

**“Analysis of Feature Films”**

**Master of Data Science**



# *Submitted By*

**Harini G(2048034)**

**DEPARTMENT OF COMPUTER SCIENCE (UG)**

**MSc PROGRAMME**

**[CHRIST (DEEMED TO BE UNIVERSITY)](https://www.christuniversity.in/)**

**Hosur Rd, Bhavani Nagar, S.G. Palya, Bengaluru, Karnataka 560029**

**Table of Content**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Topic** | **Page No** |
| 1. | Introduction | 3 |
| 2 | Scope of the Project | 4 |
| 3 | Problem statement | 4 |
| 4 | Data Collection | 5 |
| 5 | Data Description | 5-6 |
| 6 | Exploratory Data Analysis | 7-15 |
| 7 | Pre-Processing | 16 |
| 8 | Processing | 16-18 |
| 9 | Conclusion | 19 |
| 10 | Acknowledgment | 20 |
| 11 | References | 21 |

1. **Introduction**

Internet Movie Database (IMDb) is an online information base committed to a wide range of data about a broad scope of film substance, for example, movies, TV and web-based streaming shows, etc. The data which is introduced on the IMDb portal incorporates cast, creation group, director crew, individual accounts, plot outlines, random data, evaluations, fan and critics reviews.

The [IMDb dataset](https://analyticsindiamag.com/20-machine-learning-datasets-project-ideas/) contains 50,000 surveys, permitting close to 30 audits for each film. It was developed in 2011 by the researchers: Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts of Stanford University. The dataset was evenly divided into training and test sets. The training set contains 25000 reviews so as the test set.

A negative review has a score of ≤ 4 out of 10, and a positive survey has a score of ≥ 7 out of 10. Neutral studies were excluded from this dataset.

Here, we will examine the information in this dataset, how it was gathered, and give some benchmark models that gave high accuracy on this dataset. Further, we will implement the IMDB dataset using [Keras](https://analyticsindiamag.com/tutorial-on-keras-tokenizer-for-text-classification-in-nlp/) Library.

1. **Scope of the project**

* Understand the average ratings for different movie genres over the years (from 1995 to 2015). The correlation between the trends for other genres (8 different genres is considered: Animation, Comedy, Adventure, Horror, Action, Drama, Biography, and Crime).
* This will give us insight into how people's liking for the different movie genres change over time and the strength of association between trends in between different movie genres, possibly valuable insights for the critics.

1. **Problem Statement**

The answer to the following research questions will be searched for, using exploratory analysis and visualization:

1. The trend in average ratings for different movie genres: How the average ratings for a few other movie genres (namely, Animation, Comedy, Romance, Thriller, and Horror) change over time (different years, from 1995 to 2015)? How can the average ratings for different genres be compared among themselves?
2. Correlation between the trends for different genres: How are the directions for the genres correlated? For example, are the average ratings for Comedy and Crime movies positively associated with each other? What is the strength of association?

### ****Data Collection****

The raw data was collected by the researchers from the IMDb website. They searched the content information present in each of the reviews and discovered any highlights that were representative for judging whether the review was positive or negative. The studies were then evenly divided into training and test sets uploaded to their website. In each of the directories contained in the collections, there are another two directories representing pos and neg tags to partition the information through various marks. In every one of these folders, numerous TXT records contain the substance of the film survey, with each document containing one report.

1. **Dataset Description**

The Dataset has numeric and categorical variables. Dataset is well structured, and no duplicated values are present.

Table1: Dataset Details

|  |  |  |
| --- | --- | --- |
| **Sl.No** | **Criteria** | **Details** |
| 1 | Name | Feature Films |
| 2 | Type | Multivariate |
| 3 | No. of Rows | 100 |
| 4 | No. of Columns | 10 |
| 5 | Missing Values | Yes |
| 6 | Target Type | Categorical |
| 7 | Application Technique | Classification |

|  |  |  |
| --- | --- | --- |
| **Sl.No** | **Feature Name** | **Description** |
| 1 | Rank | The Rank of the film from 1 to 100 on the list of 100 most popular feature films. |
| 2 | Title | The title of the feature film. |
| 3 | Description | The description of the feature film. |
| 4 | Runtime | The duration of the feature film. |
| 5 | Genre | The genre of the feature film. |
| 6 | Rating | The IMDb rating of the feature film. |
| 7 | Metascore | The Metascore on the IMDb website for the feature film. |
| 8 | Votes | Votes cast in favor of the feature film. |
| 9 | Gross\_Earnings\_in\_Mil | The gross earnings of the feature film in millions. |
| 10 | Director | The main director of the feature film. Note, in the case of multiple directors. I'll take only the first. |
| 11 | Actor | The main actor in the feature film. Note, in the case of multiple actors. I'll take only the first. |

Table2: Feature Description

1. **Exploratory Data Analysis**

Data exploration is an approach similar to initial data analysis. A data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, Patterns in the dataset, Statistical techniques in the dataset rather than traditional data management systems.

Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more. Using interactive dashboards and point-and-click data exploration, users can better understand the bigger picture and get insights faster.

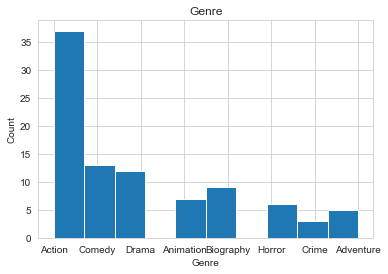


Fig1: Genre

Compared to other Genres, most people prefer to watch Action Movies, and people are very much interested and anxious to manage such types of movies.

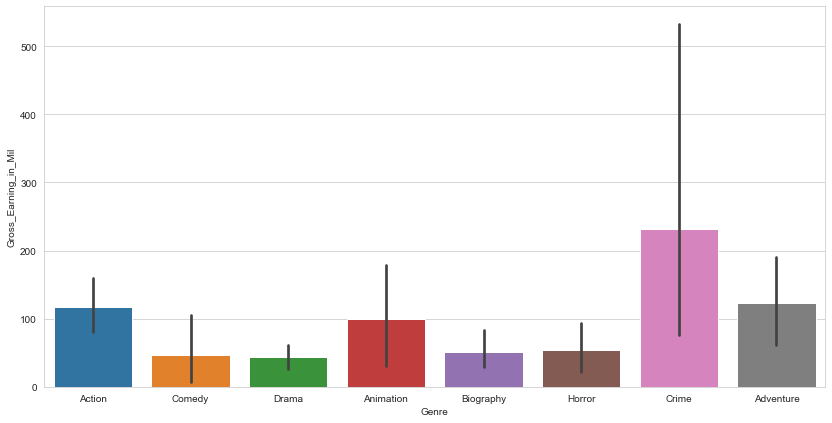


Fig2: Genre Vs. Gross Earnings

Regarding the above graph, we can have a clear idea that according to gross earnings, the crime rate movies have high importance in earnings compared to the other genres.

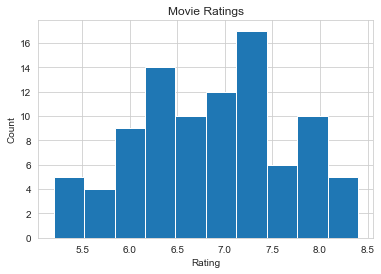


Fig3: Rating

Most of the movies are rated between the frequency of 6.0 to 7.5

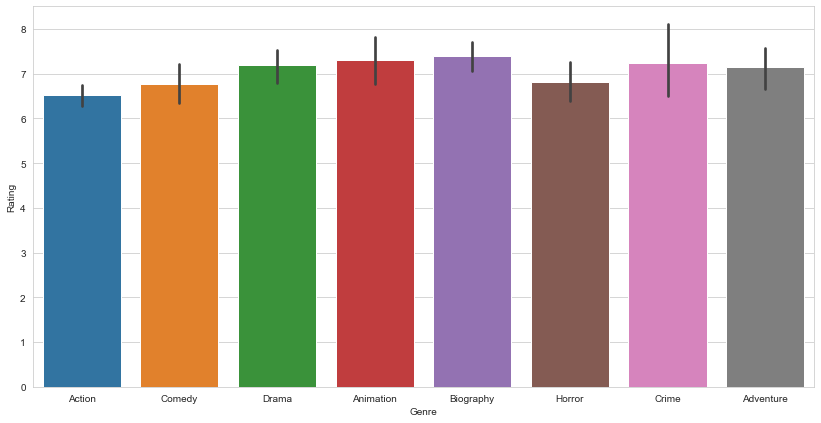


Fig4: Genre Vs. Rating

From the above graph, we can have a clear idea of the rating, Biography as high Rating compared to other Genres.

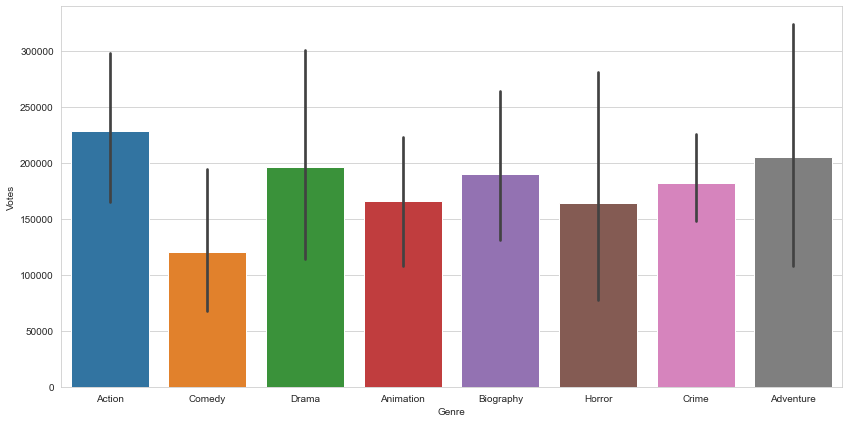


Fig5: Genre Vs. Votes

From the above graph, we can observe that Action movies are preferred concerning votes compared to other Genres, Such as Comedy, Drama, Animation, Biography, Horror, Crime, and Adventure.

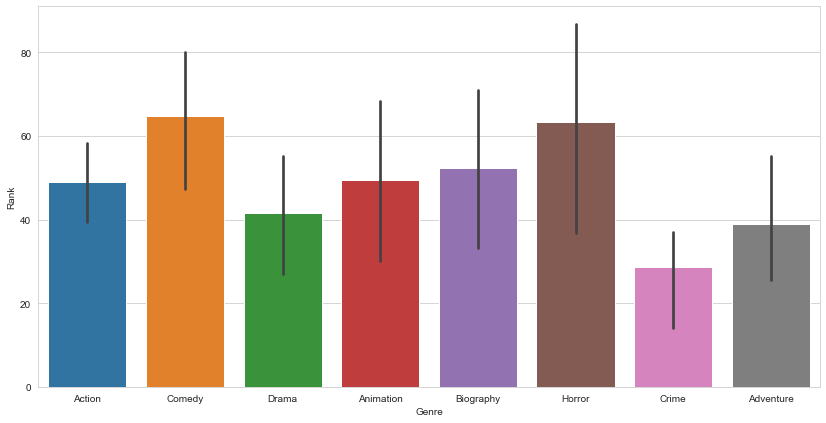


Fig6: Genre Vs. Rank

From the overhead graph, we can have a precise analysis, i.e., with respect to Rank, we can visualize that Comedy and Horror Movies were preferred.

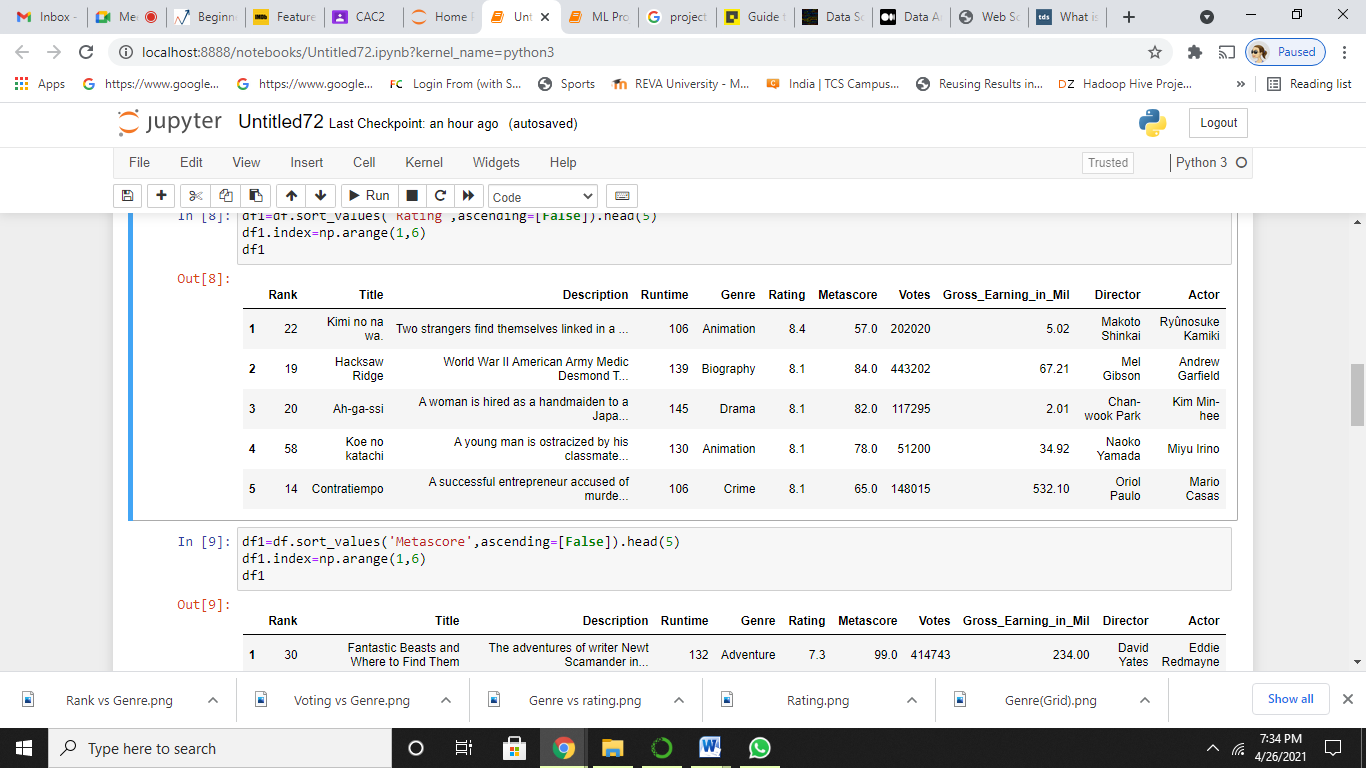


Table1: Top 5 Movies based on Rating

Here we can have a clear idea that the best five movies were shown with respect to Rating.

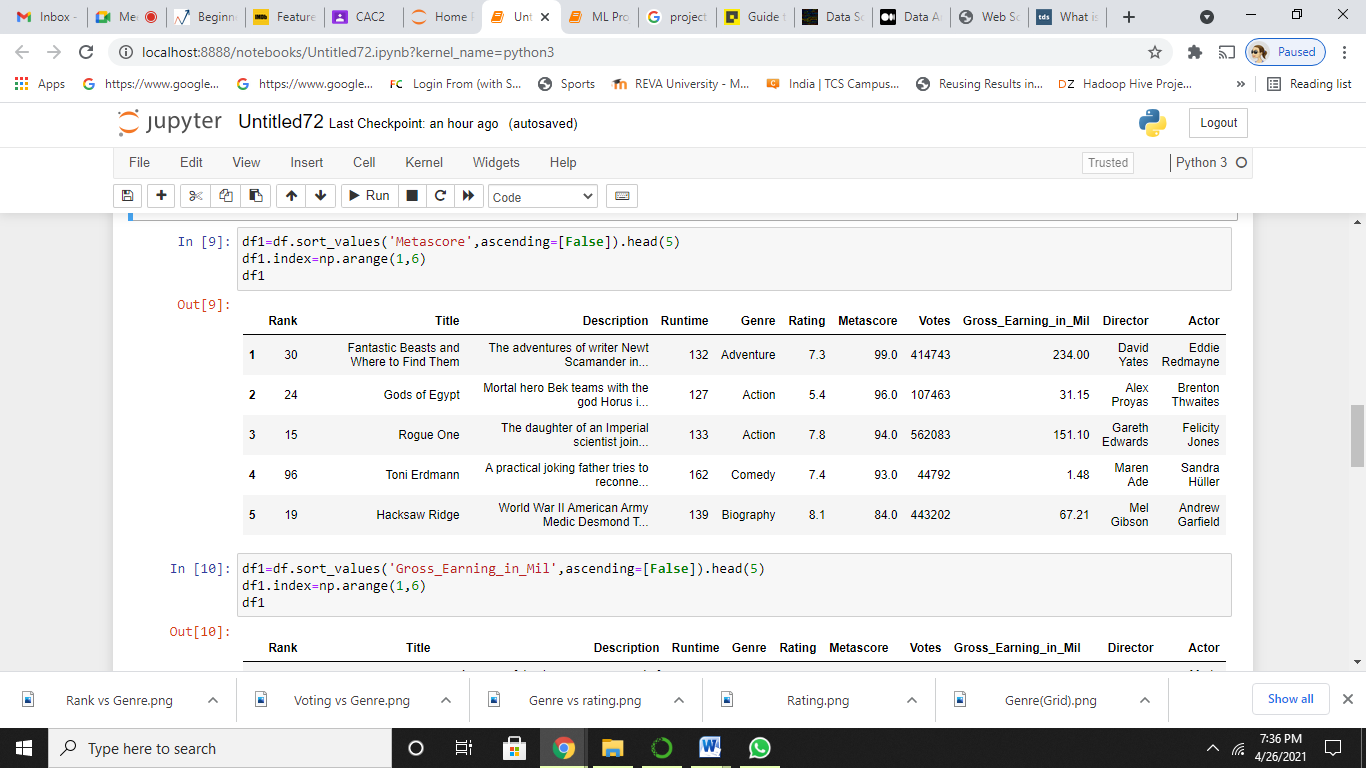


Table2: Top 5 Movies based on Metascore

With respect to Metascore, these were the top 5 best movies when compared to other films.

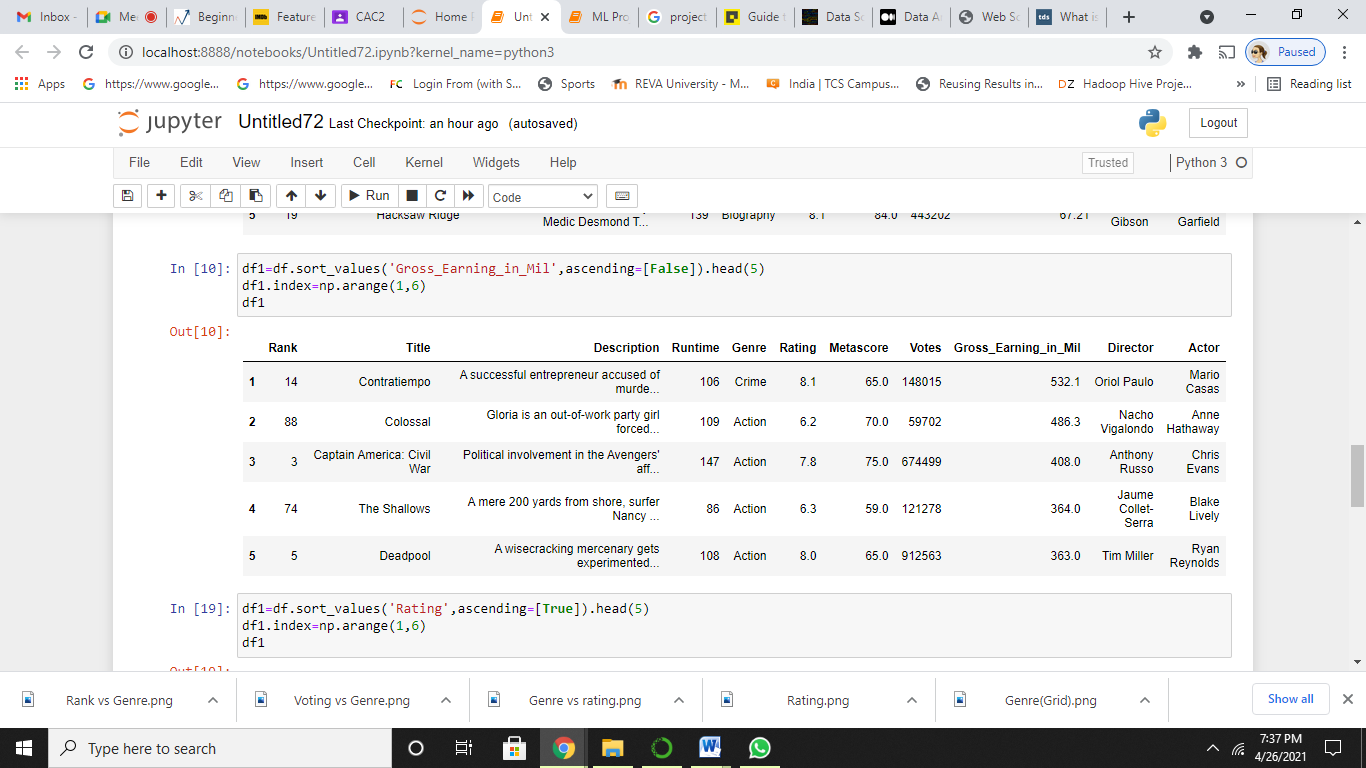


Table3: Top 5 movies based on Earnings

This list of 5 movies has high earnings when compared to the other movies.

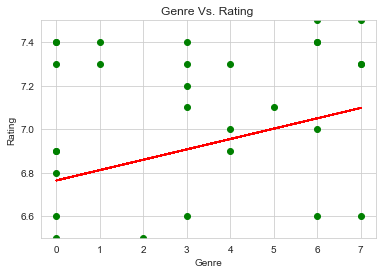


Fig7: Relation between Genre Vs. Rating

From the above graph, we could conclude an Increasing linear relationship between Genre and Rating, which indicates a Moderate correlation between them.

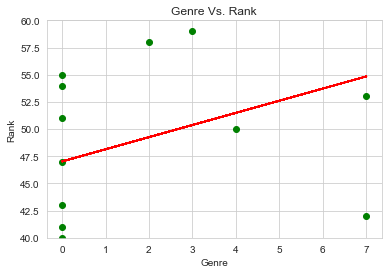


Fig8: Relation between Genre Vs. Rank

From the above graph, we could conclude an Increasing linear relationship between Genre and Rank, which indicates a Moderate correlation between them.

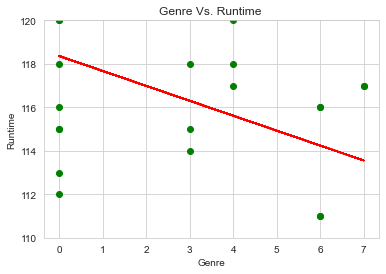


Fig9: Relation between Genre Vs. Runtime

From the above graph, we could conclude that there is Decreasing linear relationship between Genre and Runtime, which indicates a negative correlation between them.

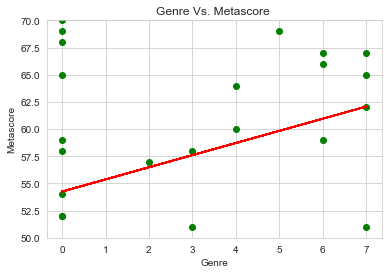


Fig10: Relation between Genre Vs. Metascore

From the above graph, we could conclude an Increasing linear relationship between Genre and Metascore, which indicates a Moderate correlation between them.

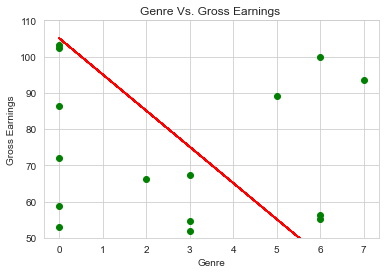


Fig11: Relation between Genre Vs. Gross Earning

From the above graph, we could conclude that there is Decreasing linear relationship between Genre and Gross Earnings, which indicates a negative correlation between them.

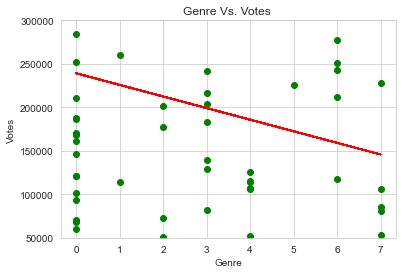


Fig12: Relation between Genre Vs. Votes

From the above graph, we could conclude that there is Decreasing linear relationship between Genre and Votes, which indicates a negative correlation between them.

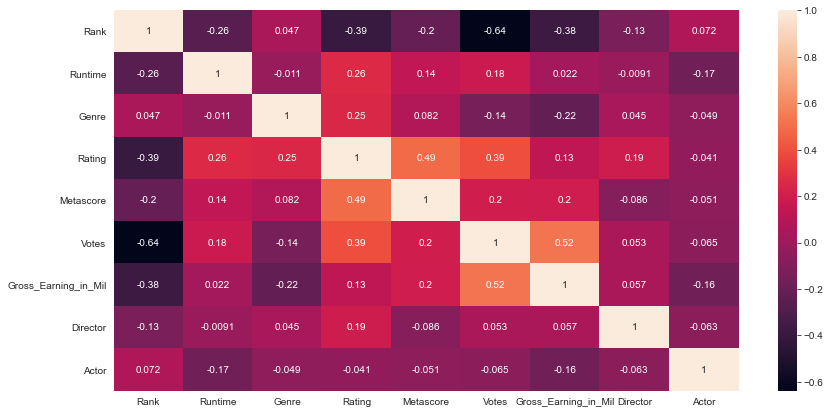


Fig14: Heat Map

* Rank and Votes have Low correlation.
* Runtime and Rating had a Moderate correlation.
* Rating and Metascore have a Good correlation.
* Votes and Gross Earning has High correlation.

1. **Pre-Processing**

The dataset had few missing values, which were further removed.

Label encoding is done for the following features, such as:

1. Genre
2. Actor
3. Director

The dataset is composed of dependent and independent variables.

Here, the independent variable is:

1. Rank
2. Runtime
3. Rating
4. Metascore
5. Votes
6. Director
7. Actor
8. Gross\_Earing\_in\_Mil

Here, the genre is considered as the dependent variable.

1. **Processing**

The data needs to be split into training and testing. Using the train\_test.split(), the data is divided into training 80% and 20% testing.

The output variable of the dataset is the Categorical variable. So, implementing classification algorithms.

**Model Implementation:**

1. Random Forest: Random forest algorithm is a supervised classification and regression algorithm. This algorithm randomly creates a forest with several trees—the more trees in the forest, the more robust the forest looks. Similarly, in the random forest classifier, the higher the number of trees in the forest, the greater is the accuracy of the results. Random forest builds multiple decision trees (called the forest) and glues them together to get a more accurate and stable prediction. The forest it builds is a collection of Decision Trees trained with the bagging method.

The model's accuracy by setting the number of trees to 500 is 0.58(58%).

Out of 19 test data observations, eight observations are missed classified.

F1 score is 0.69, and the execution time of random forest is less than 1 second.

1. SVM(Support Vector Machine): Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It is mainly used in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features), with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that best differentiates the two classes.

The accuracy of the model by setting Kernel as sigmoid, C=0.5, and Gamma as 1is 0.5(53%).

Out of 19 test data observations, nine observations are missed classified.

F1 score is 0.69, and the execution time of random forest is less than 3 seconds.

1. KNN(**k-nearest neighbors):** KNN or k-nearest neighbors is the simplest classification algorithm. This classification algorithm does not depend on the structure of the data. Whenever a new example is encountered, its k nearest neighbors from the training data are examined. Distance between two samples can be the Euclidean distance between their feature vectors.

The accuracy of the model by setting k=5 as 1is 0.47(47%).

Out of 19 test data observations, ten observations are missed classified.

F1 score is 0.67, and the execution time of random forest is less than 1 second.

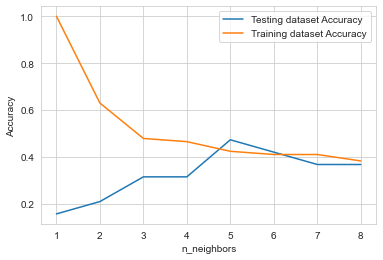
****

Fig15: KNN (Training Vs. Testing)

The value of Neighbors increases, the accuracy of the training data decreases, indicating that the model is not a good fit for the provided data.

1. **Conclusion**

In this article, we have discussed the details and implementation of the IMDb dataset using Keras Library. The model trained on the test data gave a moderate accuracy of around 60%. Additionally, we can increase the accuracy by introducing the model with more epochs.

* The people more prefer action movies.
* Crime Genre movies earn more money when compared to other Genres.
* Most of the movies are ranked between 7 to 7.5.
* Animation and Biography Genres has more Rating when compared to other Genres.
* Action movies are more popular when compared to other Genre movies.
* There is an increasing linear relationship between

1. Genre and Rating
2. Genre and Rank
3. Genre and Metascore

* There is a high negative correlation between variables

1. Rank and Votes
2. Runtime and Rank

* There is a positive correlation between variables

1. Runtime and Rating
2. Votes and Rating
3. Metascore and Rating
4. Gross Earning and Votes.

Movies have a Rating of more than 7, Runtime more than 2 hours, and Genres Action, Biography, and Animation movies are most likely selected for Best Pictures.

**Acknowledgment**

I, **Harini G** would like to thank **IMDb** for providing permission for web scrapping the information from their website.

And also we, hereby declare that the project work entitled **ANALYSIS OF FEATURE FILMS** is an original project work carried out by us, under the guidance of **Prof. Ummesalma.**

**References**

1. <http://min2bro.github.io/WebScraping/#/fragments>
2. <https://analyticsindiamag.com/guide-to-imdb-movie-dataset-with-python-implementation/>
3. <https://medium.com/analytics-vidhya/data-analysis-end-to-end-imdb-dataset-2b6d9976ebc2>
4. <https://www.analyticsvidhya.com/blog/2019/11/4-tricks-improve-deep-learning-model-performance/>
5. <https://www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/#:~:text=SVM%20(Support%20vector%20machine)%20classifier%20%E2%80%93&text=To%20know%20more%20about%20kernel,sci%2Dkit%20learn%20and%20SVM.&text=KNN%20(k%2Dnearest%20neighbours),is%20the%20simplest%20classification%20algorithm>.
6. <https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/>