CS 176 Project Report

Aadya Pawar

I. About the data

My dataset is 16 csv files containing publicly available information about movies from 1950 to 2025. Each file represents one of sixteen genres and contains columns movie_id, movie_name, year (year of release), certificate, run_time, genre, rating, description, director, director_id, star, star_id, votes, gross (in $).

II. Data Source

The data is sourced from IMdb. **IMDb** (an acronym for **Internet Movie Database**) is an online database of information related to films, television series, podcasts, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews (Wikipedia).

Description of the columns:

- movie_id - IMDB Movie ID
- movie_name - Name of the movie
- year - Release year
- certificate - Certificate of the movie
- run_time - Total movie run time
- genre - Genre of the movie
- rating - Rating of the movie (users can rate movies on scale of 10)
- description - Description of the movie
- director - Director of the movie
- director_id - IMDB id of the director
- star - Star of the movie
- star_id - IMDB id of the star
- votes - Number of votes in IMDB website (users vote for movie, can be used as a measure of popularity)
- gross(in $) - Gross Box Office of the movie

I found the 16 dataframes (csv files) in a Kaggle dataset.

III. Plan for data analysis & goals for the project.

As someone who loves watching movies and is always looking for new things to watch, I think this is a very interesting dataset for me to explore. I want to analyze the dataset to find popularity trends by ratings, votes and gross performance. This analysis could also help predict future trends by identifying patterns.

IV. Data collection & preparation

I combined all of the 16 csv files into one dataset and added a column called genre to my dataframe df where I added the genres for each entry based on the csv file.

I then cleaned the dataframe and removed unwanted values in the year column as well as removed all nan values. When I tried to make changes to my dataframe, I kept getting a slice error that occurred because I was trying to make changes to the dataframe directly and the error stated that I must use .loc and make a copy of my dataframe df. After learning how to use .loc, I dropped all the rows with NaN values in the year column specifically. I also dropped columns that were not relevant to any of my analyses.

After that, I formatted my runtimes from mins and hrs to be displayed in days hrs:mins: seconds format for ease of analysis.

I then created a separate dataframe containing all the actors names and the number of times they occurred in my dataframe df. I did the same for directors. Some entries had multiple values separated by \n so I separated them using commas in the original dataframe and added each value as a separate row in the new dataframe.

I then cleaned the genre column and identified the most popular genre for each year from 1950. I did the same for finding the top rated and most voted movie for each year

V. Data exploration

I used several different ways to explore the data . After cleaning and combing the csv files, I came up with a list of questions (detailed below) that could be explored through the data. I have explained the steps and reasons to how I explored my data in the notebook.
Questions answered by data analysis of my data:

- Number of movies each actor listed on Imdb has worked in
- Number of movies directed by each director listed on Imdb
- Most popular genre per year
- What genre has the highest run time?
- Highest rated movie of each year for the last 10 years or more
- Highest voted movie of each year for the last 10 years or more
- What is the relationship between highly rated and highly voted movies?
- How do ratings, votes and gross earnings correlate?
- Most popular actors of all time by number of movies?
- Most popular directors of all times by number of movies
- Highest grossing movie for each year
- Total gross values for each year
- What certificate movies make the most money?
- What genre of movies grosses the highest?
- Who is the best director by average rating?
- Who is the best director by average votes?
- How does rating and votes correlate?
- What is the relationship between runtime and gross earnings?

VI. Observations & Findings

Some interesting observations:

- Genre of movie with the highest runtime is biography
- Most movies are rated around 6-7
- There is a wide range of ratings, the majority of movies have moderate ratings. Votes and gross earnings, however, are dominated by a small number of movies that have very high values, indicating that only a few movies get a lot of votes or earn high gross revenue, which is typical in the movie industry where few hits make up most of the revenue and attention.
- Director analysis:
    - Lew Landers is the most popular director to work with, he has directed 232 movies. He's the most popular in terms of movies directed
    - Haruo Sotozaki is the best director in terms of the highest ratings
    - Christopher Nolan is the best director in terms of popularity, he's received the highest votes on Imdb
- The highest grossing year for the movie industry was 2015 with more than $800M in total revenue

- Most high revenue movies are PG rated, however Adventure movies make the most revenue out of all genres
- higher-rated directors (red) do not necessarily receive a higher number of votes (green), with no clear pattern of correlation observable between the two variables.
- most movies have a runtime of less than 500 minutes and there is no clear trend indicating a relationship between the length of a movie's runtime and its gross earnings

VII. Challenges

In the beginning, I ran into several errors when I was trying to work with this scale and size of data. As stated earlier, I had to learn how to use .loc effectively to avoid certain errors. I also had multiple calculation errors especially when I was attempting to calculate most voted and most rated directors and had to break that down into several smaller steps and sub parts to make sure that the calculation was accurate. I struggled with some visualizations too because of the number of parameters and needed to reevaluate my choice of visualization and plot. I also lost some of my code when I deleted a cell and reset my kernel, it automatically disconnected and did not connect but I was able to finally fix my kernel issues.

VIII. Overall Understanding & Effectiveness

I really enjoyed working on this project because there was so much that could be explored with the dataset that I had. I learned a lot about visualization techniques, dataframe cleaning, and sorting. I also explored different ways to plot things using seaborn. Additionally, I learned more about data cleaning and creating new dataframes. My key takeaway is that you should always double check the data values that you are working on because it may be inaccurate if not cleaned properly as per your requirements and parameters and result in incorrect analysis.

IX. Team

I worked by myself.

X. Code:

https://purdue.brightspace.com/d2l/le/content/948554/viewContent/15991872/View