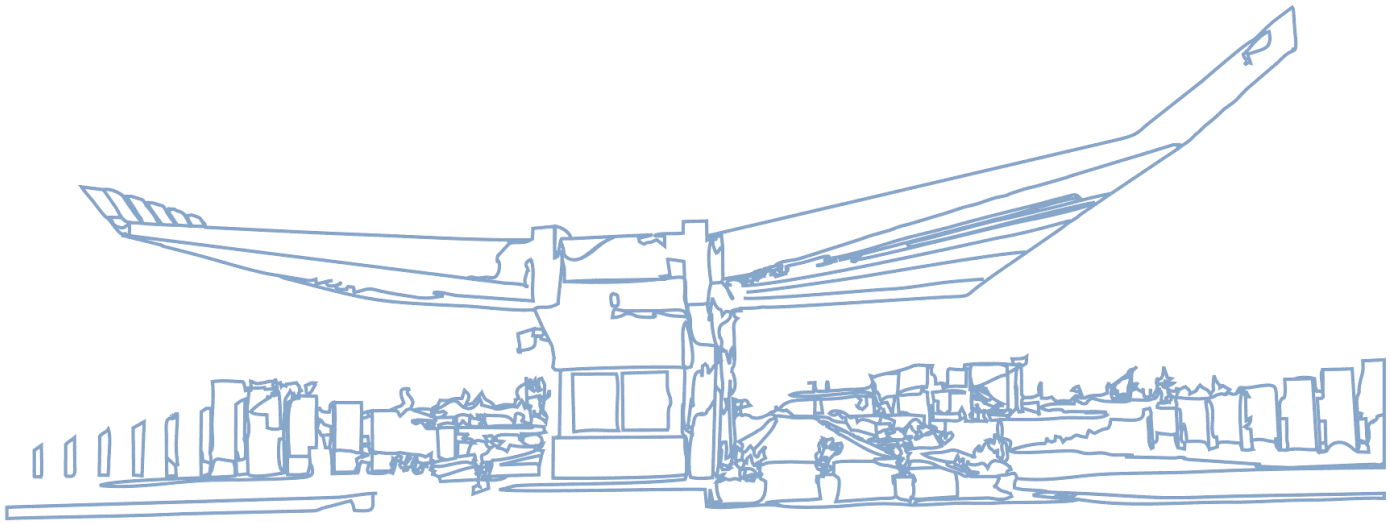
EPOKA UNIVERSITY | Computer Engineering Department

CEN 571 – Data Mining

Data Analysis with MapReduce and (Pig or Hive)



PREPARED:
Baftjar TABAKU

31.05.2020
Epoka University
Tirana, ALBANIA

ACCEPTED:
Prof.Dr. Arben Asllani

1. Title of the project

Analyze Movie Data

2. Brief description

In this project, I will analyze the IMDb movies extensive dataset, that contain information about 81,273 movies with attributes such as movie description, average rating, number of votes, genre, etc.

The rating dataset also includes 81,273 rating details from demographic perspective, and the names dataset includes 175,719 cast members with personal attributes such as birth details, death details, height, spouses, children, etc. title principals dataset includes 377,848 cast members roles in movies with attributes such as IMDb title id, IMDb name id, order of importance in the movie, role, and characters played.

The used dataset is updated in Kaggle.com 6 month ago.

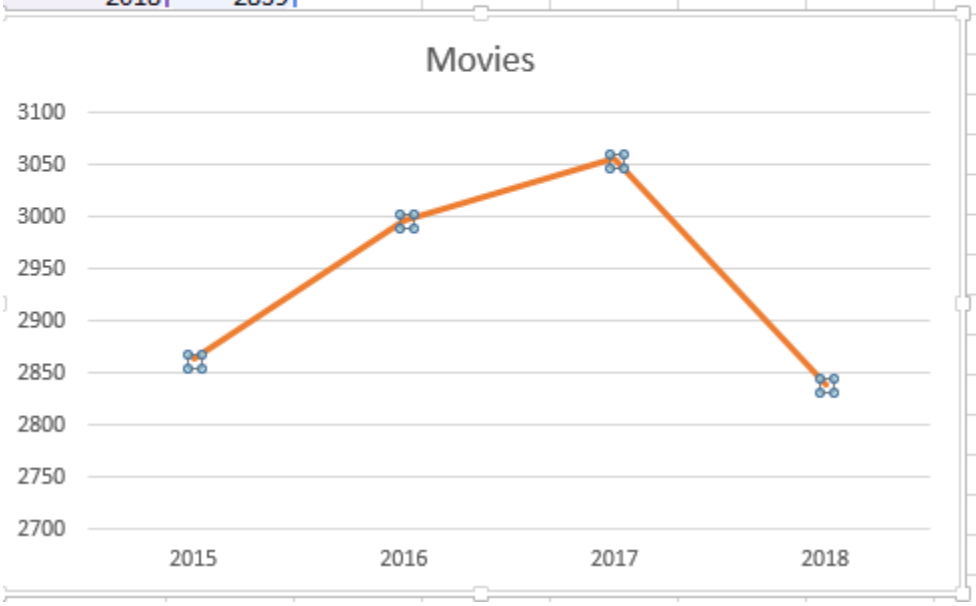
3. Project Goals

- Does the movies number get increased according to the years or decreases?
- What difference of average rating will be between movies in years (2015, 2016, 2017, and 2018) have the highest rating average, is this related with the movie quality (not display quality)?
- How much difference opinion/vote rating of males and females, a total opinion difference between them.

4. Project Findings

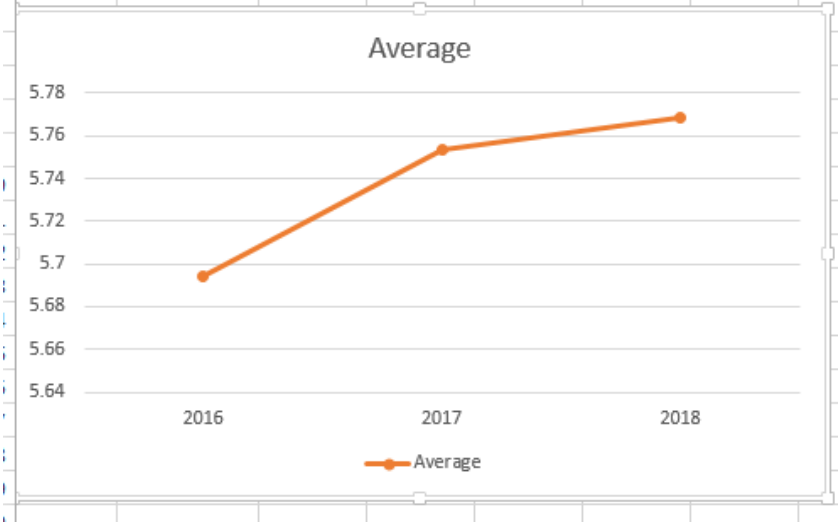
Analyzing the data the results were pretty interesting, the data sorted shows that with the passing of the time, years 2015,2016,2017,2018 we have an increased number of movies as the following graph shows;

Years	Movies
2015	2863
2016	2997
2017	3056
2018	2839



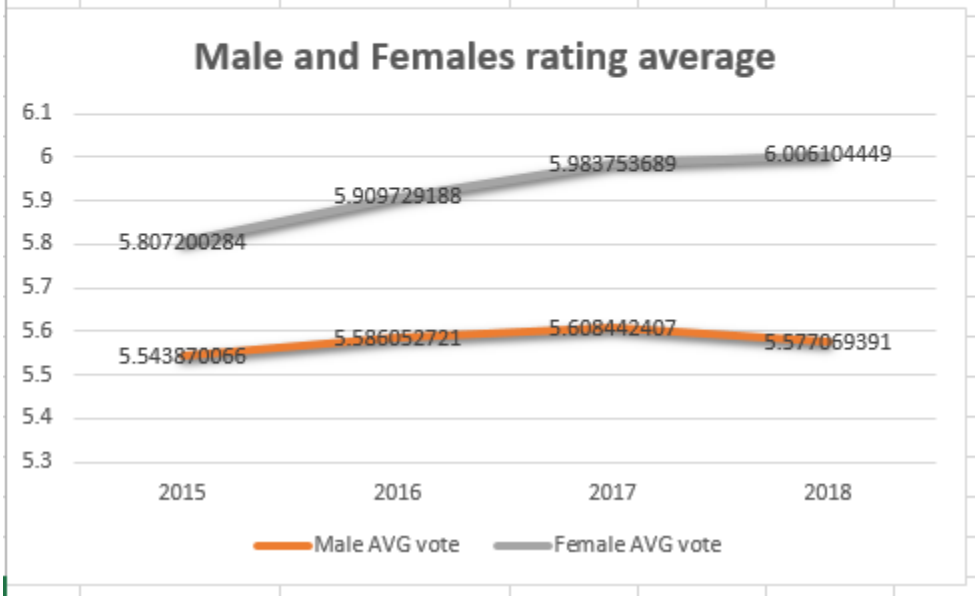
And in the 2018, we have a decreasing of the movies numbers, maybe the economic situation or many factors affect this part.

Year	Average
2016	5.694461129
2017	5.753632198
2018	5.768510034



Analyzing the data of average rating for movies between 2016 and 2018 it has an impressing result, the average averages as shown in the following graphic, there is a better increasing quality or movie components that lead the movies in a better performance as shown in the graphic.

Year	Male AVG vote	Female AVG vote
2015	5.543870066	5.807200284
2016	5.586052721	5.909729188
2017	5.608442407	5.983753689
2018	5.577069391	6.006104449



Another interesting finding is that in my analyze the females have the highest rating average, which is an interesting fact.

5. Conclusions

Datamining definitely Is the future of every kind of analyze and market data, her techniques like Hadoop or hive are very efficient for analyzing any kind of data, predicting in some cases, and many data analyzing in the future.

In the future, based on the current findings and analyzes, I would like to find the top genders that have the highest rating average votes, making more optimized visualizations and more efficient queries that analyses this data, and also giving answer to some questions like “does the gender depend on the length of the movie” or “does the movie length has is increased or decreased over the time?” and so on by performing analyzes with data mining and its tools like Hadoop, pig, hive and many more features.