Exploratory Analysis of IMDb Movies

Isabella Calfa Vieira Costa[1], *Taian Fonseca Feitosa*[2]
[1] Universidade Federal da Bahia, Salvador, Bahia, Brazil
[2] Universidade Federal da Bahia, Salvador, Bahia, Brazil

Abbreviated abstract: Many movies that break box office records are not remembered after some time, while others may not be a box office success, but are remembered for generations. In this paper were analyzed 85,855 movies available at IMDb, Internet Movies Database, to rate the best directors, producers, country, genres and others variables and understand what makes a success movie, based on movies average score on the site.

bella.calfa@gmail.com – 1, taianmeca@gmail.com – 2

3rd Conference on Statistics and Data Science
Salvador, Brazil (online)
October 28-30, 2021

Problem and Data

Problem: In the beginning of Movies Industry, directors and actors used to make movies to impress and enchant people. Years passed and this goal was turned into making money. In the USA this industry moved more than US$ 8bi/month. The question is: new movies are better than old ones? What makes a good movie?

Solution: Analyze movies reported at IMDb and their variables and scores to find out which ones impacts at the final evaluation.

Used Dataset: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset.

IMDb Dataset: Contains 85,855 movies published worldwide between 1890 and 2020.
Variables used: IMDb Title Id, Country, Production Company, Language, Genre, Date Published, Year, Month, Avg Vote, Duration, Writer and Director.

Tools: Python (3.8.5) and Packages: Matplotlib, Numpy, Pandas and Seaborn; Boxplot; Bars Graph; Line Graph; and Dispersion Graph.

UFBA Universidade Federal da Bahia

bella.calfa@gmail.com – 1, taianmeca@gmail.com – 2

3rd Conference on Statistics and Data Science
Salvador, Brazil (online)
October 28-30, 2021

Methods

Dataset Adjust:

Empty Country → Filled with modal by Production Company

Empty Language → Filled with modal by Country

Empty Director → Filled with Writer

More than 1 Country → Used first appearance

More than 1 Language → Used first appearance

More than 1 Genre → Used first appearance

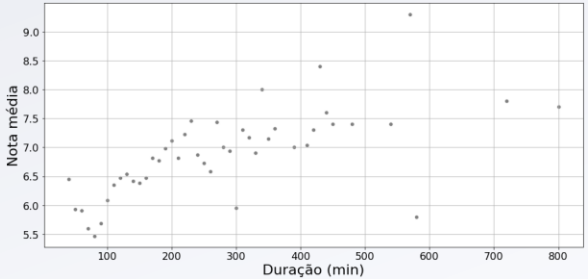
Collected Information for Analysis:

- Modal;
- Average;
- Quartiles.

General Information:

- Max. Score: 9.9;
- Min. Score: 1.0;
- 25% scores lower than 5.2;
- 75% scores lower than 6.8;
- Average score: 5.9.

Duration x Score:



Example of analysis.

bella.calfa@gmail.com – 1, taianmeca@gmail.com – 2

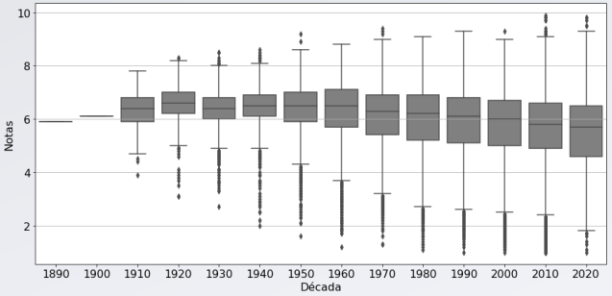
UFBA Universidade Federal da Bahia

3rd Conference on Statistics and Data Science
Salvador, Brazil (online)
October 28-30, 2021

Results and Conclusions

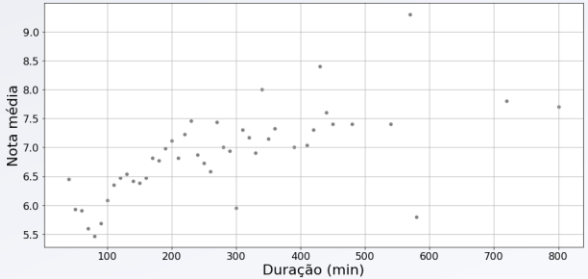
- **Country**: Produce more movies doesn't mean better ones (USA produces more movies, but the country with highest score is Azerbaijan);
- **Genre**: Animation, Music, Biography, Noir, War and Drama (in this order) receive highest rate. The opposite happens with Adult, Horror, Sci-Fi, Thriller and Fantasy;
- **Director**: There is no relation between how many movies the director works and the score of them.

Decade x Score:



Old movies scores better than new ones (median).

Duration x Score:



No strong relation.

bella.calfa@gmail.com – 1, taianmeca@gmail.com – 2

3rd Conference on Statistics and Data Science
Salvador, Brazil (online)
October 28-30, 2021