**UCD Professional Academy**



**Certificate in Introductory Data Analytics Project Report**

**Student:** Gavin Walsh

**Assignment Dataset:** Netflix Movies and TV Shows

**Assignment due date:** 23 November 2021

**Table of contents**

[1 GitHub URL 3](#_Toc87623656)

[2 Abstract 3](#_Toc87623657)

[3 Introduction 3](#_Toc87623658)

[4 Dataset 3](#_Toc87623659)

[5 Implementation Process 4](#_Toc87623660)

[5.1 Importing the Python Packages: 4](#_Toc87623661)

[5.2 Downloading and importing the dataset: 4](#_Toc87623662)

[5.3 Replacing missing values 4](#_Toc87623663)

[5.4 Histogram Function 5](#_Toc87623664)

[6 Results 6](#_Toc87623665)

[6.1 Distribution of the number of countries per Movies and TV Show Productions 6](#_Toc87623666)

[6.2 Distribution of the number of categories/genres a Movie or TV Show is listed in 7](#_Toc87623667)

[6.3 Most common categories of movies and TV shows 8](#_Toc87623668)

[6.4 Most common combination of movie and TV show categories 9](#_Toc87623669)

[6.5 Most common combinations of movies and TV shows location of production 10](#_Toc87623670)

[6.6 Cross-map of location of production and categories 11](#_Toc87623671)

[7 Insights 13](#_Toc87623672)

[8 References 13](#_Toc87623673)

# GitHub URL

# Abstract

In this report a dataset related to “Netflix Movies and TV Shows” from Kaggle.com1 was analysed. The data was analysed using the Python coding language and the PyCharm Python Integrated Development Environment (IDE). A number of Python Packages were used in the analyses such as Pandas and Matplotlib Multiple. The data analyses showed the following five key insights.

# Introduction

Netflix is the biggest platform for streaming movies and TV Shows on the planet. Netflix is available over 190 countries and as of Q4 20202 has 204 million subscribers2. Netflix was launched in 2007 and since then other major streaming services have also been launched such as Disney+, Apple TV+ and Now TV.

Analysing the dataset of movies and TV shows from Netflix can give valuable insights into how the industry works both overall and in individual countries. Insights such as where Movies and TV shows are mainly procedure and how they are categorised (genres) can be of both academic and economical interest. These insights can lead to marketing and production decisions that will bring more revenue to Netflix and also increase viewer satisfaction.

# Dataset

For this project the “Netflix Movies and TV Shows” dataset1 has been chosen. It is well-made, contains real-life and relevant information. The dataset is also large (8807 line items), this means there will be minimal error caused by outliers in the dataset. This dataset is also freely available.

# Implementation Process

## Importing the Python Packages:

Here the required Python packages are imported.

# Importing  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
  
# Plot Style  
plt.style.use('ggplot')

## Downloading and importing the dataset:

The dataset is then downloaded and imported. For this, the API command available from the “Netflix Movies and TV Shows” dataset at Kaggle.com1 was used. The code checks that the dataset is not all ready downloaded. If the dataset is not all ready downloaded, the “os.system('kaggle datasets download –d shivamb/netflix-shows')” is used.

# Checking to see if the data is downloaded already  
import os  
# If the file is not present in the working directory  
if not os.path.isfile('netflix-shows.zip'):  
 # Execute a command in console that downloads the dataset through the Kaggle API  
 # Note: API command copied and pasted from Kaggle Dataset  
 os.system('kaggle datasets download -d shivamb/netflix-shows')  
# Read the dataset from zip archive  
import zipfile

Now the archive with the dataset is in the working folder, we use “zipfile” python module to unpack the archive and get the file with data, which is then passed to pandas and is read into memory as a dataframe.

# Create an object for the archive  
data\_archive = zipfile.PyZipFile('netflix-shows.zip')  
# Get filename of the first file in the archive (This is the dataset)  
data\_filename = data\_archive.filelist[0].filename  
  
# With the opened file  
with data\_archive.open(data\_filename) as datafile:  
 # Read the file (.csv dataset) into pandas dataframe  
 # Specify the date column (its format will be inferred automatically)  
 data = pd.read\_csv(datafile, parse\_dates=['date\_added'])

## Replacing missing values

The dataset has a number of missing entries. E.g. there is 2634 entries with no director listed. For this project no analysis is done using the director name as such this has no impact on the data analyses.

For some categorises, missing data cannot be replaced therefore missing entries for the “country” column will be dropped when it is used. Potentially the missing values could be manually inputted by searching each movie and TV show individually.

There is 10 entries that have “date\_added” missing therefore “date\_added” will be substituted by “release\_year”. This is based on the reasonable assumption that date added is approximately equal to date of release.

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

missing = data.date\_added.isna()  
# Assign to corresponding places in the frame  
data.loc[missing, 'date\_added'] = data.release\_year[missing]

## Histogram Function

To visualise and analyses different distributions in the Movies and TV Shows dataset a histogram was used. The histogram was used to perform the operation of building and plotting a categorical histogram multiple times therefore a separate function was created. The separate function allows code to be reused and easily change details in a plot. Also groups least common entries in "Others" category

def plot\_hist(series, entries\_to\_plot=15, title=None):  
 # Count unique values  
 counted = series.value\_counts().sort\_values(ascending=False)  
  
 # Only keep some number of most common ones  
 to\_plot = counted[:entries\_to\_plot]  
 # If the are "leftovers", group them together and add as "Others"  
 if entries\_to\_plot < len(counted):  
 to\_plot['Others'] = counted[entries\_to\_plot:].sum()  
 # Use pandas shortcut to matplotlib barplot to display the values  
 to\_plot.plot(kind='bar', figsize=(entries\_to\_plot \* 0.4 + 3, 6), title=title)  
 # Add a label to the Y axis  
 plt.ylabel('Number of Movies and TV shows')  
 plt.show()

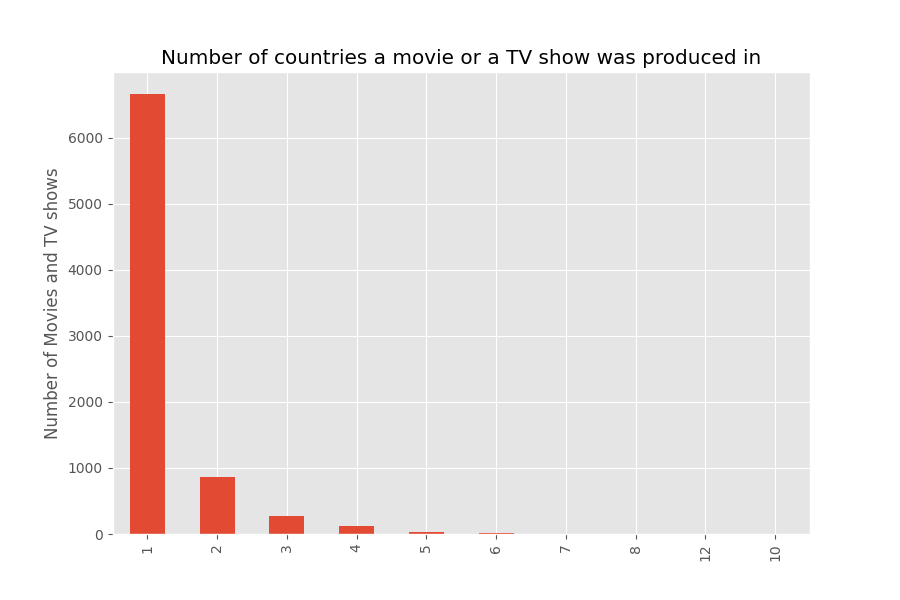
# Results

## Distribution of the number of countries per Movies and TV Show Productions

The distribution of the number of countries per Movies and TV Show Productions was analysed. Every entry is a list of countries, separated by commas. To convert that to “Number of countries” the pandas package was used to work with strings and lists. The strings were split by commas (converts them to a list of strings), the number of strings in every list was then counted.

Some values for countries are missing (831 out of 8807) therefore we drop these rows at this point. They were not dropped earlier as these rows still have other values such as categories which can be analysed also. The output type is also converted to an integer value for easier readability.

countries\_counts = data.country.dropna().str.split(',').str.len().astype(int)  
plot\_hist(countries\_counts, title='Number of countries a movie or a TV show was produced in')



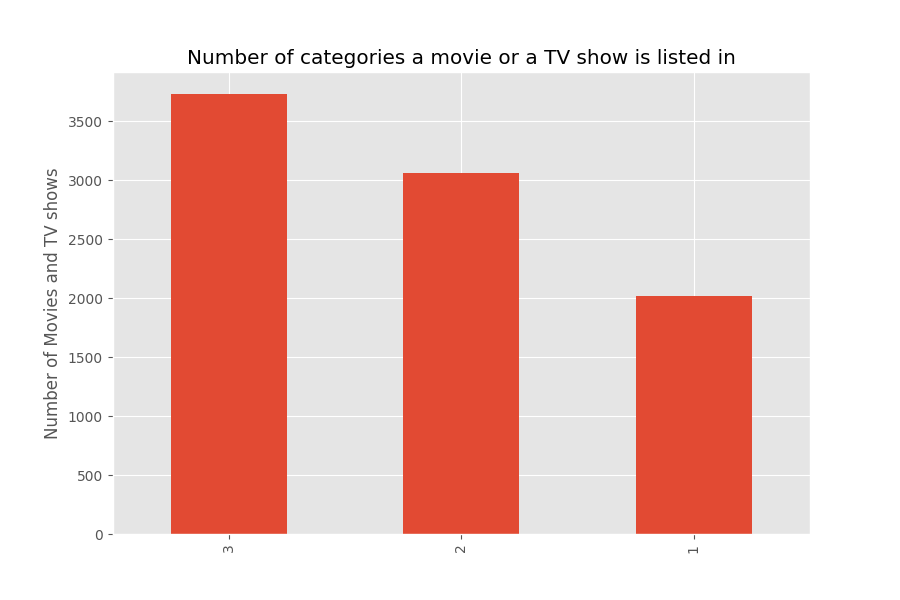
### Insight 1

As can be seen from the graph above, the absolute majority of movies and TV shows are produced in a single country. A notable number of movies and TV shows are produced in two or three countries. It is extremely rare that movies and TV shows are produced in four or more countries. The 831 out of 8807 missing entries would not impact these results majorly as the number of movies and TV shows produced in one country versus two is exponentially higher. The above histogram shows that, the number of countries a movie of TV show is produced in is exponentially distributed.

## Distribution of the number of categories/genres a Movie or TV Show is listed in

The distribution of the number of categories a Movie or TV Show is listed in was analysed. Every row in the dataset has categories/genres listed (“listed in”) as such there is no missing values. The same analysis approach is taken as in Section 6.1 above.

categories\_counts = data.listed\_in.dropna().str.split(',').str.len().astype(int)  
plot\_hist(categories\_counts, title='Number of categories a movie or a TV show is listed in')



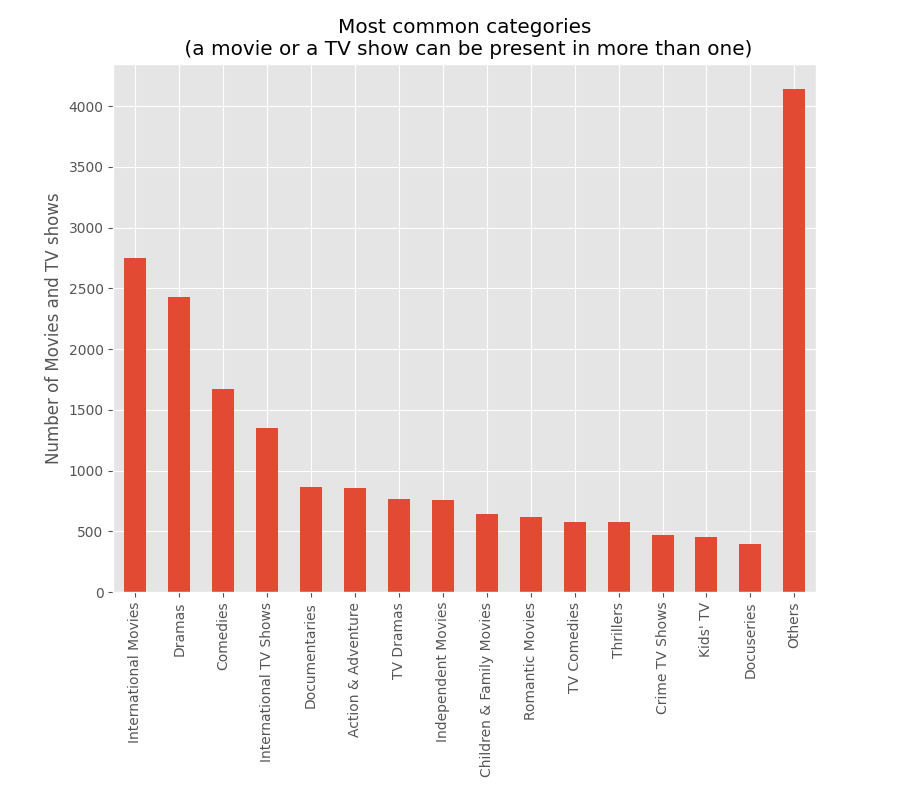
### Insight 2

All movies and TV shows are listed in at least one category/genres, with the majority being listed in three categories.

## Most common categories of movies and TV shows

The most common categories of movies and TV shows were analysed. To find the most common categories, every string is converted into a list of categories by splitting. In order to have a series of singular categories, the explode function is then used on the series, this is used to transform each element of a list-like to a row. The explode function “flattens”/ “unrolls” the series.

categories = data.listed\_in.str.split(', ').explode()  
plot\_hist(categories, title='Most common categories\n (a movie or a TV show can be present in more than one)')



### Insight 3

The most common category for movies is “International Movies”. “Dramas” is the second most comment movie category. The most common category for TV shows is “International TV Shows” followed, again, by “TV Dramas”. “Comedy” is also the third most present for both.

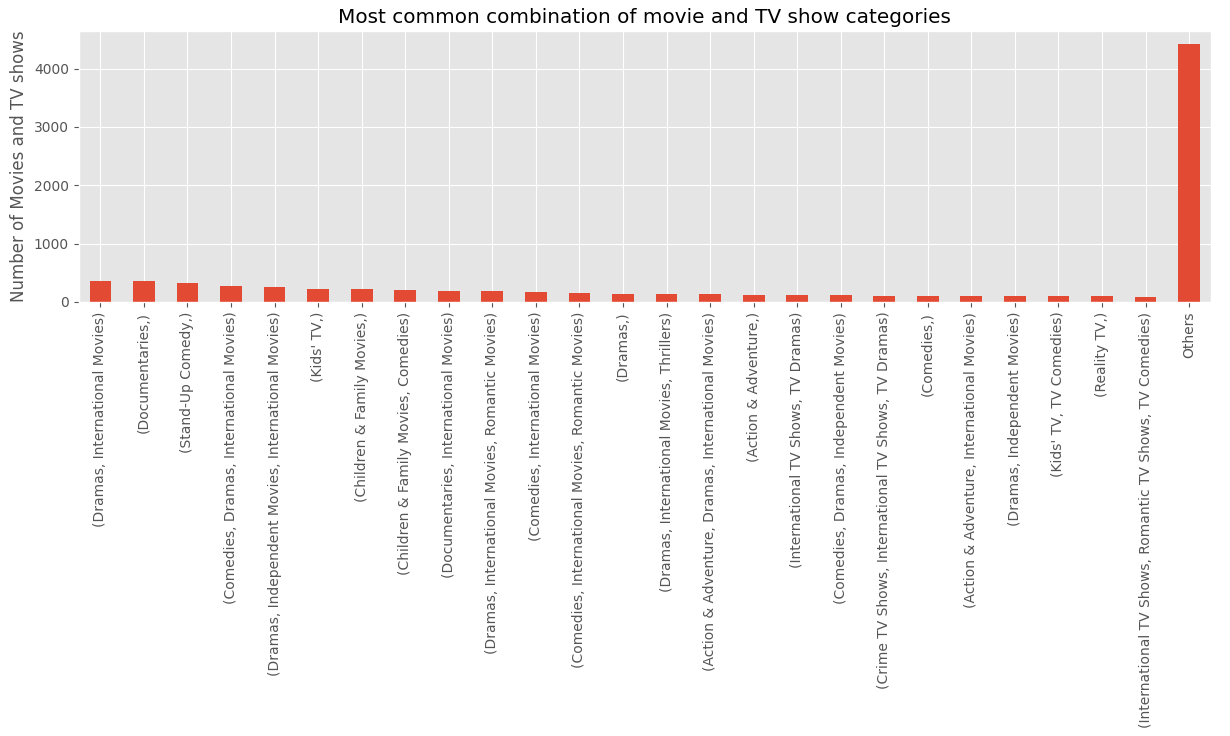
Therefore, the most common categories are same for both movies and TV shows. The most common categories are as follow; “International”, “Drama”, “Comedy”

The most common categories, start differing for movies and TV shows at the 4th most common category. “Documentaries” for movies and “Crime” for TV shows.

## Most common combination of movie and TV show categories

In Section 6.4 the overall distribution of movies and TV shows categories has been analysed. The most common combination of movie and TV show categories will be analysed also. Similarly to Section 6.3, the first step is repeated but the series is not exploded. Since the built-in pandas method for counting values in a series requires values to be hashable (immutable), the lists have to be converted to a “tuple”. The values of a tuple cannot be modified.

categories = data.listed\_in.str.split(', ').apply(tuple)  
plot\_hist(categories, 25, title='Most common combination of movie and TV show categories')

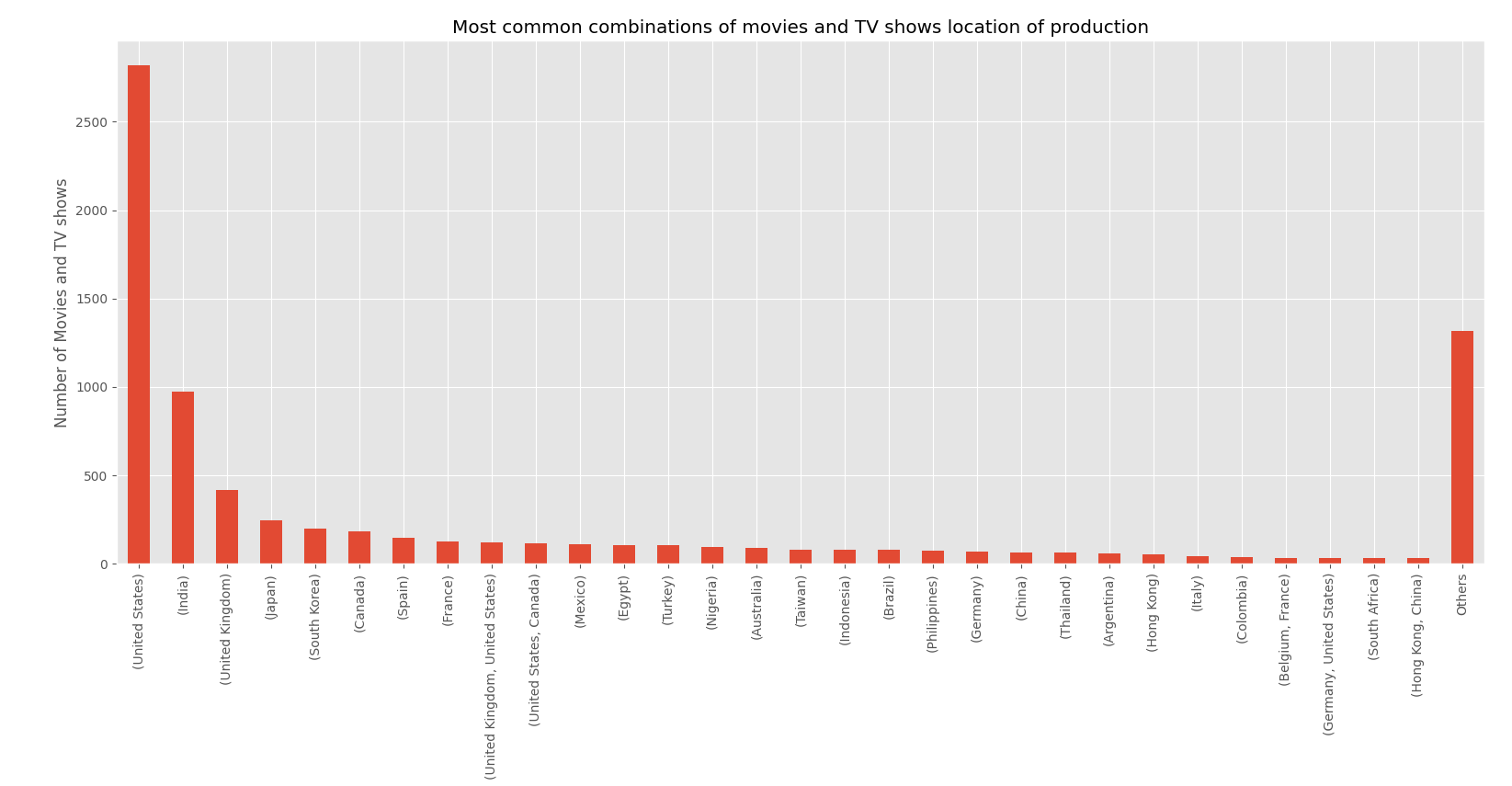


The most common combination of movies and TV shows categories is “Dramas” and “International”. This is not a new insight given that the in Section 6.3 above we have shown that the most common categories “International”, “Drama” are both the same for movies and TV shows . It can be seen that movies and TV shows seem to be distributed more or less equally between different combinations of categories. Therefore there is no major insight to be gained from this analysis.

## Most common combinations of movies and TV shows location of production

The most common combination of movies and TV shows location of production will be analyzed. The processing here is very similar to that above in Section 6.4, except for the use of “dropna” since some country values are missing. The “frozenset” is used instead of “tuple” in order to tread “USA, UK” and “UK, USA” as a single entry (i.e. we want to ignore the order of countries listed).

countries = data.country.dropna().str.split(', ').apply(frozenset)  
plot\_hist(countries, 30, title='Most common combinations of movies and TV shows location of production')



### Insight 4

Among the thirty most movies and TV shows location of production combinations there are only five that are actually combinations (the rest are singular countries). The two most represented are collaborations of English-speaking countries (either USA and UK or USA and Canada) and rank relatively high at positions 9–10. We also see that 4 out of 5 combinations list countries that share a language be it English, French, or Mandarin.

## Cross-map of location of production and categories

A cross-map of location of production (country) and categories will be analysed. The visulation outline above in Section 6.1-6.5 will be combined on a 2D histogram with countries on one axis and categories on the other one.

A data frame with only two columns is created; countries and categories. This is done by merging two frames by index. Both columns are then exploded (i.e. converting the frame from the “wide” to the “long” format). The index is also reset. The columns are also renamed for easier readability.

data2 = pd.merge(data.country.dropna().str.split(', ').to\_frame(),  
 data.listed\_in.str.split(', ').to\_frame(),  
 left\_index=True, right\_index=True)  
  
data2 = data2.explode('country').explode('listed\_in').reset\_index().drop(columns='index')  
data2 = data2.rename(columns={'country': 'Country', 'listed\_in': 'Category'})

There are 127 unique countries in this new frame. Displaying them all would make the plot too overloaded and almost unreadable, therefore the cross-map will be limited to only a few most commonly present countries.

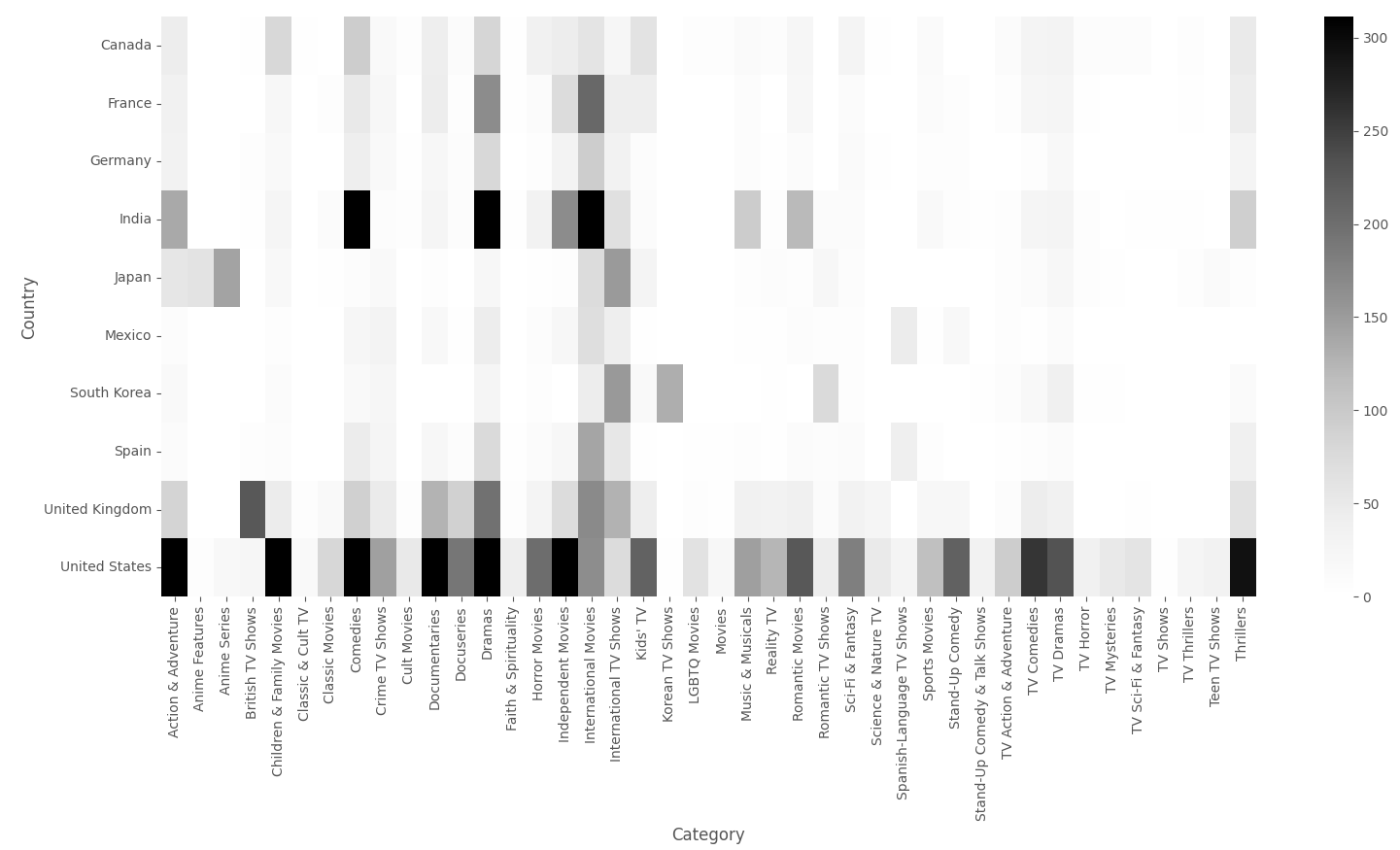
A list is created of the most common countries. Then each row in the new frame is iterated over to check if the county is present in this list. If it is, we save its index to the list of rows we will delete later. After the looping is done, we drop the rows with uncommon countries.

The 2D histogram is built with the pandas method crosstab. It produces a new (third) frame that has [most common] countries as rows and categories as columns. The seaborn method heatmap to display the calculated histogram.

unique\_countries = data2.Country.value\_counts()  
print(unique\_countries)  
most\_popular\_countries = data2.Country.value\_counts().index[:10]  
to\_delete = [] # holds indices of rows that will be deleted later  
for row in data2.index: # iterate over every row  
 # if row has an uncommon country  
 if data2.loc[row, 'Country'] not in most\_popular\_countries:  
 to\_delete.append(row) # mark for deletion  
  
# delete "uninteresting" rows  
data2 = data2.drop(to\_delete)  
# calculate the heatmap  
countries\_categories\_heatmap = pd.crosstab(data2.Category, data2.Country)  
# display the heatmap  
plt.figure(figsize=(21, 4))  
# grey colormap seems to give the best picture  
# `robust` means that quantiles for colormap are selected so that  
# outliers don't influence that much (USA is overrepresented, for example)  
sns.heatmap(countries\_categories\_heatmap.T, cmap='Greys', robust=True)  
plt.show()

### Insight 5

Categories are not evenly distributed for different countries. Clear preferences can be identified for example, “Anime Series” are overrepresented in Japan compared to the rest of the countries, and “Musicals” are much more popular in India than in other countries.



# Insights

* The absolute majority of movies and TV shows are produced in a single country.
* Even if a project is produced in more than one country, it is countries that share a language. Truly intercultural projects are extremely rare.
* Any movie or TV show is listed in at most 3 categories, with the higher number of categories being more common.
* The most common genres are “International”, “Drama”, and “Comedy” for both movies and shows, but after that, the genre preference starts diverging.
* Categories are not evenly distributed for different countries, some countries have unique genre preferences. Clear preferences can be identified for example, “Anime Series” are overrepresented in Japan compared to the rest of the countries, and “Musicals” are much more popular in India than in other countries.

# References

1 Kaggle.com. 2021. Kaggle: Your Machine Learning and Data Science Community. [online] Available at: <https://www.kaggle.com/> [Accessed 10 November 2021].

2Visual Capitalist. 2021. Which Streaming Service Has the Most Subscriptions?. [online] Available at: <https://www.visualcapitalist.com/which-streaming-service-has-the-most-subscriptions/> [Accessed 10 November 2021].