# **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.

# 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

1. Assess Data:

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

1. 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'])

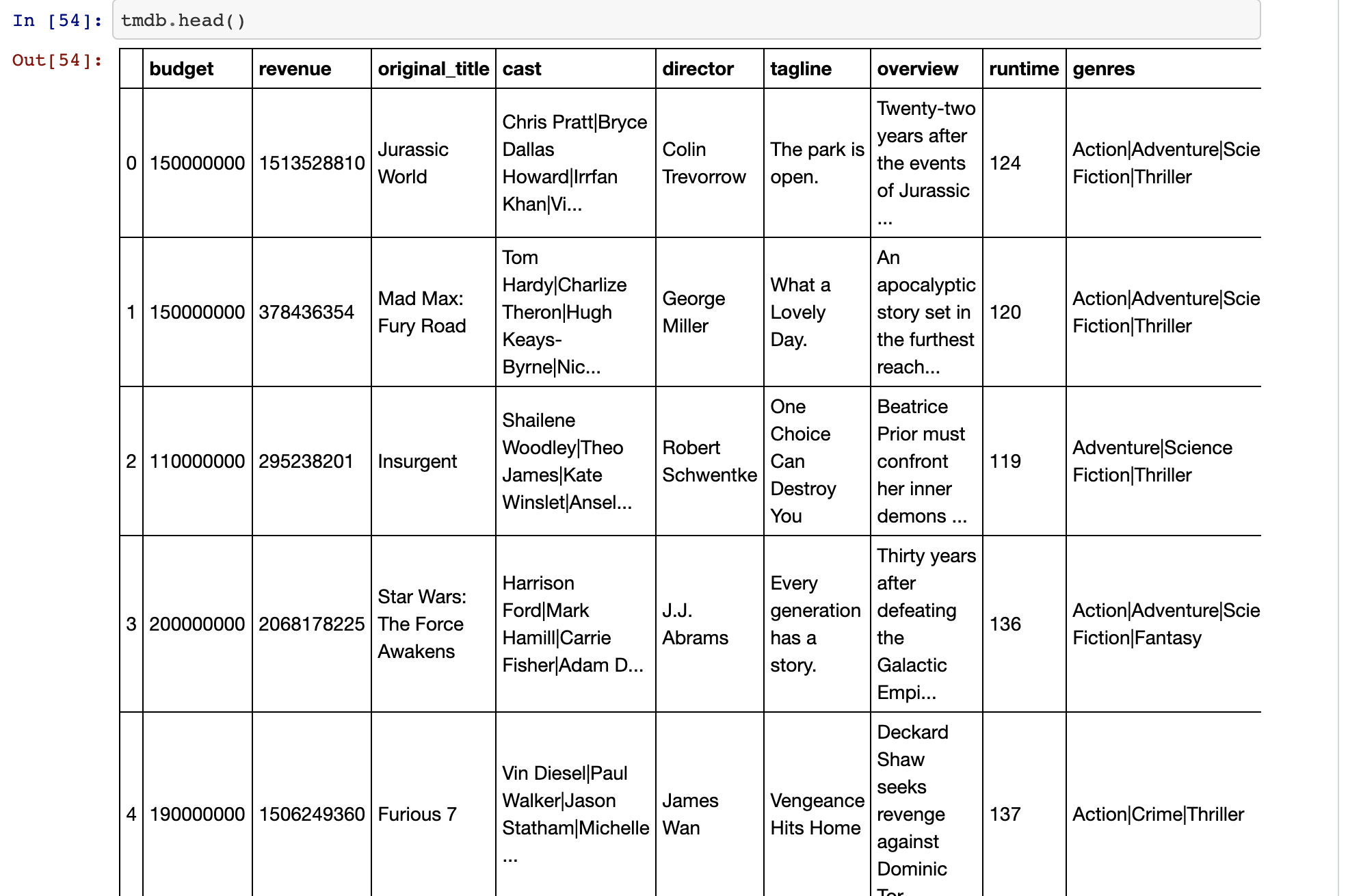
*# change datat type of rev budj*

tmdb['budget']=tmdb['budget'].apply(np.int64)

tmdb['revenue']=tmdb['revenue'].apply(np.int64)

tmdb.info()

The cleaned dataset will be as below screenshot:



### 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 $

**References:**

GitHub:

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