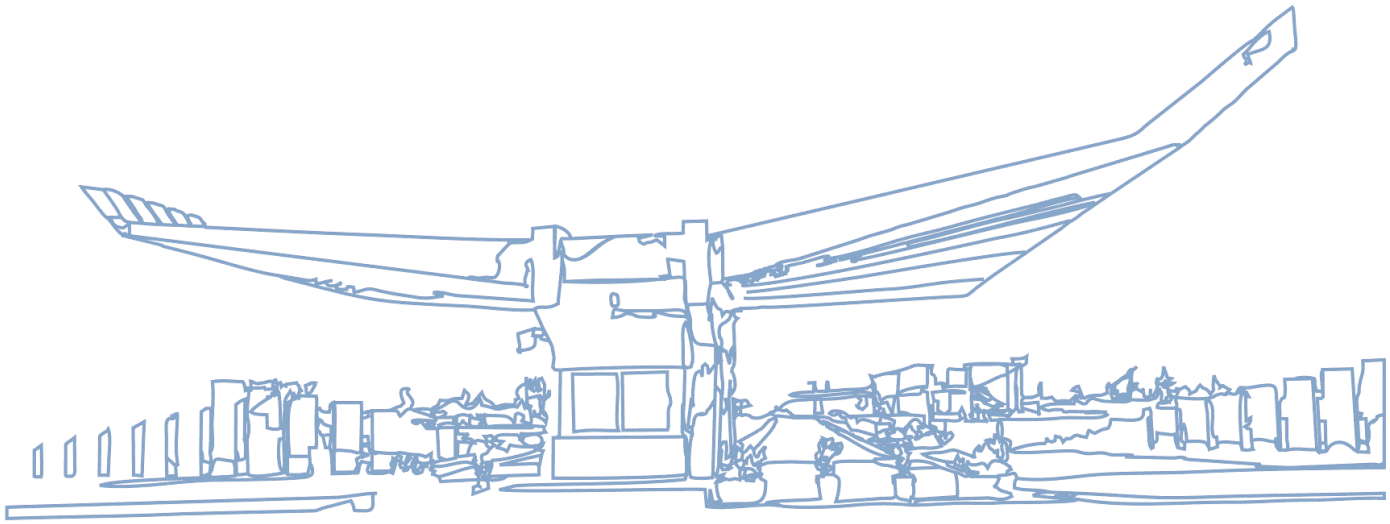
EPOKA UNIVERSITY | Computer Engineering Department

CEN 571 – Data Mining

Data Mining with Spark



PREPARED:
Baftjar TABAKU

21.06.2020
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.

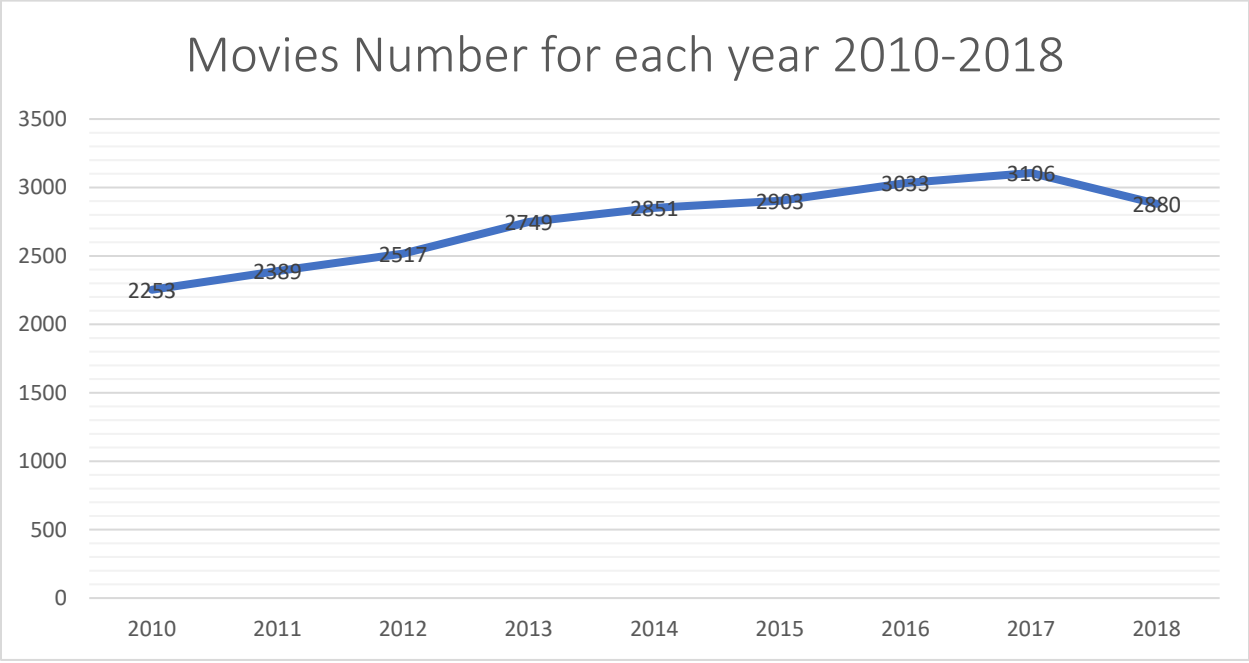
3. Project Goals

- Does the movies number get increased according to the years or decreases?
- Comparing the ratting average over the years 2010-2018
- Males and females' average votes over the years 2010-2018
- Making an analyze related with average ratting and movies number over the years 2010 -2018

4. Project Findings

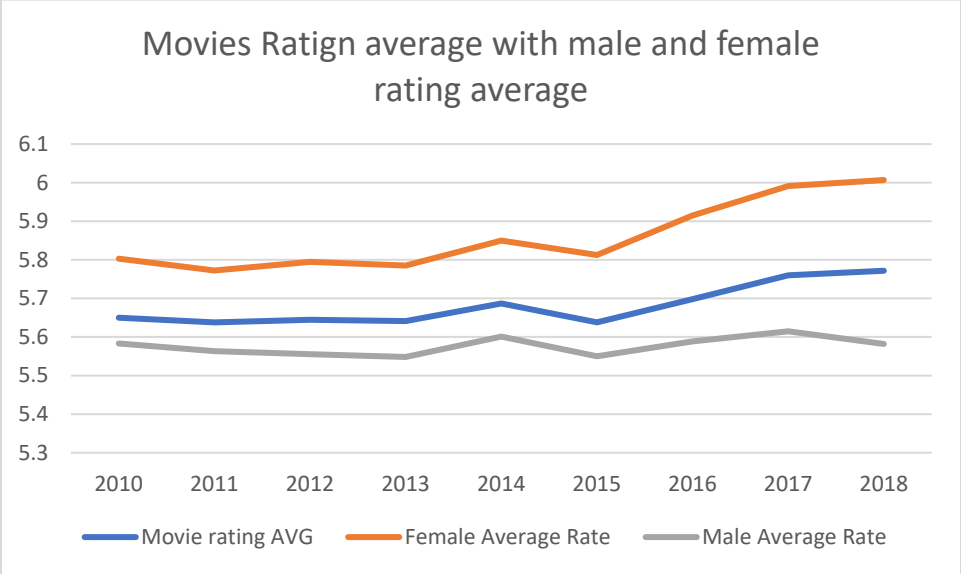
Trying to see the number of movies over the years 2010 - 2018, how it goes,

Year	Movies Numb
2010	2253
2011	2389
2012	2517
2013	2749
2014	2851
2015	2903
2016	3033
2017	3106
2018	2880

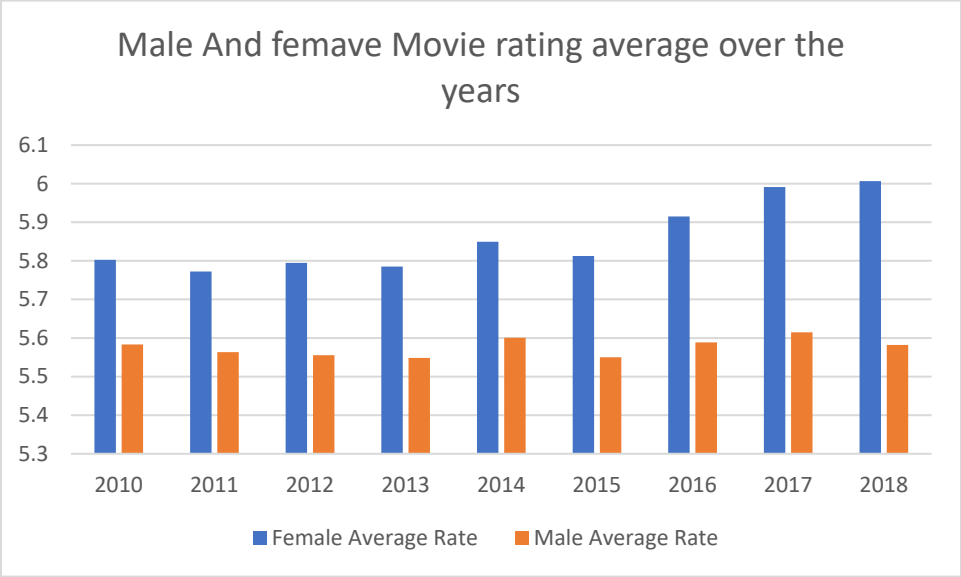


Analyzing the data of average rating for movies between 2010 and 2018 it has an impressing result, the average averages as shown in the following graphic, there is a better increasing quality or movie components that lead the movies in a better performance as shown in the graphic.

Year	Movie rating AVG	Female Average Rate	Male Average Rate
2010	5.650022191	5.802798757	5.583222367
2011	5.637840095	5.772331521	5.5631645
2012	5.644775525	5.794874854	5.555462851
2013	5.641360492	5.785334791	5.548235721
2014	5.687022097	5.849684217	5.600876885
2015	5.637857389	5.812512931	5.550017223
2016	5.698186615	5.915031384	5.588789978
2017	5.759787508	5.99126652	5.614745652
2018	5.771666663	6.006782612	5.582118057



Year	Female Average Rate	Male Average Rate
2010	5.802798757	5.583222367
2011	5.772331521	5.5631645
2012	5.794874854	5.555462851
2013	5.785334791	5.548235721
2014	5.849684217	5.600876885
2015	5.812512931	5.550017223
2016	5.915031384	5.588789978
2017	5.99126652	5.614745652
2018	6.006782612	5.582118057



Another interesting finding is that on 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 combined with Spark and spark applications 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 spark, that was perfect by combining it with Hadoop and it gives better and more precise results than pig ,hive and many more tools that I’ve used.