What can we learn from user-movie ratings?

Niloofar Zarif, Lucie Polakova, Deepansha Chhabra

niloofar.zarf@gmail.com, lucie.polakova.18@ucl.ac.uk, deepanshachhabra21@gmail.com

Abstract

Recommender systems have been developed for years for different purposes including movie recommendation. Recommending movies to users is a billion-dollar industry where even small increases in recommendation accuracy and performance efficiency can lead to huge business benefits. Some may believe using machine learning and neural networks is the ultimate solution to building good recommender systems but in practice we have seen that belief being untrue. Recommender experts have recently realized having a good body of knowledge about the data and user behaviour is essential to building better solutions. On this project we are going to explore and extract knowledge from the publicly available movie datasets. We hope the results of this work can provide an insight which can be used for designing movie recommender systems.

Introduction

In the past decade, large companies such as Netflix, Amazon, Apple and Disney have invested a lot of resources and effort into their streaming services [3]. In 2020 the number of subscriptions to online movie streaming services surpassed 1.1 billion and consumers spent around 80$ billion on streaming content while this number was only 12$ billion for movie theatres [1]. These numbers and records show how important it is to design movie recommendation systems that serve the users as effectively as possible. Designing such systems without knowing user preferences would be a shot in the dark. Different users certainly have different tastes and preferences but there will be patterns in the data that can point towards the most common preference criteria as well as introduce the temporal or regional user preferences.

Movie streaming companies are well aware of the complicated nature of the recommendation task. Therefore, unlike many other areas, large datasets have been made publicly available by companies that own them. These datasets are the results of companies logging interactions within their systems. For example, Netflix and Amazon Prime have been very open in releasing data and updating it every year. This gives the wider public opportunity to learn about how users act on these platforms.

As a prominent movie streaming service, Netflix has more than 209 million users as of July 2021 [2]. In 2006 Netflix started a competition called the Netflix Prize. The competition was open to anyone around the globe and the purpose was for the participants to design the best collaborative filtering algorithm to predict user ratings for movies. Back then collaborative filtering algorithms were the dominant solution for recommender systems. These days collaborative filtering has been dragged aside by deep learning models. Yet what the Netflix prize competition meant for the public was that a large amount of data was published by Netflix. The culture of publishing data has continued and to this day we still can get our hands on large datasets of user-movie interactions from Netflix.

For this project, we will use Netflix data to extract knowledge and discover patterns that can be found within the dataset and further used by the target audience. Assuming the rating users give to each movie is the dependent variable, we are going to see how the other independent variables including genre, duration and country of origin affect the rating. We will seek meaningful patterns in the large body of data we have and will try to use adequate visualization techniques to present them in a clear dashboard.

We hope for the results of this work to be used by people trying to design and build recommender systems and make them able to come up with more effective solutions. Another group of users of this project are the managers of smaller and independent movie theatres who are trying to know more about users' tastes to use that knowledge while making business decisions.

The topic was chosen based on our common interest in recommendation algorithms and the motivation to enhance our understanding of how movie data can be analysed.

Related Work

Before deep learning models took over, for most of the recommendation systems, researchers were keen to know more about the characteristics of their dataset when making recommendations. This made them able to design and build recommenders that could give better recommendations based on the nature of the problem and the patterns that could be seen within the dataset. Back then we still did not have enough computing power to use deep learning for movie suggestions, methods like Collaborative filtering was being used [7][8]. At the same time Content Based filtering was a popular method for building recommendations too [4][9][10]. When such methods were being used for movie recommendation, knowing the data and the patterns that existed was important. Therefore, thorough analysis of data and preferably visualizing it was crucial [11].

As of the mid-2010s we had enough computing capabilities to have deep learning and neural networks in action. Deep learning models for image classification and object detection were being developed rapidly. Yet making practical deep learning recommender models was a bit more complicated. But the time came and by late 2010s production-level recommender models that were using deep learning came into existence [12][13]. Nowadays, large deep learning recommender models are being used along with older methods such as collaborative filtering to produce the best result with high-performance to meet the application level latency agreements [14][15].

For a while, people assumed having enough computing resources, we could give them a large body of data during training and use the model efficiently for making recommendations later by doing inference on the trained model. This approach proved to fail very soon. The datasets grew exponentially in size and complexity and it was very important to produce results with high accuracy. Even a 0.01% increase in error can affect the user experience [15]. As a result, the trained models became excessively large such that they could not fit in memory anymore. Those models were not practical to be used for online use-cases anymore. This is why we see smaller neural networks working along with filtering methods adventing these days [14][15]. Now that filtering is again a player in the game, analysing data becomes important once more. So once again knowing more about data becomes important. The knowledge extracted from data can be used in the filtering stages of modern recommender models.

Exploring and visualizing movie datasets is not limited to recommender models though. Business analysis or making investment decisions can be another motivation. Computer scientists and visualization enthusiasts may also do it out of curiosity. Some have already done exploratory visualizations on movie datasets that are publicly available [16]. Some of these exploratory visualization projects have used the publicly available Netflix dataset [17] and some use other datasets such as TMDB [18].

With business purposes in mind for visualization, there are projects and papers focusing on profitability and how it relates to different film properties [6]. There are a few solutions where data have been analyzed with the motive of making a contribution towards building a global brand and making a sustainable long term plan in terms of the production of movies. The relationships between different parameters of the dataset have been shown by scatter plots, bar plots, kernel density estimate plots, histograms box plots, linear model plots, heatmaps etc., using programming tools like Python Libraries and simple and efficient tools like Tableau [5]. Based on a few currently available analytical solutions, it has been observed that a particular genre movie may give the highest return per investment but is rated low and hence, in turn, does not bring in high revenue. These analyses have proven to be effective tools assisting the movie production teams to get an insight into the audience's interests. Thus, our idea of identifying the correlations between user rating and other movie properties aims to fill in the gaps in the available research and provide an alternative tool when the aim is not pure financial profit.

Data and Task Abstraction

Domain

As mentioned, big data analysis is gaining more popularity in the media industry as well and some of the major providers actively seek public involvement in optimising the existing algorithms on working with this data. This project, therefore, aims to provide a clear overview of the different patterns in the publicly available data by learning more about user preferences. The target users are independent computer scientists and engineers that do not have the time and resources to conduct robust customer research and big data analysis, but want to contribute to the optimization of the recommendation algorithms and design new and better solutions. Laic movie enthusiasts, such as indie movie theatre managers, might use these visualisations to understand the movies with high potential despite not having a large budget and being heavily marketed.

Task

The main task of the visualisation is to discover patterns that can be found within the datasets and clearly show the different relationships between user rating and other film properties (genre, country of origin, year, duration, original language etc.). As a result, the properties that have the highest effect on user rating will be identified.

Data

Data for this project are sourced from Kaggle and the final dataset will be obtained by linking multiple related datasets to gain a wide variety of properties for each attribute. In total, we will be working with a dataset of 1,100 items (movies) and the average rating will be obtained from a dataset containing 26 million items (ratings). The attributes we intend to use in our analysis are both categorical (year, country of origin, genre, original language) and ordered (rating, runtime, number of votes). Specifically, the levels of categorical attributes are as follows:

- Original language: 9
- Genre: 19
- View rating: 9
- Release month: 12
- More than one language spoken: 2

The ranges for ordered attributes are:

- Budget: 0-400,000,000
- Revenue: 0-1,900,00,000
- Year: 1930-2020
- Rating: 1-5
- Runtime: 70-210
- Number of votes: 1-5,639
- View rating: 1-9
- Popularity: 0-60
- Awards received: 0-130
- Awards nominated for: 0-120

Some of the data will be aggregated for analysis. For example budget and revenue will be divided into 15-20 clusters so it can be treated as categorical data.

Solution

The final solution of this project will consist of an interactive dashboard with overview of the different movie properties and multiple static visualisations, each addressing a different property and how it relates to user rating.

We would be able to extract useful information with the help of these visualizations, such as which genre is rated the highest, trend of changes in ratings based on different genres, whether the budget is directly or inversely related to the ratings given to the particular genre movies, the trend of ratings among different genres based on release dates, whether a particular genre was famous over a short or long period of time, how revenue is affected by the movie ratings, how language affects the popularity of different genres in different areas and how it affect the ratings.

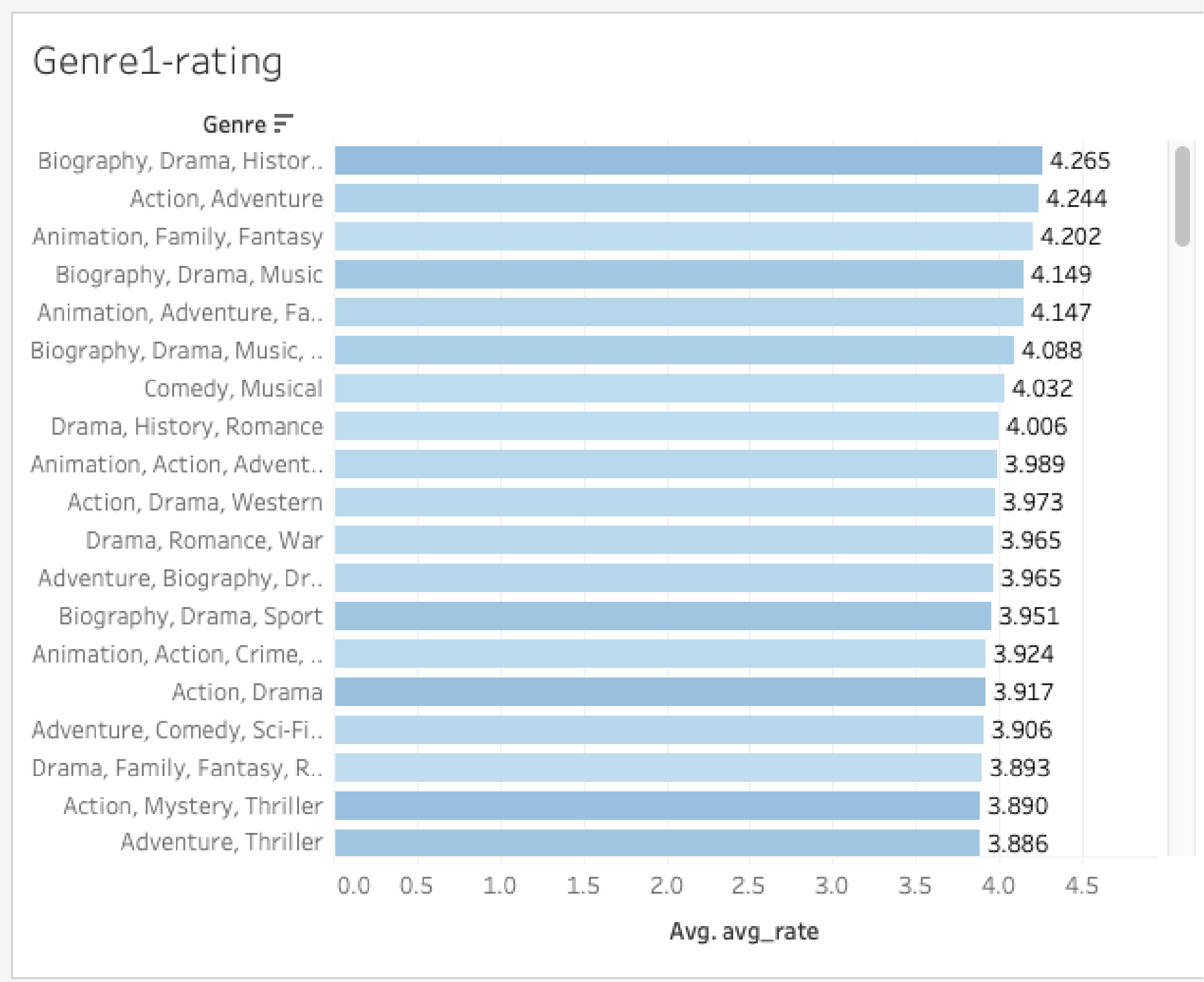
Hitherto, the goal is to allow to unravel valuable aspects and information from the visualizations and hence contribute to the movie industry in terms of building better models and making better decisions.

Potential idioms and design choices include:

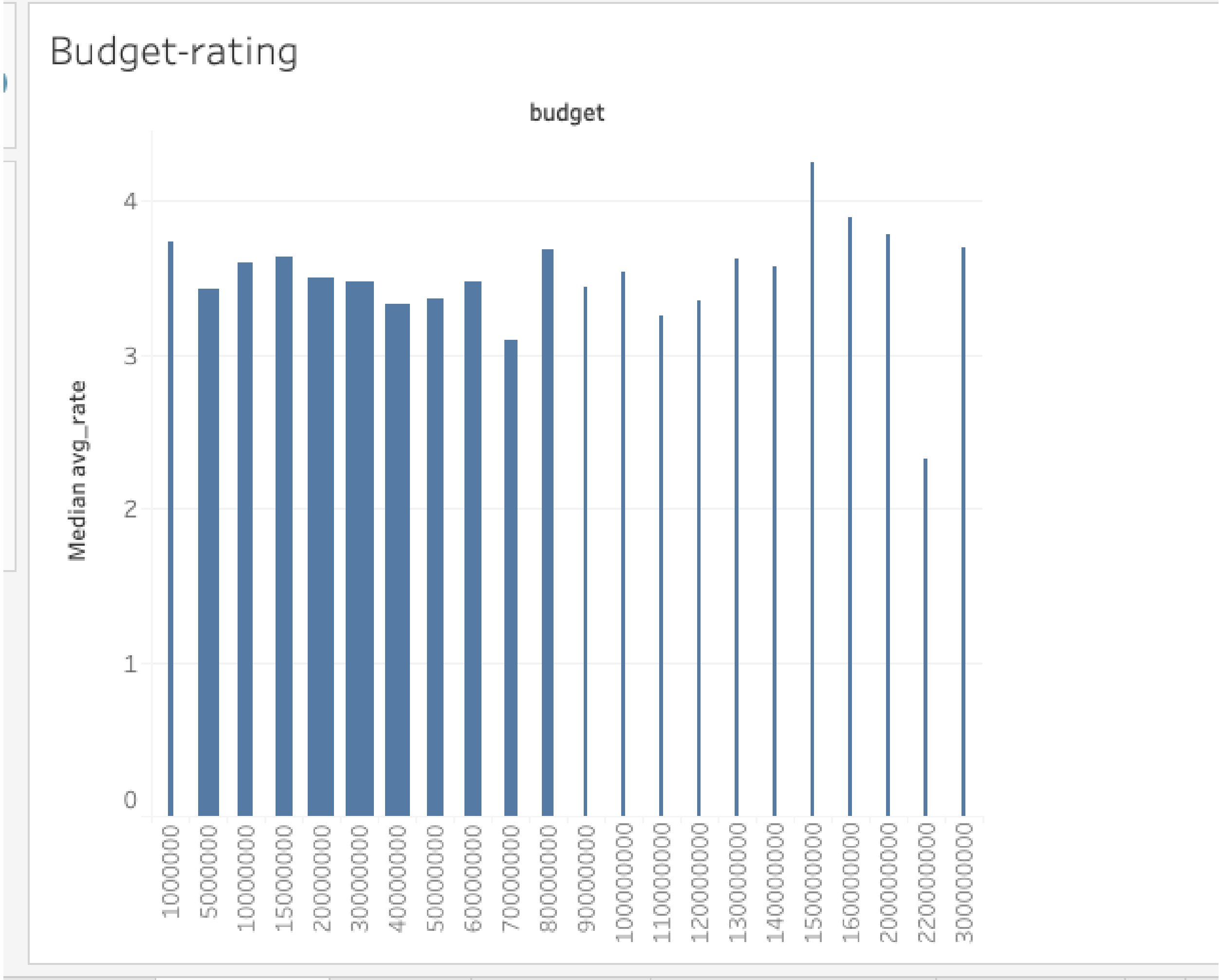
- Bar graph - to show correlation coefficients between user rating and different properties
- Heat map - clustered rating and the different properties to show an overview of the relationships

- Line graph - to show change of rating over time for different genres/other properties and identify the latest trends
- Box/jitter plot - to show the distribution of user ratings based on years/genres/etc.

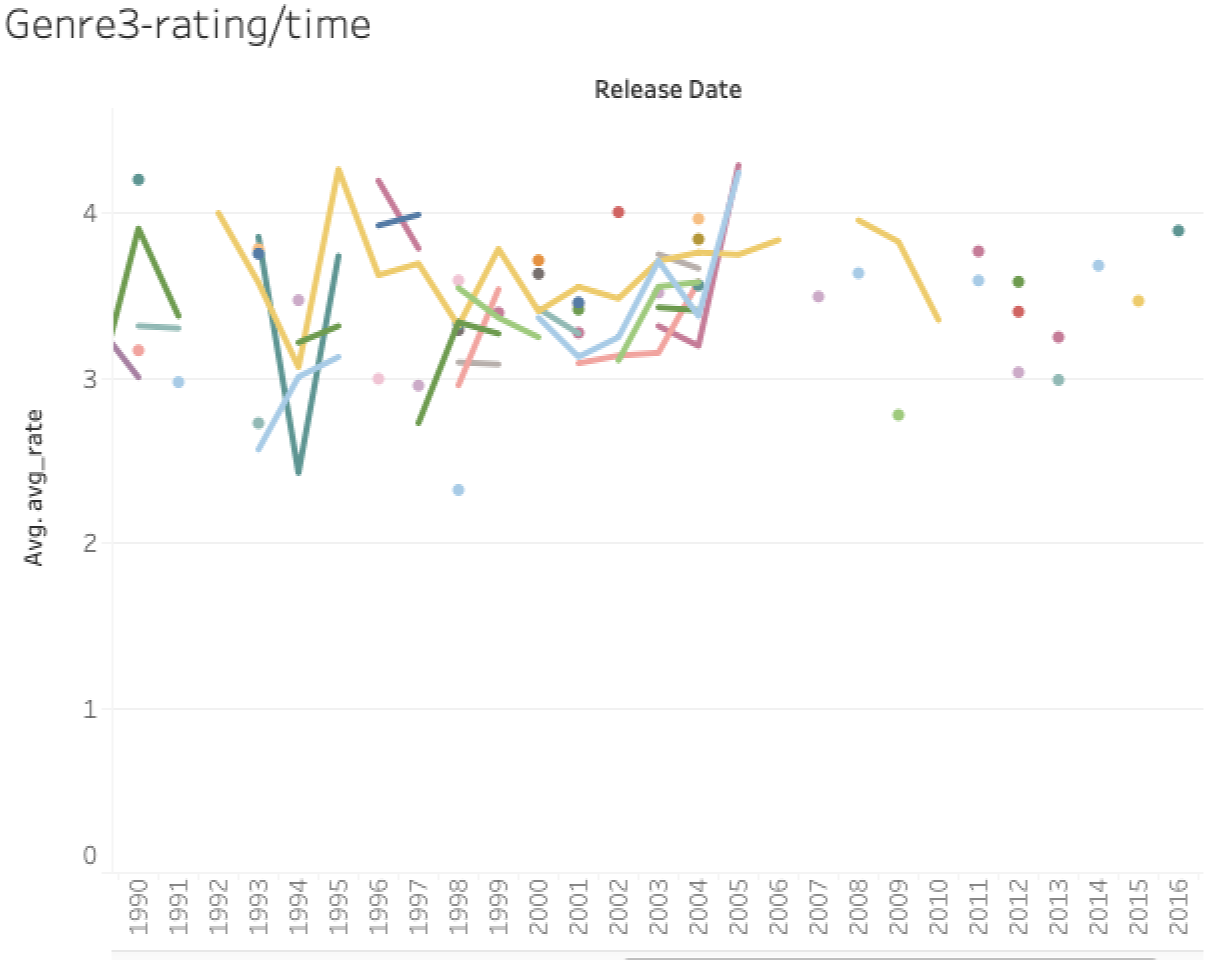
Below screenshots represent the work in progress on the individual idioms that will be then combined into the dashboard.



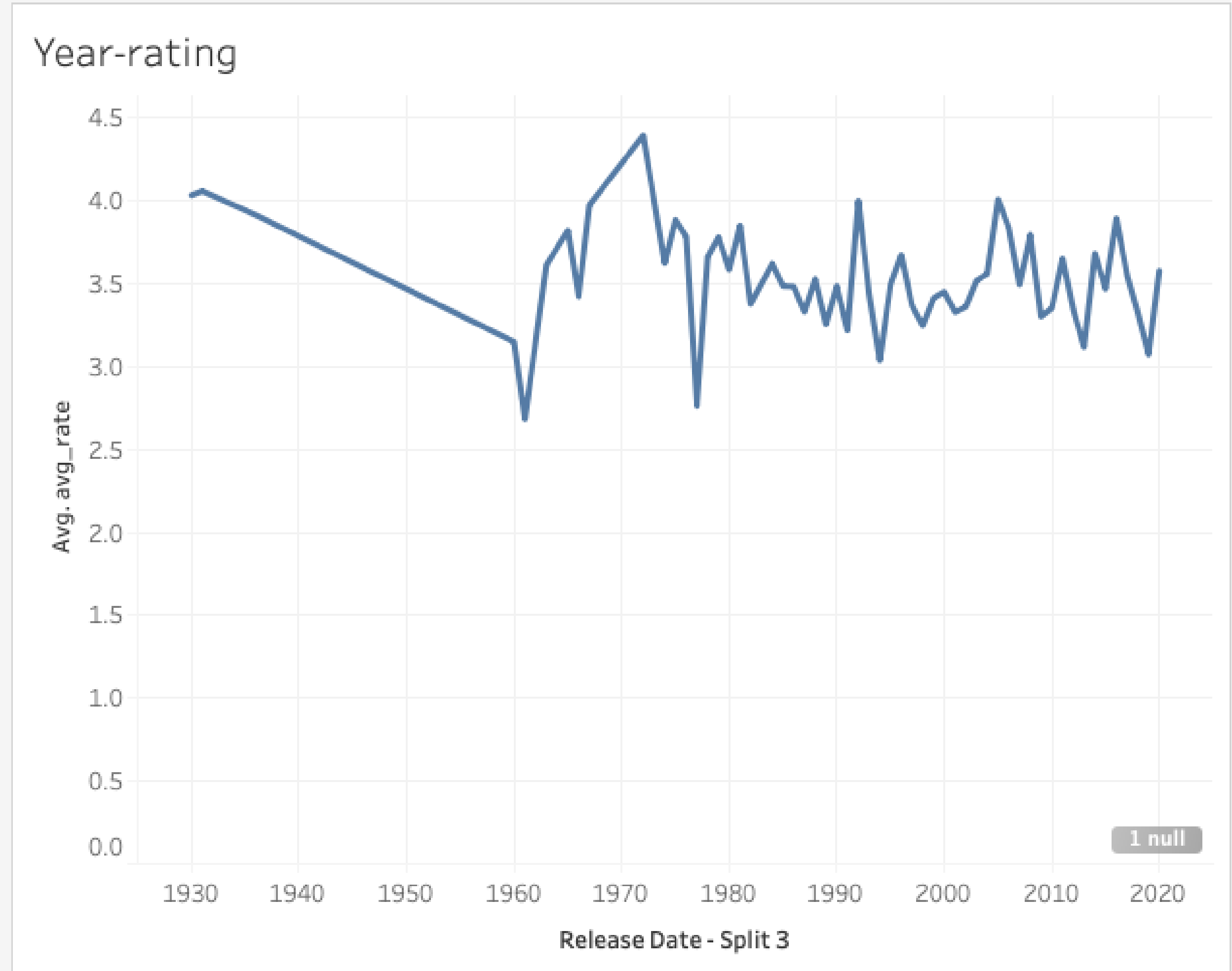
Screenshot 1: Average ratings for different genre combinations.



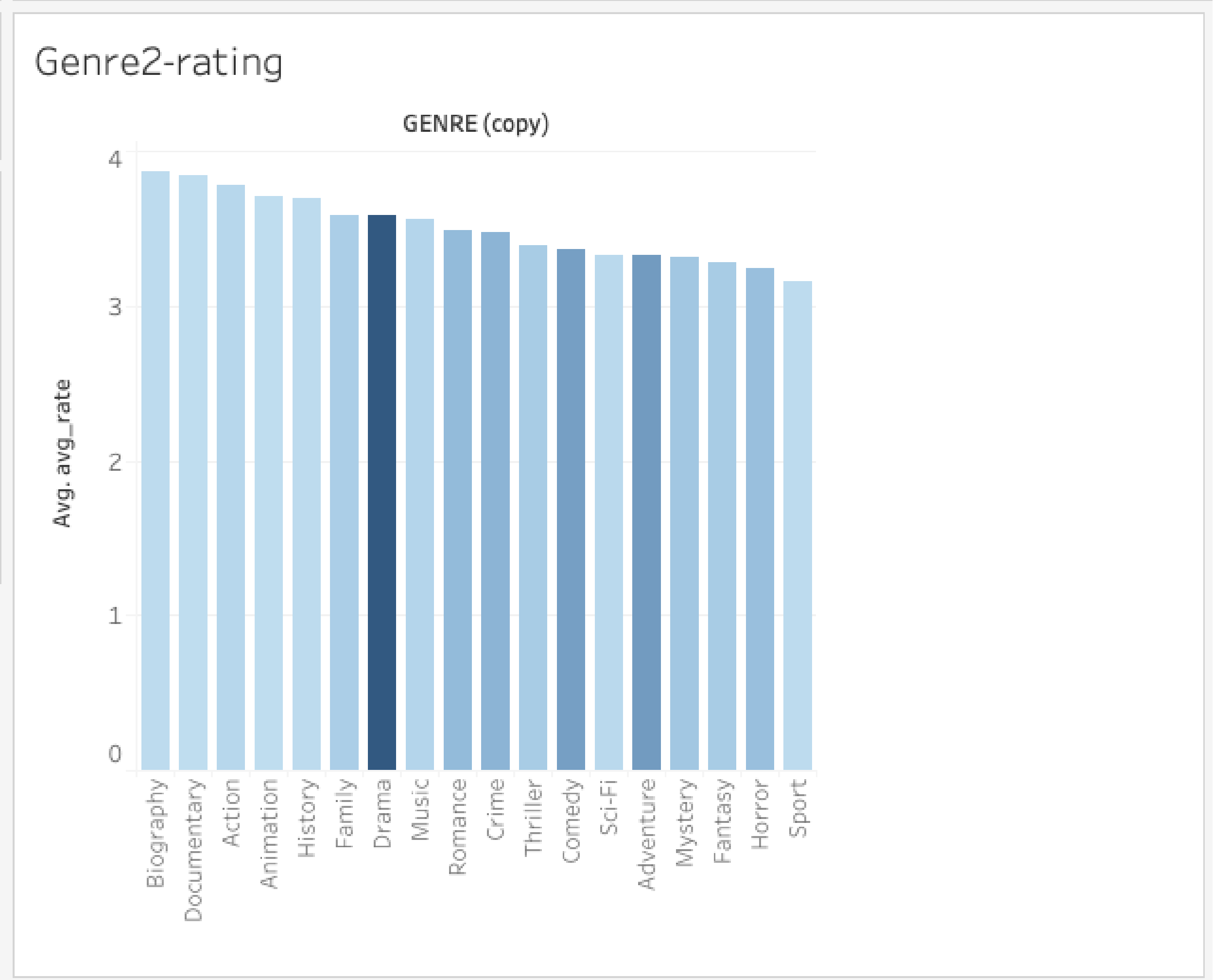
Screenshot 2: Median average rate for different budget investments.



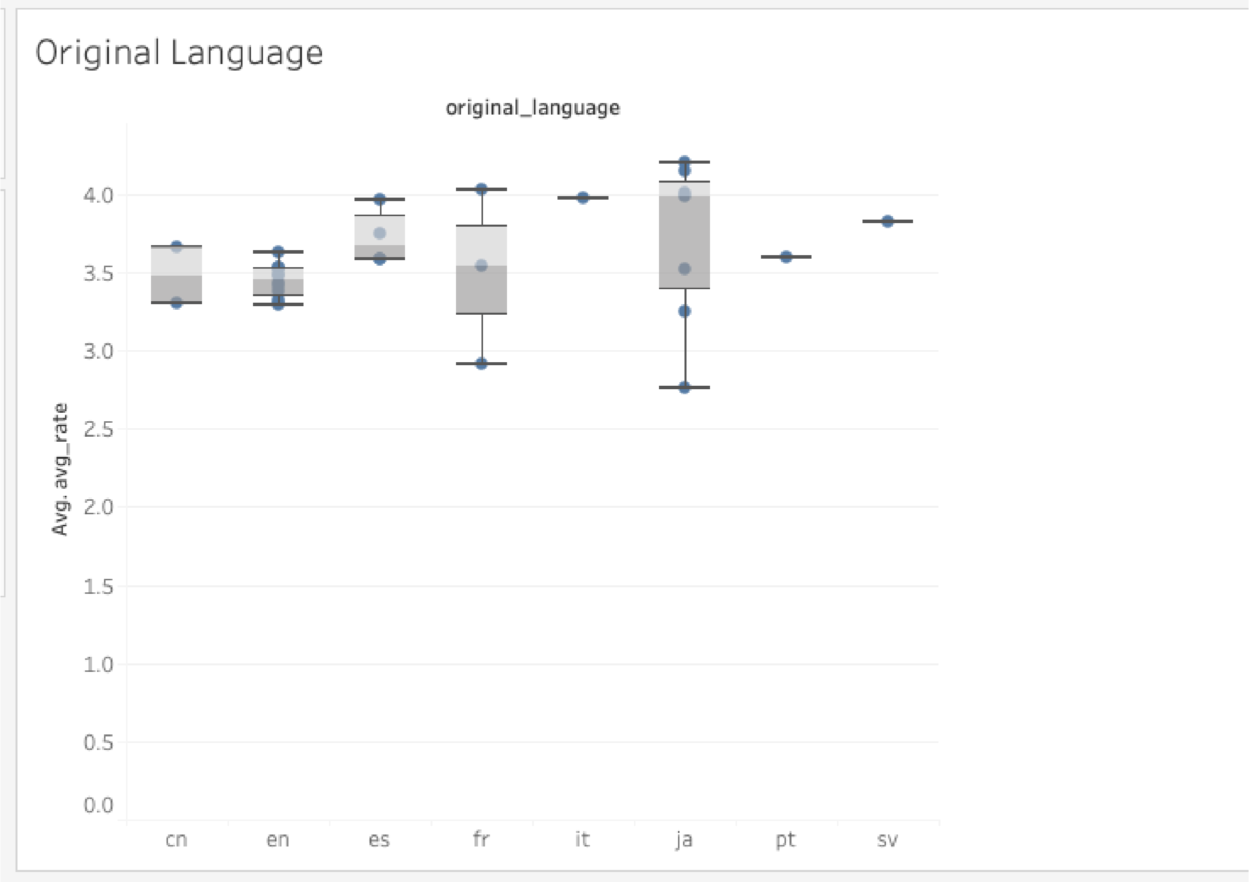
Screenshot 3: Average ratings over the years for the different genres which are differentiated using color channel.



Screenshot 4: Average rating for genre over a period of time/for various years.



Screenshot 5 : Average ratings for different genres.



Screenshot 6: Average ratings for different genres based on the different languages.

Additional channels will be used to enhance the user experience and interaction with the visualisations. Given that a lot of our data is categorical, color and its properties (e.g. hue, luminance) might be used to distinguish between the different categories.

Potential features also include the ability to manipulate view of the visualisation such as allowing a temporary selection of items and properties that will be displayed and included in the dashboard.

Implementation

Datasets in a CSV format will be loaded into a pandas data frame in Python and will be cleaned and linked together based on the movie names. The combined datasets will be then further processed in Tableau Prep to ensure they are ready for analysis and visualisation. All visualisations will be designed using Tableau Desktop.

Scenarios

- Indie movie theatre manager is looking for what movies he should stream during the upcoming season in order to please their demanding and critical clientele. By interacting with our visualisation dashboard, he can clearly see the genre that has been trending in the past couple of years, which countries might be generally producing good movies and what runtime might most likely lead to satisfied customers.

- Independent computer engineer decided to work on a new machine learning recommendation system in order to contribute to an open call of a movie streaming company. The large amount of data available means that he has enough input, however, knowing what properties he should prioritize would remarkably shorten the development period. By interacting with our solution and selecting the different properties to display, he can clearly see the most significant starting points for his project.

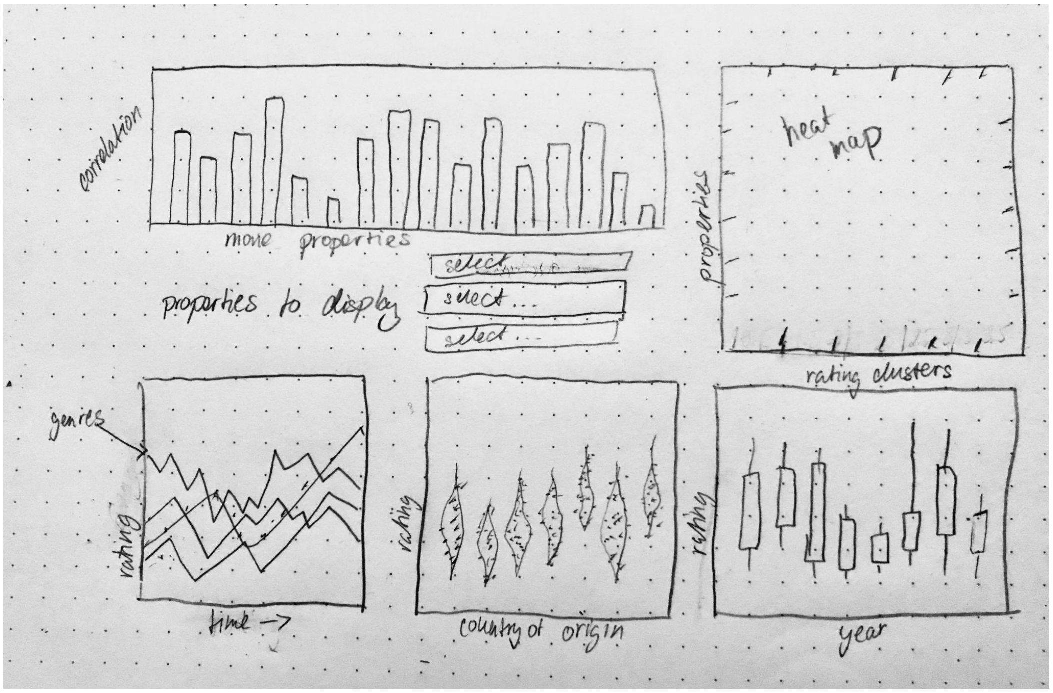


Figure 1: Low-fidelity prototype of final dashboard

Milestones

October 13: Preliminary Work

- (4 hours, All) Prepare project pitches, exploration and discussion on data source
- (5 hours, All) Meetings to discuss project direction

October 21: Project proposal

- (5 hours, all) Our first step is choosing the best dataset we can find and clean it such that it will be ready for our further use
- (15 hours, All) Develop and edit project proposal

November 10: - Preliminary data exploration, data preparation and vis finalisation

- (15 hours, Niloo) Cleaning data and linking relevant datasets to create a final dataset for analysis.
- (15 hours, Niloo) aggregation of rating records based on movie name
- (15 hours, Lucie and Deepansha) Further exploration of vis options, finalising the visualisations used, sketching/designing low-fidelity prototypes
- (2 hours All) Group discussion of prototypes and design finalisation

November 16: Update - Create first version of vis software using finalized idioms

- (4 hours, Lucie) Knowledge extraction phase, preparing data in Tableau Prep, preliminary statistical analysis, seeking meaningful patterns and useful knowledge
- (8 hours, Niloo) Finalise related works section for the final paper
- (10 hours, Deepansha and Lucie) Begin data analysis and correlation overview in Tableau.
- (10 hours, All) Write update report

November 24: Visualisation design

- (15 hours, Deepansha) Design 2 idioms on movie properties
- (20 hours, Lucie) Design 4 idioms on movie properties

December 3: Finalize vis dashboard

- (20 hours, All) Final refinements and bug fixes
- (20 hours, All) Developing other features of the vis and final dashboard (multiple views together with chosen idioms, interactivity to support selection/filtering)

December 15: Final presentation

- (10 hours, Deepansha) Slides
- (3 hours, All) Rehearsal, demo

December 17: Final paper

- (20 hours, All) Documenting and reporting the results, finalisation of the paper, formatting and citation check

Discussion, Future Work, and Conclusion

TBD

Bibliography

[1] Mendelson, Scott. “Streaming Subscriptions Top One Billion: A New Normal, Or A Temporary Disruption?” *Forbes*, 2021. *Forbes*, https://www.forbes.com/sites/scottmendelson/2021/03/18/streaming-tops-one-billion-subscribers-a-new-normal-or-a-temporary-disruption/?sh=37e626c8465a. Accessed 18 10 2021.

[2] Kats, Rimma. “Netflix statistics: How many subscribers does Netflix have?” *Insider Insights*, 2021, https://www.insiderintelligence.com/insights/netflix-subscribers/. Accessed 18 10 2021.

[3] Stodart, Leah, and Ashley Keegan. “Best streaming sites for movies.” *Mashable*, 2021, https://mashable.com/roundup/best-movie-streaming-sites. Accessed 18 10 2021.

[4] Sunilkumar, Chaurasia Neha. "A review of movie recommendation system: Limitations, Survey and Challenges." ELCVIA Electronic Letters on Computer Vision and Image Analysis 19, no. 3, 2020: 18-37.

[5] Lee, Jeremy. “Exploratory Data Analysis With Movies - Towards Data Science*”*. *Medium*, 2020, September 8 https://towardsdatascience.com/exploratory-data-analysis-with-movies-3f32a4c3f2f3. Accessed 20 10 2021

[6] Panchal, Kishan and Swalin, Alvira. “MSDS 622 Final Project”. *GitHub*, 2021, August 19. https://github.com/k7p/dataviz_project. Accessed 20 10 2021

[7] Subramaniyaswamy, V., R. Logesh, M. Chandrashekhar, Anirudh Challa, and V. Vijayakumar. "A personalised movie recommendation system based on collaborative filtering." *International Journal of High Performance Computing and Networking* 10, no. 1-2 (2017): 54-63.

[8] Wu, Ching-Seh Mike, Deepti Garg, and Unnathi Bhandary. "Movie recommendation system using collaborative filtering." In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 11-15. IEEE, 2018.

[9] Pazzani, Michael J., and Daniel Billsus. "Content-based recommendation systems." In *The adaptive web*, pp. 325-341. Springer, Berlin, Heidelberg, 2007.

[10] Soares, Márcio, and Paula Viana. "Tuning metadata for better movie content-based recommendation systems." *Multimedia Tools and Applications* 74, no. 17 (2015): 7015-7036.

[11] Ahmed, Adel & Batagelj, Vladimir & Fu, Xiaoyan & Hong, Seok-Hee & Merrick, Damian & Mrvar, Andrej. (2007). Visualisation and analysis of the internet movie database. Asia-Pacific Symposium on Visualization. 17-24. 10.1109/APVIS.2007.329304.

[12] Naumov, Maxim, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang et al. "Deep learning recommendation model for personalization and recommendation systems." *arXiv preprint arXiv:1906.00091* (2019).

[13] Cheng, Heng-Tze, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson et al. "Wide & deep learning for recommender systems." In *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7-10. 2016.

[14] Moreira, Gabriel de Souza P., Sara Rabhi, Ronay Ak, Md Yasin Kabir, and Even Oldridge. "Transformers with multi-modal features and post-fusion context for e-commerce session-based recommendation." *arXiv preprint arXiv:2107.05124* (2021).

[15] Gupta, Udit, Samuel Hsia, Mark Wilkening, Javin Pombra, Hsien-Hsin S. Lee, Gu-Yeon Wei, Carole-Jean Wu, and David Brooks. "RecPipe: Co-designing Models and Hardware to Jointly Optimize Recommendation Quality and Performance." *arXiv preprint arXiv:2105.08820* (2021).

[16] Singhal, Shashank. “Data Visualization -- Netflix Data set*”*. *Medium*, https://medium.com/analytics-vidhya/data-visualization-netflix-data-set-d4fa2da97253. Accessed 15 11 2021

[17] Rastogi, Kashish. “Performing EDA on Netflix Dataset with Plotly”. *Analytics Vidhya*, https://www.analyticsvidhya.com/blog/2021/09/performing-eda-of-netflix-dataset-with-plotly/. Accessed 15 11 2021

[18] Panchal, Kishan. “Exploring Movie Data with Interactive Visualizations” . *Towards Data Science,* https://towardsdatascience.com/exploring-movie-data-with-interactive-visualizations-c22e8ce5f663 . Accessed 15 11 2021