NETFLIX CASE STUDY

Key Objective: To analyze the data and generate insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries.

Initial exploration of the data

The data provided has a total of 8807 rows and 12 columns. The column names along with the count of non-null values are as follows:

```
In [12]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

Also, lets looks at the descriptive stats for all the columns:

```
df.describe(include = 'all')
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8807.000000	8803	8804	8807	8807
unique	8807	2	8807	4528	7692	748	1767	NaN	17	220	514	8775
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	NaN	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	1	19	19	2818	109	NaN	3207	1793	362	4
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014.180198	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.819312	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	NaN	NaN	NaN	NaN

The above table shows the number of unique values in each column along with the top most value and its corresponding frequency. Release_year shows that the content on

NETFLIX has content which is as old as 1925 and as latest as 2021. Further, the range of the Netflix data that is available to us is from January 1st, 2008 till September 25th, 2021.

For each column, let us look at some key metrics for us to have a general understanding:

1. show_id - Unique ID for every Movie / Tv Show

2. type - Identifier - A Movie or TV Show

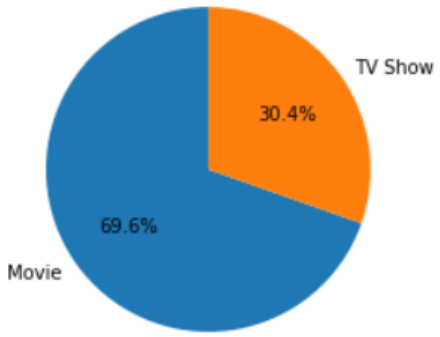
So, type is a categorical variable and hence we look at the value_counts for each category present in it.

```
df['type'].value_counts()
```

```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

We understand that there are a total of 6131 movies and 2676 TV shows which are added on the Netflix channel.

```
plt.pie(type_counts, labels= type_counts.index, autopct='%1.1f%%', startangle=90)
plt.show()
```



We can say that Netflix has roughly 70% movie content and 30% series/TV show content.

3. title - Title of the Movie / Tv Show

It has all the unique values in it. Just like show_id, title is also the unique key of the dataset.

4. director – This column has a total of 4528 unique entries out of a total of 6173 non-null entries with Rajiv Chilaka directing the maximum number of movies/TV shows.

```
df['director'].value_counts()
```

```
Rajiv Chilaka                      19
Raúl Campos, Jan Suter             18
Marcus Raboy                       16
Suhas Kadav                        16
Jay Karas                          14
                                   ..
Raymie Muzquiz, Stu Livingston      1
Joe Menendez                        1
Eric Bross                          1
Will Eisenberg                      1
Mozez Singh                         1
Name: director, Length: 4528, dtype: int64
```

Also, an observation can be made that the names of the directors are nested in case there are multiple directors for a movie/TV show.

Hence, this column will require both missing value treatment and unnesting.

5. cast – Actors involved in the movie/show

cast is a string variable with 7982 non-null entries and 7692 unique entries. Also, just like in director column, we have nested data, cast has nested data as well. To take a brief look, lets take a look at the value_counts.

```
df['cast'].value_counts()
```

```
David Attenborough
19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil
14
Samuel West
10
Jeff Dunham
7
David Spade, London Hughes, Fortune Feimster
6

..
Michael Peña, Diego Luna, Tenoch Huerta, Joaquin Cosio, José María Yazpik, Matt Letscher, Alyssa Diaz
1
Nick Lachey, Vanessa Lachey
1
Takeru Sato, Kasumi Arimura, Haru, Kentaro Sakaguchi, Takayuki Yamada, Kendo Kobayashi, Ken Yasuda, Arata Furuta, Suzuki Matsu
o, Koichi Yamadera, Arata Iura, Chikako Kaku, Kotaro Yoshida     1
Toyin Abraham, Sambasa Nzeribe, Chioma Chukwuka Akpotha, Chioma Omeruah, Chiwetalu Agu, Dele Odule, Femi Adebayo, Bayray McNwiz
u, Biodun Stephen                                                1
Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malkeet Rauni, Anita Shabdish, Chittaranjan Tr
ipathy                                                           1
Name: cast, Length: 7692, dtype: int64
```

Hence, cast will need both missing value treatment as well as unnesting.

6. country – Country where the movie/show was produced

Country is string variable with 7976 non null values and 748 unique values. **Majority content on Netflix is from United States, followed by that from India, and then from UK and Japan. This is an important piece of information as the type of shows that Netflix should upload will depend on the people who are watching it and will therefore be a function of the average age of people or the typical demographical characteristics of these countries.**

```
df['country'].value_counts()
```

```
United States                                2818
India                                         972
United Kingdom                                419
Japan                                         245
South Korea                                   199
                                             ...
Romania, Bulgaria, Hungary                      1
Uruguay, Guatemala                              1
France, Senegal, Belgium                        1
Mexico, United States, Spain, Colombia          1
United Arab Emirates, Jordan                    1
Name: country, Length: 748, dtype: int64
```

We will treat this column for missing values but won't be doing the unnesting, as the countries which appear in nested form hardly have any content coming from them. Hence, we can ignore those cases.

7. date_added – Date it was added on Netflix

This column is of string type as observed from the info table with 8797 non-null entries. For any analysis on time, we will need to change this column to datetime format. Furthermore, we will need to extract the year from the date as well as the month to carry out any year on year and month on month analysis. Also we notice that this column has few missing values. Before making these changes to the columns, we will do suitable imputations for the missing values.

8. release_year - Actual Release year of the movie/show

This column has no null values and is an int variable.

```
df['release_year'].value_counts()
```

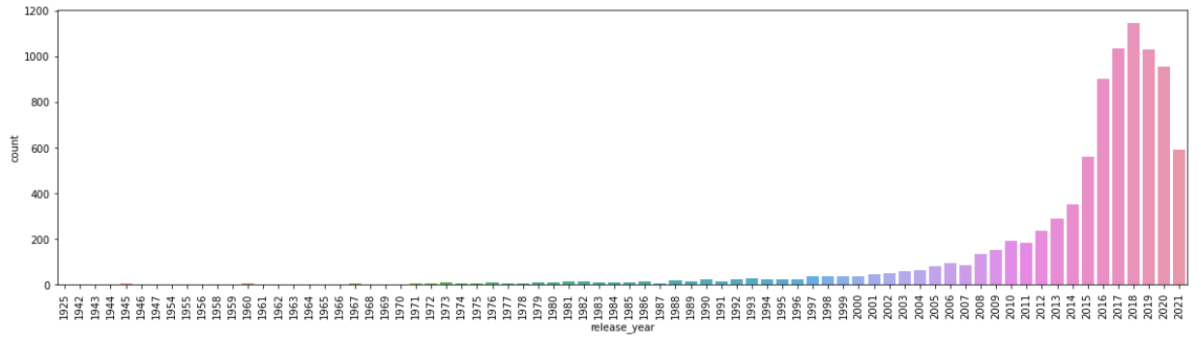
```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
        ...
1959       1
1925       1
1961       1
1947       1
1966       1
Name: release_year, Length: 74, dtype: int64
```

By looking at the above table, we observe that maximum movies that Netflix has have been released in 2018, followed by 2017 and 2019.

Also, it is worthwhile to look at the min and the max of this columns shown in the describe table. It shows that the content on NETFLIX is as old as 1925 and as latest as 2021.

Let us look at the corresponding barplot to see the year for this variable.

```
fig = plt.figure(figsize = (20,5))
sns.countplot(x= 'release_year', data = df)
plt.xticks(rotation = 90)
plt.show()
```



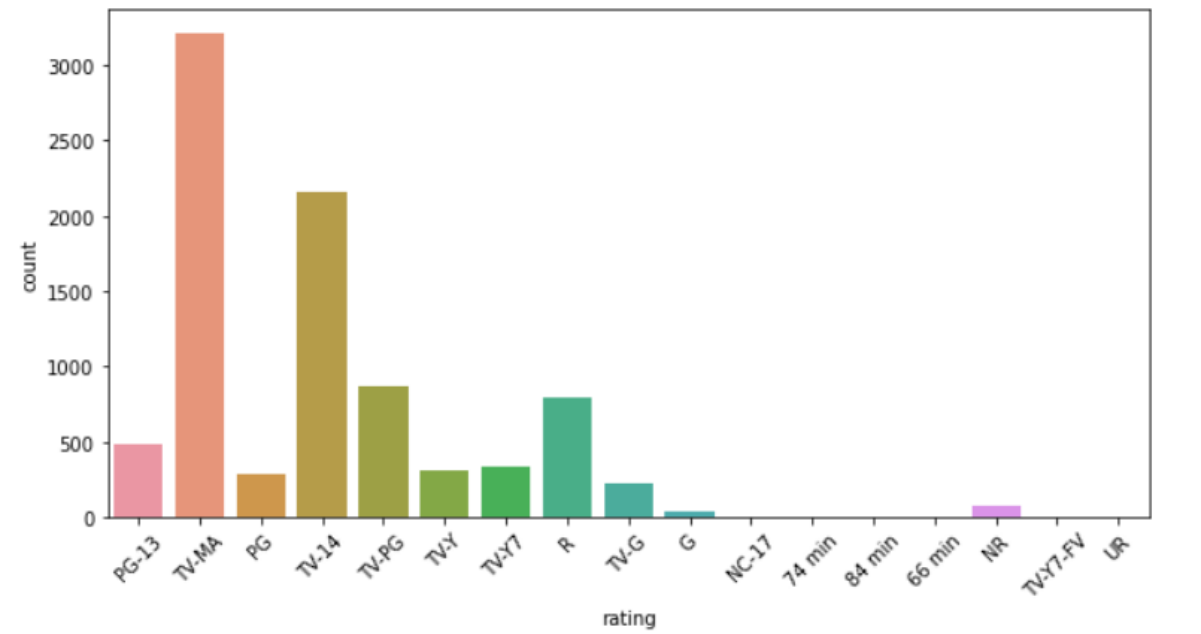
We observe that Netflix has maximum latest content and very less, selective old content. It has highest content released in year 2018, with contents from (2018+/-3yrs) adding to the majority.

9. rating – TV Rating of the movie/show

This is a categorical variable with 17 unique values. It has mostly non- null values. Only 3 values are none, which can be easily imputed with the mode value.

Creating a barplot for the same, we get,

```
fig = plt.figure(figsize = (10,5))
sns.countplot(x= 'rating', data = df)
plt.xticks(rotation = 45)
plt.show()
```



TV Rating	Meaning
TV-MA	Mature Audiences - not suitable under 17 years
TV-14	Unsuitable for ages under 14
TV-PG	Under parental guidance, suitable for children of all ages
TV-Y	Recommended for children under 6 years
TV-G	For all ages
TV-Y7	For ages 7 and up
TV-Y7-FV	For ages 7 and up - with fantasy violence
PG	Parents Guidance
PG-13	Parents Strongly Cautioned
R	Restricted
G	General Audiences
NR	Not rated

From the above barplot and the corresponding table, we can see that Netflix has maximum content for mature audiences, followed by those suitable for 14 years and above and Restricted content. Also, 74 min, 84 min and 66 min seem to have been a wrong entry in this column as can be seen from below.

```
df.loc[df['rating']=='74 min']
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	DA_year	DA_month
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	74 min	NaN	Movies	Louis C.K. muses on religion, eternal love, gi...	2017.0	Apr

```
df.loc[df['rating']=='84 min']
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	DA_year	DA_month
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	84 min	NaN	Movies	Emmy-winning comedy writer Louis C.K. brings h...	2016.0	Sep

```
df.loc[df['rating']=='66 min']
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	DA_year	DA_month
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	2016-08-15	2015	66 min	NaN	Movies	The comic puts his trademark hilarious/thought...	2016.0	Aug

The values for duration have been wrongly populated in this field. Also, all the above three can be imputed with the same value as that of the missing values.

10. duration - Total Duration - in minutes or number of seasons

This column has different categories for movies and for seasons. The duration for movies is mentioned in minutes and TV shows are mentioned in seasons. We will have to be cautious of this fact while analysing the duration for the respective category. It has 3 null values, which is getting reflected in the above rows and we can suitably treat these three rows.

11. Listed_in – Genre

It is a string variable with non-null values. It has nested values. Hence, it needs to be unnested.

```
df['listed_in'].value_counts()
```

```
Dramas, International Movies                            362
Documentaries                                           359
Stand-Up Comedy                                         334
Comedies, Dramas, International Movies                  274
Dramas, Independent Movies, International Movies        252
                                                       ...
Kids' TV, TV Action & Adventure, TV Dramas                1
TV Comedies, TV Dramas, TV Horror                         1
Children & Family Movies, Comedies, LGBTQ Movies          1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows        1
Cult Movies, Dramas, Thrillers                            1
```

12. description - The summary description

It is a string value. This column can be ignored as the analysis for the same is beyond the scope of this study.

Data pre-processing

From the above, we have seen that there are 4 kinds of issues present in our data.

1. Missing values – Columns have to be treated for missing values in such a manner that they don't get biased. For this, we have decided that the most practical approach would be to impute the missing values in director and cast column with 'missing'. This is done to make the analysis workable and fairer, also it would avoid any bias that could be created if we impute the same with any mode or any single value (as estimation of the same is difficult). The rest of the columns have been imputed with their respective mode values.

2. Nested data – Nested data is present in director, cast, country and listed_in columns. We will do the unnesting process for all of these columns.

3. Datatype of some columns – date_added column has all the dates in the object(str) format. Hence, we need to convert the same to datetime format for our analysis. Also, suitable treatment needs to be done for the rows that were extracted above.

4. Duration – the duration column is in str format. Therefore, it is unsuitable for calculation of any averages, sum etc. Hence, we need to split it and then change its datatype.

The following table lists down the actions that need to be taken to pre-process the data for further analysis

S.no.	Column Name	Data type	Non-null values	Count of Null values	Missing value Treatment	Unnesting required	Other Action
1	Show_id	Object	8807	0	NA		
2	Type	Object	8807	0	NA		
3	Title	Object	8807	0	NA		
4	Director	Object	6173	2634	Impute By 'Missing'	Unnesting required	

5	Cast	Object	7982	825	Impute By 'Missing'	Unnesting required	
6	Country	Object	7976	831	Imputation by mode	Unnesting required	
7	date_added	Object	8797	10	Imputation by mode		Convert column to datetime and add year and month column
8	Release_date	int64	8807	0	NA		
9	Rating	Object	8803	4	Imputation by mode		
10	Duration	Object	8804	3	Identified the 3 rows and impute by values wrongly present in the rating field.		Keep the numerical values of the columns only and change it to int
11	Listed_in	Object	8807	0	NA	Unnesting Required	
12	Description	Object	8807	0	NA		

We have completed all the pre-processing except unnesting. The new dataset has the following info:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      8807 non-null   object
 4   cast          8807 non-null   object
 5   country       8807 non-null   object
 6   date_added    8807 non-null   datetime64[ns]
 7   release_year  8807 non-null   int64
 8   rating        8807 non-null   object
 9   duration      8807 non-null   int32
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
 12  DA_year       8807 non-null   int64
 13  DA_month      8807 non-null   object
dtypes: datetime64[ns](1), int32(1), int64(2), object(10)
memory usage: 929.0+ KB
```

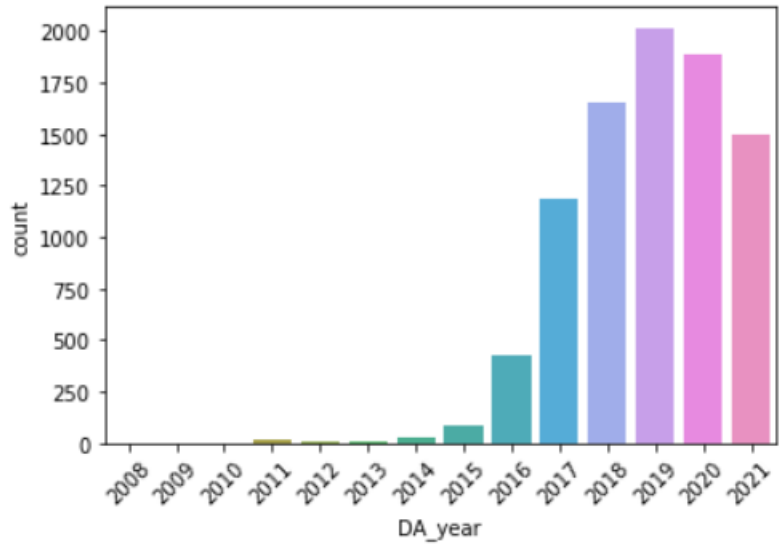
At this stage, we can perform certain univariate, bivariate and multivariate analysis for all the columns except director, cast and listed_in. This is because, once the unnesting happens, the structure of our data will change and that may cause duplication during aggregations of these columns.

Some Visual Analysis

Univariate Analysis

1. Countplot for the date_added column w.r.t each year

```
sns.countplot(x = 'DA_year', data = df)
plt.xticks(rotation = 45)
plt.show()
```



This shows that maximum content is added in 2019. Also, when we find the min and max of the date_added column, we get the range of Netflix data that is available to us.

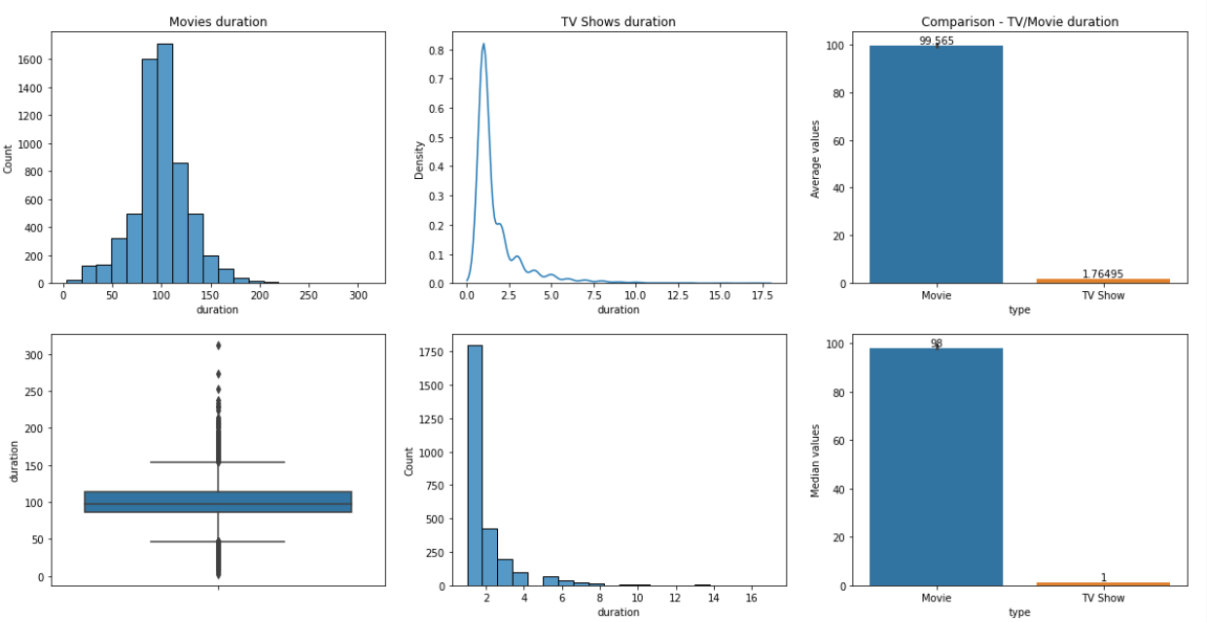
```
df['date_added'].min()
```

```
Timestamp('2008-01-01 00:00:00')
```

```
df['date_added'].max()
```

```
Timestamp('2021-09-25 00:00:00')
```

2. To check the distribution of the duration field wrt movies and TV shows. We create different datasets for movies and TV shows and also compare them together in the following subplots



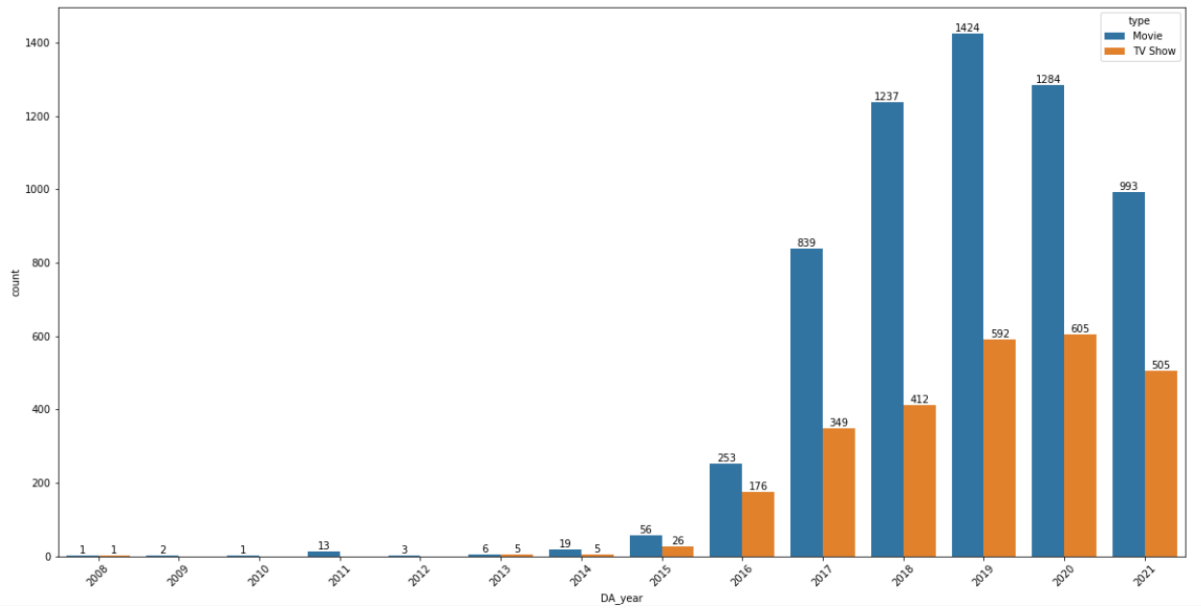
From the above, we conclude that:

1. Average length of movies is around 99-100 mins, and the median is almost the same at 98 mins.

2. Maximum TV shows have a single season only, but the average length of TV shows is 1.76 ~ 2 seasons long. This suggests that the shows which have more than 2 seasons must definitely be popular, which is why Netflix must have added it to its platform. We will explore this later.

3. To explore if there is a changing trend in the type of content added on Netflix over the years

```
fig = plt.figure(figsize = (20,10))
ax = sns.countplot(x= 'DA_year', data = df, hue = 'type')
for i in ax.containers:
    ax.bar_label(i,)
plt.xticks(rotation = 45)
plt.show()
```



This clearly shows that there has been an increasing trend in terms of the number of TV shows that are added to the Netflix platform each year.

Another metric that we must consider exploring is the difference between the release date of movies/TV shows and the date when it is added on the Netflix portal in the past recent years. For this, we have considered only a part of the dataset containing movies released post 2015 (as a cutoff to mark recent years).

```
df_cut = df.loc[df['release_year']>= 2015]
df_cut.head()
```

```
df_cut['lag'] = df_cut['DA_year']-df_cut['release_year']
df_cut['lag'].value_counts()
```

```
 0    3216
 1    1562
 2     684
 3     394
 4     207
 5      98
 6      41
-1      12
-2       1
-3       1
Name: lag, dtype: int64
```

This indicates that post 2015, 3216 movies have been added to Netflix in the same year as it has been released, however, more concerning is the fact the a considerable number of movies have been added to Netflix after a considerable lag period of atleast an year. This is definitely an area where Netflix must improve. People like to watch new releases as soon as possible. Therefore, to retain old subscribers and to attract new subscribers, Netflix must work on adding the content with a lag of not more than 2-3 months of the time the movie/TV Show is getting released.

Un-nesting of the data and merging of datasets

As previously discussed, the following columns have nested data:

1. Cast

2. Director

3. Country

4. Listed_in

For each of these columns, we have run the following code to create a dataframe which has one column as Title and the other column as the un-nested data.

```
c_split = df['country'].apply(lambda x: str(x).split(',')).to_list()
df_c = pd.DataFrame(c_split,index = df['title'])
df_c = df_c.stack()
df_c = pd.DataFrame(df_c)
df_c.reset_index(inplace = True)
df_c.drop('level_1', axis = 1, inplace = True)
df_c.columns = ['title', 'country']
```

To just give a peep on the resultant dataframe:

```
df_c.head()
```

	title	country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	United States
3	Jailbirds New Orleans	United States
4	Kota Factory	India

Likewise, for each of the nested columns, we have generated a dataframe.

Post this, we have created a temporary staging table where we have removed the old nested columns.

```
df_st1 =df
df_st1.drop(['cast','country','director','listed_in'], axis=1, inplace = True)
```

And merged all the data frames one after the other using a left join on the master table (refers to the table containing all the original columns from the dataframe without the nested columns).

```
# Merge df_new to df_st1 (Adding cast info)
```

```
df_st2 = df_st1.merge(df_new, on = 'title', how = 'left')
df_st2.shape
```

```
(64951, 11)
```

```
# merge df_dir to df_st2 ( Adding director information)
```

```
df_st3 = df_st2.merge(df_dir, on = 'title', how = 'left')
df_st3.shape
```

```
(70812, 12)
```

```
# merge df_c to df_st3 ( Adding country information)
```

```
df_st4 = df_st3.merge(df_c, on = 'title', how = 'left')
```

```
df_st4.shape
```

```
(89415, 13)
```

```
# merge df_gen to df_st4 (Adding genre information)
```

```
final = df_st4.merge(df_gen, on = 'title', how = 'left')
```

```
final.shape
```

```
(202065, 14)
```

Post the merge, have dropped duplicate rows, if any. The resultant data set has 202058 rows and 14 columns.

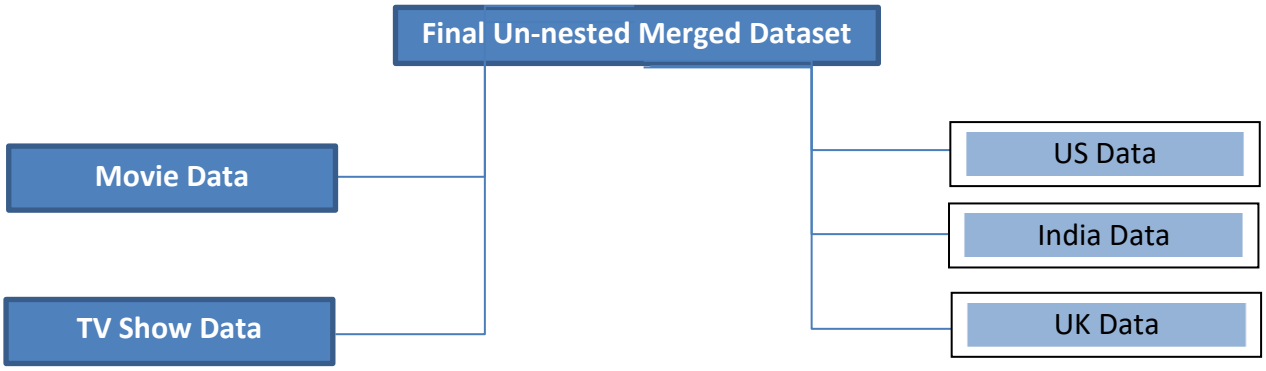
Data Analysis

While arriving at the final dataset, we ran various value-counts and unique functions on the unnested subset of dataframes, the result of which is summarised below:

Column1	cast		directors		listed_in		country	
TOP 10 value by counts	missing	825	missing	2634	International Movies	2624	United States	4042
	Anupam Kher	39	Rajiv Chilaka	22	Dramas	1600	India	1008
	Rupa Bhimani	31	Jan Suter	18	Comedies	1210	United Kingdom	628
	Takahiro Sakurai	30	Raúl Campos	18	Action & Adventure	859	United States	479
	Julie Tejwani	28	Marcus Raboy	16	Documentaries	829	Canada	271
	Om Puri	27	Suhas Kadav	16	Dramas	827	Japan	259
	Shah Rukh Khan	26	Jay Karas	15	International TV Shows	774	France	212
	Rajesh Kava	26	Cathy Garcia-Molina	13	Independent Movies	736	South Korea	211
	Andrea Libman	25	Martin Scorsese	12	TV Dramas	696	Spain	181
	Paresh Rawal	25	Jay Chapman	12	Romantic Movies	613	France	181
No. of Unique Values	39297		5121		73		197	

From here, we see that Netflix has content featuring 39297 different actors, 5000+ different directors, 73 different genres across movies and TV shows and maximum content produced by United States followed by India and UK. Apart from that, the table lists down the top 10 actors/directors/genres/countries by count.

The above observations are on an aggregated Netflix level and calculated individually for the respective columns. It is worthwhile to access similar matrix and more across columns and also across movies and TV and across countries. For this we have created the subsets of our final dataset in the following manner:



Analysis by type

For the movies and TV Show data, we get a similar matrix:

MOVIE DATA									
Column1	cast		directors		listed_in		country	
TOP 10 value by counts	missing	475	missing	188	International Movies	2624	United States	2804
	Anupam Kher	38	Rajiv Chilaka	22	Dramas	1600	India	927
	Om Puri	27	Jan Suter	18	Comedies	1210	United States	388
	Rupa Bhimani	27	Raúl Campos	18	Action & Adventure	859	United Kingdom	382
	Shah Rukh Khan	26	Suhas Kadav	16	Documentaries	829	Canada	187
	Boman Irani	25	Jay Karas	15	Dramas	827	France	155
	Paresh Rawal	25	Marcus Raboy	15	Independent Movies	736	United Kingdom	152
	Julie Tejwani	24	Cathy Garcia-Molina	13	Romantic Movies	613	France	148
	Akshay Kumar	23	Martin Scorsese	12	Children & Family Movies	605	Canada	132
	Rajesh Kava	21	Youssef Chahine	12	Thrillers	512	Spain	129

TV SHOW DATA									
Column1	cast		directors		listed_in		country	
TOP 10 value by counts	missing	350	missing	2446	International TV Shows	774	United States	1238
	Takahiro Sakurai	24	Ken Burns	3	TV Dramas	696	United Kingdom	246
	Yuki Kaji	17	Alastair Fothergill	3	International TV Shows	577	Japan	174
	Junichi Suwabe	17	Gautham Vasudev Menon	2	TV Comedies	461	South Korea	164
	Ai Kayano	17	Iginio Straffi	2	Crime TV Shows	399	United States	91
	Daisuke Ono	14	Joe Berlinger	2	Kids' TV	388	Canada	84
	David Attenborough	14	Jung-ah Im	2	Romantic TV Shows	338	India	81
	Takehito Koyasu	13	Rob Seidenglanz	2	British TV Shows	253	Taiwan	70
	Yoshimasa Hosoya	13	Shin Won-ho	2	Docuseries	221	France	64
	Yuichi Nakamura	13	Stan Lathan	2	Anime Series	176	Australia	56

These two tables show how the actors/directors/genre and the country producing the TV shows differ amongst themselves. For eg: Netflix has majority of movie and TV content coming from US while the next best is India for movies and UK for TV Shows.

In terms of genre, however, TV and Movie data shows diversification across genres but overweight on dramas and international movies/shows. This may work well for Netflix, but Netflix may consider diversifying a bit more in terms of genre, may be adding some shows or movies in different regional languages across the globe to attract a more wider subscriber base across the globe. Also, Netflix must add movies with subtitles so that good content across the globe can be watched by anybody anywhere, without the language barrier.

Another metric to look at is how many movies/TV Shows are being added to Netflix by each month:

For this, the comparison is as follows:

Column1	Movies		TV Shows	
TOP 10 value by counts	Jul	565	Dec	266
	Apr	550	Jul	262

	Dec	547	Sep	251
	Jan	546	Aug	236
	Oct	545	Jun	236
	Mar	529	Oct	215
	Aug	519	Apr	214
	Sep	519	Mar	213
	Nov	498	Nov	207
	Jun	492	Jan	202
	May	439	May	193
	Feb	382	Feb	181

The typical holiday season in US, UK and typically around the world is in winters during Christmas. They sometimes begin in end of November and continue till January first week. In India, holiday seasons are in June-July and during Diwali (Oct/Nov). Considering the fact that Netflix has maximum content coming from US, India and UK, it must align its timelines of adding new and exciting content for the target audience around their holiday season. However, as we can see the content is added more or less in the same fashion, in fact it is lower in November. Maximum movies in different countries are typically released around the holiday season to attract more business, and Netflix must tap the same opportunity.

We have further gone to the extent of analysing movies/TV shows that are added to Netflix on a weekly basis. We have tried to understand the top genre (along with its corresponding counts) that Netflix is adding to its kitty each week. The results for the same are elaborate and cannot be shown here (space constraint) but are available in the code file.

Here we again find that Netflix is adding majority of international movies/TV shows week after week and there is very little diversity observed. This drives us to think that considering both the sides of catering to a larger/ wider viewership across the world and the holiday seasons that come across, Netflix must definitely consider more diversification across genres. It needs to better plan when, how much and what kind of content needs to be added. May be a little less of International movie/TV Show will work, but it must invest in more diversification. That will help in boosting overall profitability of the company, by reducing costs, creating more customer satisfaction and attracting new subscribers. Also, it must deploy marketing strategies to communicate the same to the target audience.

Analysis Based on Country

We know that best actors and best directors in any country are a brand in themselves. They are able to pull crowds towards their movies just by their name and presence. For this reason, we have looked at the top 10 actors, directors, actor-director pair and actor and type (movie/TV show) for top 3 countries i.e. US, India and UK.

US DATA (Top ten unique values by counts)

cast		directors		cast-type pair			cast - director pair		
missing	561	missing	1349	missing	TV Show	239	missing	missing	241
Rupa Bhimani	25	Rajiv Chilaka	17	Rupa Bhimani	Movie	23	Fortune Feimster	missing	15
Andrea Libman	22	Marcus Raboy	16	Adam Sandler	Movie	20	Rupa Bhimani	Rajiv Chilaka	15
Fred Tatasciore	21	Suhas Kadav	15	Julie Tejwani	Movie	19	Julie Tejwani	Rajiv Chilaka	15
Julie Tejwani	21	Jay Karas	15	Rajesh Kava	Movie	17	Rajesh Kava	Rajiv Chilaka	15
Adam Sandler	20	Jay Chapman	12	Jigna Bhardwaj	Movie	16	Jigna Bhardwaj	Rajiv Chilaka	15
Rajesh Kava	19	Martin Scorsese	12	Andrea Libman	Movie	15	Vatsal Dubey	Rajiv Chilaka	14
Vincent Tong	18	Steven Spielberg	11	Fred Tatasciore	Movie	15	Swapnil	Rajiv Chilaka	12
Jigna Bhardwaj	18	Don Michael Paul	10	Alfred Molina	Movie	15	Mousam	Rajiv Chilaka	12
Fortune Feimster	16	Shannon Hartman	9	Molly Shannon	Movie	14	Vincent Tong	missing	11

INDIA DATA (Top ten unique values by counts)

cast		directors		cast - type pair			cast - director pair		
Anupam Kher	Movie	3 missing	85	Anupam Kher	Movie	36	missing	missing	18
Om Puri	Movie	David Dhawan	9	Om Puri	Movie	26	Anupam Kher	David Dhawan	6
Shah Rukh Khan	Movie	Anurag Kashyap	7	Shah Rukh Khan	Movie	25	Alok Nath	Sooraj R. Barjatya	5
Boman Irani	Movie	Ram Gopal Varma	7	Boman Irani	Movie	25	Julie Tejwani	Rajiv Chilaka	4
Paresh Rawal	Movie	Sooraj R. Barjatya	6	Paresh Rawal	Movie	25	Rajesh Kava	Rajiv Chilaka	4
Akshay Kumar	Movie	Ashutosh Gowariker	6	Akshay Kumar	Movie	23	Salman Khan	Sooraj R. Barjatya	4
missing	Movie	Anees Bazmee	6	missing	Movie	20	Mohnish Bahl	Sooraj R. Barjatya	4
Naseeruddin Shah	Movie	Imtiaz Ali	6	Naseeruddin Shah	Movie	20	Rajpal Yadav	Priyadarshan	4
Kareena Kapoor	Movie	Rajkumar Santoshi	6	Kareena Kapoor	Movie	20	Asrani	Hrishikesh Mukherjee	3
Amitabh Bachchan	Movie	Priyadarshan	6	Amitabh Bachchan	Movie	20	Rupa Bhimani	Rajiv Chilaka	3

UK DATA (Top ten unique values by counts)

cast		directors		cast-type pair			cast - director pair		
missing	97	missing	267	missing	TV Show	39	missing	missing	43
David Attenborough	17	Alastair Fothergill	4	David Attenborough	TV Show	13	David Attenborough	missing	13
Michael Palin	14	Edward Cotterill	4	Michael Palin	Movie	9	Terry Jones	missing	5
Eric Idle	12	Martin Campbell	3	John Cleese	Movie	9	Eric Idle	missing	5
Terry Jones	12	Orlando von Einsiedel	3	Brendan Gleeson	Movie	8	Michael Palin	missing	5
John Cleese	12	Tom Hooper	3	Helena Bonham Carter	Movie	8	David Attenborough	Alastair Fothergill	4
Terry Gilliam	11	Terry Gilliam	3	Terry Gilliam	Movie	7	Brendan Coyle	missing	4
Helena Bonham Carter	9	Vince Marcello	3	Terry Jones	Movie	7	Terry Gilliam	missing	4
Jim Broadbent	8	Blair Simmons	3	Eddie Marsan	Movie	7	Molly Ringwald	Vince Marcello	3
Brendan Gleeson	8	Jerry Rothwell	3	Judi Dench	Movie	7	missing	Jerry Rothwell	3

It is understandable that a sizeable information regarding the cast and directors is absent for each country. Nonetheless, it is noteworthy that some of the best names in the film industry irrespective of Hollywood or Bollywood are not seen in the top 10 list of Netflix, which is little unexpected.

Hence, Netflix must take care of noting the details of the cast and directors of movies/TV show in their database as well as include movies with top rated actors and directors in some of it main business regions. This logic is even more applicable for regions where Netflix has a small subscriber base.

Recommendations for Netflix

Netflix is one of the best video streaming service across the globe for movies and TV shows. It has a very good collection of movies and TV shows. However, as we studied through the data provided for the period between 2008 till 2021, there were certain areas where we thought Netflix must work on….

1. It has been observed that more than 70% of the content on Netflix is meant for adult audiences or for 14 years and above. Therefore, to make Netflix more appealing to the masses and to cater to all age groups, it must have a more balanced spread.

2. Secondly, it was observed that for movies released during latest years, i.e. for periods post 2015, a considerable chunk of movies are being added after a lag of an year or more. In order to attract new subscribers and also to retain existing ones, this lag time must be shortened to as low as 2-3 months only.

3. Netflix must optimise on the kind/genre of movies it is adding to its kitty and not become overweight on any one genre to be able to cater to a wider viewership across the world. Currently, it is overweight on international movies/TV shows and dramas. It must invest in certain regional content in other countries apart from US, India and UK to tap these markets. Eg: French, Japanese content with subtitles.

4. Netflix must add new, more diverse content more towards the holiday season and can also offer some discount packages for attracting new customers across different ages, tastes and preferences.

5. It has been observed that some of the popular names in the film and TV industry are absent when we look at the top cast/directors for a country. This is again an opportunity for Netflix to optimise its resources in order to attract the crowd.

6. The number of TV shows that have been added to Netflix annually is on an increasing trend. Netflix can continue this trend depending on the customer preferences.