Project #2: TMDB movie data

Date: Sep20-2019
Name: Sami Adham
Program: Data Analysis Nanodegree

Project Task:

The project is to show abilities to investigate dataset by using dataset analysis process.

Introduction:

The dataset was taken from the movie database (TMDB) it is like a movies hub of all the movies over the years with all information about each move. In this analysis we will analyze the profitability of file industry over the years. and to do that we will focus on the following.

- Revenue
- Budget or Cost
- Profit
- Genre
- Release Date
- Release Year
- Title of Movies

We want to understand is film industry is profitable or not. If yes, where can I suggest to invest your money.

Goals:

1. Apply Data Analysis Process methodology.
2. How to gather, assess and clean data and tried to extract useful information
3. How to explored data and get intuition regarding data
4. Abilities to work with visualization to emphasis your point of view
5. Try to optimize python code abilities.

Tools:

1. python through Jupyter notebook (ANACONDA) to take a chance to practice in this first project for future projects.
2. Upload work in github." https://github.com/SamiAdham/TMDb-movie-data/blob/master/investigate-a-dataset-%5BSami%20Adham%5D.ipynb"
3. Microsoft Word to prepared 'PDF'report

Project Steps:

- Step#1: generate Questions that help me to analysis dataset
- Step#2: Data Wrangling
  - Gathering data from TMDB.CVS
  - Assess Dataset
  - Cleaning Data
- Step#3: Exploratory Data Analysis
  - Answer Question
  - Create some calculated factor that support our tasks
- Step#4: Conclusion and comment

Step#1: Generate Questions that help me to analysis dataset:

1-highst and lowest revenue
2-highst and lowest budget
3-highst and lowest net profit
4-longest and shortest runtime
5-What is an average runtime
6-What is an average profit
7-What is an average revenue
8-What is an average budget
9-relationship between profit and budget over the years (ROI)
10-Top genres over the years
11-Top cast over the years

Step#2: Data Wrangling

A) Assess Data:

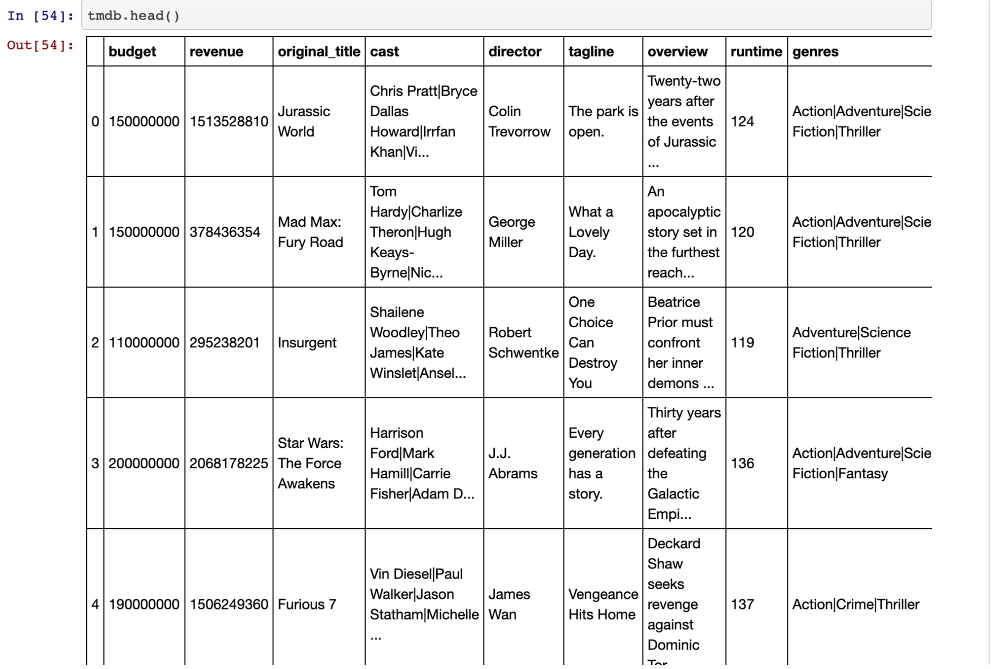
- `tmdb.head() # to see What do we need from dataset and drop unnecessary columns`
- `tmdb.info() # to look at data type`

B) Cleaning Data:

```
#1-Removing Unused features
del_column=['id', 'imdb_id', 'popularity' ,'homepage','keywords','homepage','production_companies','vote_count','vote_average','budget_adj','revenue_adj']


tmdb=tmdb.drop(del_column,axis=1)
print('There is {} Column in the TMDB'.format(len(tmdb.columns)))
tmdb.head()

#2- Remove zero's and nan from data as we need it. There is no revenue or budget = 0
tmdb['budget']=tmdb['budget'].replace(0,np.NAN)
```

```
tmdb['revenue']=tmdb['revenue'].replace(0,np.NAN)
tmdb.dropna(inplace=True)
tmdb.info()

#3.1- chech duplicate
print('There is {} duplicated rows in the TMDB'.format(sum(tmdb.duplicated())))

#3.2  We need to remove duplicated row by using drop
tmdb.drop_duplicates(inplace=True)
print('There is {} duplicated rows in the TMDB'.format(sum(tmdb.duplicated())))

#4- change data type and format
tmdb['release_date']=pd.to_datetime(tmdb['release_date'])

# change datat type of rev budj
tmdb['budget']=tmdb['budget'].apply(np.int64)
tmdb['revenue']=tmdb['revenue'].apply(np.int64)
tmdb.info()
```

The cleaned dataset will be as below screenshot:



```
In [54]: tmdb.head()
Out[54]:
```

	budget	revenue	original_title	cast	director	tagline	overview	runtime	genres
0	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	Twenty-two years after the events of Jurassic ...	124	Action|Adventure|Scie Fiction|Thriller
1	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	An apocalyptic story set in the furthest reach...	120	Action|Adventure|Scie Fiction|Thriller
2	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	One Choice Can Destroy You	Beatrice Prior must confront her inner demons ...	119	Adventure|Science Fiction|Thriller
3	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	Every generation has a story.	Thirty years after defeating the Galactic Empi...	136	Action|Adventure|Scie Fiction|Fantasy
4	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	Vengeance Hits Home	Deckard Shaw seeks revenge against Dominic Tor	137	Action|Crime|Thriller

Step#3: Exploratory Data

Answering Question by analyzing Data that related to question. Please open my git hut link for more information regarding this step.

Step#4: Conclusion and comment

Findings:

1-Revenue could not ever be 0's so we need to remove it

2- The most movies that generate revenue is [Avatar] and lowest is [Mallrats]

3- The most moves that cost high budget is [The Warrior's Way] and lowest is [Lost & Found]

4- The most movies that is profitable over the years is [Avatar] and lowest is [The Warrior's Way]

5- The longest runtime movies was [Carlos] and shortest movies is [Mickey's Christmas Carol]

6- Averages:

-Average runtime is: [109.12290033594626]
-Average profit is: [75,118,992.06]
-Average revenue is: [113,833,739.16]
-Average budget is: [38,714,747.10]

7- Top 5 Genres are:

-Drama
-Comedy
-Thriller
-Action
-Adventure

8- Top 5 most successful cast are:

-Robert De Niro
-Samuel L. Jackson
-Nicolas Cage
-Matt Damon
-Tom Hanks

Conclusion and Opinion:

1. The Return of investment (ROI) in the last 15 years increased significantly as shown in the ROI graph.
2. At the beginning of Movies industry, the business was struggled maybe that because of lack of technology and using costly materials to create scenes.
3. Around 2009 the cost become stable and start to decreased (Technology in film making become easier and cheaper)
4. In genres section we exclude immature file maker to get better top genres by remove all movies with revenue less or equal to 40M $

Limitations:

1. In revenue and budget data I find some number which is imposable such as revenue 5 dollars. It would be better to understand some hint about these number to take decision should i delete it or leave it.

2. In case we exclude low revenue movie the genres will change and that will affect our decision which is where should i invest my money.
3. Currency exchange rate is not considered at this dataset regarding revenue and budget.

References:

GitHub:
https://github.com/SamiAdham/TMDb-movie-data/blob/master/investigate-a-dataset-%5BSami%20Adham%5D.ipynb”