FEBRUARY 10, 2022

TMDB PROJECT

DATA ANALYSIS

UDACITY PLATFORM
By/Mostafa Gamal

Questions

1- The most used genre
2- Highest rating movie name and director name
3- Highest budget
4- Highest revenue
5- Highest profit and producer name
6- Highest popularity
7- Longest and shortest movie
8- Most used original language
9- The total money spent on movies every year
10- The total money came from movies every year
11- Highest rating every year
12- Highest popularity every year
13- Correlation between budget and revenue
14- Histogram of budget of movies
15- The count of each genre over the years
16- The count of each language over the years
17- Top 10 popular movies
18- Top 10 longest movies
19- Top 10 rated movies
20- Top 10 rated movies directors
21- Top 10 high profit movies
22- Top 10 high profit producers
23- Top 10 most appearing actors

Wrangling

Gather data

- Data downloaded from website
- https://www.kaggle.com/tmdb/tmdb-movie-metadata

Assess

Data consists of two files

1- tmdb_5000_credits.csv: contains 4 columns, two of them will be used to relate to the second file and the other two columns contain crew and cast data for movies
2- tmdb_5000_movies.csv: Contains id and movie name columns to relate to the first file and other columns containing budget, revenue, rating ,etc.
3- Some data points are missing however there other data for the same movie is found so these rows can't be deleted but will provide less value than complete data
4- Date column type needs to be changed

Clean & organizing data

- Data organizing will consists of merging two csv files based on movie id to get one data set which we will work on.
- We extracted year from each date and put it in a column to group data by later.
- Many columns contains data as string items, so to analyze them we will extract data by two main methods
    1. Cast and crew columns: We need to extract the following items (Actor_1, Actor_2, Director, Producer)
        - So first we transformed cast str items to list of dictionaries and made a new column (cast_dict) then loop in the first and second dictionaries of each row(if found) to get actors names and place them in two new columns (Actor_1, Actor_2)
        - For crew column the order of crew was not the same so we deal with it as str then used find method to find keywords like “Producer” and “Director” then get the following word regardless of word length using ‘ “ ‘ as stop station then placed them in two new columns (director, producer)
    2. For genres column: we needed to get all the genres for every movie so we used different approach
        - Gather a list of all genres found in column and put them in dictionary with values equal zeros
        - Iterate through every row of genre column and then iterate with every possible genre from the dictionary within the row and add 1 to the key value if found in the row
        - Finally, we have a dictionary of all genres and frequency of each.

EDA

Addressing the mentioned questions, they fall in four main categories

1- The highest parameters all over the years
2- Parameters change over the years
3- Top 10 characteristics over the years
4- Correlation for financial analysis and histogram

Category one : The highest parameters all over the years

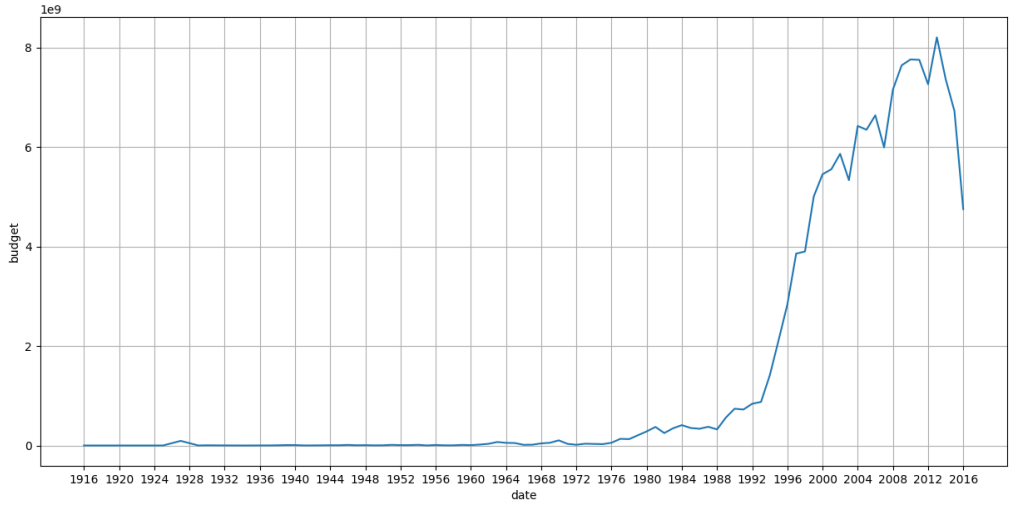
Code output:

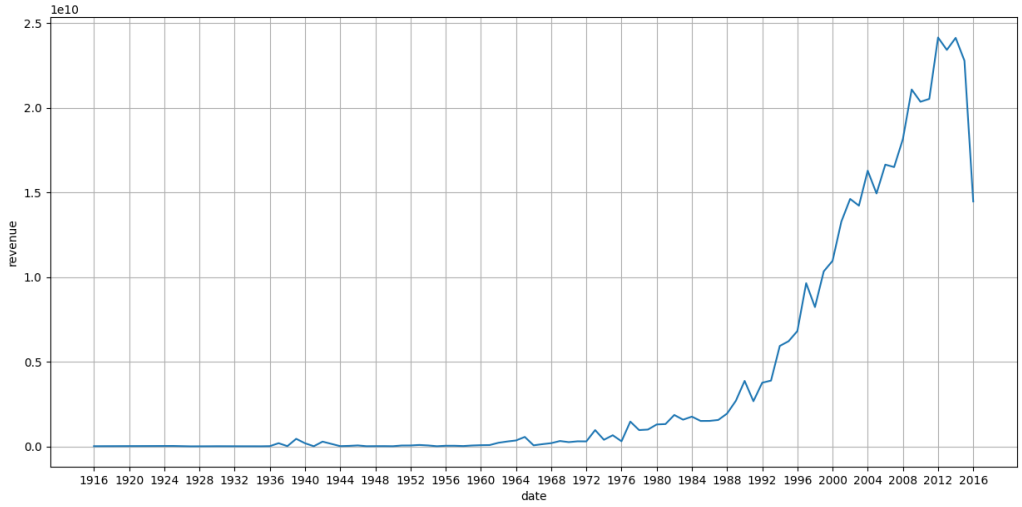
1. The most used genre over the years is: **Drama**
2. The highest rated movie ever is The Shawshank Redemption with rating: 8.5 for director: Frank Darabont
3. The highest revenue ever is 2,787,965,087 for the movie: Avatar
4. The highest budget ever is 380,000,000 for the movie: Pirates of the Caribbean: On Stranger Tides
5. The highest profit ever is 2,550,965,087 for the movie: Avatar for producer: James Cameron

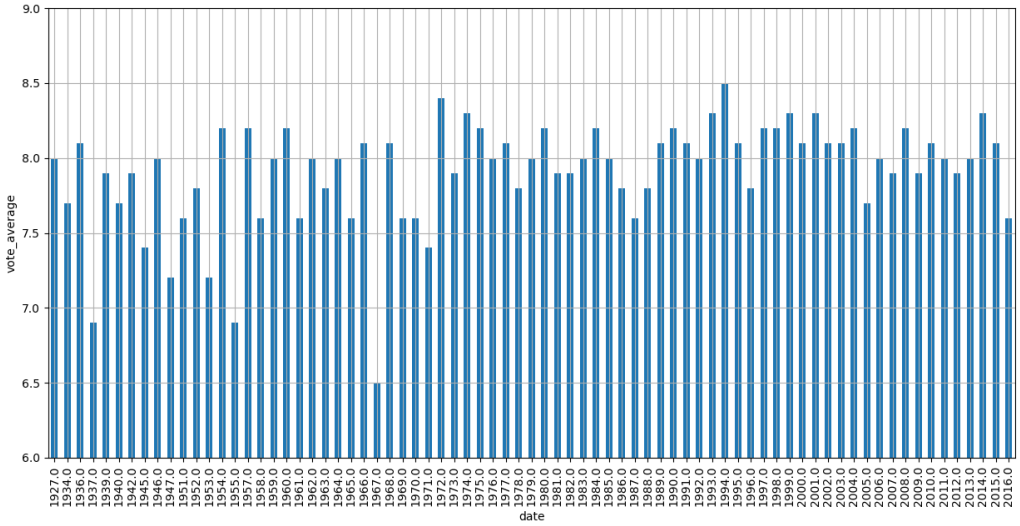
6. The highest popularity ever is 875.581305 for the movie: Minions
7. The longest movie ever is 338.0 min for the movie: Carlos
8. The shortest movie ever is 14.0 min for the movie: Vessel
9. The most original language ever is en

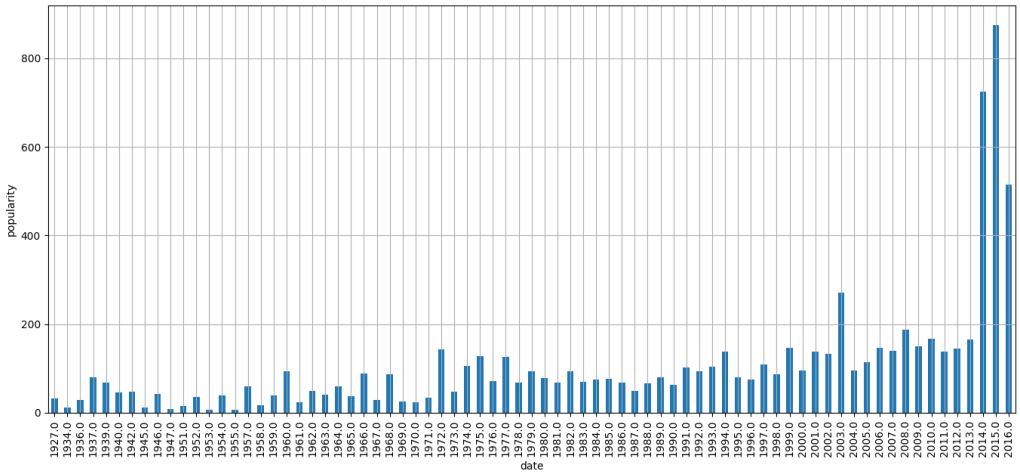
Category Two: Parameters change over the years

(Budget, Profit and revenue units are dollars and multiplied as shown in axis)

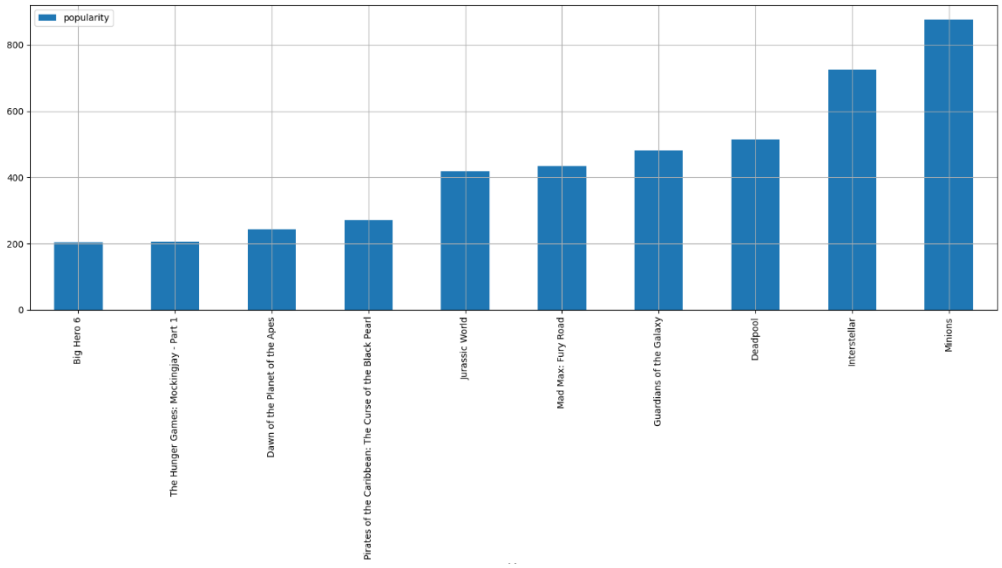


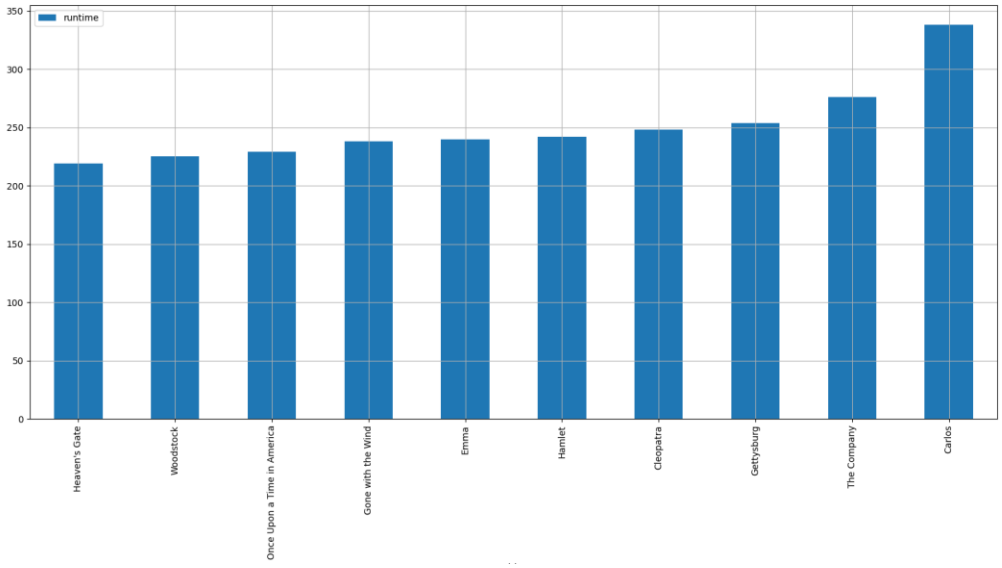


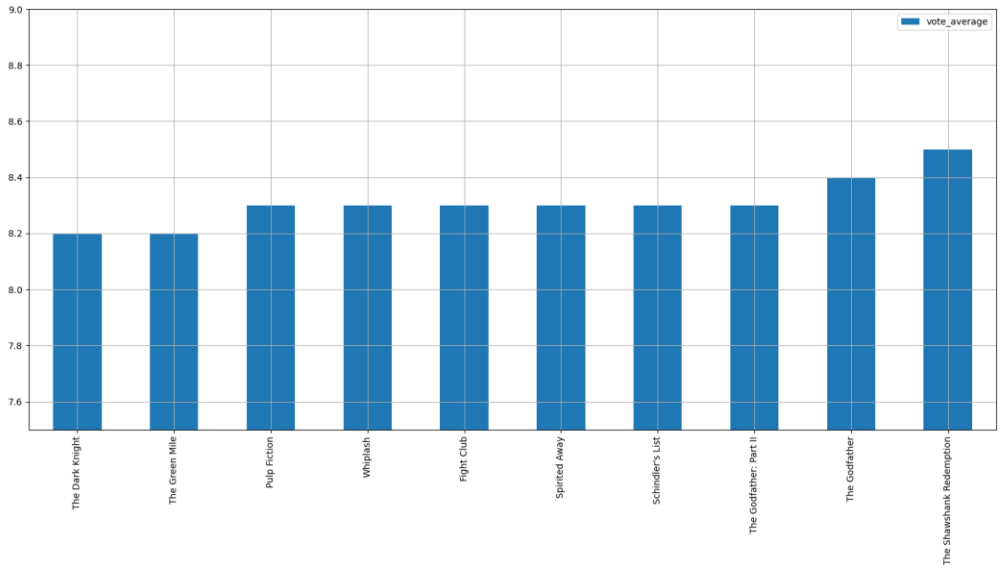

vote_average
9.0
8.5
8.0
7.5
7.0
6.5
6.0
date

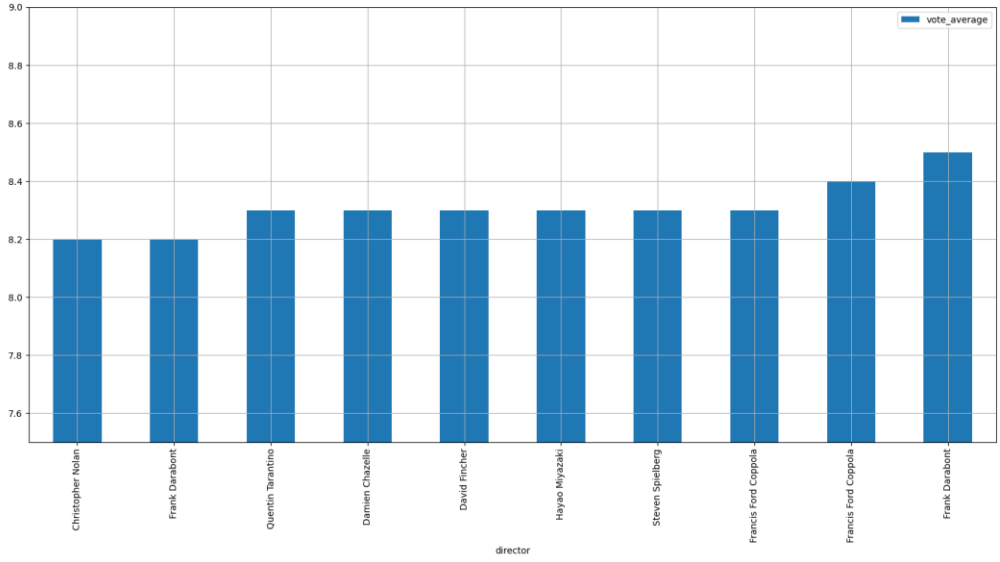

popularity
800
600
400
200
0
date

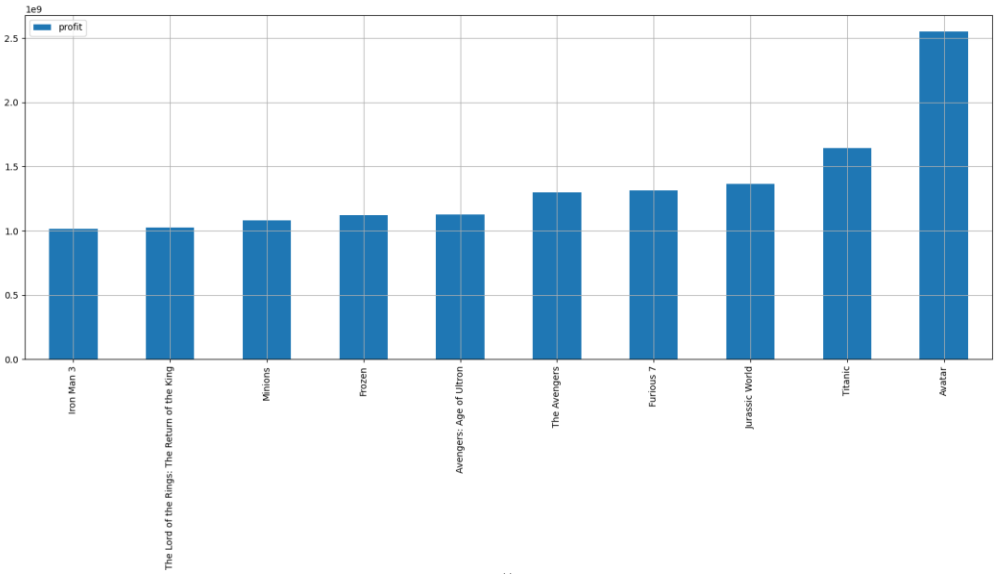
Category three: Top 10 characteristics over the years

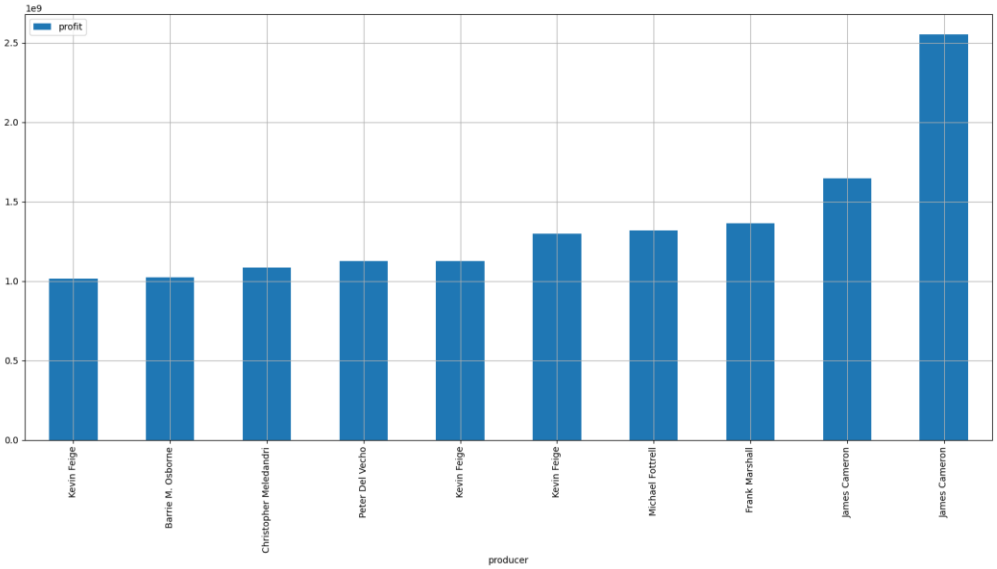


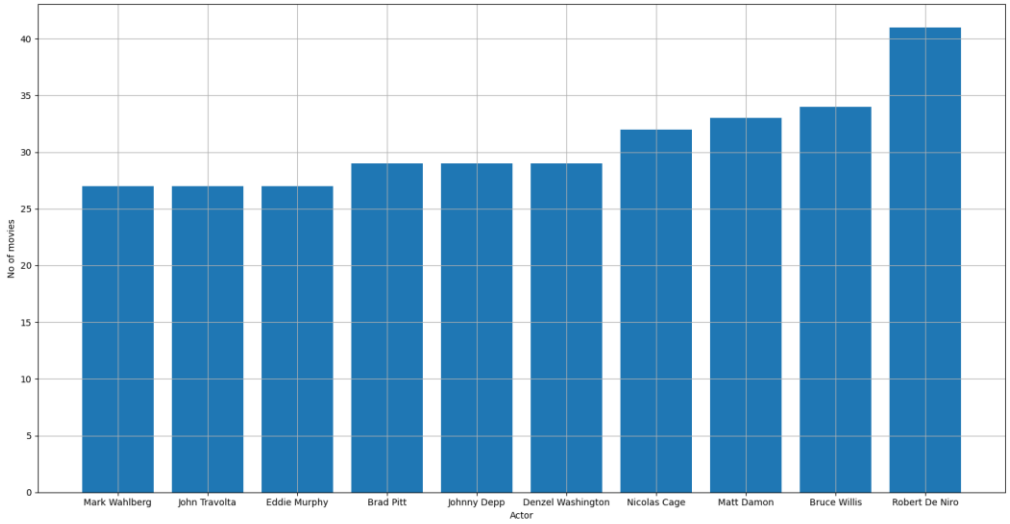



vote_average
9.0
8.8
8.6
8.4
8.2
8.0
7.8
7.6
The Dark Knight
The Green Mile
Pulp Fiction
Whiplash
Fight Club
Spirited Away
Schindler's List
The Godfather: Part II
The Godfather
The Shawshank Redemption

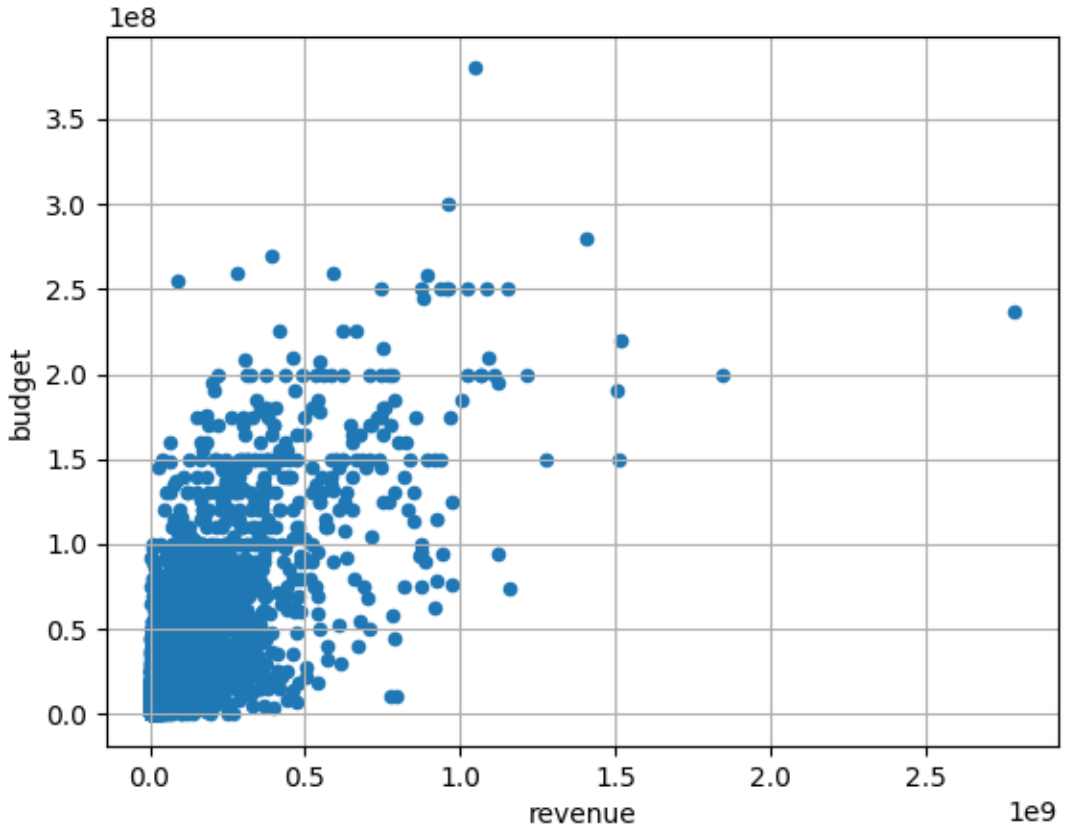

vote_average
9.0
8.8
8.6
8.4
8.2
8.0
7.8
7.6
Christopher Nolan
Frank Darabont
Quentin Tarantino
Damien Chazelle
David Fincher
Hayao Miyazaki
Steven Spielberg
Francis Ford Coppola
Francis Ford Coppola
Frank Darabont
director

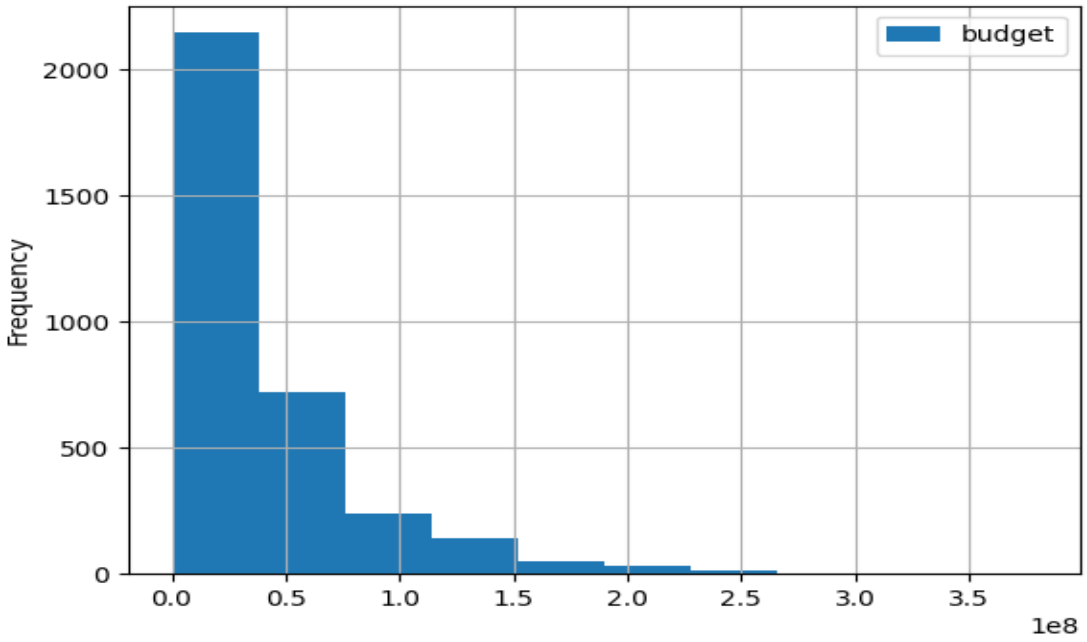

1e9
profit
2.5
2.0
1.5
1.0
0.5
0.0
Iron Man 3
The Lord of the Rings: The Return of the King
Minions
Frozen
Avengers: Age of Ultron
The Avengers
Furious 7
Jurassic World
Titanic
Avatar


1e9
profit
2.5
2.0
1.5
1.0
0.5
0.0
Kevin Feige
Barrie M. Osborne
Christopher Meledandri
Peter Del Vecho
Kevin Feige
Kevin Feige
Michael Fottrell
Frank Marshall
James Cameron
James Cameron
producer


No of movies
0
5
10
15
20
25
30
35
40
Mark Wahlberg
John Travolta
Eddie Murphy
Brad Pitt
Johnny Depp
Denzel Washington
Nicolas Cage
Matt Damon
Bruce Willis
Robert De Niro
Actor

Category Four: Correlation for financial analysis and histogram





Conclusions

Category one

Drama genre is found in most movies may be because it is suitable to many audience ages unlike animation and action movies

Highest profit movie is Avatar may be because it was one of his kind back there

The most used original language is English and it was expected as one of the most used languages

Category two

Movies industry witnesses massive improvements in both budget and revenue aspects due to marketing campaigns in addition to the decreased value of money over the years

Max vote of movies every year is fluctuating and generally increased over the last decade.

Popularity of movies increased over the last years may be due to availability of movies everywhere

Category three

Popular movies are not the same genres and there a considerable difference between first and last movie.

Longest movies are generally with low popularity as not many people tend to spend a long time watching single movie.

Top rated movies have little standard deviation in rating values and mostly drama movies

Most rated movies related to old directors may be due to their experiences in the field.

James Cameron was brilliant for directing the two most profit movies over the years.

The most appeared actors in movies are generally old but note that not all old actors are popular in movies

Category Four

The more money you spend on a movie, the more money you will gain as the correlation show positive correlation between revenue and budget

Budget spend on movies falls most frequently in the range of 0-0.4*10^8 dollars

Communication

This step will be done by uploading the project files to UDACITY platform to be reviewed by out mentors