Contents

Dataset Analyzed

Overview

The dataset chosen for this analysis is the “tmdb.csv” dataset.

It contains over 10,000 records of movies from the 1960s to 2015 from the IMDB movies website.

Data Columns Description and Meaning

The dataset contains the following columns:

- **id**: The unique number that identifies each row or record
- **imdb_id**: The unique number assigned by IMDB
- **popularity**: This is a number that represents the popularity of the movie among people
- **budget**: This is the amount of money spent in making the movie in the year it was made
- **revenue**: This is the amount of money made from sales of the movie
- **original_title**: This is the title of the movie
- **cast**: This gives us the names of the actors and actresses
- **homepage**: This gives the URL of the movie
- **director**: This gives the name(s) of the director(s) of the movies
- **tagline**: This is a catchphrase that is associated with the movie
- **keywords**: This is a set of words that are associated with the movie
- **overview**: This gives the synopsis of the movie
- **runtime**: This gives the total time the movie runs for
- **genres**: This column tells us what genre the movies are; drama, action, adventure, and so on.
- **production_companies**: This gives the production company or companies responsible for producing the movie
- **release_date**: This gives the date when the movie was released
- **vote_count**: This gives the total number of people that voted for or against the movie in IMDB
- **vote_average**: This gives the average rating of the movie out of 10.0
- **release_year**: This gives the year the movie was released
- **budget_adj**: This gives the budget for making the movie adjusted for inflation
- **revenue_adj**: This gives the revenue made from movie sales, adjusted for inflation

Questions Posed

The questions posed for the analysis of this dataset are:

1. Relationship between budget and popularity.
2. Which years were the largest budgets spent on movies
3. Which years were the largest revenues made?
4. Which release months are associated with revenue
5. The season/quarter in a year, when most movies are released

Investigating the Dataset

Question 1:

- First of all, a heatmap depicting the correlation of the features in the dataset was constructed to give a general feel for how the various features in the dataset are related
- Based on this first visual, it was observed that movie budget and popularity had a moderately high correlation of about 0.6. This also happened to give some intuition about one of the questions posed i.e. Question 1.
- Next, a scatter plot of the budget against popularity is plotted to visualize this relationship even better.

Question 2:

- Another interesting question was "Which years were the largest budgets spent on movies?"
- To answer this question, a bar graph plotting all the years against the sum total of the movie budgets used up was plotted.
- Perhaps an even more interesting question is "What were the top 10 years that had the highest budgets?"
- A bar graph was also plotted to visualize this

Question 3:

- In a similar manner, a bar graph is plotted to visualize the top 10 years with the largest revenues

Question 4:

- A dot plot showing which months had the larger spread of movies as well as the revenues generated for each movie is plotted

Question 5:

- A series of pie charts showing the quarter or season when movies tend to be released more.

Data Wrangling and Cleaning

Data Cleaning

The Data Cleaning process took the following steps:

1. The columns for Homepage, tagline, keyword, budget, revenue, and overview were dropped. The 'budget' and 'revenue' columns were dropped because they had not been adjusted for inflation unlike 'budget_adj' and 'revenue_adj'
2. Dealing with missing values:
    a) Directors: check which movies have no directors and see if you can find the directors else input unknown director
    b) Genres: check which movies have no genres and see if you can input them else input "unknown genres"
    c) Production companies: "unknown production company"
    d) Dealing with original_titles with bad names. Remove unwanted characters.

Data Wrangling

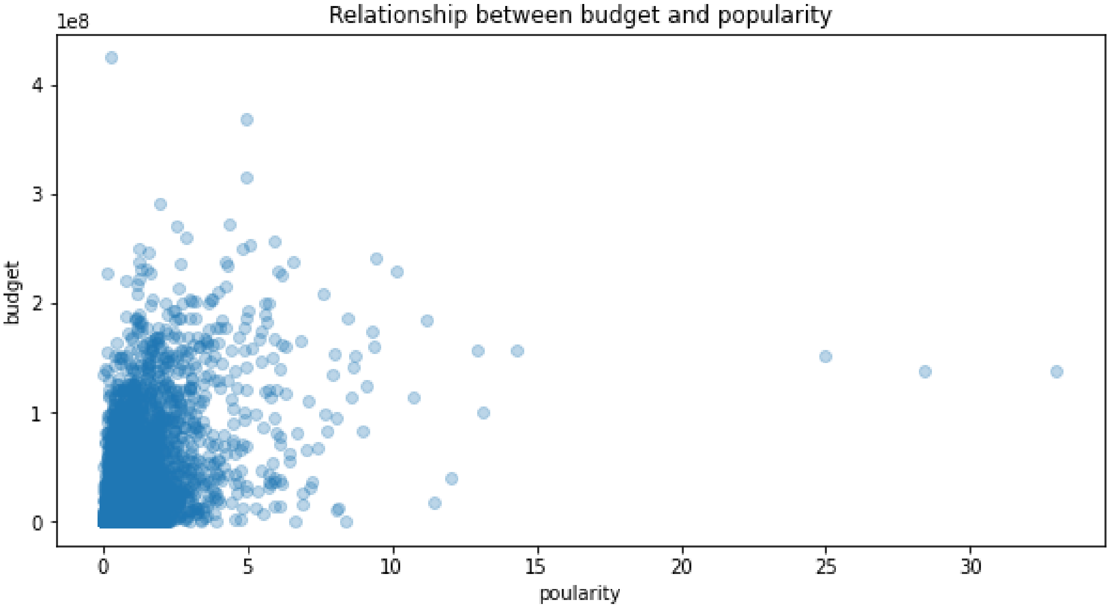
The Data wrangling process took the following steps:

1. The column for budget_adj was renamed to budget and revenue_adj to revenue.
2. Created a grossing and percent_grossing column. The grossing column showed the profit made off the movies, while percent_grossing showed by what factor the initial budget was recouped
3. Convert all string values to lower case.
4. Break cast, director, genres, and production_companies into fragments.
5. Dealing with dates:
    a) Check if release date year == release year: This is to ensure consistency of the data
    b) Create a “month” column
    c) Create a “day” column
    d) Create a “quarter” column

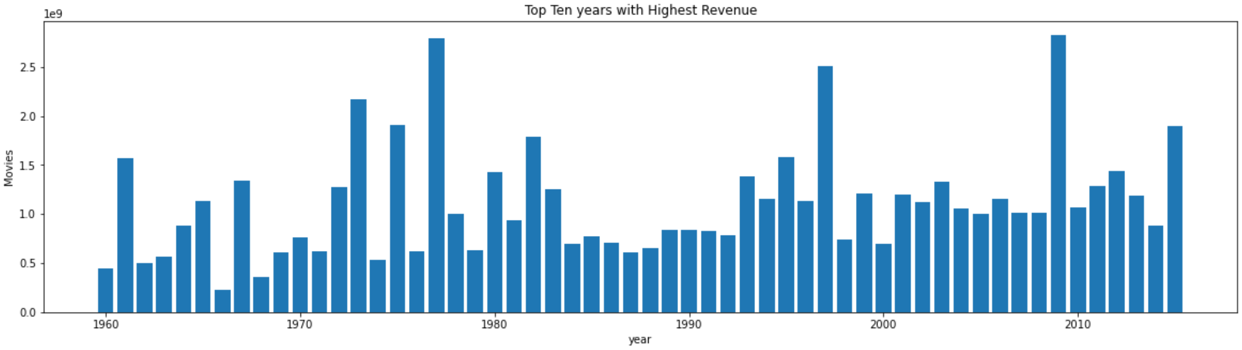
Summary Statistics and Plots

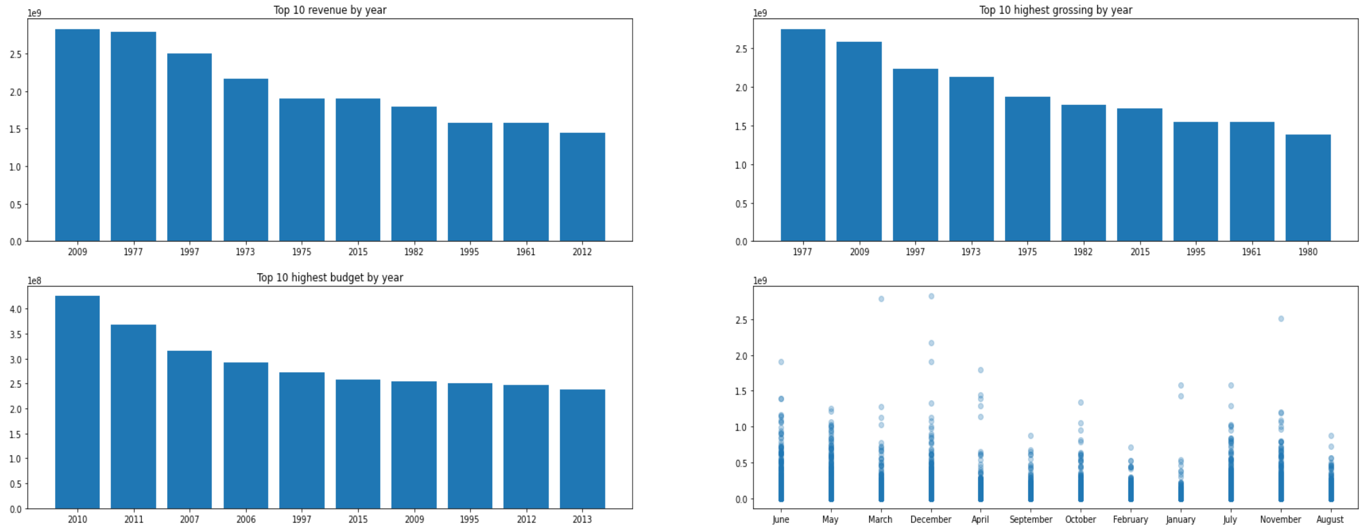
The summary of the analysis is given below:

1. There is a positive correlation between the movie Budget and popularity. This implies that a movie with a large budget could be quite popular but a causal relationship between the two variables cannot be firmly established at this time because correlation does not imply causation.



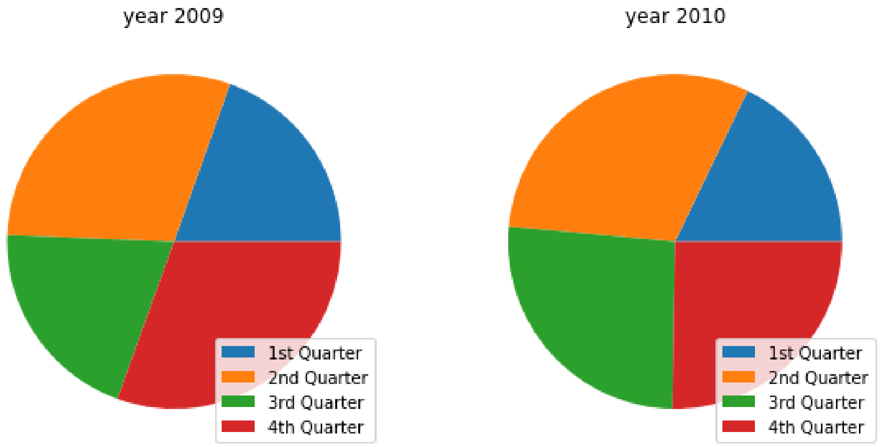
2. The years 2009, 1997, and 1973 have the highest revenue and highest grossing.

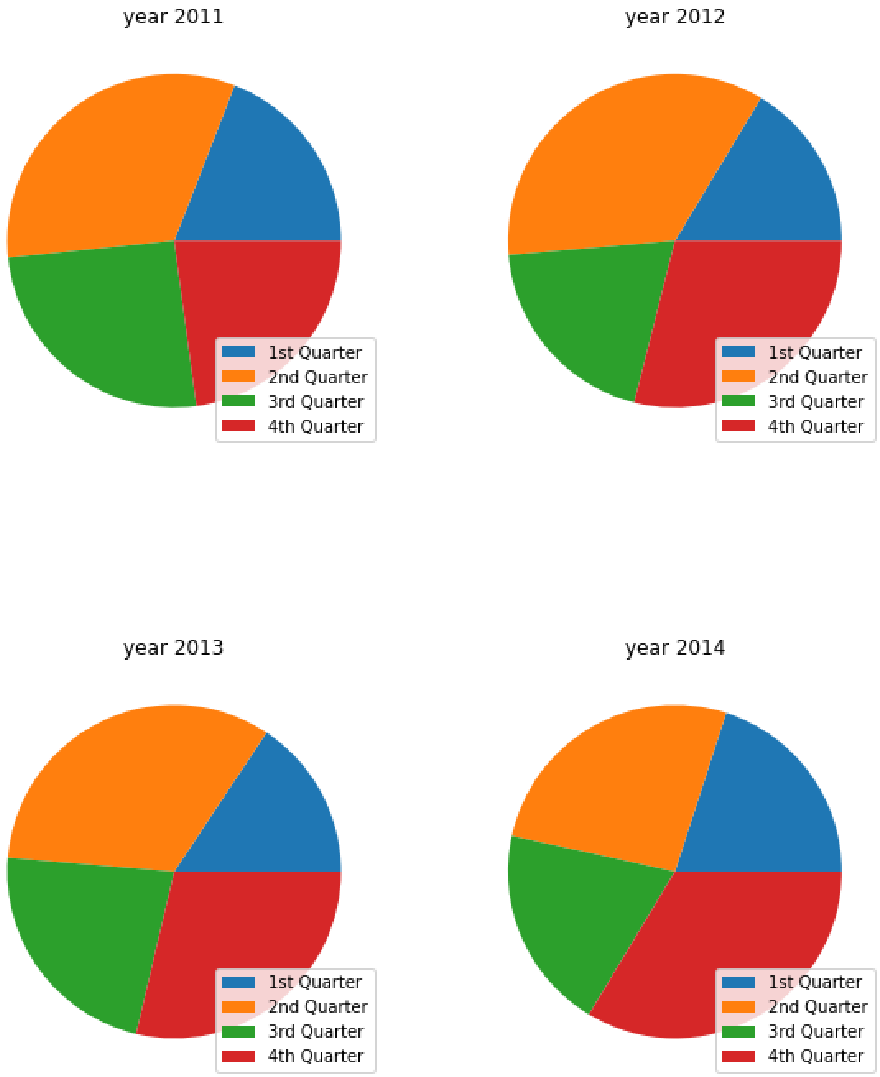




3. Budgets for 2010, 2011, 2007, and 2006 were the largest, yet they are not in the top 10 years for revenue generated and highest-grossing as seen above. This could imply that a large movie budget does not necessarily guarantee a large gross or revenue, although this claim needs to be properly investigated.

4. Movies tend to be released more in the 2nd and 4th quarters of the year according to the IMDB database, and this could be a reason why revenue generated is also the highest at those periods of the year. (See next page for the complete visual)




year 2011
year 2012
year 2013
year 2014
🟦 1st Quarter
🟧 2nd Quarter
🟩 3rd Quarter
🟥 4th Quarter