Semester Project Report

Bidya Bhattarai

Department of Business Analytics

Beacom School of Business

University of South Dakota

BADM 505: Business Analytics Fundamentals

4th December 2023

Introduction:

In this digital streaming service era, Netflix is a of the dominated entertainment industries. It was founded on 1997 which has been one of the frontrunners in streaming movies and TV shows. The primary objectives of this analysis are to find out the popularity and performance of various movies and TV shows, to identify the top genres, highest rating, duration, and seasons. For this analysis, I used python which has been a great tool to analyze and visualize the data. This programming language has an extensive library such as Pandas, NumPy, and Matplotlib which plays a crucial role to extract a meaning insight.

Problem Description

In this dataset, the challenging part is to comprehend the consumers preference, content performance and linkage between other variables to improve the content suggestions according to the locations. This analysis aims to address the following question:

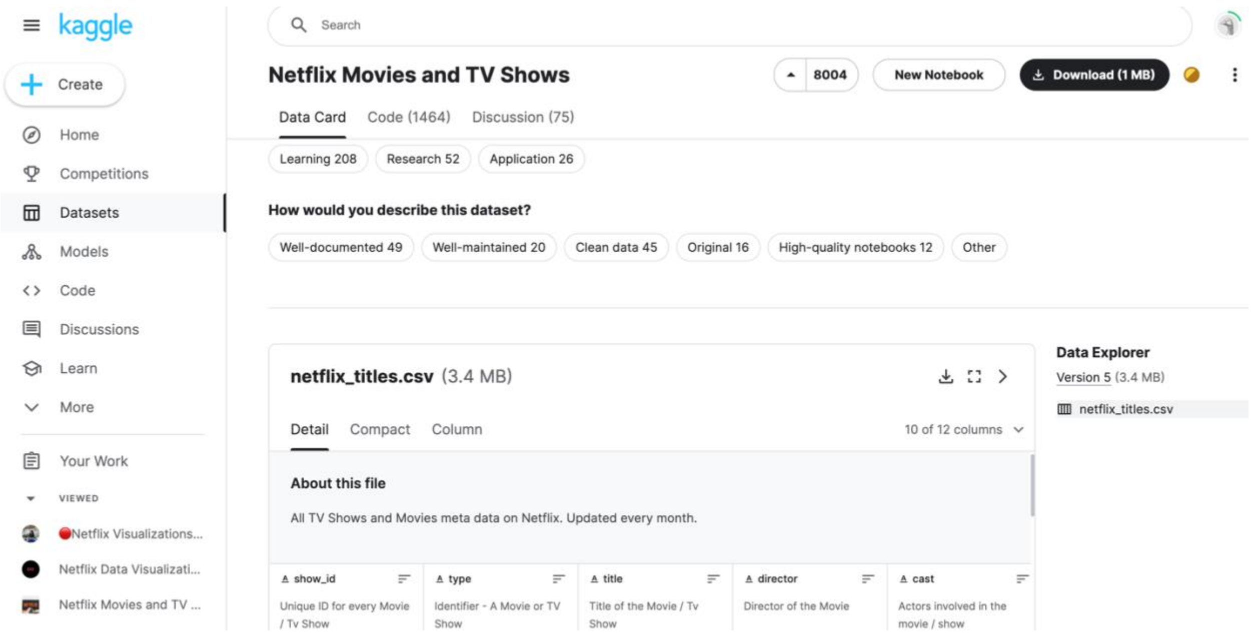
1. What are the engagement levels of movies and TV shows?
2. What genres are the most popular?
3. What is the trend of movie released in terms of year?
4. Identify the number of each rating in terms of movies and TV shows.

Dataset Description

This dataset was extracted from Kaggle.com. Here's the link of dataset:

https://www.kaggle.com/datasets/shivamb/netflix-shows

Also, for detailed one, below is the screenshot of where I got the data.



Netflix Movies and TV Shows

How would you describe this dataset?

netflix_titles.csv (3.4 MB)

About this file

All TV Shows and Movies meta data on Netflix. Updated every month.

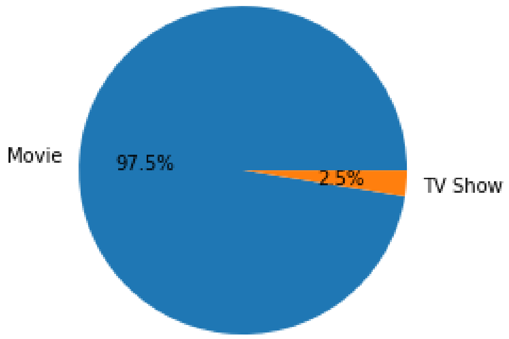
show_id	type	title	director	cast
Unique ID for every Movie / Tv Show	Identifier - A Movie or TV Show	Title of the Movie / Tv Show	Director of the Movie	Actors involved in the movie / show

Here, in this dataset, there are 12 data types (variables), 6234 rows including null values. After removing outliers, there are 3774 values on each column. There are categorical and quantitative data types where the variables named Title, Genre and Country falls in categorical and release year and duration lies in quantitative variables.

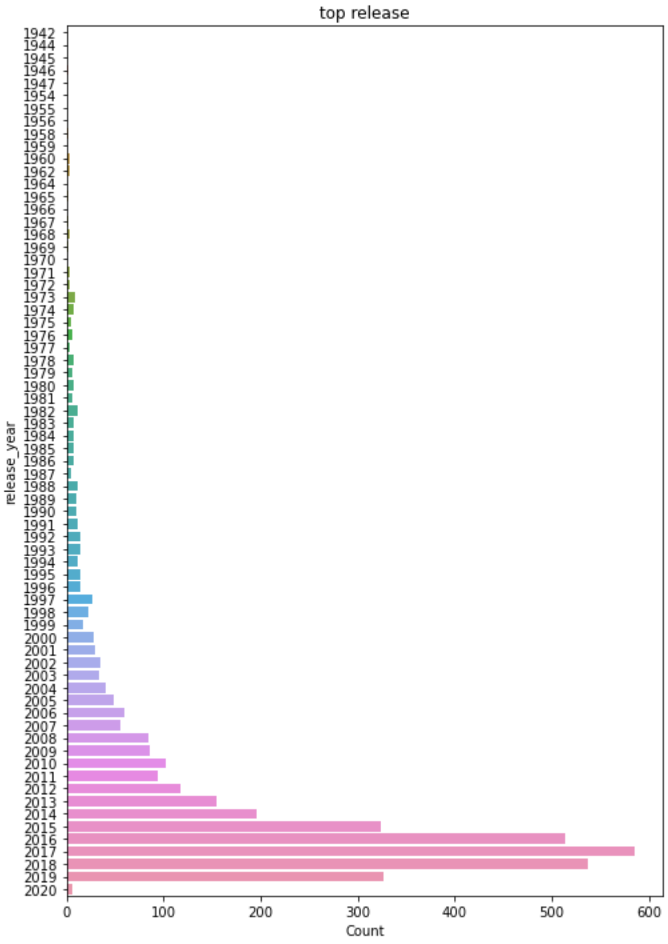
Data Visualization and Interpretation of Results

For the analysis of this dataset, we visualize the different variables in bar chart, pie chart, scatter plot, horizontal bar chart and histogram.

In this dataset, we visualize the proportion of movies and TV shows in the form pie-chart. Here, we can see the 97.5% of movies and 2.5% of TV shows which indicates that they streamed more movies as compared to TV shows.

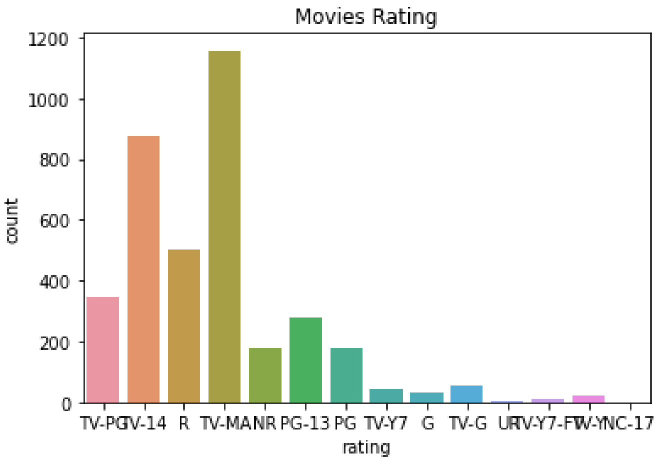


Now, we extract the release year information, here we used the horizontal bar chart. This chart reflects the number of movies and TV shows released in the given year.



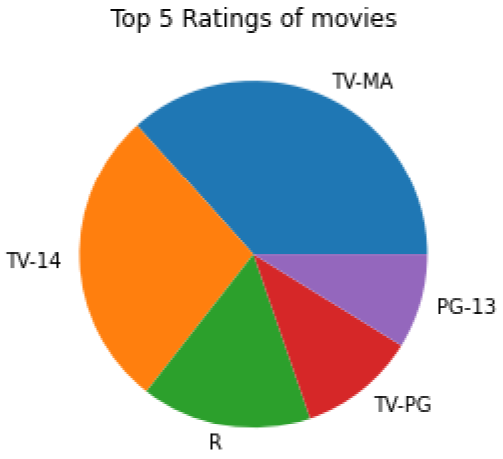
From the chart, we can conclude that, the majority of movies and TV shows was released on 2017.

In addition, here we visualize the count of movies by ratings. We use the count plot method to display the information.



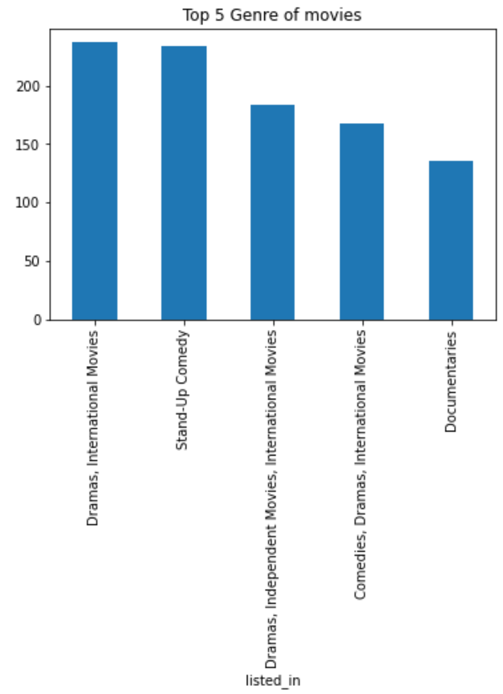
Based on the accompanying chart, the majority of the movies have the 'TV-MA' rating. A television program intended only for mature audiences is rated "TV-MA" by the TV Parental Guidelines. The second-largest category is "TV-14," which says content which is inappropriate for children under the age of 14. R category movie is the third largest in the list which denotes content is not suitable for people under 17 by Motion Picture Association of America.

The top 5 ratings of movies are displayed in pie chart.



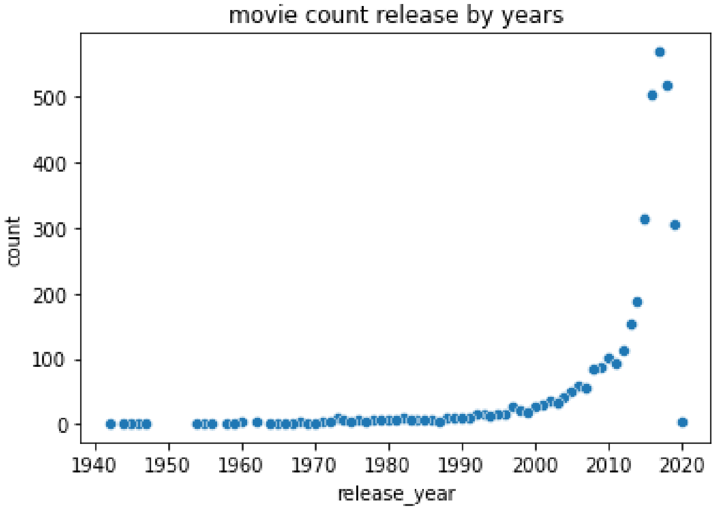
From this chart, we get to know that the dataset has the movie of the rating TV-MA and TV-14 in a highest number.

Also, we analyze the top 5 genre of movies listed in dataset. From this information, company can do cross promotions by collaborating with other companies to promote content with the most popular.

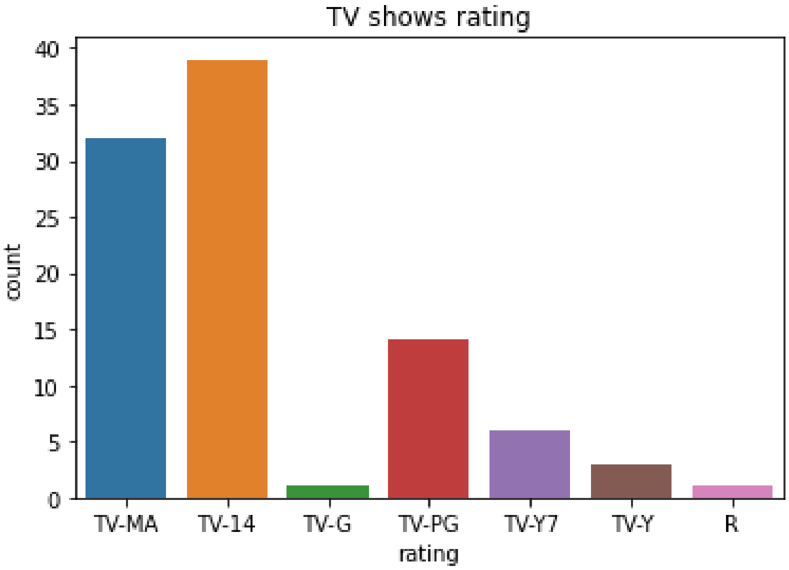


Dramas is the top genre of movie where stand-Up comedy is the second popular one. From this analysis, company can make a strategy on content promotion and allocated resource accordingly.

I did the analyzation of movie count released by years. From this scatter plot, we can conclude that highest number of movies was released in 2017 whereas 2020 shows the lowest number.

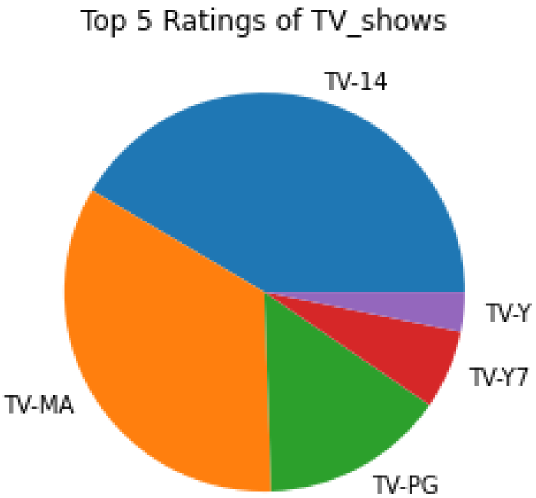


For the TV shows also, we analyze it with ratings.

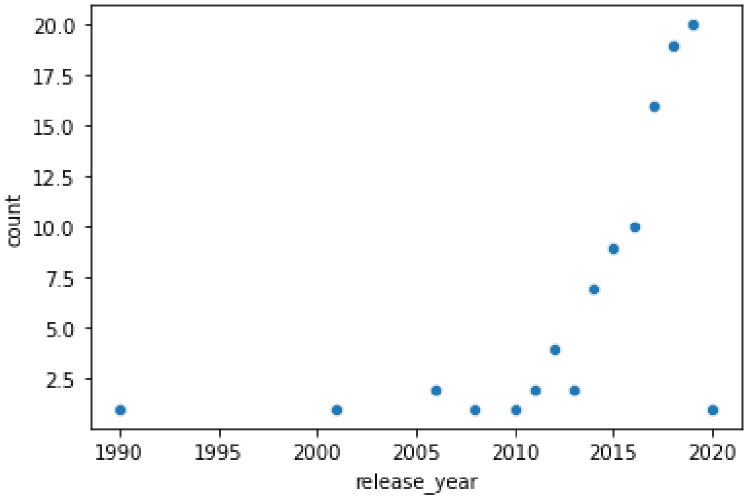


From the above chart, we can conclude that, most of the TV shows has the TV-14 ratings, and the rating of R and TV-G has the least shows.

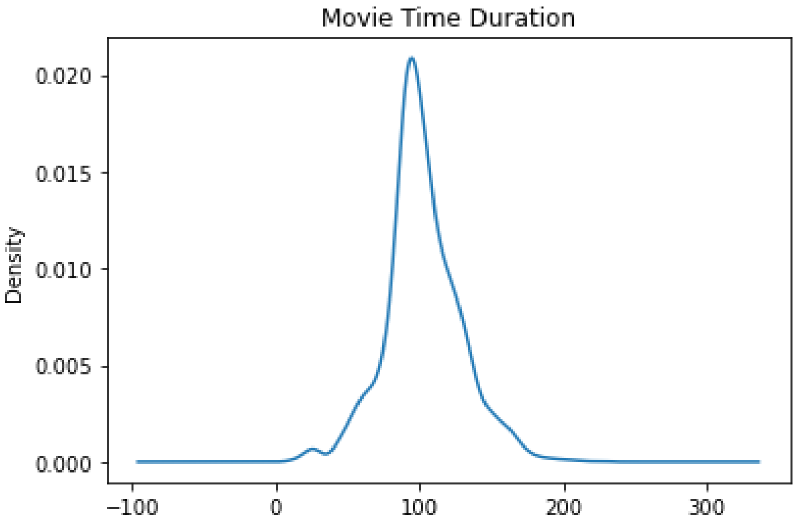
Here I analyze the top 5 ratings of TV shows are displayed in pie chart. From this chart, we get to know that the dataset has the movie of the rating TV-MA and TV-14 in a highest number. Most of the movies are suitable for children above 13 years old. TV-Y ratings Tv shows is in the top 5th which is suitable for all ages.



I did the analyzation of TV shows count released by years in the form of scatterplot. From this scatter plot, we can conclude that highest number of TV shows was released in 2017 whereas 2020 shows the lowest number.

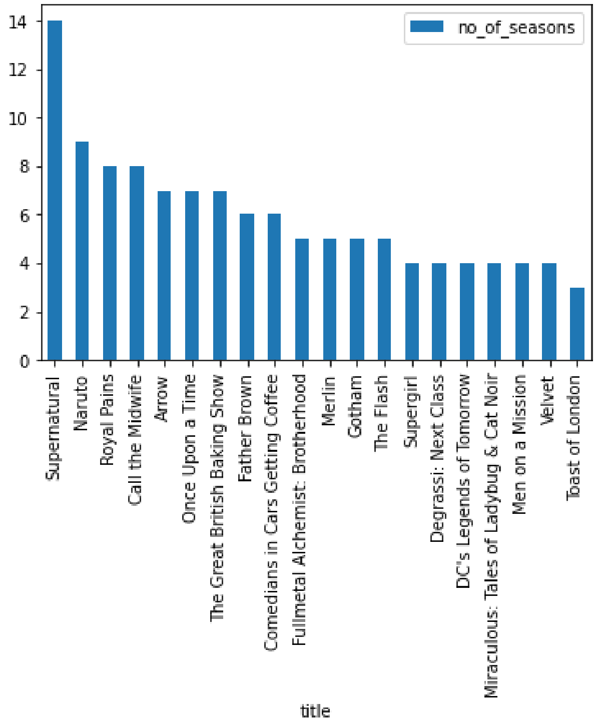


Furthermore, I analyzed the movie duration time. Here I used the kernel density estimate (KDE) function to display in plot.



Here, the above chart displays that a significant number of movies are of 75-120 mins which is the fair time to watch a movie.

So, for the TV shows, they have the duration in season form. Here, we sorted and visualize the top 20 TV shows by number of seasons.



In the given chart, we can see supernatural TV shows has the highest number of seasons which is 14 and Toast of London has the lowest among top 20.

Results

From the analysis of this dataset, I got to know the viewers preference based on the nature of content. From the analyzing chart of release years of movies and TV shows, we can see the trends of how the consumer changes their preference overtime and how the production priorities shifted. From the ratings patterns, it reveals that which type of movies and tv shows are mostly showing in Netflix and which genre has what kind of ratings.

Challenges and Improvements

This dataset has the limited number of quantitative values which make difficult to show the correlation between two variables. For example, if we have the ratings in the number form, we could show the relationships between duration and ratings that might give us some results about how the duration of content fluctuate the ratings. If this dataset has the more numeric values, we could visualize it clearly. Also, for the genre, the details are quite lengthy which makes the visualization unclear and for the country wise analysis, if we have more numeric data on how the content are released, we could show the relation with country and genre or ratings.

Conclusion

In conclusion, this analysis provided us a valuable insight about the temporal trends and users preference patterns. This analysis helps the company in data-driven decision making in an industry level. The key findings of this dataset are the user engagement influenced by the genre, country of production and duration. Netflix can use these insights to plan content creation cycles and target the actual audience. Also, dataset reflects the wide range of movies and tv shows from different country which allows this company to attract the audience from the global market and become a leading digital streaming service.

Code to Solve the problem:

```
# -*- coding: utf-8 -*-
"""
Created on Sun Dec  3 08:27:14 2023

@author: bidyabhattarai
"""



############ Importing Necessary Libraries ###############


import pandas as pd
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

import matplotlib.pyplot as plt

######## Loading the dataset ################

df = pd.read_csv('netflix_titles.csv')
df.head()

############## Displaying Column Information #################
df.columns, df.columns.__len__()
```

```
################ Checking the Null values in the dataset ##############

df.count()


##############Creating the copy of dataframe ###############


df_copy = df.copy()


############# Removing rows with null values #################

df_copy.dropna(inplace=True)

df_copy.count()

df_copy.info()



############ Visualizing the count of movies and TV shows ################

c = df_copy.copy().value_counts('type').reset_index()

c.columns = ['Type', 'count']

print(c)



######## plotting a pie chart for the distribution of movies and TV shows#####

plt.pie(c['count'], labels=c['Type'], autopct='%1.1f%%')

plt.show()
```

```
######## Extracting release year information ################

top_release=df_copy[['release_year']]

####### sorting and visualizing the count of releases by year ###############
############################################################################

top_release = top_release.sort_values(by='release_year', ascending=False)
top_release
plt.figure(figsize=(8, 12))
sns.countplot(data=top_release, y='release_year')
plt.xlabel('Count')
plt.ylabel('release_year')
plt.title('top release')
plt.show()



#################### creating dataframe of movies#######

net_movies=df_copy[df_copy['type']=='Movie']
```

```
####### Visualizing the count of movies by rating  ######

sns.countplot(data=net_movies, x='rating', )
plt.title('Movies Rating')
plt.show()

############# Extracting and visualizing the top movie ratings ##############
rating = net_movies.value_counts('rating')
print(rating)
rating[:5].plot(kind='pie')
plt.title('Top 5 Ratings of movies')
plt.show()



############# Extracting and visualizing the top movie genres ##############

genre = net_movies.value_counts('listed_in')
print(genre)
genre[:5].plot(kind='bar')
plt.title('Top 5 Genre of movies')
plt.show()
```

```
######### movie count released by years ##############

a = net_movies.copy().value_counts('release_year', ascending=False).reset_index()

a.columns = ['release_year', 'count']

print(a)

sns.scatterplot(data=a, x='release_year',y='count')

plt.title('movie count release by years')

plt.show()


########## creating dataframe of tv shows seprately ######


TV_shows=df_copy[df_copy['type']=='TV Show']


################## count rated tv shows #####################


sns.countplot(data=TV_shows, x='rating', )

plt.title('TV shows rating')

plt.show()


############# Extracting and visualizing the top movie ratings ##############

rating = TV_shows.value_counts('rating')

print(rating)

rating[:5].plot(kind='pie')

plt.title('Top 5 Ratings of TV_shows')
plt.show()


###### TV_shows count released by years#####################

a = TV_shows.copy().value_counts('release_year', ascending=False).reset_index()

a.columns = ['release_year', 'count']

print(a)

sns.scatterplot(data=a, x='release_year',y='count')

plt.show()



#######################################################

######## removing 'min' from the movie duration column ######


net_movies['duration']=net_movies['duration'].str.replace(' min','')

net_movies['duration']=net_movies['duration'].astype(str).astype(int)

net_movies['duration']



##############Plotting a kernel density estimate for movie time duration######


net_movies['duration'].plot(kind='kde')

plt.title('Movie Time Duration')
```

```
#####################Extracting the number of seasons from TV shows data######

features=['title','duration']

durations= TV_shows[features]

durations['no_of_seasons']=durations['duration'].str.replace(' Season','')

durations['no_of_seasons']=durations['no_of_seasons'].str.replace('s','')

durations['no_of_seasons']=durations['no_of_seasons'].astype(str).astype(int)

t=['title','no_of_seasons']

top=durations[t]

############### Sorting and visualizing the top 20 TV shows

############### by number of seasons##############

top=top.sort_values(by='no_of_seasons', ascending=False)

top20=top[0:20]
```

```
top20.plot(kind='bar',x='title',y='no_of_seasons')
```