INVESTIGATING A DATASET

By Nandhitha

1. **A note specifying which dataset I analyzed :**

   I chose to use the IMDb Dataset obtained from Kaggle. This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

   ● Certain columns, like ‘cast’ and ‘genres’, contain multiple values separated by pipe (|) characters.

   ● There are some odd characters in the ‘cast’ column. Don’t worry about cleaning them. You can leave them as is.

   ● The final two columns ending with “_adj” show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.

2. **A statement of the question(s) I posed:**
   - What is the most popular genre with which movies are made?
   - How many movies ended up in profit? And how many were in loss??
   - Does the budget or the release date/year or the popularity or the runtime or the vote average or the combination of all influence the movie's outcome?
   - Is it possible to visualize the ratio of high budget movies to low budgets?
   - What will be the number of latest movies included in the dataset?
   - Who might be the most popular or frequently casted actor?
   - Which movie made the highest profit and which made the lowest profit of all?

3. **A description of what I did to investigate those questions:**
   - I started with extracting details about the dataset such as investigating the number of rows and columns, datatype of each column, presence of duplicate values, etc.
   - After knowing completely about the dataset, I wrangled the data so that it will be suitable for further analysis.
   - I created bar plots and scatter plots to understand the correlation between certain columns in the data.
   - I also used some functions such as idxmax() to find out answers for my questions
4. **Documentation of data wrangling I did:**
   - Remove columns that are not needed for the analysis, such as the imdb_id, budget, revenue, homepage, keywords, overview. Budget and revenue can be removed as will be using the budget adjusted and revenue adjusted columns.

     ```
     imdb.drop(['imdb_id', 'budget', 'revenue', 'homepage', 'keywords', 'overview',
     'director','tagline'], axis=1, inplace=True)
     ```

- Drop duplicated rows..

  ```
  imdb.drop_duplicates(inplace=True)
  ```

- Drop the rows where the budget or revenue adjusted value is equal to 0 or not filled.

  ```
  imdb.drop(imdb[(imdb.budget_adj == 0)].index, inplace=True)
  ```

- Drop the row that have no genre type mentioned and production companies information.

  ```
  imdb.dropna(inplace=True, subset=['genres', 'production_companies'])
  ```

- Covert the datatype of 'budget_adj' and 'revenue_adj' from float to int.

  ```
  cols = ['budget_adj', 'revenue_adj']

  imdb[cols] = imdb[cols].applymap(np.int64)
  ```

- Rename the necessary columns to ensure comfortable working with tha data.

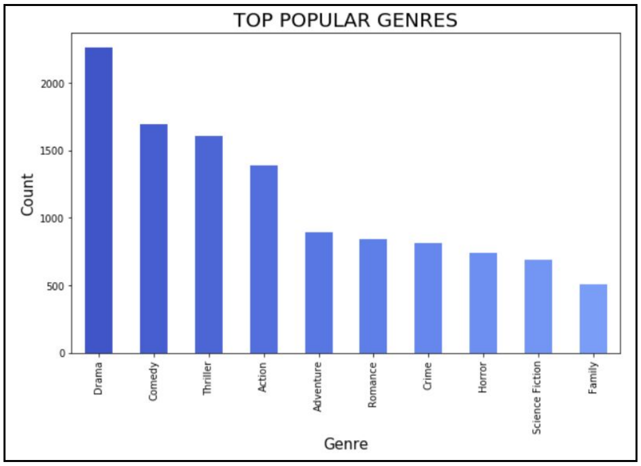
  ```
  imdb.rename(columns = {'budget_adj' : 'budget_in_$', 'revenue_adj' :
  'revenue_in_$'}, inplace = True)
  ```

- Change the release_date into datetime datatype.

  ```
  imdb['release_date'] = pd.to_datetime(imdb['release_date'], format='%m/%d/%y')
  ```

5. Summary statistics and plots communicating my final results

- The most popular genre is 'DRAMA' followed by 'COMEDY', as per the information given in the dataset.



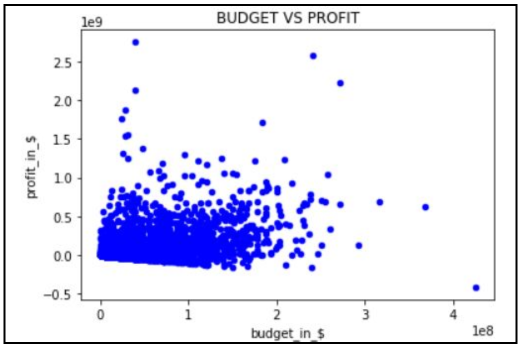
- In the dataset, 2757 movies makes profit whereas 2275 movies ended in loss.
- From the plots and visuals, we can infer the below:

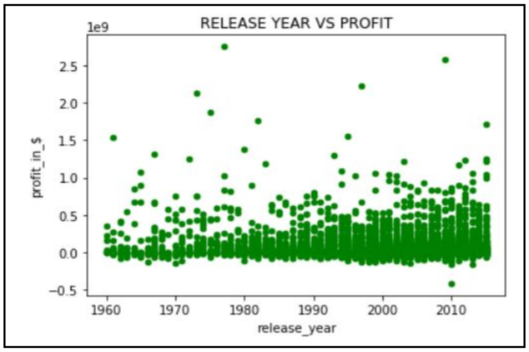
  -- It is not always true that higher the budget of the movie, the more profit it will make.

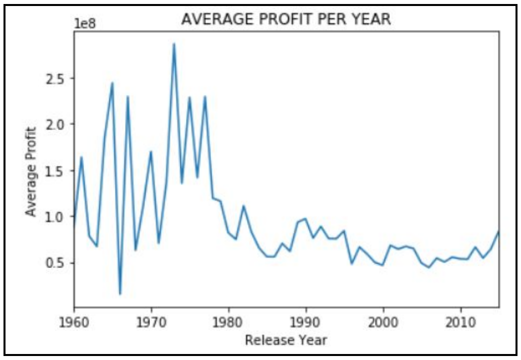
  -- It also seems to us that popularity does not affect the profit much.

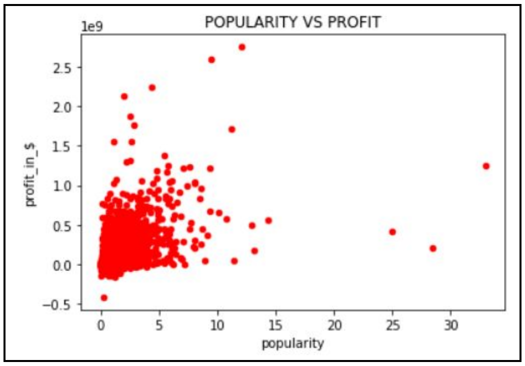
  -- The movies released during 1970 and 1980 are noted to contribute more money.

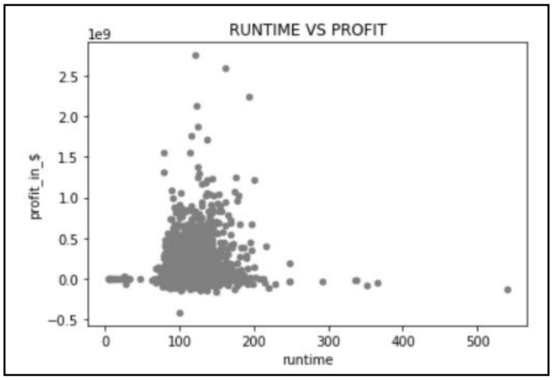
  -- Most of the movie duration is less than 200 minutes. Moreover, it is the runtime duration for which the movies made more money.



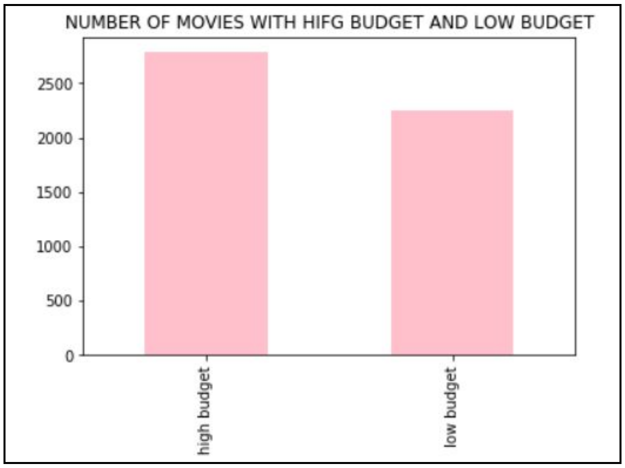




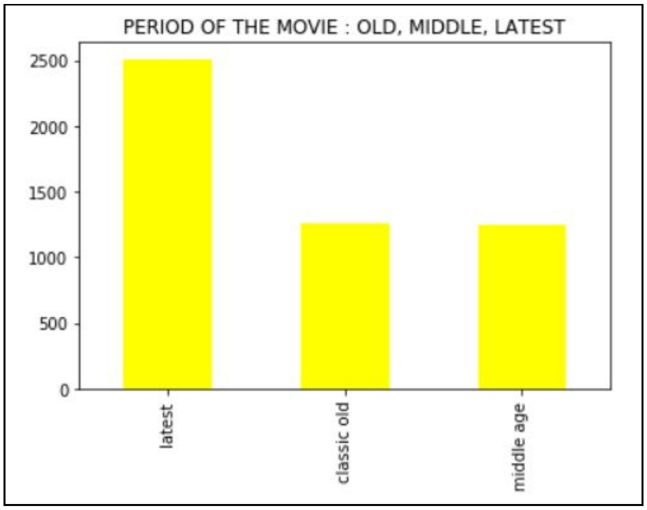




- After visualization we have found that 2278 movies were produced at high budget and 2245 movies were produced at low budget



- There are 2511 latest movies, 1248 middle age movies and 1264 classic old movies.



- 'Tom Cruise' is the actor who has done his part in more number of movies, followed by the actors 'Tom Hanks'.

- Movie with highest profit is found to be 'Star Wars' and movie with lowest profit is found to be 'The Warrior's Way'.