# Relational Data Modelling in Data Science

(Assignment 1)

Student(s): Nikita Kapoor

Instructor: Sean Chester

Course: CSC501 Data Models and Algorithms

Program: Masters in Applied Data Science

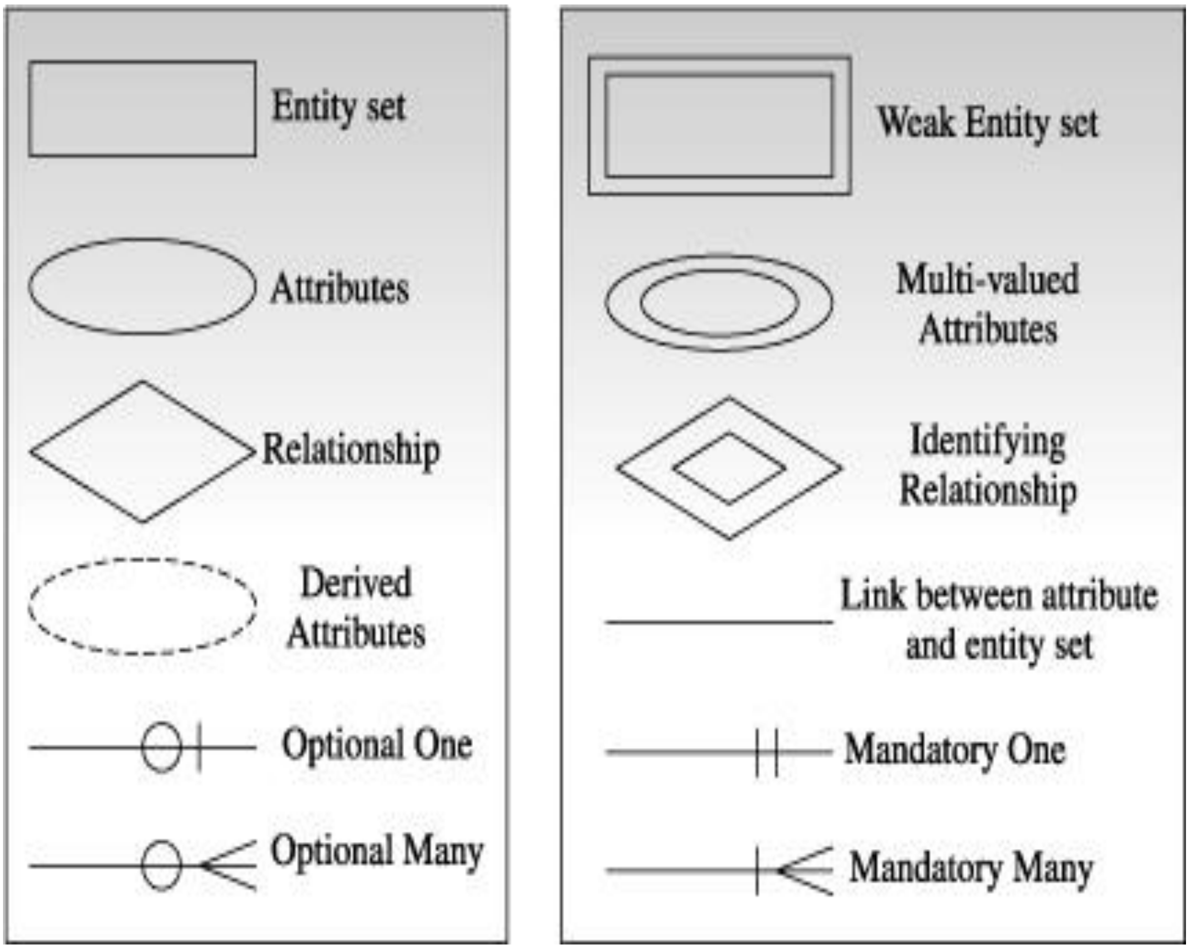
Date Due: 26 September 2019

## ABSTRACT

In this report, visualizations on Movielens 20M dataset have been done, doing so we revealed a lot of interesting insights of the history of cinematography like trends of genres over the years, ratings provided by the users etc. Python has been used as the language for writing the code and dataset has been managed using PostgreSql. For plotting graphs we have used Matplotlib and Pandas & Numpy library are used for cleaning and manipulation of data. Relational model has been used for solving the challenges faced during implementation.

# Table of Contents

## NOMENCLATURE



Source: https://www.oreilly.com/library/view/database-systems-concepts/9788177585674/9788177585674_ch06lev1sec5.h

# 1. INTRODUCTION

This report provides information obtained from visualization done on the Movielens 20M dataset. This report will pay particular attention on the challenges faced while working on the dataset and how we solved those using relational model and other techniques. This report will also talk about the insights that we received from visualizations done on the dataset.

# 2. Technology Used

1) PostgreSQL: For creating the Relational Database.
2) Anaconda Jupyter Notebook: Platform for Python and doing visualizations using Python’s Matplotlib library. Pandas and Numpy are used for data cleaning and manipulation.

# 3. VISUALISATIONS AND DATA MODELLING

Visualisations gives us a clear and a better view of the information, which sometimes appear as raw and unproductive. Through visualisations, we can extract useful insights and inferences about the data in the form of bar-graphs, histogram, line plots, and scatter-plots etcetera. It even becomes more useful and insightful if we do the data-cleaning first. Here comes the role of Data modelling. Data modelling is, as the name states, modelling of the raw data so that it becomes easy to do the data visualisation. Here, in our project, we are mainly concerned about modelling the data into Relational data models.

## 3.1 MODELLING THE DATA

At first, we modelled the raw data into relational data models. As mentioned, we used PostgreSQL for managing the data in the database. We had a total of six data sets, or six Comma Separated Files. We analysed each one of them and did the data cleaning. Fortunately, in our case, all the data sets were already normalized and didn't need any further cleaning.

But, we did some manipulations with the data to make the things more clear. We separated out the movie release years from the title of the movie which proved to be useful for the data visualisation part. We also extracted out the dates from the timestamps mentioned in the "ratings.csv" and "tags.csv". Furthermore, we applied some join and group-by operations on the data to extract out some useful insights from the data. Last, but not the least, was the "movies.csv" table which had the movieId, title and genres. But the genres were made a multi-valued attribute which contained multiple genres for a movie. So, we split the attribute into multiple atomic attributes such as action, adventure, comedy, romance, fantasy etc. But, in the ER diagram of the following database, we did not mention each and every genre individually, rather we just made the genre as the multi-valued attribute due to the space constraints.

Now, we hereby give the table definition :-

genomescore(movieId , tagId, relevance):

- genometags (tagId, tag):
- movies (movieId, title, year, action, adventure, animation ….):
- tags (userId, movieId, tag, year, month):
- rating (userId, movieId, rating, timestamp, year, month):
- links (movieId, imdbId, tmdbId):

By looking at the above table definitions, it is evident that how the table definition have been manipulated to simplify the data visualisation process. Hence, relational modelling have proved to be a wonderful tool in simplifying the data visualisations process. Although, it is just the one aspect of the relational modelling and hence shouldn't be considered as the final verdict.

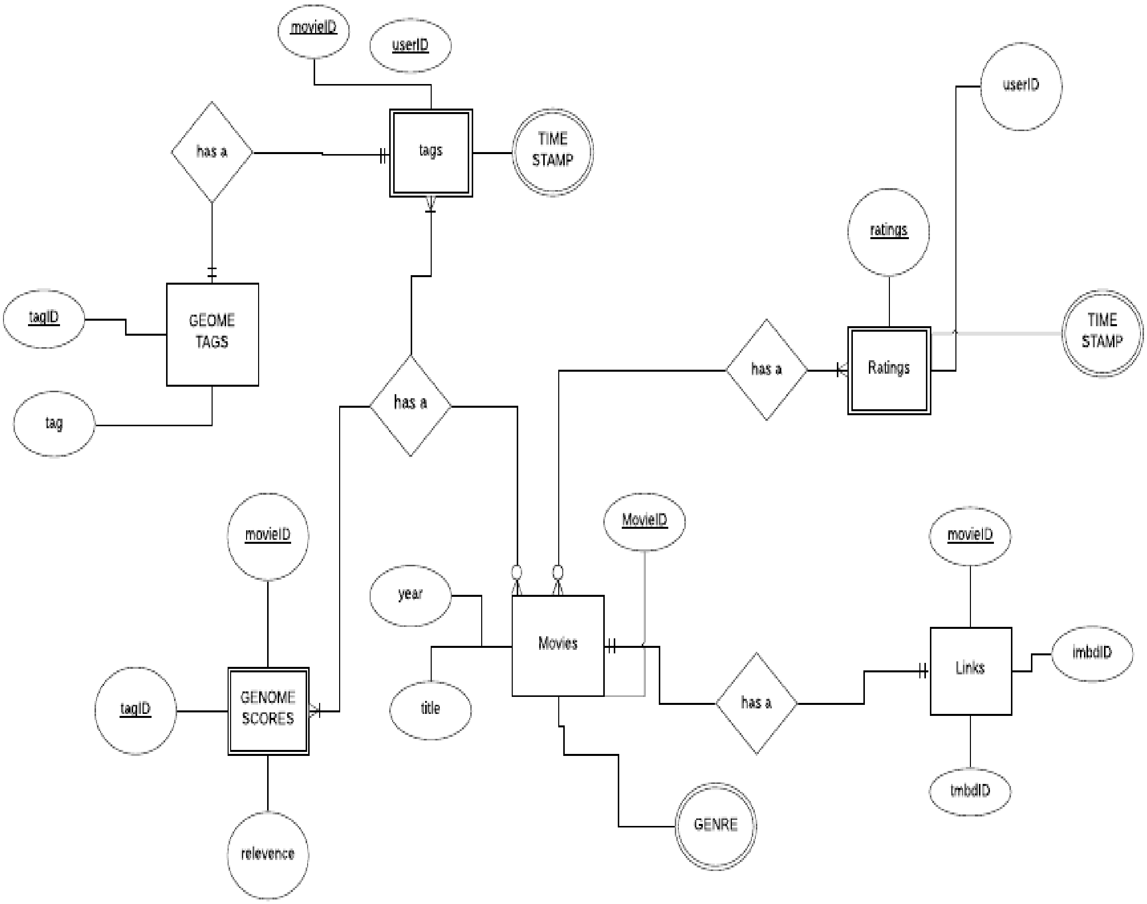


**Fig 1: ER DAIGRAM**

The Entity relationship diagram above clearly demonstrates the relationship between entities and the attributes of the various entities.

### 3.2 TRANSFORMING THE DATA INTO VISUALISATIONS

After the data modelling, we extracted out some useful, unintuitive insights about the data. So we plotted the graphs accordingly.

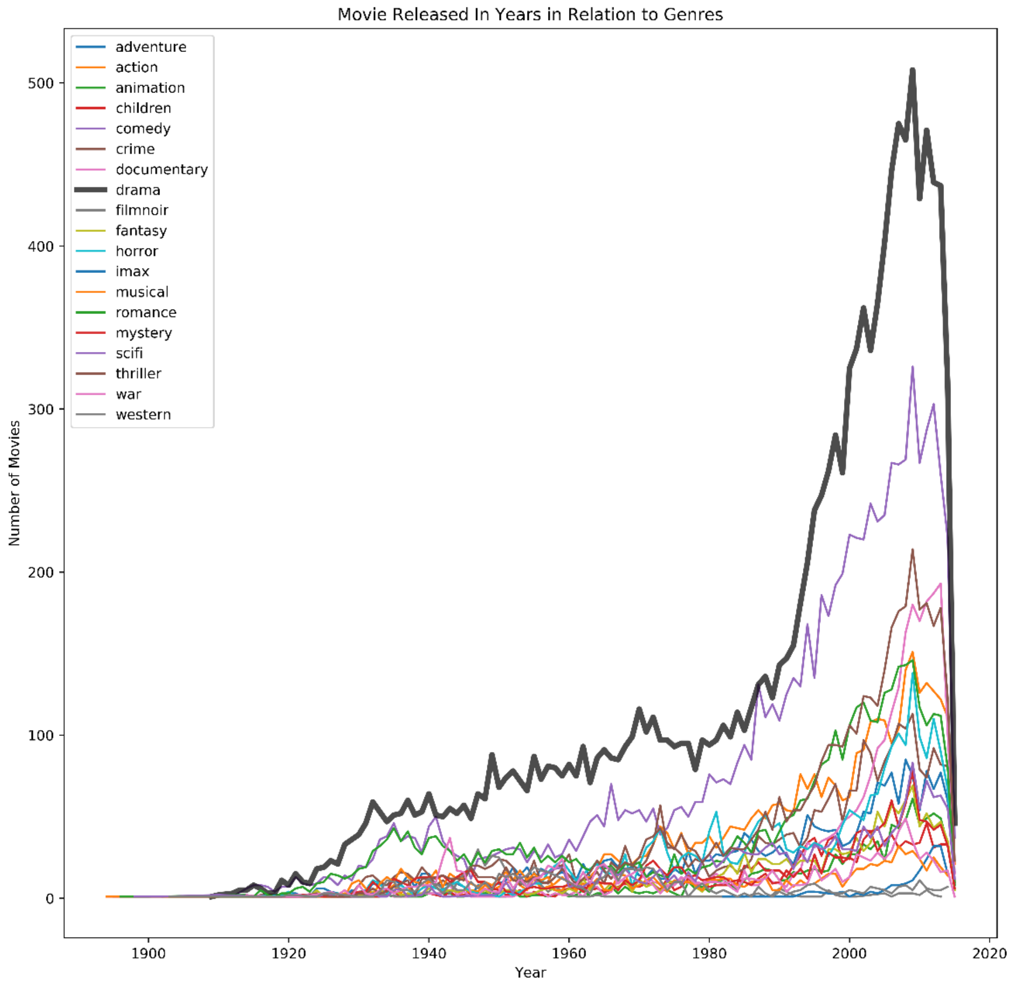


**Fig 2: Trend of genres over a century**

This graph is between the number of movies per genres released over a century. From this graph we can see that the movies released with genre “Drama” over the years were maximum throughout the century. This indicates the popularity and the demand of drama genre among the people. Not falling far behind was the “Sci-fi” genre which competed with the drama throughout the span of 100 years. However, the western movies struggled to create the demand in the industry as this genre movies released were the least throughout the years. The movies released of every genre decreased badly around 2015.

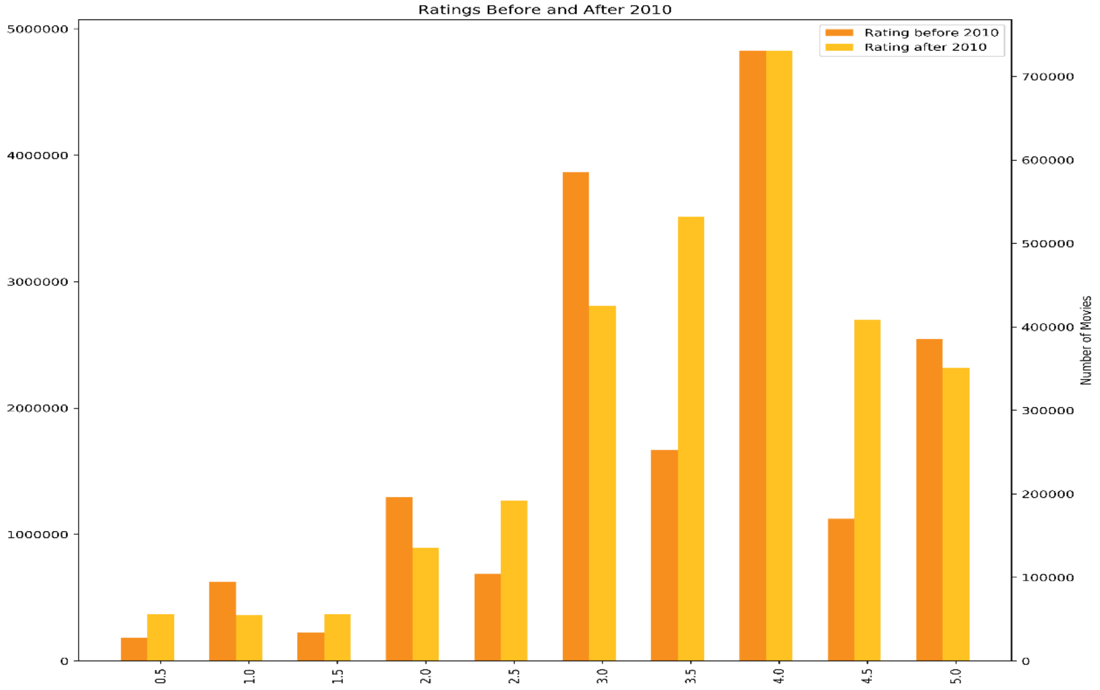


**Fig 3: Ratings before and after 2010**

From this graph we saw that the maximum number of movies were rated 4 .0 before and even after the year 2010. And that equal number of movies were rated 4.0.

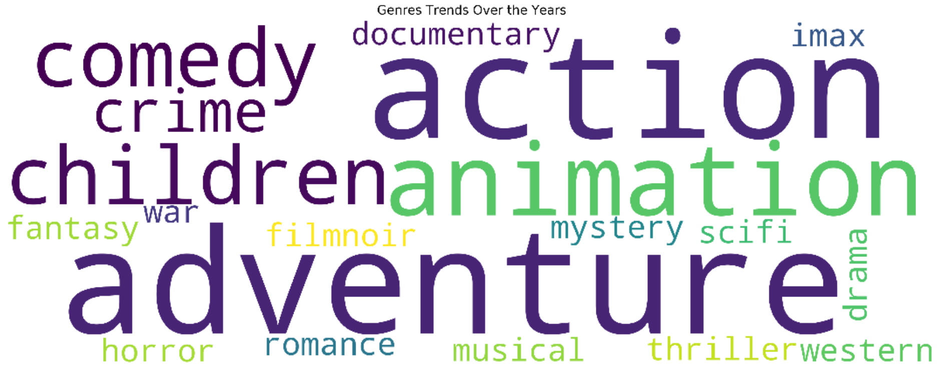
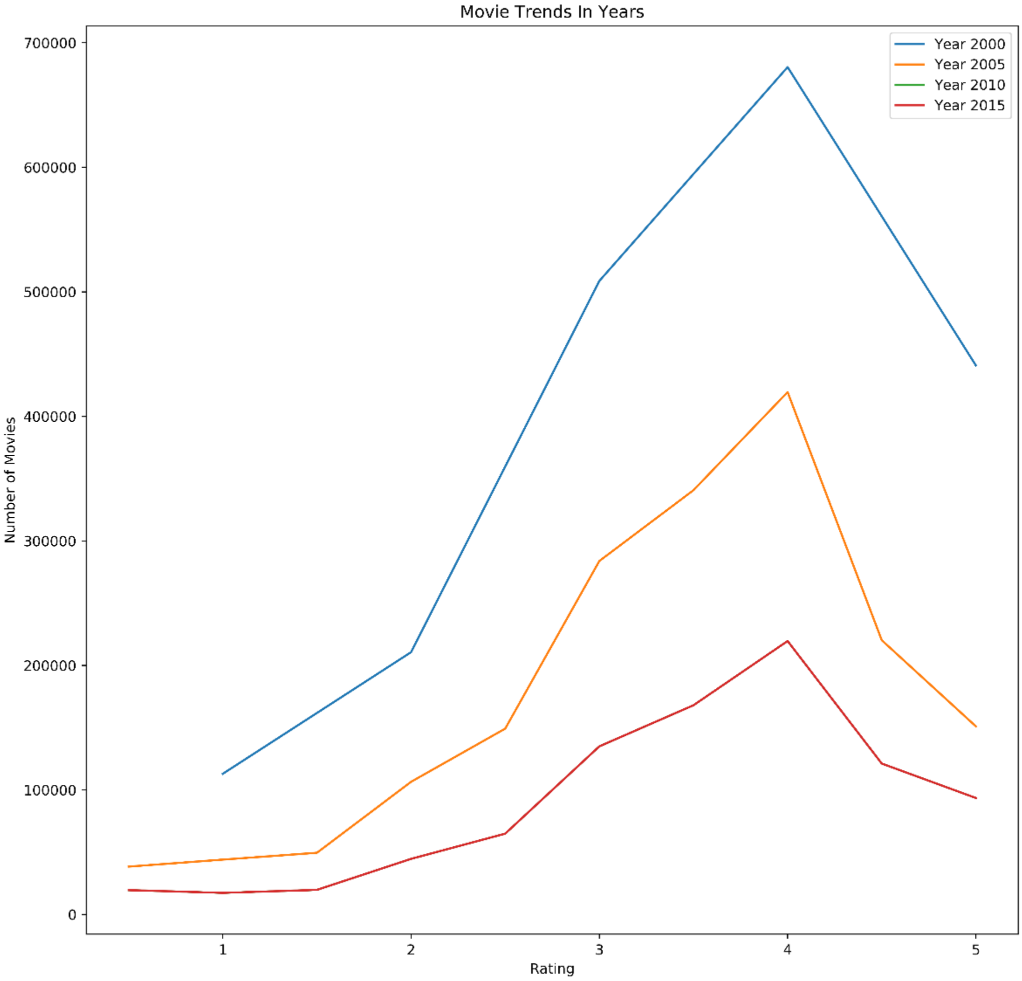


**Fig 4: Genres trend over the years**



**(Year 2010 and 2015 has overlapping result)**

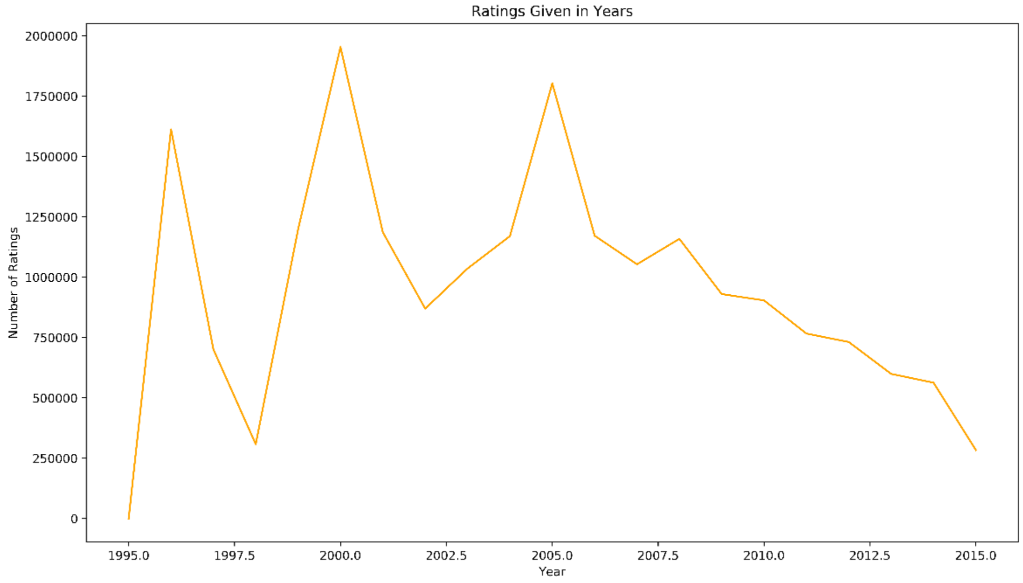


**Fig 6: Ratings given in years**

This graph depicts the number of ratings over the years. It is evident from the graph that the number of ratings increased from 1997 to 2000 and decreased from 2000 year with some fluctuations till 2015 and increased from year 2015.

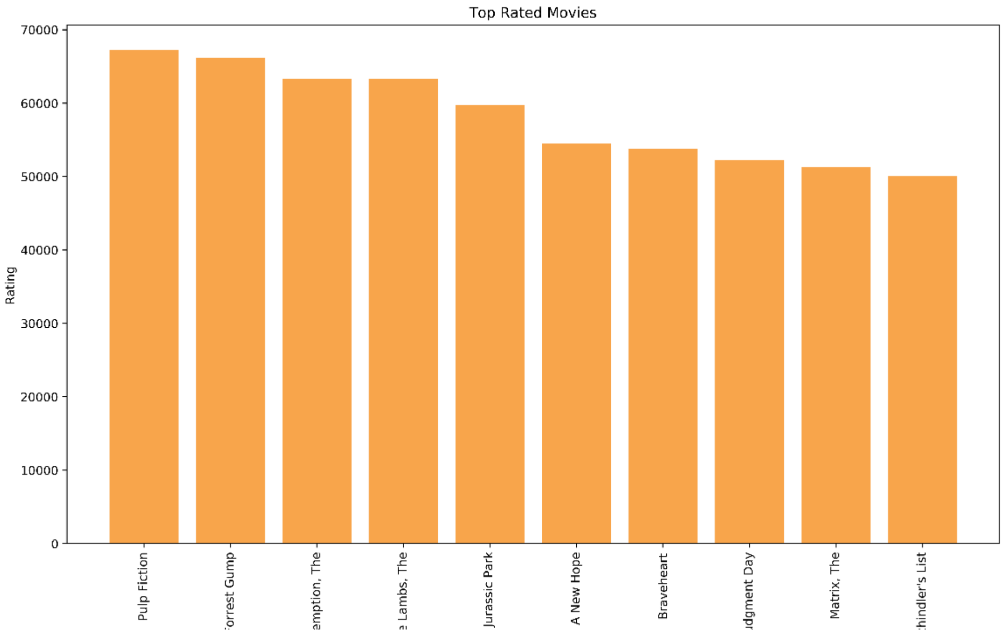


**Fig 7: Top rated movies**

From the graph we can see that the movie with movieid 356 is the top-rated movie with 318 movieid with almost the same percent of rating.

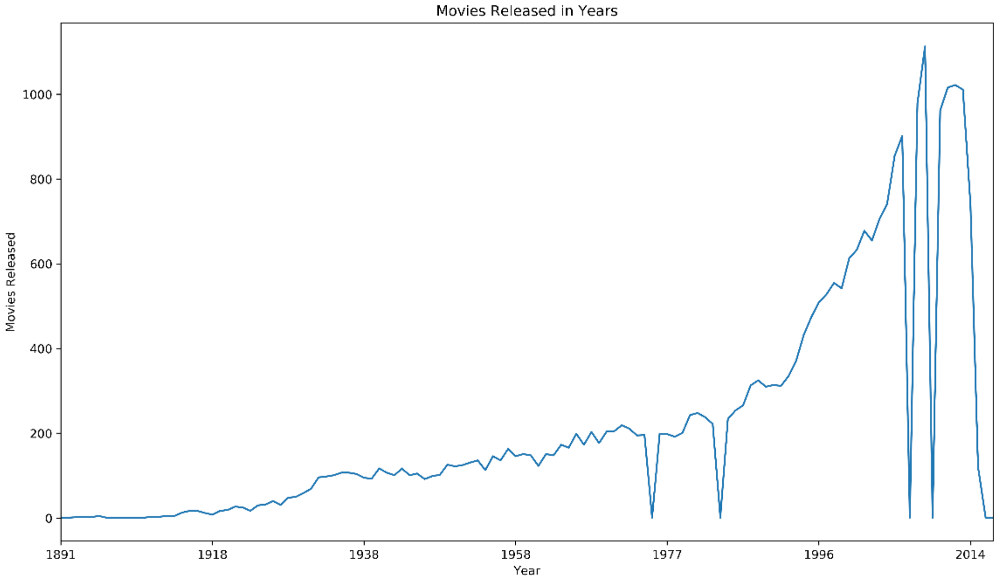


**Fig 8: Movies released over the years**

From the graph we can see that the number of movies released, gradually increased over the years 1902 to around 2007 and had a great downfall by the year 2007 and again saw a steep inclination in the span of a year or two. After the year 2015, no of movies released have been seeing a great descent till date. That's probably because of the rise of Netflix and Amazon prime which have been releasing one of the most top rated web-series.

# 4. References

All references should be listed here in alphabetical order as illustrated below:

Reeves, T.C., & Laffey, J.M. (1999). Design, assessment, and evaluation of a problem-based learning environment in undergraduate engineering. *Higher Education Research and Development Journal, 18*(2), 219-232.

Woud, H. K., & Stapersma, D. (2002). Design of Propulsion and Electric Power Generation Systems, London: IMarEST.

Use American Psychological Association (APA) referencing style both for in-text citation and the list of references at the end of the report. The University of Tasmania Library provides comprehensive information on referencing using the APA style of referencing.

http://utas.libguides.com/content.php?pid=27520&sid=199805

References intext should contain both author and date. The following example illustrates an in-text citation and reference list style according to the APA style.

Some research (Woud & Stapersma, 2002) suggests that …

… there is strong evidence of this in the literature (e.g., Reeves and Laffey, 1999).

Remember failure to appropriately acknowledge the ideas of others constitutes academic dishonesty (plagiarism), a matter considered by the University of Tasmania as a serious offence.

# APPENDIX A

This section is optional. Appendices are used for very detailed or lengthy sections of information. Information placed in an Appendix is usually supplementary to, or supportive of the discussion in the body of the report, but usually it is not critical to the main points being made in the report.

Appendices are referenced using letters, i.e. Appendix A, Appendix B, etc. Each appendix always starts on a new page.