Final Project Paper

MovieLens: An Interactive Movie Data Visualization

Alexander Bogdanowicz[*1] Jose Ricardo Chacon Rodriguez[*2]

Exploring the most relevant movies of the last 30 Years through data analysis.

Abstract – In this project, we were able to build four graphical views using the d3 JavaScript library and simple HTML and CSS which allow the user to have an interactive experience as they explore data from the most relevant movies of the last 30 years. This unique platform enables the users to understand factors like revenue, release date and genre co-occurrence through a minimalist UX for a seamless interaction.

Keywords : MovieLens, IMDB, data visualization, d3

1 Introduction

Cinematography is the top-level medium of entertainment for mass audiences in contemporary society. This is an impressive feat for any medium and it ' s fascinating how the state of the art for producing and consuming films has changed ever since its inception at the beginning of the 20th century.

Like many other entertainment mediums, film-making and its related industries at large generate impressive amounts of data. From production costs and ticket sales to specific data on movies and their categorical classifications to their impact and reception by audiences across the world, data is interwoven throughout the medium and provides a rich platform for analysis. Raw data however, in and of itself, is a rather unrefined mineral; its real utility more often is derived from the product of a rigorous preprocessing effort that provides the basis for a clever depiction of the form, readily accepted by the human visual-perceptual system (Visualization Analysis and Design, CRC Press, 2014.).

On this basis, we attempt to enlist a two-phased approach towards the analysis and visualization of the MovieLens dataset found at

"https://grouplens.org/datasets/movielens"

1) Understanding the nuance of the dataset, cleaning, pre-processing, gathering further data where needed and drawing initial conclusions from the dataset and 2) Developing effective visualizations rooted in the science of Information Visualization, that will both provide for an easier data discovery and learning experience as well as augment the process of drawing conclusions on the effectiveness of our initial analysis.

2 MovieLens Overview

In this section, we explain the work done beforehand with the data used, how we preprocessed it, what our intentions were when we analyzed it, the motivation and design principles behind the UX design for the website and how the system is structured.

2.1 Data Used

For this project we used the dataset by its same name MovieLens and also pulled in various other datasets from IMDB and Box Office Mojo to complement the data. As a consequence, the data preprocessing part of the project was undertaken with even more care as there was a need to clean up any inconsistencies between datasets and be aware that any integration potentially carries underlying biases.

Being aware of the potential complications and future inferences that can be extracted from our integrated data is fundamental to the scientific method and as such we are proud to present full transparency to our readers. The dataset collected by the GroupLens Research group, MovieLens, consists of 4 tables, Links, Movies, Ratings, and Tags. We summarize each table and their features as well as additional

[*1]Department of Engineering and Computer Science - Data Science
[*2]Engineering and Computer Science- Computer Science

data sources in Table 1. Our analysis of this dataset and corresponding web scraping scripts can be found in a python-based Jupyter Notebook accompanying this paper.

Table 1

Name	Features	Source
Links	9,742	GroupLens Research
Movies	9,742	GroupLens Research
Ratings	100,836	GroupLens Research
Tags	3683	GroupLens Research
BoxOfficeMojo	13,153	Scrapped Website

Number of features and sources for the data used

2.2 Data Preprocessing

The data preprocessing for this project, as aforementioned, was one of the most important and time-consuming due to the level of scrutiny that we wanted for the project. The first step was extracting the complementary datasets and format them in a way that would make sense for us.

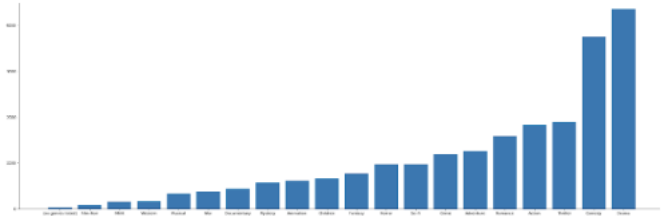
Next up, we handled all the cleaning and processing of data through python notebooks. Fortunately, the frameworks provided through Python ' s repertoire of libraries for data analysis greatly eased the process (namely: numpy, pandas and matplotlib).

We initially began with simple descriptive statistical techniques to get a feel for the MovieLens dataset. It was immediately clear that certain aspects of the dataset would need to be ignored, such as the 3,658 tags, of which almost 50% were unique, and which corresponded to only a minute group of users. Perhaps these would be better suited for a future NLP project corresponding to the dataset.

In addition to this textual data, the dataset also feature genre classifications for each movie. As movies typically feature many genres, as the dataset showed, we also began to ask questions relating to the relationships between genres and their true "equality". For example, a genre that exists concurrently with other genres might be less "dominant" than a genre that appears mostly on its own.

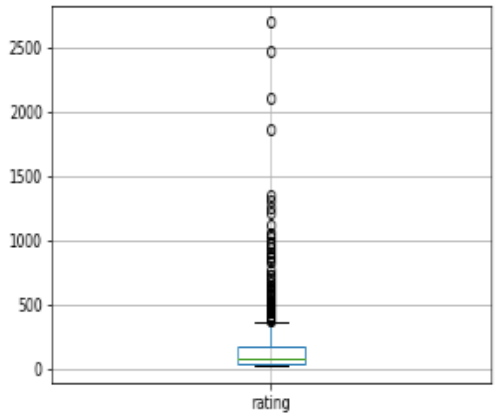
We depict the absolute counts of genre appearances in films throughout the dataset lifespan in Figure 1.

Figure 1



As Figure 1 shows, we found that the distributions amongst user ratings (the absolute amount of ratings) was highly skewed to the right, a unique circumstance that warranted further consideration.

Figure 2



Users who contributed large amounts of absolute ratings might correspond to "hubs", that is, individuals that are representative of a cluster of other users and thus we feared excluding such users might inhibit further analysis of user groups. At the same time, if these users were "generalists" in the film space, including them might inhibit our analysis of user groups.

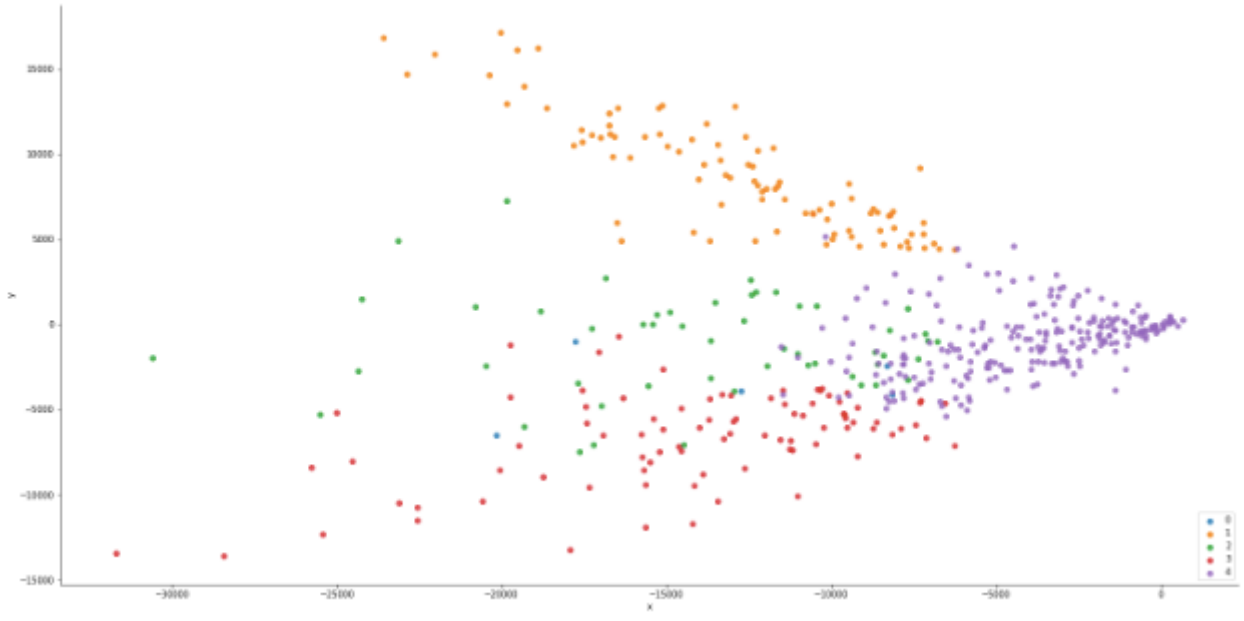
To remedy our concerns, we decided to perform an initial Singular Value Decomposition of the User-Movie Ratings Matrix, without removing outliers and plot our results, with color initially corresponding to users most viewed singular genres. The result can be seen in Figure 3.

Figure 3

It is rather clear that, genre categories, at least in our technique of using the most popularly occuring singular categories of the 18 unique genres featured in the dataset are not reflective of user groups, at least relating to these two factors of the correlation matrix of Users to Movies.

We also noticed, at this stage, that the outliers present in our study were, in fact making up the population of peripheral users around naturally clustered groups and therefore decided that our analysis would be more fruitful by removing these outliers. We proceeded to repeat SVD on the cleaned User-Movie matrix, and plot the results of our K-means analysis in Figure 4.

Figure 4



Finally, we decided to pursue further information as to the release schedule of films and the evolution of the box office over time. We thus scraped the Box-OfficeMojo website for their top films ranked by revenue over the period represented by our Movie Rating dataset (i.e. 1989 - 2018).

data pre-processing was performed using Python libraries for data analysis (e.g. pandas, numpy, matplotlib, sklearn, and scipy). Additional steps were taken to format data into more presentable forms and we also took into consideration the best data formats to import data into d3.js for visualization purposes.

2.3 System Structure

Our website consists of four main views: P-Coord, Industry Snap, Clustering, Recommendation. Each one of these views has been integrated into the same HTML file which allows for a cohesive experience.

2.4 Data Analysis

We proceeded to discover ways in which the data might be best represented. We specifically took into consideration aspects of interactive Information Visualization such as selection, exploration, filtering, and connection, (Toward a Deeper Understanding of the Role of Interaction in Information Visualization) placing particular emphasis on the "five E's" of Information Visualization (i.e. effectiveness, efficiency, engagement, tolerance for error, and ease of use) and on which marks and channels would best represent certain data. Eventually, we settled on a web-based solution through which users can navigate four views, each tackling a different relationship among our dataset. We proceeded therefore to an analysis of our views and their purpose, as well as a critical analysis of their strengths and weaknesses. The top 5 Movie specific characteristics of Season of Release, Date of Release, Rank, and Gross Revenue. The idea of the parallel coordinates plot is rather simple in nature and has seen various interactive adaptations (e.g. fisheye, highlights, brushing etc.).

We employed a highlighting mechanism that, upon hovering over each data-line, would expand the line and change its color, as well as revealing the specific information for the movie such as movie-title, such that the magnitude channels (e.g. position on a common scale) and identity channel (e.g. color of selected), would best convey the information and trends of the dataset.The chart also has a relatively high data-ink ratio.

The second view is an attempt to bring together a snapshot of the film industry over a specific period of time. We decided to deploy a timeline at the bottom of our view, which featured a time-series line-chart of gross industry revenues over time, which immediately made clear the seasonality of the box-office. We positioned a brush which allowed for a selection in time which would then translate to the other charts featured in our view. We include a bubble-chart that is visually appealing in it's transitions as it employs the unique design features and physical characteristics of the force-directed graphs in the D3 design package. Whereas we are aware that area is not the best means of conveying relative magnitude, we wanted to show the transition over time of studios (e.g. growth of the circles over time) and provide for the absolute values in studios gross revenue over the time period upon hovering. We also implemented a bar-chart that reacts to the desired time frame of the brush and displays the ordered top 10 performing films. Combined, these three charts decompose industry revenues into their industry, studio, and film levels and are a fun and interactive experience.

Throughout the visualization process, we also paid particular attention to the Gestalt Laws, which breakdown specific attentive meanings to the viewer into pre-attentive and post-attentive features. Whereas rules such as continuity and proximity corresponded to our Parallel Coordinates chart, our third view paid particular attention to the rules of symmetry. We applied the K-means clustering algorithm and were interested in how well the clustering actually represented the genre tastes of a particular user. We therefore implemented an interactive bubble-chart to help the user visualize the clusters with bubbles separated by color and a gravitational force and thus taking advantage of proximity and continuity, sized according to the number of ratings a particular user gave. We used a radial-axis chart to compare a given user against their cluster in the realm of top-5 genre preferences. The chart takes advantage of symmetry as an effective means of communicating whether the clusters are good approximations of user genres. When a user hovers over a specific circle in a cluster, the radial-axis chart updates to reflect the current cluster and user selection relative to the clusters top-5 most watched genres.

Lastly, we wanted to gauge the most-popular genres by co-occurrence in a palpable manner, one which would highlight the relative magnitudes of co-occurrences as well as convey the dominance of particular genres over others. The chord-diagram is also directional in nature, and as the MovieLens dataset features genres labeled in direction of dominance we were able to visualize this relationship particularly well with the help of the chord-diagram. Whereas the chart is slightly less interactive in nature, it allows the user to gather which genres co-occur the most with others, and their relative magnitudes.