Project: Investigate a Dataset (TMDB-Movies)

After going through all the datasets, I chose to analyse the first dataset , namely TMDB-Movies.

As this seemed as an interesting topic to start off with, I decided to take it up.

After going through the dataset and the sample questions, I came up with the following questions which I thought were necessary to be answered:

1. Which genres were most popular from year to year ?
2. Is the any relationship between the type of genre and votes ?
3. How are revenue generated and the rating the film received related ?

A description of how I investigated these questions.

First I imported all the packages I would be needing during my analysis.
Next I loaded the dataset to be investigated.

Printed the dataset to check the structure and length of the data set.
Next I cleaned the data by converting the different columns to their appropriate data types ex. To datetime type.

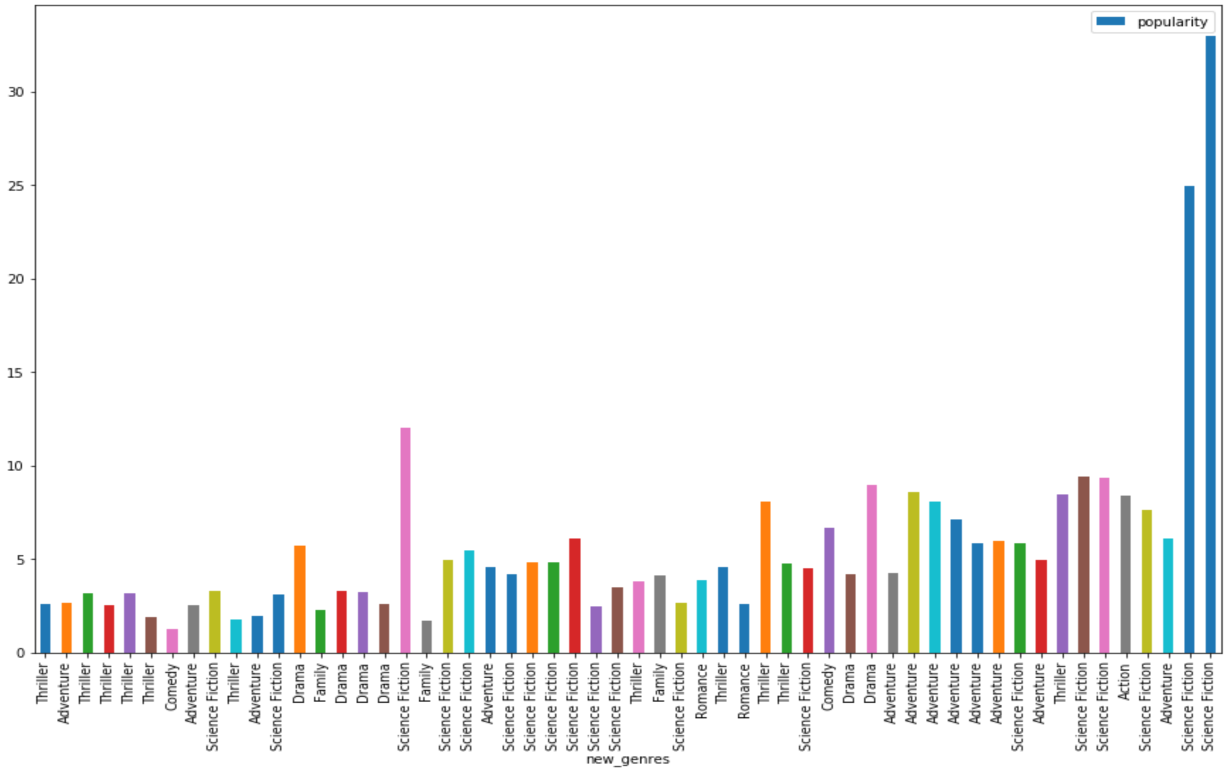
Then deleted all the duplicate data, handled the important missing data and dropped the other missing data.

After cleaning, I went on to address the three questions I posed.

1. Which genres were most popular from year to year ?

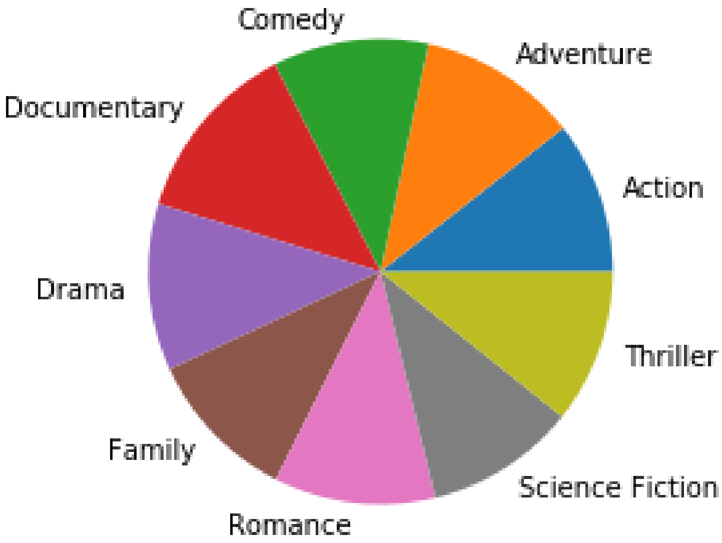
   For this I filtered out the columns I needed, here the genres, release year and popularity.
   Then I using ***groupby*** and ***idmax()*** I got release year and the most popular genre for that year.

   Then using plot() function I plotted ***popularity*** and ***genres*** for the given years.
   Below is the graph generated:



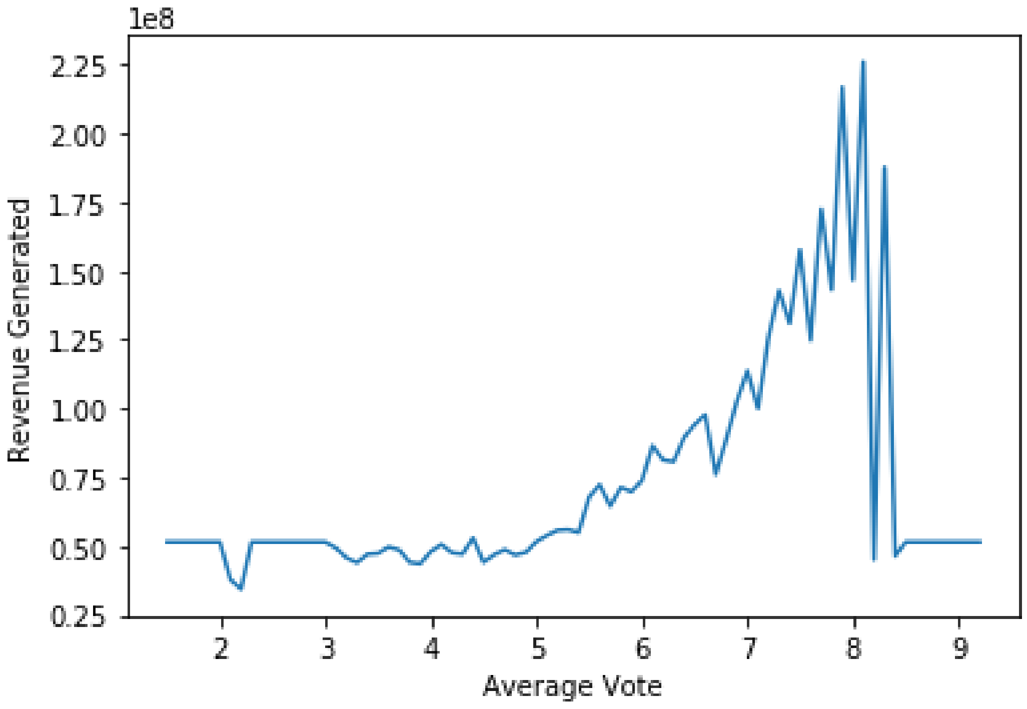
2. Is the any relationship between the type of genre and votes ?

   First filtered out the columns required ie, 'vote_average' and 'new_genres'.
   Then I grouped and found out the mean of each category of genre.
   Using pie chart I plotted the graph between vote average and genres.
   Below is the pie chart generated.



3. How are revenue generated and the rating the film received related ?

   Grouping revenue by vote average and finding mean.
   Then plotted a simple line graph to check for any correlation between the two parameters:



Conclusions

From the bar graph, it is clear that the genre of "Science Fiction" has gained more popularity over the years.

From the second graph which is a pie chart, we can conclude easily that there is no relation as such between genres type and their voting pattern, since all the different genres have similar voting averages.

From the last line graph, it is clear that the movies which receive higher votes have done well at the profit end as well and have generated higher revenues.

Resources used:

- https://pandas.pydata.org
- https://www.geeksforgeeks.org
- https://matplotlib.org
- https://mode.com/python-tutorial/