysteries                    After crossing paths at
a party
```

Question 1:

Find the counts of each categorical variable both using graphical and nongraphical analysis.

```
       a. For Non-graphical Analysis:

       b. For graphical analysis:
```

1)Non-Graphical Analysis: Value counts and unique attributes

2)Visual Analysis - Univariate, Bivariate after pre-processing of the data

3)For continuous variable(s): Distplot, countplot, histogram for univariate analysis

4)For categorical variable(s): Boxplot

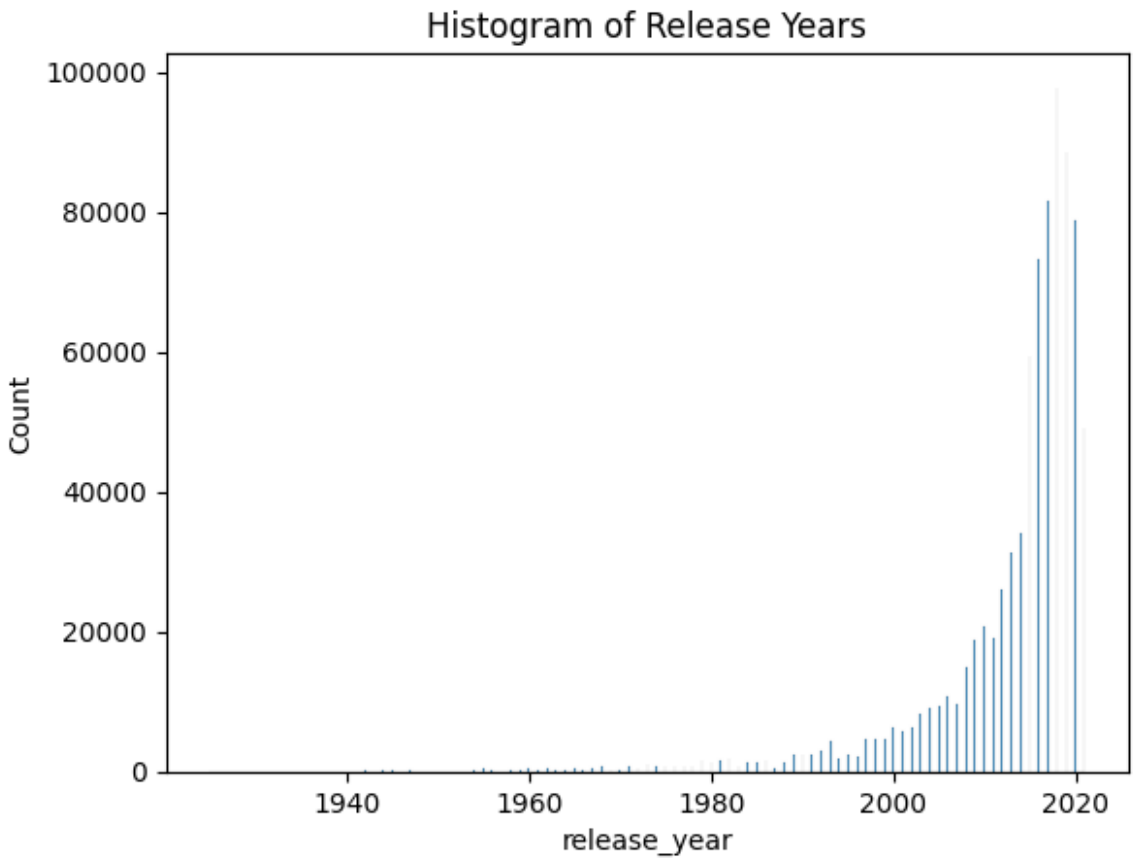
5)For correlation: Heatmaps, Pairplots

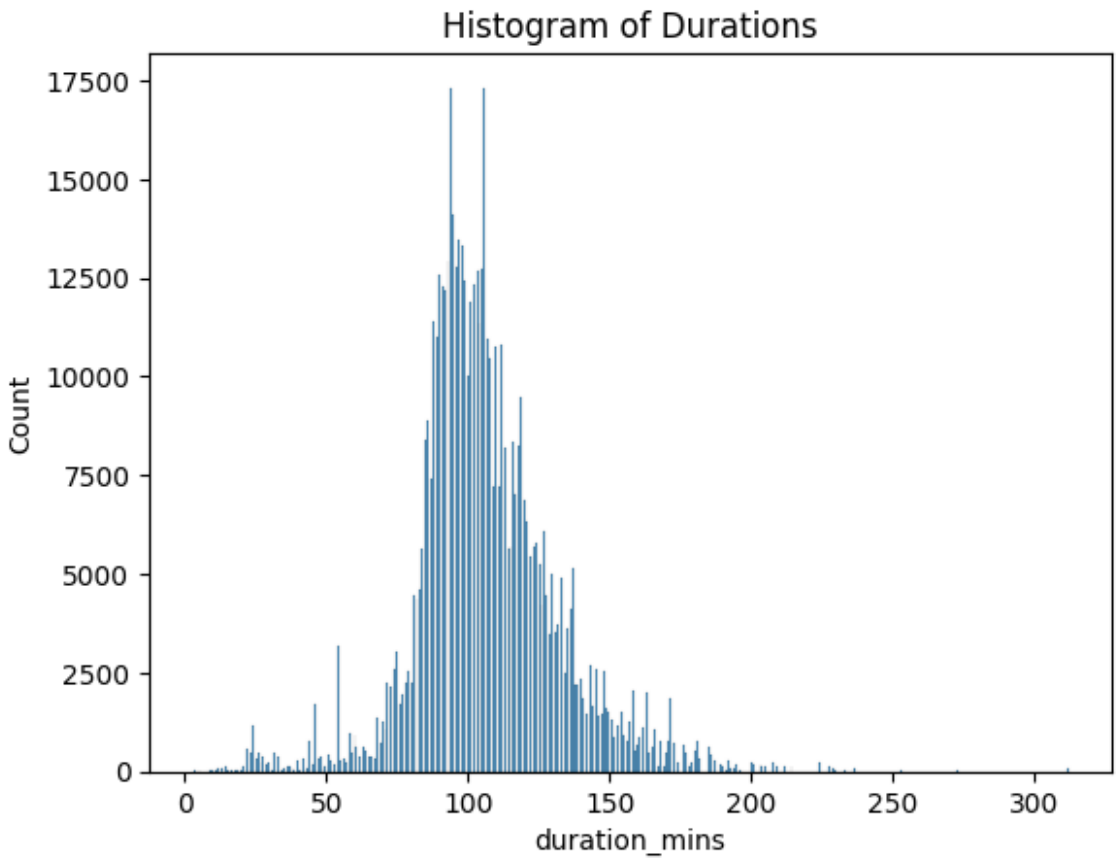
6)Missing Value & Outlier check

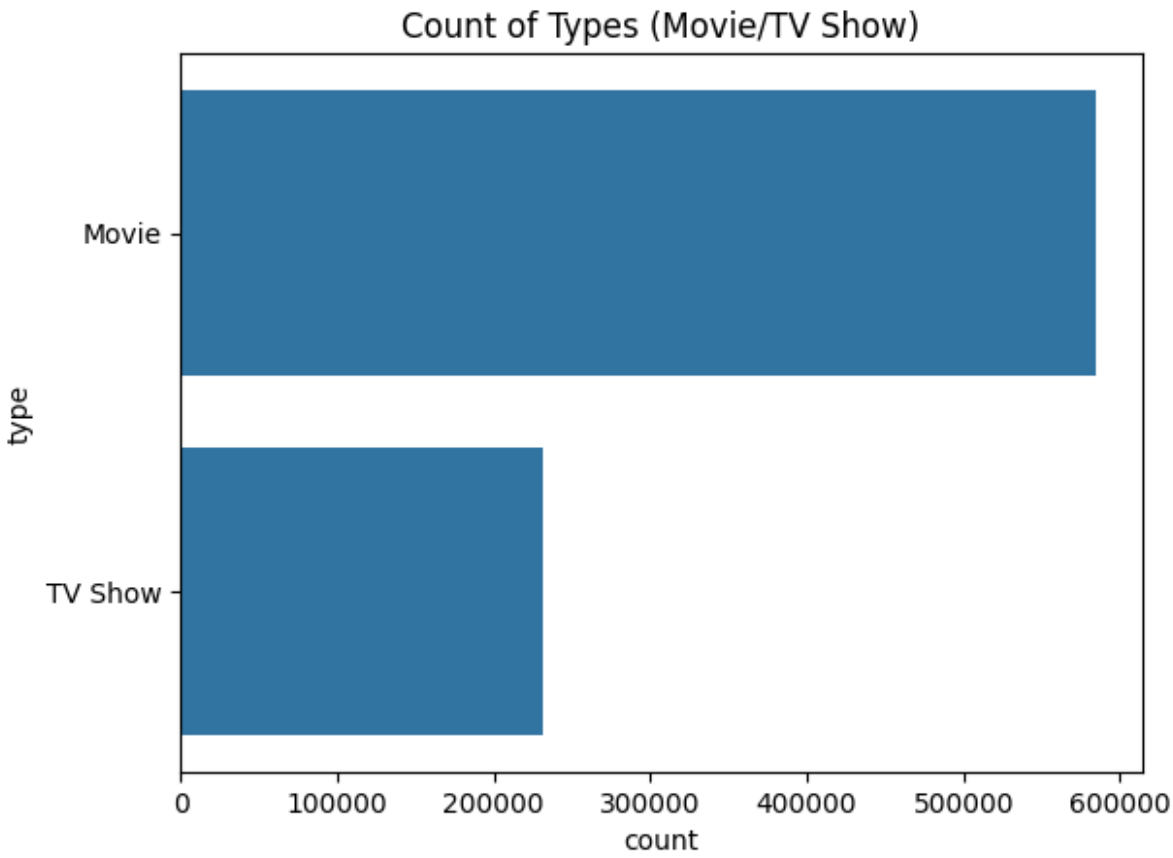
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns



# Identify categorical columns
categorical_columns=['show_id','type','director', 'cast', 'country',
'date_added', 'release_year', 'rating', 'duration', 'listed_in',
'description']


# Non-graphical analysis: value counts for each categorical column
for col in ['type', 'rating', 'duration']:
    print(data[col].value_counts())

# Graphical analysis:
```

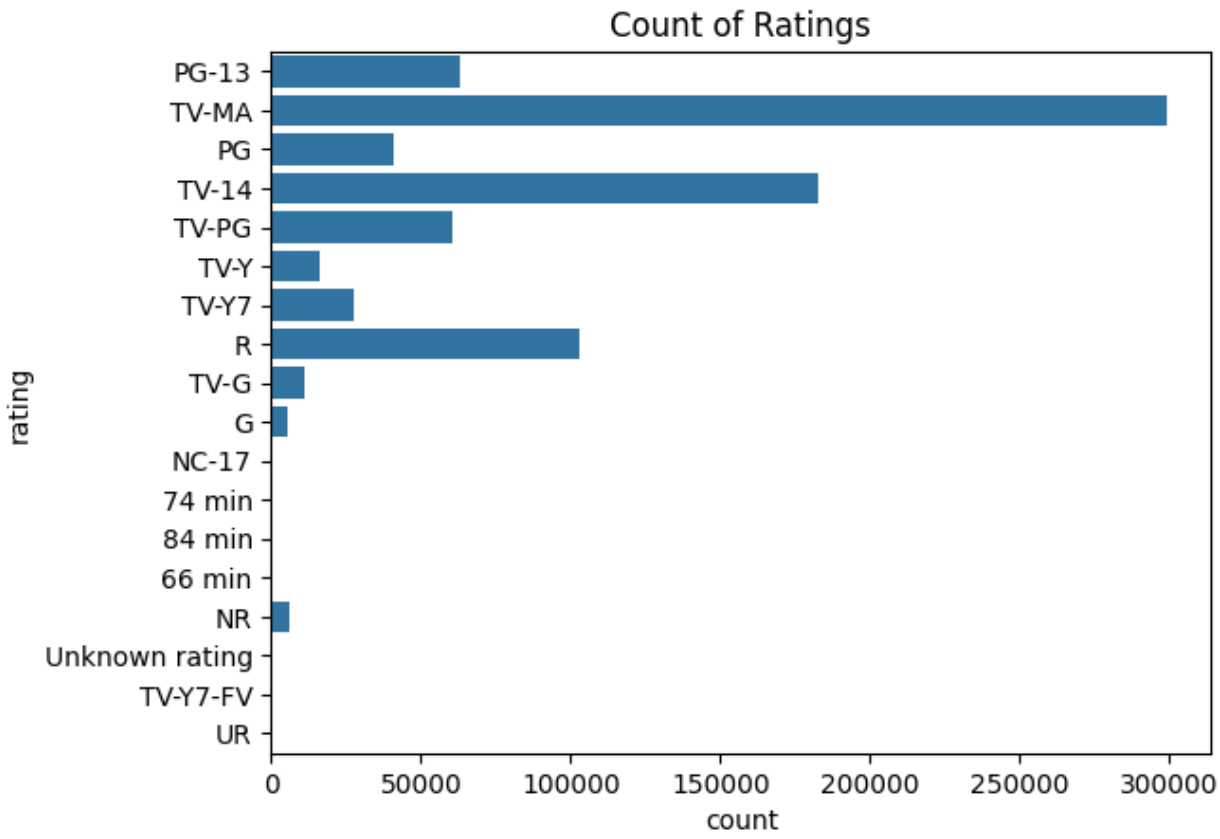
```
# Assuming 'release_year' and 'duration' (in minutes) are continuous
variables

sns.histplot(data['release_year'].dropna(), kde=False)
plt.title('Histogram of Release Years')
plt.show()

# Convert 'duration' to numeric if it's in the format 'XX min'
data['duration_mins'] = pd.to_numeric(data['duration'].str.replace('
min', ''), errors='coerce')
sns.histplot(data['duration_mins'].dropna(), kde=False)
plt.title('Histogram of Durations')
plt.show()


# Count plot for categorical variables
sns.countplot(y='type', data=data)
plt.title('Count of Types (Movie/TV Show)')
plt.show()

sns.countplot(y='rating', data=data)
plt.title('Count of Ratings')
plt.show()

# Assuming we have numerical data for correlation;

valid_numeric_cols = data.select_dtypes(include=[np.number]).columns
sns.heatmap(data[valid_numeric_cols].corr(numeric_only=True),
annot=True)
plt.title('Heatmap of Correlation')
plt.show()

# For pairplot, ensure to drop NaN values or fill them as needed
sns.pairplot(data[valid_numeric_cols].dropna())
plt.show()


# Missing values
print(data.isnull().sum())

# Outlier check with a boxplot example for duration in minutes
sns.boxplot(x='duration_mins', data=data)
plt.title('Boxplot of Duration in Minutes')
plt.show()



Movie      585034
TV Show    231723
Name: type, dtype: int64
TV-MA             298987
```

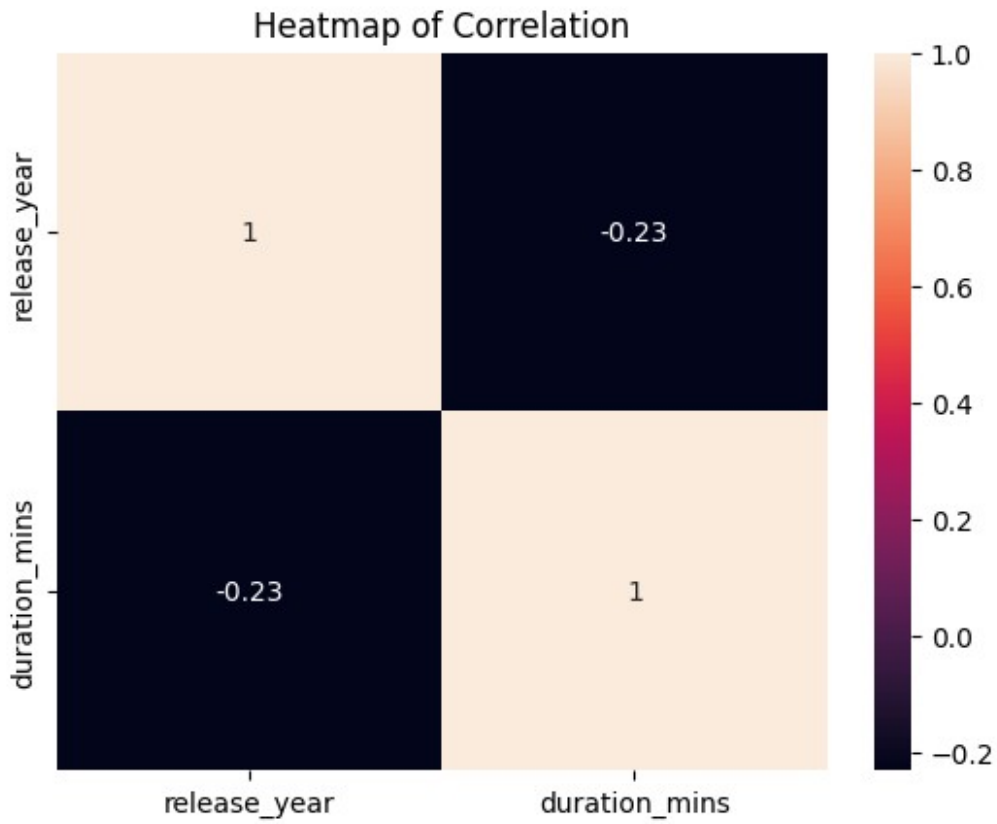
```
TV-14              182436
R                  103046
PG-13               62808
TV-PG               60689
PG                  40914
TV-Y7               27586
TV-Y                16253
TV-G                11154
NR                   5960
G                    5396
NC-17                 636
UR                    304
TV-Y7-FV              296
Unknown rating        268
74 min                 12
84 min                  6
66 min                  6
Name: rating, dtype: int64
1 Season     146342
2 Seasons     39855
3 Seasons     18904
94 min        17324
106 min       17322
              ...
10 min            8
11 min            8
9 min             8
5 min             6
8 min             4
Name: duration, Length: 221, dtype: int64
```

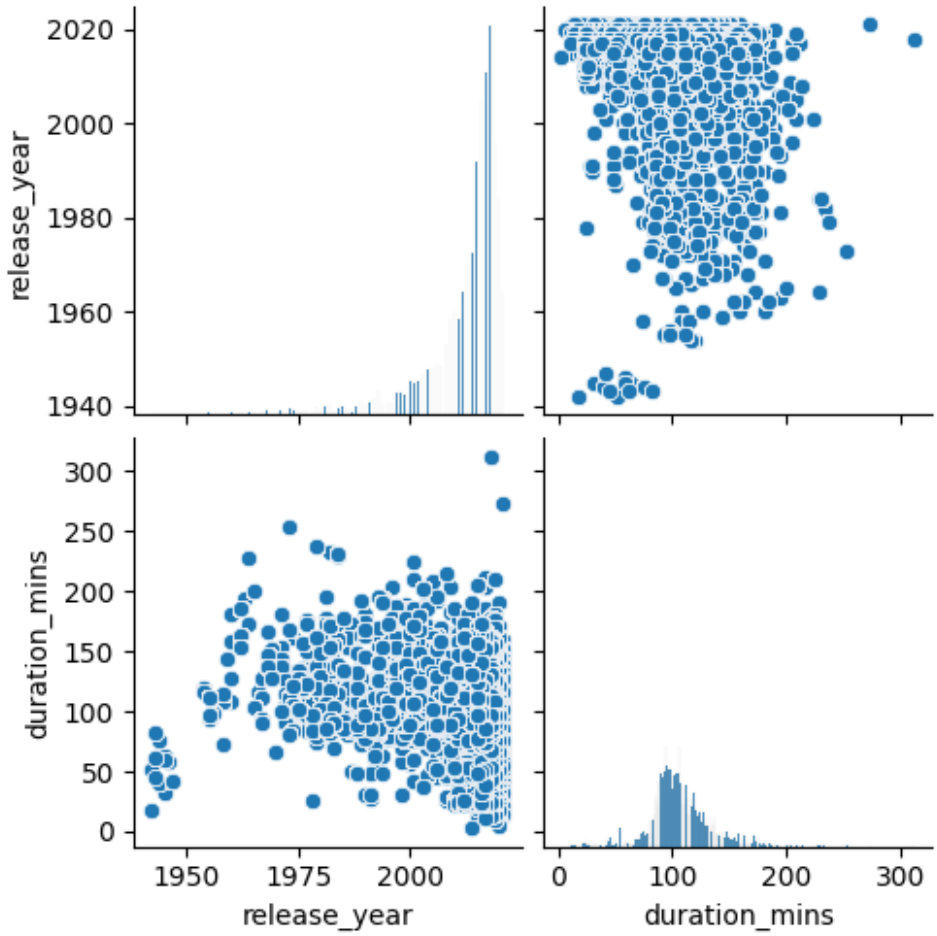

Histogram of Release Years
100000
80000
60000
40000
20000
0
Count
1940
1960
1980
2000
2020
release_year


Histogram of Durations
Count
17500
15000
12500
10000
7500
5000
2500
0
0
50
100
150
200
250
300
duration_mins

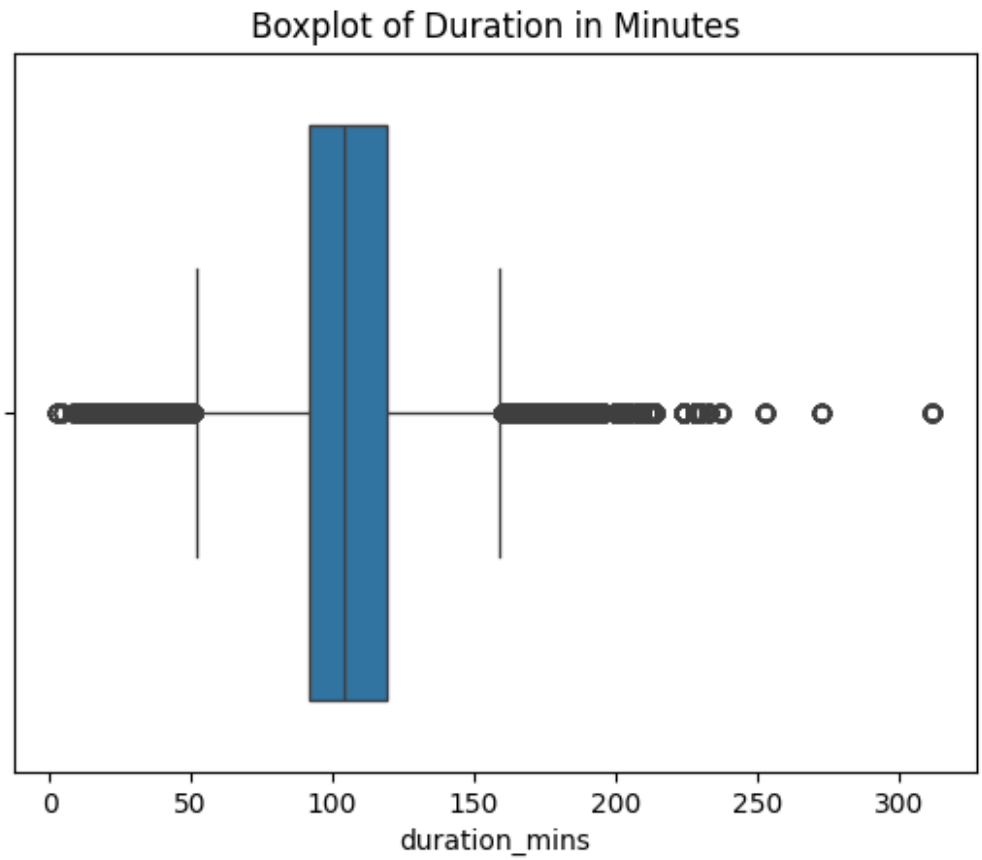

Count of Types (Movie/TV Show)
type
Movie
TV Show
0
100000
200000
300000
400000
500000
600000
count


Count of Ratings
PG-13
TV-MA
PG
TV-14
TV-PG
TV-Y
TV-Y7
R
TV-G
G
NC-17
74 min
84 min
66 min
NR
Unknown rating
TV-Y7-FV
UR
rating
0
50000
100000
150000
200000
250000
300000
count


Heatmap of Correlation
release_year
duration_mins
1
-0.23
-0.23
1
release_year
duration_mins
1.0
0.8
0.6
0.4
0.2
0.0
−0.2




```
show_id              0
type                 0
title                0
director             0
cast                 0
country              0
date_added           0
release_year         0
rating               0
duration             0
listed_in            0
description          0
duration_mins    231747
dtype: int64
```



6. Insights based on Non-Graphical and Visual Analysis

6.1 Comments on the range of attributes

Type: The dataset contains a significantly higher number of movies than TV shows, indicating a possible focus on movie content.

Rating: The most common ratings are TV-MA and TV-14, suggesting that content is largely aimed at mature audiences. The presence of ratings like '74 min', '84 min', and '66 min' in the rating column indicates potential data quality issues; these seem like they should be part of the 'duration' column.

Duration: The 'duration' column contains both the length of movies (in minutes) and the number of seasons for TV shows, making it a mixed-type column. The mode of '1 Season' suggests a large number of single-season TV shows.

6.2 Comments on the distribution of the variables and relationship between them

The **histogram** of release years shows a dramatic increase in content added in recent years, which could reflect either a growth in production or an increase in acquisitions by the platform.

The **bivariate** plots show that there's no clear correlation between the release year and the duration of the content. The scatterplot for release years and duration shows a wide distribution without a discernible pattern.

The **boxplot** for duration indicates a median duration of around 100 minutes, which is typical for movies. The outliers suggest there are some significantly longer-duration pieces of content.

6.3 Comments for each univariate and bivariate plot Histogram of Release Years: Indicates that more content has been released or added to the platform in recent years.

Scatterplots: There seems to be no direct relationship between release year and duration, indicating a consistent variety of content lengths over the years.

Boxplot of Duration: There's a concentration of movie durations around the median, with outliers suggesting that there are movies much longer than typical.

7. Business Insights

There is a growing trend of content being added to the platform, particularly in the last decade.

Mature-rated content dominates the platform, suggesting that the audience demographic may skew older.

The majority of TV shows are limited to one season, which could indicate a strategy of testing a wide variety of content to see what gains popularity.

8. Recommendations

Quality Assurance: Review and correct data quality issues, particularly in the 'rating' and 'duration' columns.

Content Strategy: Given the prevalence of mature content, consider balancing with more family-friendly options to widen audience reach.

Season Renewal Analysis: Analyze the performance of single-season TV shows to determine the potential for multi-season renewals based on viewer engagement and retention.

Question 2:

Comparison of tv shows vs. movies.

a. Find the number of movies produced in each country and pick the top 10 countries.

b. Find the number of Tv-Shows produced in each country and pick the top 10 countries.

Answer

Top 10 countries by number of movies produced are

United States-2407,

India-933,....

Spain-133

Top 10 countries by number of TV shows produced: United States-859,

Unknown country-398,....

France-64

```
# a. Number of Movies Produced in Each Country - Top 10
movies = data[data['type'] == 'Movie']
movies_count_by_country = movies.groupby('country')
['title'].nunique().sort_values(ascending=False).head(10)
print("Top 10 countries by number of movies produced:")
print(movies_count_by_country)

# b. Number of TV Shows Produced in Each Country - Top 10
tv_shows = data[data['type'] == 'TV Show']
tv_shows_count_by_country = tv_shows.groupby('country')
['title'].nunique().sort_values(ascending=False).head(10)
print("\nTop 10 countries by number of TV shows produced:")
print(tv_shows_count_by_country)
```

```
Top 10 countries by number of movies produced:
country
United States      2407
India               933
Unknown country     450
 United States      390
United Kingdom      385
Canada              192
 France             157
 United Kingdom     155
France              151
Spain               133
Name: title, dtype: int64

Top 10 countries by number of TV shows produced:
country
United States      859
Unknown country    398
United Kingdom     247
Japan              175
South Korea        168
 United States      93
Canada              84
India               81
Taiwan              70
France              64
Name: title, dtype: int64
```

Question 3

3. What is the best time to launch a TV show?

a. Find which is the best week to release the Tv-show or the movie. Do the analysis separately for Tv-shows and Movies

b. Find which is the best month to release the Tv-show or the movie. Do the analysis separately for Tv-shows and Movies

Answer

Based on the analysis, the best time to release both movies and TV shows on Netflix is in the first week of January.

The data indicates a significant peak in releases during this period, which suggests that it is an optimal time for new content to capture audience attention, likely due to holiday viewership patterns.

To capitalize on this, Netflix should consider aligning major releases with the start of the new year.

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

# Drop rows where 'date_added' is NaT (not a time)
data.dropna(subset=['date_added'], inplace=True)

# Extract week and month from 'date_added'
data['week_of_year'] = data['date_added'].dt.isocalendar().week
data['month'] = data['date_added'].dt.month

# Function to find top counts for a given time period and type
def find_top_counts(df, time_column, content_type, top_n=5):
    filtered_df = df[df['type'] == content_type]
    return filtered_df.groupby(time_column)
['title'].count().sort_values(ascending=False).head(top_n)

# Best weeks for releasing Movies and TV Shows
print("Best weeks for releasing Movies:")
print(find_top_counts(data, 'week_of_year', 'Movie'))
print("\nBest weeks for releasing TV Shows:")
print(find_top_counts(data, 'week_of_year', 'TV Show'))

# Best months for releasing Movies and TV Shows
print("\nBest months for releasing Movies:")
print(find_top_counts(data, 'month', 'Movie'))
print("\nBest months for releasing TV Shows:")
print(find_top_counts(data, 'month', 'TV Show'))
```

```
Best weeks for releasing Movies:
week_of_year
1     194480
53     60040
52     37997
Name: title, dtype: int64

Best weeks for releasing TV Shows:
week_of_year
1     70450
53    30189
52    15026
Name: title, dtype: int64

Best months for releasing Movies:
month
1    292517
Name: title, dtype: int64

Best months for releasing TV Shows:
month
1    115665
Name: title, dtype: int64
```

Question 4

Analysis of actors/directors of different types of shows/movies.

a. Identify the top 10 directors who have appeared in most movies or TV shows.

b. Identify the top 10 directors who have appeared in most movies or TV shows.

Answer

a)Top 10 actors appearing in most movies/TV shows: cast

Unknown cast 845

Anupam Kher 39

Rupa Bhimani 31,...

Paresh Rawal 25

b)Top 10 directors appearing in most movies/TV shows: director

Unknown director 2662

Rajiv Chilaka 22

Raúl Campos 18,... Youssef Chahine 12

```
# a. Identify the top 10 actors who have appeared in the most movies
or TV shows
top_actors = data.groupby('cast')
['title'].nunique().sort_values(ascending=False).head(10)
print("Top 10 actors appearing in most movies/TV shows:")
print(top_actors)

# b. Identify the top 10 directors who have appeared in the most
movies or TV shows
top_directors = data.groupby('director')
['title'].nunique().sort_values(ascending=False).head(10)
print("\nTop 10 directors appearing in most movies/TV shows:")
print(top_directors)
```

```
Top 10 actors appearing in most movies/TV shows:
cast
Unknown cast          845
 Anupam Kher           39
 Rupa Bhimani          31
 Takahiro Sakurai      30
 Julie Tejwani         28
 Om Puri               27
 Rajesh Kava           26
Shah Rukh Khan         26
 Andrea Libman         25
 Paresh Rawal          25
Name: title, dtype: int64

Top 10 directors appearing in most movies/TV shows:
director
Unknown director       2662
Rajiv Chilaka            22
Raúl Campos              18
 Jan Suter               18
Marcus Raboy             17
Suhas Kadav              16
Jay Karas                15
Cathy Garcia-Molina      15
Martin Scorsese          12
Youssef Chahine          12
Name: title, dtype: int64
```

Question 5:

Which genre movies are more popular or produced more

Answer

From the image, it appears that genres such as

"International", **"Dramas"**, **"Comedies"**,** "TV"**,** "Movies"**,** "Action"**,** "Romantic"**, and** "Thrillers"** are among the most prominent.

```
from wordcloud import WordCloud

all_genres = ' '.join(data['listed_in'].dropna())

# Generate a word cloud image
wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = None,
                min_font_size = 10).generate(all_genres)

# Display the generated Word Cloud
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```


Movies Comedies
Comedies International
Dramas Dramas
Children Family
Movies Horror
Sci Fi
Action Adventure
Movies Independent
Movies International
Dramas Independent
Shows TV
Fi Fantasy
Independent Movies
Comedies Children
Adventure International
Music Musicals
Kids' TV
TV Comedies
Dramas International
Movies Action
Thrillers Thrillers
Shows International
Horror Movies
Movies Children
Fantasy Sci
Shows Crime
Adventure Action
TV Dramas
TV Kids'
International TV
Shows Romantic
TV Shows
TV Action
Comedies Dramas
Movies Dramas
Crime TV
Comedies Comedies
Dramas TV
Movies Thrillers
Movies Classic
International Movies
Movies Romantic
Musicals Music
TV Mysteries
Teen TV
Family Movies
Romantic Movies
Spanish Language
Anime Series
Cult Movies
Language TV
Romantic TV
Dramas Action
Adventure Children
Features Children
Comedies Action
Movies Documentaries
Fantasy Thrillers
Series Anime
Comedy Talk
Comedies Sci
Sports Movies
Docuseries Science
LGBTQ Movies
Thrillers International
Spirituality Children
Classic Movies
Fantasy Dramas
Docuseries Docuseries
Fantasy Children
British TV
Thrillers Comedies
Shows Korean
Dramas Teen
TV Science
Shows Teen
Shows Spanish
Dramas Sports
Movies Crime
Adventure Dramas
Series Kids'
Thrillers Action
Nature TV
Features Sci
Dramas Documentaries
Movies Anime
Musicals Romantic
Korean TV
Thrillers Independent
TV Horror
Mysteries International
TV International
Adventure Thrillers
Adventure Cult
Adventure Comedies
Series International
Comedies Independent
Dramas Sci
Adventure Classic
Dramas Kids'
Comedies Cult
Movies Music
TV British
TV Reality
Dramas Crime
Fantasy Comedies
Thrillers Horror
Thrillers Dramas
Shows Anime
Adventure Sci
Features Anime
Mysteries TV
Movies Cult
Documentaries Music
Documentaries Sports
Spirituality Faith
Dramas Children
Documentaries International
Adventure Anime
Dramas Comedies
Faith Spirituality
Science Nature
Comedies Crime
Anime Features
Fantasy TV
Shows British
Movies LGBTQ
Comedy Stand
TV Sci
Fantasy International
Dramas Faith
Movies Kids'
Shows Stand
TV Korean
TV Anime
Features International
Reality TV
Mysteries Crime
Movies Stand
Musicals Children
Comedies Kids'
Talk Shows
Movies Sports
Musicals Comedies
Series Crime
Documentaries Documentaries
Movies Movies
Musicals Documentaries
Shows Reality
Horror International
Fantasy Action
Classic Cult
Thrillers Crime
TV Thrillers
TV Classic
TV TV
Shows Docuseries
Stand
Adventure Romantic
Dramas Romantic
Adventure Independent
Thrillers TV
Dramas Music
Thrillers British
Cult TV
Features Sci
Movies Anime
Dramas Classic
TV Crime
Thrillers
Comedy
Dramas