Exploratory Data Analysis of Movies

```
!apt-get install -qq -y -t /
!apt-get update -qq -y
!apt-get install -qq -y -t /
!apt-get install -qq -y -t /
!apt-get install -qq -y -t /
!apt-get install -qq -y -t /
!apt-get install -qq -y -t /
```

```
!pip install -qq -U -t /usr/local/lib/python3.6/dist-packages/dist -U --no-deps
!cp -r /usr/local/lib/python3.6/dist-packages/dist/* /usr/local/lib/python3.6/dist-packages
!rm -r /usr/local/lib/python3.6/dist-packages/dist
```

starting with https://www.kaggle.com/ryanschaefer/movie-analysis-and-visualization-pt0a

```
!mkdir -p ~/.kaggle/
!echo '{"username":"ryanschaefer","key":"a8e0e2c7d1ccb7ccb7d2d0d0d0d0d0d0"}'
!chmod 600 ~/.kaggle/kaggle.json
!kaggle -v
!kaggle competitions download -c tmdb-movie-metadata
!ls -lh *.zip
!unzip -qq -o *.zip
!ls -lh *.csv
```

```
Kaggle command_line_tool 1.5.12
Warning: Your Kaggle API key is readable by other users on this machine!
Downloading tmdb-movie-metadata.zip, 0 bytes compressed
Downloading tmdb-movie-metadata.zip, 0 bytes compressed
total 0
-rw-r--r-- 1 root root 0 Jan 1 1970 tmdb-movie-metadata.zip
total 0
-rw-r--r-- 1 root root 0 Jan 1 1970 tmdb-movie-metadata.zip
```

Make the column names uniform

```
!mv credits.csv credits.csv
!mv keywords.csv keywords.csv
!mv tmdb_5000_movies.csv tmdb_5000_movies.csv
!ls -lh *.csv
```

```
-rw-r--r-- 1 root root 0 Jan 1 1970 credits.csv
-rw-r--r-- 1 root root 0 Jan 1 1970 keywords.csv
-rw-r--r-- 1 root root 0 Jan 1 1970 tmdb_5000_movies.csv
```

Getting all current dependencies up-to-date and set up. Can also use: dfrankow/kaggle-sans-notebook

```
!mv tmdb_5000_movies.csv tmdb_5000_movies.csv
!mv credits.csv credits.csv
!mv keywords.csv keywords.csv
```

```
-rw-r--r-- 1 root root 0 Jan 1 1970 credits.csv
-rw-r--r-- 1 root root 0 Jan 1 1970 keywords.csv
-rw-r--r-- 1 root root 0 Jan 1 1970 tmdb_5000_movies.csv
```

Write-ups:

```
!pip install -qq -U -t /
!pip install -qq -U -t /
```

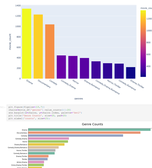
```
Requirement already up-to-date: /usr/local/lib/python3.6/dist-packages/dist (0.0.0)
Requirement already up-to-date: /usr/local/lib/python3.6/dist-packages/dist (0.0.0)
```

```
!ls -lh *.csv
```

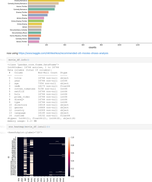
```
-rw-r--r-- 1 root root 0 Jan 1 1970 credits.csv
-rw-r--r-- 1 root root 0 Jan 1 1970 keywords.csv
-rw-r--r-- 1 root root 0 Jan 1 1970 tmdb_5000_movies.csv
```

```
!head -n 1 credits.csv
!head -n 1 keywords.csv
!head -n 1
```

```
plt.figure(figsize=(10,20))
sns.countplot(y='director', data=netflix_movies, order=netflix_movies['director'].value_counts().index[0:100])
plt.title('Most movies made by director')
plt.show()
```




```
plt.figure(figsize=(10,20))
sns.countplot(y='director', data=netflix_movies, order=netflix_movies['director'].value_counts().index[0:100], hue='type')
plt.title('Most movies made by director')
plt.show()
```



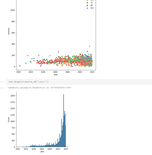

```
plt.figure(figsize=(10,20))
sns.countplot(y='country', data=netflix_movies, order=netflix_movies['country'].value_counts().index[0:100])
plt.title('Genre Counts')
plt.show()
```

see also https://www.kaggle.com/ahmedmohamed2054/netflix-movies-shows-analysis

```
netflix_df.info()
```

```
<class> DataFrame.info(verbose=True, buf=None, max_cols=None, memory_usage=None, show_counts=None, null_counts=None)
<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: 1.1 MB
```

```
sns.heatmap(netflix_df.isnull())
```



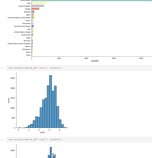

```
netflix_df.nunique()
```

```
show_id       8807
type             2
title         8807
director      4528
cast          7692
country        748
date_added    1767
release_year    74
rating          14
duration       220
listed_in      514
description   8775
dtypes: int64    1
       object   11
dtype: int64
```

```
netflix_df.shape
```

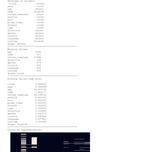
```
(8807, 12)
```

```
sns.countplot(netflix_movies['type'])
```




```
sns.countplot(netflix_movies['release_year'])
```

```
sns.countplot(netflix_movies['rating'])
```




```
sns.countplot(netflix_movies['duration'])
```

```
netflix_df.duplicated().sum()
```

```
0
```

```
netflix_df.drop_duplicates(inplace=True)
```

```
netflix_df.shape
```

```
(8807, 12)
```

```
netflix_df.isnull().sum
```

```
<bound_method_>
```

```
netflix_df.isnull().sum()
```

```
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
dtypes: int64    1
       object   11
dtype: int64
```

```
netflix_df.dropna(inplace=True)
```

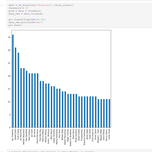
```
netflix_df.shape
```

```
(8807, 12)
```

```
netflix_df.isnull().sum()
```

```
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
dtypes: int64    1
       object   11
dtype: int64
```

```
sns.countplot(netflix_movies['type'], hue='rating')
```

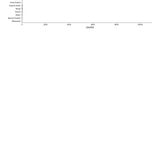



```
sns.countplot(netflix_movies['type'], hue='release_year')
```

```
sns.countplot(netflix_movies['type'], hue='duration')
```

```
sns.countplot(netflix_movies['type'], hue='listed_in')
```

```
sns.countplot(netflix_movies['type'], hue='description')
```




```
netflix_df.groupby('type').count()
```

```
      show_id  title  director  cast  country  date_added  release_year  rating  duration  listed_in  description
type
Movie    6131   6131      4528  5653     5691        6131          6131    6128      6128       6128         6128
TV Show  2676   2676      1645  2329     2285        2666          2676    2675      2676       2676         2676
```

```
netflix_df.groupby('type').count()['show_id']
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=True)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).head(5)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).head(10)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(5)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(10)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(15)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(20)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(25)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(30)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(35)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(40)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(45)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(50)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(55)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(60)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(65)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(70)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(75)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(80)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(85)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(90)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(95)
```

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

```
netflix_df.groupby('type').count()['show_id'].sort_values(ascending=False).tail(100)
```

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

```
sns.countplot(netflix_movies['country'], hue='type')
```



```
sns.countplot(netflix_movies['type'], hue='director')
```



```
sns.countplot(netflix_movies['type'], hue='cast')
```



```
sns.countplot(netflix_movies['type'], hue='country')
```



```
sns.countplot(netflix_movies['type'], hue='date_added')
```



```
sns.countplot(netflix_movies['type'], hue='release_year')
```


Movie
TV Show
1925
1928
1929
1930
1931
1932
1933
1934
1936
1937
1938
1939
1940
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
3100
3200
3300
3400
3500
3600
3700
3800
3900
4000
4100
4200
4300
4400
4500
4600
4700
4800
4900
5000
5100
5200
5300
5400
5500
5600
5700
5800
5900
6000
6100
6200
6300
6400
6500
6600
6700
6800
6900
7000
7100
7200
7300
7400
7500
7600
7700
7800
7900
8000
8100
8200
8300
8400
8500
8600
8700
8800
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
25000
26000
27000
28000
29000
30000
31000
32000
33000
34000
35000
36000
37000
38000
39000
40000
41000
42000
43000
44000
45000
46000
47000
48000
49000
50000
51000
52000
53000
54000
55000
56000
57000
58000
59000
60000
61000
62000
63000
64000
65000
66000
67000
68000
69000
70000
71000
72000
73000
74000
75000
76000
77000
78000
79000