Project - Investigate TMDb movie data set

Summary:

This project Uses the TMDB(The Movie Database) data set to determine common characteristics of the most profitable films in the data base. The central idea being what factors are indicators that a film will be successful financially based on the top 80% of profitable films. This data set contains information about roughly 11,000 films (10,866). The following information for each film is provided (column headings): id, imdb_id, popularity, budget revenue original_title cast, homepage, director, tagline, keywords, overview, runtime, genres, production_companies, release_date, vote_count, vote average, release_year, budget_adj, revenue_adj. Notes about the columns: the 'cast', 'genres', and 'keywords' columns have multiple pipe (|) delimited values, the two columns ending in _adj show the budget and revenue are amounts stated in 2010 dollars.

This project will explore what variables are associated with high profit movies. Variables of interest from the data set will be director and runtime. In addition calculated variables of interest will be release_month (derived from release_year), release_DOW (DOW = day of week) (derived from release_year), release date of month(DOM) (derived from release year). All variables of interest will be forward looking, I.E. variables that would be under the control of the film maker prior to the start of production, as such things such as vote_count and vote_average are irrelevant for the purposes of this study as this is information post release and outside a predictive analysis.

Methodology:

1. load Tmdb data base as a csv file
2. Drop columns of no interest.
3. Check for duplicated rows in data set
4. Drop any rows determined to be duplicates
5. Determine if the basis for the calculated column are missing any values
6. add calculated column (net_revenue = revenue_adj - budget_adj). This column will indicate total profit.
7. Create a new dataframe consisting of the movies in the top 80 percent of profitability
8. Inspect new dataframe and clean as necessary.
9. Create a new calculated column for day of the week, day of the month, and month number the movie was released on.
10. plot results for variables of interest
11. Report insights.

1.Load data

2. Inspect data and drop columns of no interest.

Upon inspection the column heads are:imdb_id, popularity, budget, revenue,original_title,cast, homepage, director, tagline, keywords,overview,runtime,genres,production_companies,release_date,vote_count,vote_average,release_year,budget_adj,revenue_adj

After review. The data that is of limited use is divided into 3 types:

1) Data that is outside the film makers control: popularity, vote_count, vote_release. As these are facors measured post release,
2) Data that is irrelevant to analyzing common factors in profitability as by the very nature of the data (an inherent limitation) the data is unique and thus ill-suited to aggregation such as imbdID, homepage, tagline.
3) Date that is pointless to measure give the scope of the investigation: release_year (how would a film maker in 1991 travel 10 years forward in time knowing 2021 was a year for profitable movies?),
Next, drop columns with variables of no interest: imdb_id, popularity, budget, revenue, homepage, tagline, overview, production_companies , vote_count, vote_average, release_year

3. Inspect for Duplicates

The only variable of interest where duplicates are a concern is 'original_title', there is a possibility of duplicate titles with unique films but if the release date is different then each movie is unique. A calculated field will be created that is a concatenation of title and release date and any possible duplicate returns inspected for duplicates, once possible duplicate rows are identified, and duplicate rows dropped this column will be dropped. Upon inspection there are rows that are truly duplicates

4. Delete row at rowindex number 2090

5. Inspect the data set for missing values in budget_adj and revenue_adj columns, as this investigation is about common elements in movies based on net revenue and rows with zero value in both will be irrelevant.

1) get a count for rows with a zero value and a zero value in the revenue_adj column.
2) drop rows with a zero in both budget_adj and revenue_adj columns
3) for rows with only one zero in each replace the zero with the average value for that column.

6. Create calculated column for net_revenue (profit)

7. Create new dataframe

The top 20% of rows by (sorted by net revenue) will be identified and a new data frame created. This will be the data set of interest for this investigation.

Create a new dataframe sorting the data frame by 'net' descending 2) add a column starting at 1 and incrementing by 1 for each row 3) determine the count for the top 20% of rows (using the Pareto Principle that 80% of outcomes come from 20% of causes
(https://en.wikipedia.org/wiki/Pareto_principle) 4) using the 20% count take that subset to create a new data frame movieDB_pp.

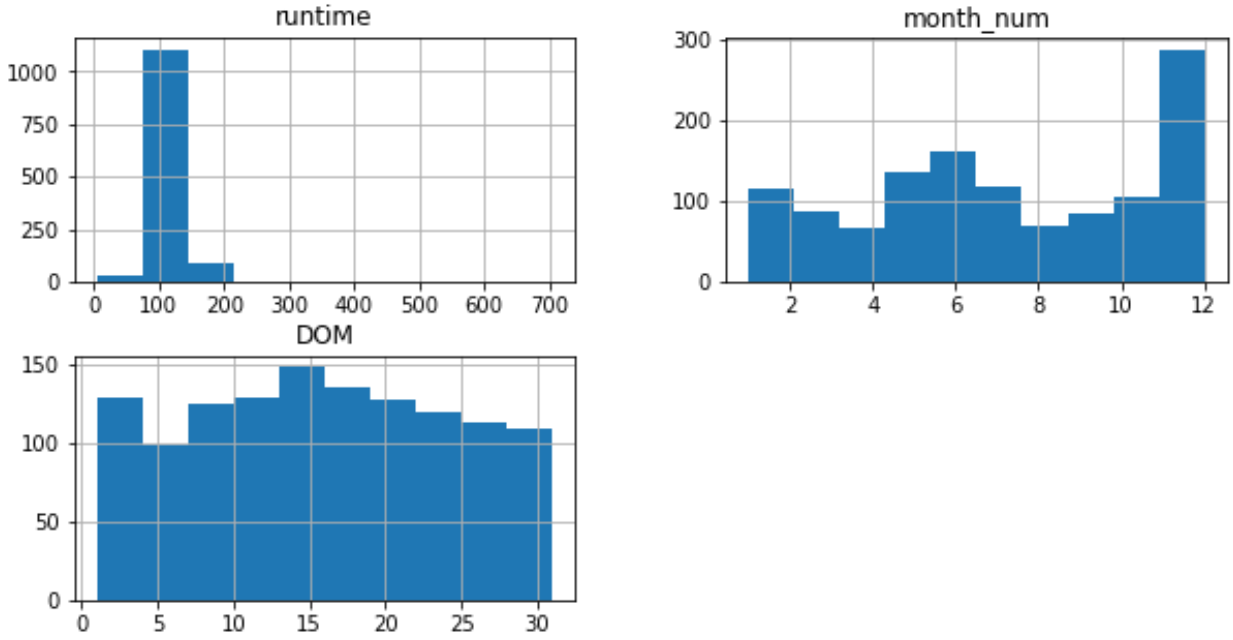
8. Inspect new dataframe and clean as necessary.

1) Inspect new dataframe for null value, unless there is a null count greater than 10% of values in a column, the count will be irrelevant in the next stage of analysis.
2) Given the small number of nulls there is no need to further cleanse data

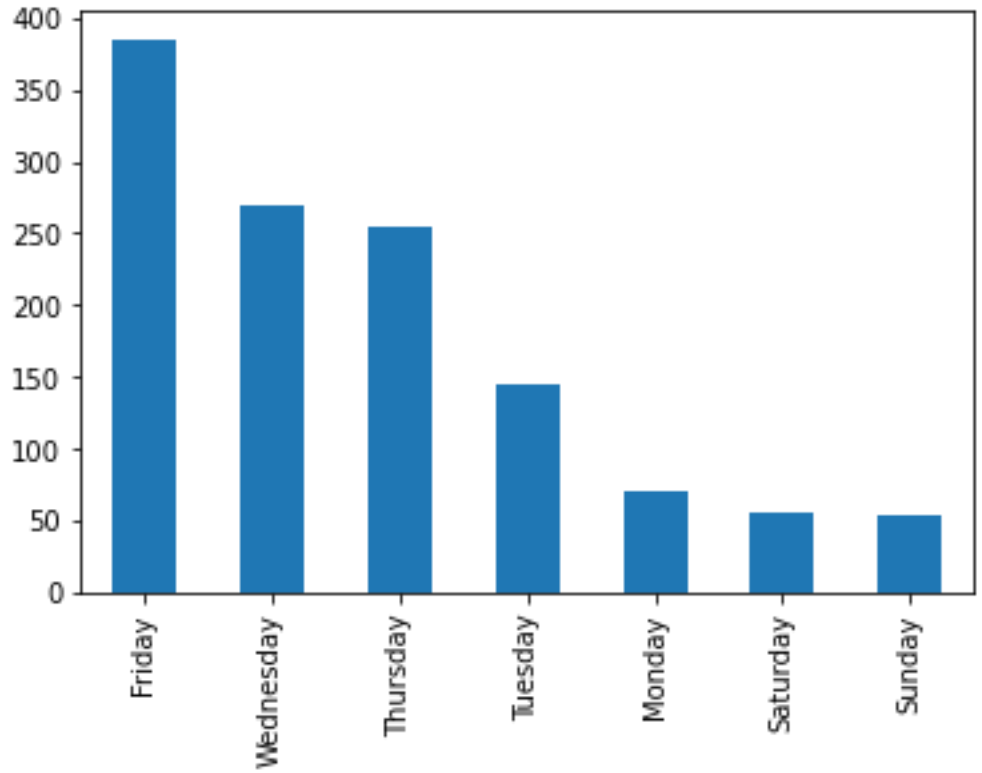
9. Create calculated columns for month number, day of week and date of month.

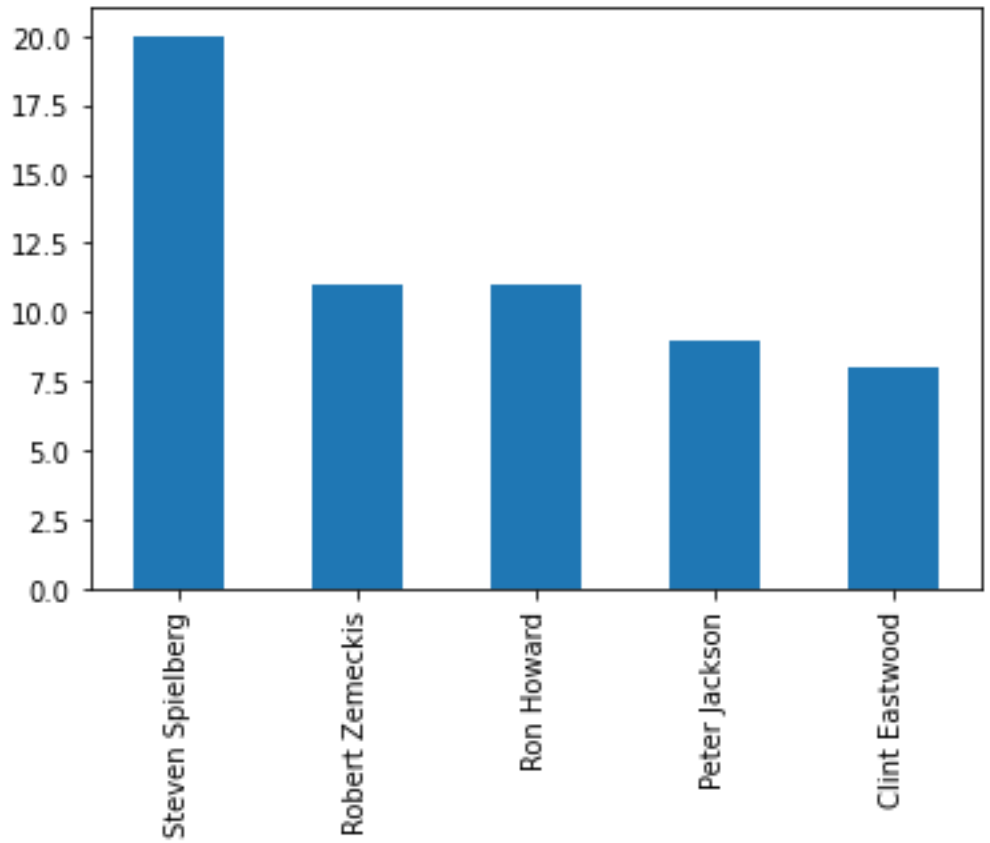
10. Visualize Date

1) Create histograms for appropriate data types ('runtime','month_num','DOM')



2) Create value counts (column types :day of week, director) visualization


400
350
300
250
200
150
100
50
0
Friday
Wednesday
Thursday
Tuesday
Monday
Saturday
Sunday


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Steven Spielberg
Robert Zemeckis
Ron Howard
Peter Jackson
Clint Eastwood

11. Conclusion/limitations/suggestions. 4 factors are common for the most profitable movies: 1) runtime of roughly 100 minutes, 2) December release, 3) Friday release day, and 4) Steven Spielberg being the director.

Limitations in the data:

1) Production company is irrelevant as the focus of the study was factors within a production companies' control (such as a Monday release day or September release month), these factors are seen through the lenses of a generic production company.
2) Date of month is inconclusive. Based on the date there does not appear to be a cleare best date of month.
3) 3)Some values are irrelevant to this study as they concern factors that happen after release of the movie: imdb_ID, popularity, vote_count, vote*average. and as thus cleary out of the control of a production company*
4) *Release* year is irrelevant as while a film maker has control over the day of week of release or month of release, release year release discretion is not a practical matter

Suggested further investigations: It would be valuable to revist this study if it were possible to capture net revenue by country, would the same top level insights hold up on a per country basis (maybe Tuesday is the best day to release a film in Italy).