Most Rated Genres

Mohammed Innat Mahmud

Abstract

IMDB Movie Dataset. Our research question is **What types of movies genres user viewed and rated most than other movies genres ?** We take some general approach to solve it. We successfully find the relationship between genres and ratings value including average ratings value with concern launch year.

Motivation

We want find the most rated genres in the given movies data set. Our goal is to find what types of movies user like most and rated most as well. We have to merge two data set for this purpose and doing some operation on it to get the desire result.

If we're asked what genres people like most and vote them as well. Here we try to give the answer.

.

Dataset(s)

We are using the following dataset : **IMDB Movie Dataset**

This dataset describe 5-star ratings and free-text tagging activity from , a move recommendation service. It contains 20000263 ratings and 465564 tags applications across 27278 movies. These data were created by 138493 user between January 09 1995 and March 31 , 2015.

The data are contained in six files links.csv , movies.csv , ratings.csv and tags.csv etc. In this project we will use only two but massive csv files which is movies.csv and ratings.csv.

Data Set can be get from this site : http://grouplens.org/datasets/

Data Preparation and Cleaning

We are fully concern on preparing and cleaing our data after acquiring it from the source. Befor we start further operation like statistical or merge or plot or something like that we know we have to prepare our data first. Real time data are often corrupted with unexpected value or null value, to get rid of these problem we need to operate some data cleaning operation.

We look through first is there any null value , if so drop those rows containing null values. This is how we prepare our data set , most like cleaing the corrupted data or missing data.

We took two data frame for our operation and clean it well. In movies data set , we have title and genres columns , from their we need to extract the launching year for each movies and create a new column name year and after that we delete the title column as it’s then not necessary any more. Then taking average ratings from ratings data frame we merge the both data set and group them by with movie Id.

We again then clean the new created data set and make sure there is no null or missing value in our data set.

Research Question

Our research question is **What types of movies genres user viewed and rated most than other movies genres ?**

Methods

For our data analysis we use mostly pandas module also matplotlib module for ploting purposes. The research question that we want to find answer lead us to use some of most effective function in pandas.

The methods is a general approach nothing rocket science but effective enough to give the answer of our research question.

We created a new data set from movies and ratings data set , where we get some essential columns such as ‘MovieId’ , ‘Title’ , ‘Genres’ , ‘Ratings’ … Now , from Title column , we extract the value of Launching year of each movies and made a

New column name as ‘Year’ . We did it we can visualize the co-relation of ratings values of movies with its launching year. We compared between ratings and genres , and see a significant result. Among of all genres , ‘Drama’ has the height in number , so we found a general idea that Drama movies are rated mostly. We plot pie to see these on diagram.

We also observe the ratings value of movies with its concern launching year and visualize it by plotting.

Findings

We can visualize that drama genres are most in number and rated more than other movies genres.

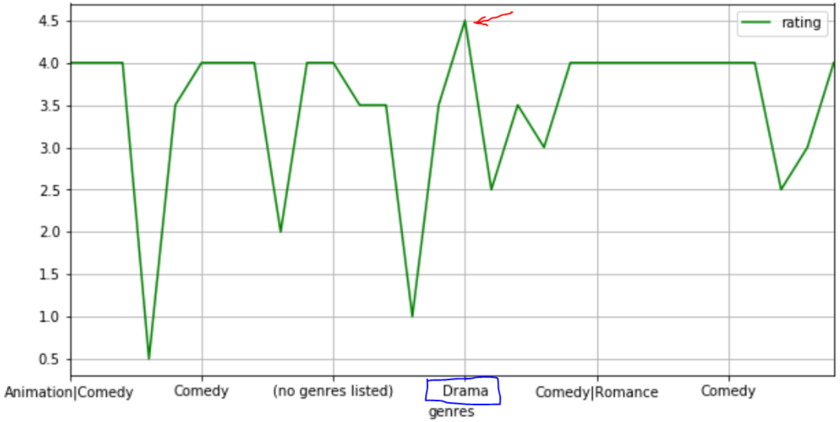

Fig : Ploting ratings VS genres . Drama genres tend to high than other movie genres

Visualize in pie plot with percentile.

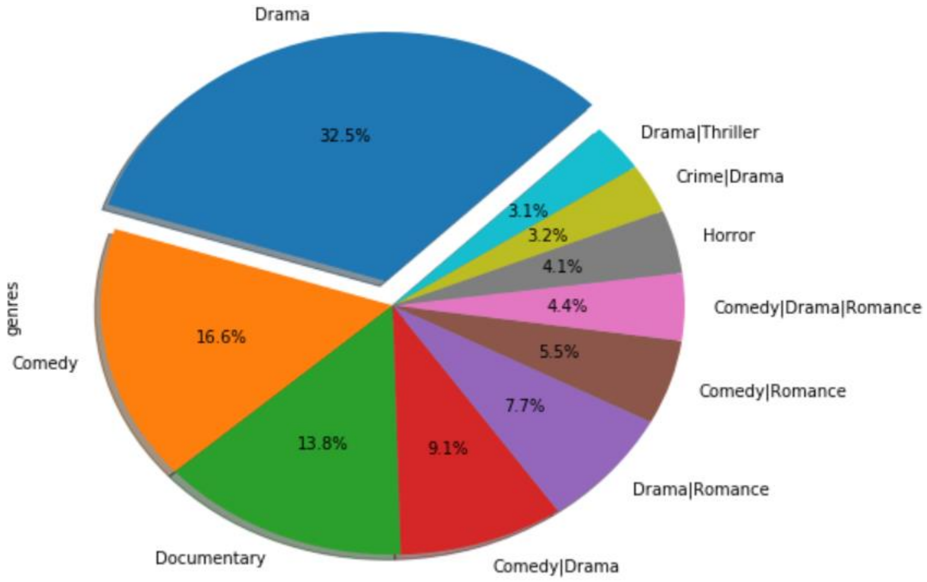

Fig : Ploting the most frequent genres , here which is Drama

Rating value of each movies with concern launching year. Also the average ratings of value in ratings columns.

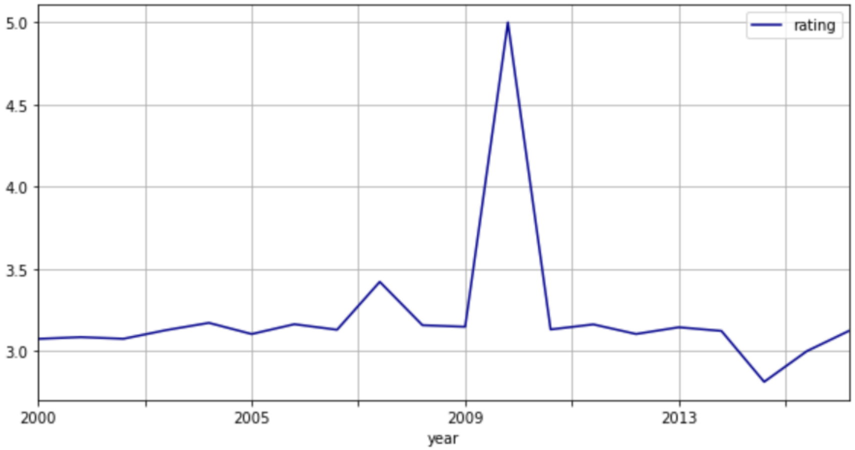

Fig : Average Movie Ratings over Time

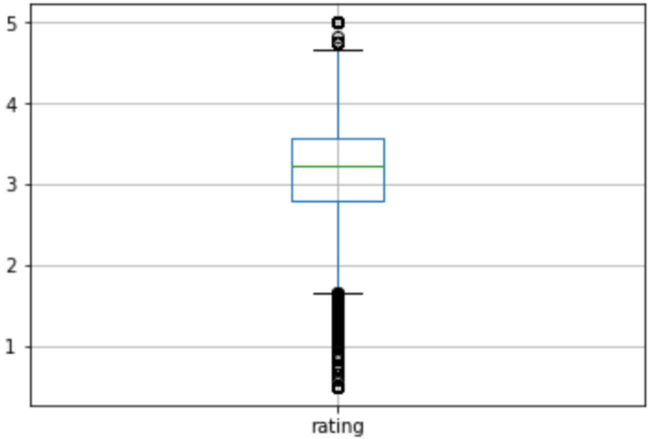

Fig : Visualize Rating in Box Plot

Limitations

The whole dataset frequently updated so , new data generate every time and variation is obvious . To generate a predictive model we need more data to find out the hidder pattern of it to predict the model.

Conclusions

We try to answer our research question by exploring our merge data and find some interesting result from our analysis.We found that drama genres are release more than other genres and viewd most and so rated most as well.

Acknowledgements

Where did you get your data? Did you use other informal analysis to inform your work? Did you get feedback on your work by friends or colleagues? Etc. If you had no one give you feedback and you collected the data yourself, say so.

We are using the following dataset : **IMDB Movie Dataset** http://grouplens.org/datasets/ . And yes , I used some other informal analysis to inform my work. Taking about feedback , unfortunately I don't have no one to get feedback from now.

References

http://grouplens.org/datasets/

http://pandas.pydata.org/pandas-docs/stable