Name: Rishu Singh

Date: 17th June 2019

Place: New Delhi, INDIA

UDACITY

Data Analysis Nanodegree

Project 02:

Investigate a Dataset (tmdb-movie)

Overview:

To complete my Data Analysis project I selected TMDB movies dataset.

This data set contains 10 about 10 thousand movie collection with 21 columns of each.

Goals:

1. Importing tmdb-movie csv file
2. Remove unwanted columns
3. Checking for duplicated entries and removing if found any
4. Changing release date format
5. Replacing zeros with NAN in runtime, budget and revenue columns and drop those entries.
6. Changing format of revenue and budget.

Tools Used:

1. Python: For calculating moving average and plotting line chart.
2. ANACONDA - Jupyter Notebook: For writing python code and making observations.
3. Excel: Having a look at the data and writing project.

Packages imported:

1. Pandas
2. Numpy
3. Matplotlib
4. Datetime

STEP 1 -

Importing tmdb_movie csv file

```
movie = pd.read_csv('tmdb-movies.csv')
```

STEP 2 - Remove unwanted columns

```
cols = ['id', 'imdb_id', 'popularity', 'budget_adj', 'revenue_adj', 'homepage', 'keywords', 'overview', 'production_companies', 'vote_count', 'vote_average']

# drop columns

movie = movie.drop(cols,1)
```

STEP 3 - Checking for duplicated entries and removing if found any

```
# checking duplicate entries
    movie.duplicated().sum()

# Delete duplicate entry
    movie.drop_duplicates(inplace=True)
```

STEP 4 - Changing release date format

```
    movie['release_date'] = pd.to_datetime(movie['release_date'])
```

STEP 5 - Replacing zeros with NAN in runtime, budget and revenue columns and drop those entries.

```
# Replacing zeros with NAN\
    list_cols = ['runtime', 'budget', 'revenue']
    movie[list_cols] = movie[list_cols].replace(0, np.NAN)

# Dropping NAN rows
    movie.dropna(inplace=True)
```

STEP 6 - Changing format of revenue and budget.

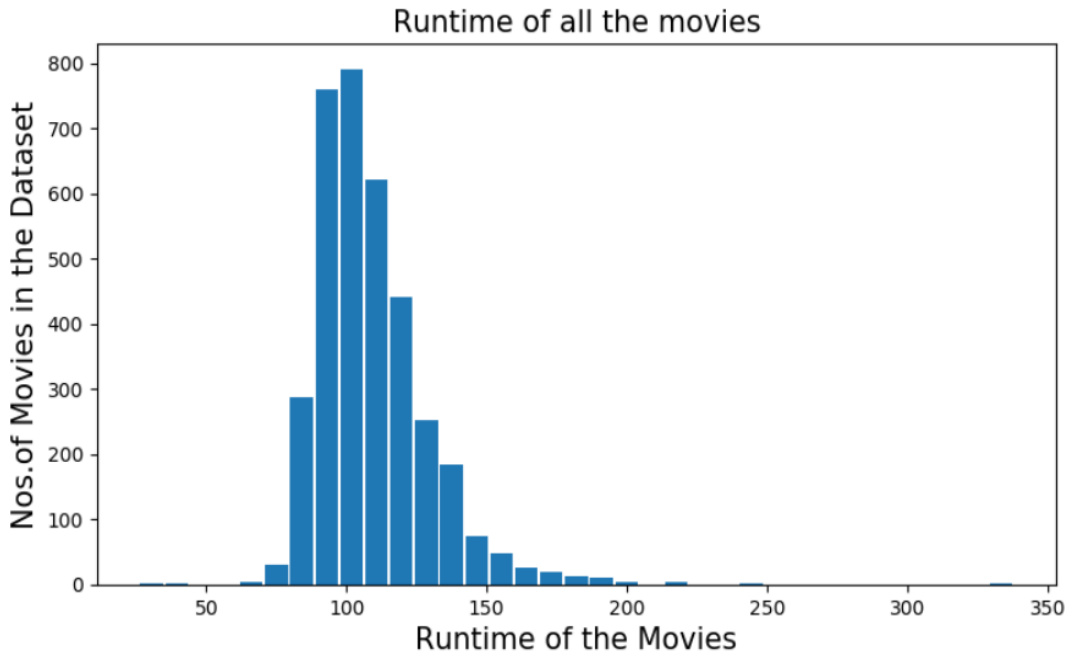
```
# changing datatype fo revenue and budget
    columns = ['budget', 'revenue']
    movie[columns] = movie[columns].applymap(np.int64)

# verify changes
    movie.dtypes
```

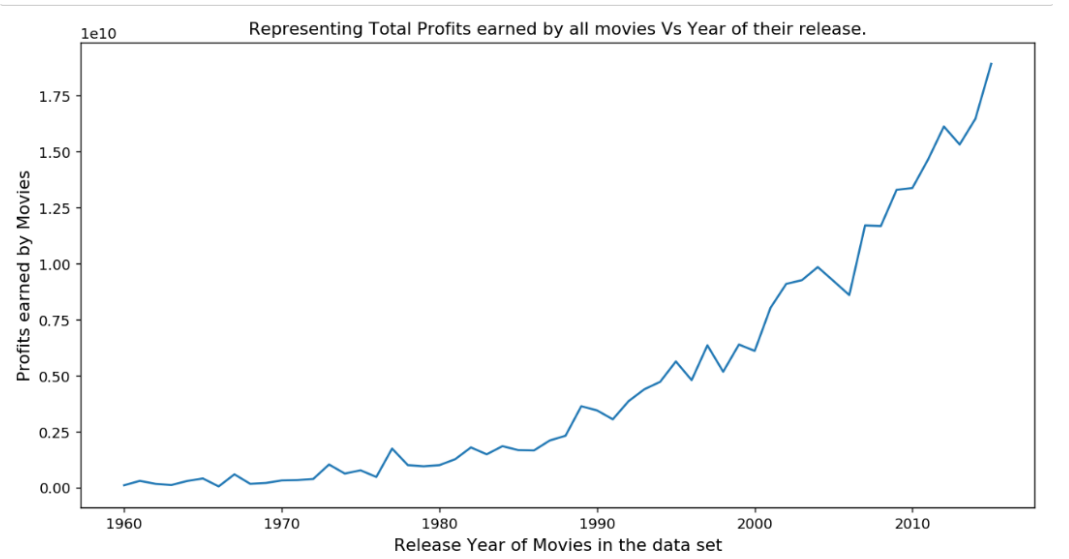
As now the Data is clean and trim for further process.

These are the questions came in my mind regarding this dataset.

1. Calculating the profit of each movie
2. Movies which had most and least profit
3. Movies with largest and lowest budgets
4. Movies with most and least earned revenue
5. Movies with longest and shortest runtime
6. Average runtime of the movies
7. Year of release vs Profitability - To find which year made the highest profit?
8. Most Successful Genres
9. Most Frequent Cast
10. Average Budget of the movies
11. Average Revenue earned by the movies
12. Average duration of the movies

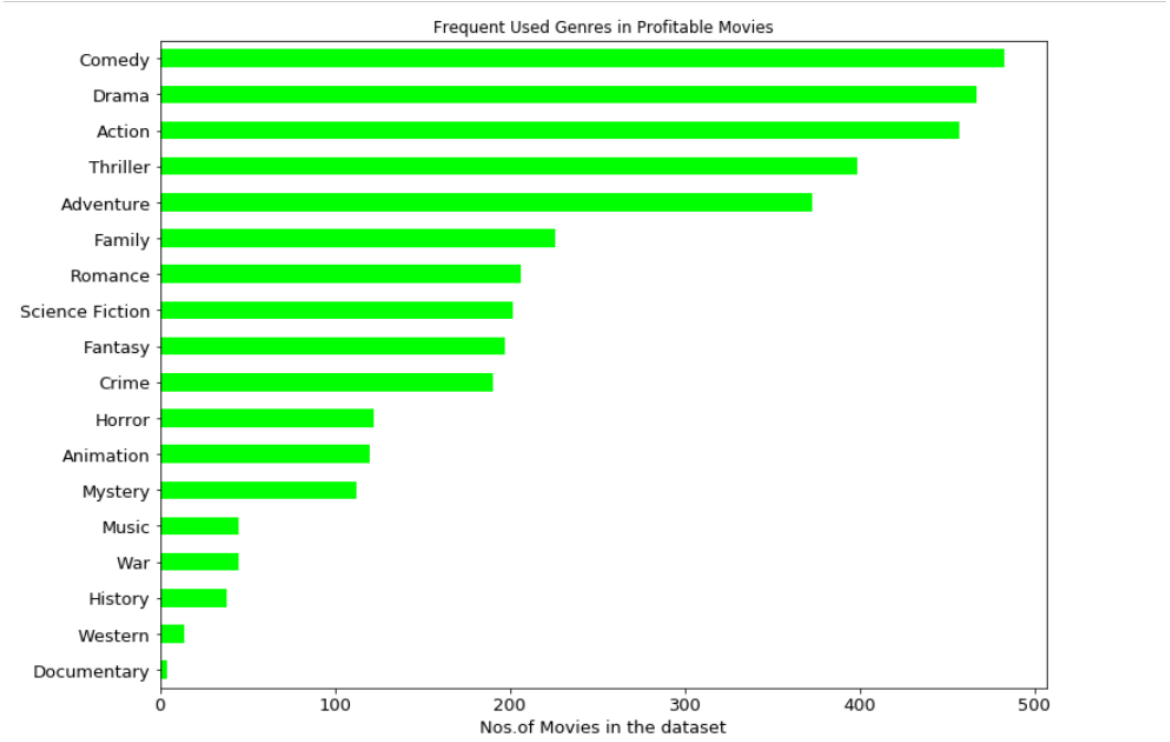
Sharing quick matplotlib results screenshot





[32]:
```
# To find which year made the highest profit
profits_year.idxmax()
```

Out[32]: 2015



Final Conclusion:

On the basis of the above analysis we can conclude following:

1. Average Budget must be around 60 million dollar.

2. Average duration of the movie must be 113 minutes.

3. Any one of these should be in the cast :Tom Cruise, Brad Pitt, Tom Hanks, Sylvester Stallone,Cameron Diaz

4. Genre must be : Action, Adventure, Thriller, Comedy, Drama.

By doing all this the movie might be one of the hits and hence can earn an average revenue of around 255 million dollar.