Investigate_a_Dataset

November 3, 2020

0.1 Project 2 : Investigate a Dataset (TMDB Movie Analysis)

0.2 Table of Contents

## Introduction

In this project, I have analysed TMDB dataset, which is available on Kaggle. This dataset contains information related to arround 10,000 movies collected from TMDB (www.themoviedb.org).It includes information about movie's viewer's rating,budget, revenue, genres, production companies, director, casting, keywords associated with movies, popularity of the movies and runtime.

This dataset can help to understand various factors like profitability, the trend around runtime, popularity over the years, popular genres for the profitability, connection between popularity ratings and profit; reveal information like profitable directors, casts and production companies over the span.

I am focusing on answering the following questions for this Movie dataset:

Q1. List of Generic questions based on the datset that can be answered are:

1. Which movie had the highest and lowest profit?
2. Which movie had the highest and lowest budget?
3. Which movie had the highest and lowest revenue?
4. What is the average runtime of all movies?
5. Which duration movies are most liked by the audiences according to their popularity?

Q2. List of Questions that can be answered based on the Profit of movies making more then 25M Dollars:

1. What is the average budget of the movie?
2. What is the average revenue of the movie?
3. Which are the most frequent cast involved?
4. Which are the successful genres?

```
In [1]: # Load your data and print out a few lines. Perform operations to inspect data
```

```
#importing packages
import numpy as np
import pandas as pd
import csv
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

## Data Wrangling

During this step, we are going to import the csv file and display its main information. We will perform the following steps:

- Display the info of the dataset and get the idea of the size, number of records and number of columns
- Familiarize with the dataset and find any unusal values

```
In [2]: # After discussing the structure of the data and any problems that need to be
        # cleaned, perform those cleaning steps in the second part of this section.

        #importing csv dataset

        df = pd.read_csv('http://d17h27t6h515a5.cloudfront.net/topher/2017/October/59dd1c4c_tmdb
        print("Original TMDB Dataset contains (Rows,Columns) : ",df.shape)
        df.shape
```

```
Original TMDB Dataset contains (Rows,Columns) :  (10866, 21)
```

```
Out[2]: (10866, 21)
```

```
In [3]: df.head(2)
```

```
Out[3]:        id    imdb_id  popularity     budget     revenue      original_title  \
        0  135397  tt0369610   32.985763  150000000  1513528810      Jurassic World
        1   76341  tt1392190   28.419936  150000000   378436354  Mad Max: Fury Road

                                                        cast  \
        0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
        1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...

                                homepage          director             taglne  \
        0  http://www.jurassicworld.com/  Colin Trevorrow   The park is open.
        1    http://www.madmaxmovie.com/    George Miller  What a Lovely Day.

               ...                                              overview runtime  \
        0      ...       Twenty-two years after the events of Jurassic ...     124
        1      ...       An apocalyptic story set in the furthest reach...     120

                                                 genres  \
0  Action|Adventure|Science Fiction|Thriller
1  Action|Adventure|Science Fiction|Thriller

                               production_companies release_date vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15       5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15       6185

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08

[2 rows x 21 columns]
```

```
In [4]: #Explore the information of the dataset

        print("Quick glance at the dataset for some statistical values: \n\n")
        df.describe()
```

```
Quick glance at the dataset for some statistical values:
```

```
Out[4]:                   id    popularity        budget       revenue       runtime  \
        count   10866.000000  10866.000000  1.086600e+04  1.086600e+04  10866.000000
        mean    66064.177434      0.646441  1.462570e+07  3.982332e+07    102.070863
        std     92130.136561      1.000185  3.091321e+07  1.170035e+08     31.381405
        min         5.000000      0.000065  0.000000e+00  0.000000e+00      0.000000
        25%     10596.250000      0.207583  0.000000e+00  0.000000e+00     90.000000
        50%     20669.000000      0.383856  0.000000e+00  0.000000e+00     99.000000
        75%     75610.000000      0.713817  1.500000e+07  2.400000e+07    111.000000
        max    417859.000000     32.985763  4.250000e+08  2.781506e+09    900.000000

                 vote_count  vote_average  release_year    budget_adj   revenue_adj
        count  10866.000000  10866.000000  10866.000000  1.086600e+04  1.086600e+04
        mean     217.389748      5.974922   2001.322658  1.755104e+07  5.136436e+07
        std      575.619058      0.935142     12.812941  3.430616e+07  1.446325e+08
        min       10.000000      1.500000   1960.000000  0.000000e+00  0.000000e+00
        25%       17.000000      5.400000   1995.000000  0.000000e+00  0.000000e+00
        50%       38.000000      6.000000   2006.000000  0.000000e+00  0.000000e+00
        75%      145.750000      6.600000   2011.000000  2.085325e+07  3.369710e+07
        max     9767.000000      9.200000   2015.000000  4.250000e+08  2.827124e+09
```

```
In [5]: #Check column names and datatypes

        print("Check columns and their data types: \n\n")
        df.info()
```

```
Check columns and their data types:
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
budget                  10866 non-null int64
revenue                 10866 non-null int64
original_title          10866 non-null object
cast                    10790 non-null object
homepage                2936 non-null object
director                10822 non-null object
tagline                 8042 non-null object
keywords                9373 non-null object
overview                10862 non-null object
runtime                 10866 non-null int64
genres                  10843 non-null object
production_companies    9836 non-null object
release_date            10866 non-null object
vote_count              10866 non-null int64
vote_average            10866 non-null float64
release_year            10866 non-null int64
budget_adj              10866 non-null float64
revenue_adj             10866 non-null float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

1 Data Cleaning and Further analysis

1.0.1 Observations based on accessing the TMDB dataset file

1. The columns ′id′, ′imdb_id′,′budget_adj′, ′revenue_adj′, ′homepage′, ′keywords′, ′tagline′,′overview′ are not relevant for the analysis, so we can remove them.
2. Lets delete the one duplicated row that we have in our dataset.
3. There are lots of movies where the budget or revenue have a value of ′0′ which means that the values of those movies has not been recorded. So we need to discard this rows, since we cannot calculate profit of such movies
4. The ′release_date′ column must be converted into date format.
5. Convert budget and revenue column to int datatype.
6. Replace runtime value of 0 to NAN, Since it will affect the result..
7. The dataset has not provided the currency for columns we will be dealing with hence we will assume it is in dollars.
8. Even the vote count is not same for all the movies and hence this affects the vote average column.
9. There are some invalid characters in cast and keywords, in our analysis of top keywords and casts, those are not creating any issues, so we have not cleaned those columns.

First Step : The columns 'id', 'imdb_id','budget_adj', 'revenue_adj', 'homepage', 'keywords', 'tagline','overview' are not relevant for the analysis, so we can remove them.

```
In [6]: # Columns that needs to be deleted
        deleted_columns = [ 'id', 'imdb_id','budget_adj', 'revenue_adj', 'homepage', 'keywords',

        # Drop the columns from the database
        df.drop(deleted_columns, axis=1, inplace=True)

        # Lets look at the new dataset
        df.head()
```

```
Out[6]:    popularity     budget     revenue                original_title  \
        0   32.985763  150000000  1513528810                Jurassic World
        1   28.419936  150000000   378436354            Mad Max: Fury Road
        2   13.112507  110000000   295238201                     Insurgent
        3   11.173104  200000000  2068178225  Star Wars: The Force Awakens
        4    9.335014  190000000  1506249360                     Furious 7

                                                        cast          director  \
        0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...   Colin Trevorrow
        1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...     George Miller
        2  Shailene Woodley|Theo James|Kate Winslet|Ansel...  Robert Schwentke
        3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...       J.J. Abrams
        4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...         James Wan

           runtime                                      genres  \
        0      124  Action|Adventure|Science Fiction|Thriller
        1      120  Action|Adventure|Science Fiction|Thriller
        2      119         Adventure|Science Fiction|Thriller
        3      136   Action|Adventure|Science Fiction|Fantasy
        4      137                      Action|Crime|Thriller

                                        production_companies release_date  vote_count  \
        0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15        5562
        1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15        6185
        2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15        2480
        3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15        5292
        4  Universal Pictures|Original Film|Media Rights ...       4/1/15        2947

           vote_average  release_year
        0           6.5          2015
        1           7.1          2015
        2           6.3          2015
        3           7.5          2015
        4           7.3          2015
```

Let's see the number of entries in our dataset now.

```
In [7]: # Store rows and columns using shape function.
        rows, col = df.shape

        #since rows includes count of a header, we need to remove its count.
        print('We have {} total rows and {} columns.'.format(rows-1, col))
```

```
We have 10865 total rows and 13 columns.
```

```
In [8]: # Find duplicates in the row
        sum(df.duplicated())
```

```
Out[8]: 1
```

Second Step : Lets delete the one duplicated row that we have

```
In [9]: # Drop duplicate rows but keep the first one
        df.drop_duplicates(keep = 'first', inplace = True)

        # Store rows and columns using shape function.
        rows, col = df.shape

        print('Now we have {} total rows and {} columns.'.format(rows-1, col))
```

```
Now we have 10864 total rows and 13 columns.
```

Third Step : There are lots of movies where the budget or revenue have a value of '0' which means that the values of those movies has not been recorded. So we need to discard these rows,as profit cannot be calculated

```
In [10]: # Columns that need to be checked.
         columns = ['budget', 'revenue']

         # Replace 0 with NAN
         df[columns] = df[columns].replace(0, np.NaN)

         # Drop rows which contains NAN
         df.dropna(subset = columns, inplace = True)

         rows, col = df.shape
         print('We now have only {} rows.'.format(rows-1))
```

```
We now have only 3853 rows.
```

Fourth Step : The 'release_date' column must be converted into date format.

```
In [11]: # Changing the format of dates:

         df['release_date'] = pd.to_datetime(df['release_date'], errors = 'ignore')
         print("Year range - ", df.release_year.min(), "to" ,df.release_year.max())
```

```
Year range -  1960 to 2015
```

Fifth Step : Convert budget and revenue column to int datatype.

```
In [12]: # Columns to convert datatype of
         columns = ['budget', 'revenue']

         # Convert budget and revenue column to int datatype
         df[columns] = df[columns].applymap(np.int64)

         # Lets look at the new datatype
         df.dtypes
```

```
Out[12]: popularity                        float64
         budget                              int64
         revenue                             int64
         original_title                     object
         cast                               object
         director                           object
         runtime                             int64
         genres                             object
         production_companies               object
         release_date               datetime64[ns]
         vote_count                          int64
         vote_average                      float64
         release_year                        int64
         dtype: object
```

Sixth : Replace runtime value of 0 to NAN, Since it will affect the result.

```
In [13]: # Replace runtime value of 0 to NAN, Since it will affect the result.
         df['runtime'] = df['runtime'].replace(0, np.NaN)

         # Check the stats of dataset
         df.describe()
```

```
Out[13]:         popularity        budget       revenue      runtime   vote_count  \
         count  3854.000000  3.854000e+03  3.854000e+03  3854.000000  3854.000000
         mean      1.191554  3.720370e+07  1.076866e+08   109.220291   527.720291
         std       1.475162  4.220822e+07  1.765393e+08    19.922820   879.956821
         min       0.001117  1.000000e+00  2.000000e+00    15.000000    10.000000
         25%       0.462368  1.000000e+07  1.360003e+07    95.000000    71.000000
         50%       0.797511  2.400000e+07  4.480000e+07   106.000000   204.000000
         75%       1.368324  5.000000e+07  1.242125e+08   119.000000   580.000000
         max      32.985763  4.250000e+08  2.781506e+09   338.000000  9767.000000

                vote_average  release_year
         count   3854.000000   3854.000000
```

```
mean         6.168163   2001.261028
std          0.794920     11.282575
min          2.200000   1960.000000
25%          5.700000   1995.000000
50%          6.200000   2004.000000
75%          6.700000   2010.000000
max          8.400000   2015.000000
```

## Exploratory Data Analysis

We will now compute statistics and create visualizations with the goal of addressing the research questions that we posed in the Introduction section.

1.0.2 Research Question 1.1 Which movie had the highest and lowest profit?

Before deep diving into answering the questions, lets first add a column for **Profit** in our dataset.

```
In [14]: # To calculate profit, we need to substract the budget from the revenue.
         df['profit'] = df['revenue'] - df['budget']

         # Lets look at the new dataset
         df.head()
```

```
Out[14]:    popularity     budget     revenue                original_title  \
         0   32.985763  150000000  1513528810                Jurassic World
         1   28.419936  150000000   378436354            Mad Max: Fury Road
         2   13.112507  110000000   295238201                     Insurgent
         3   11.173104  200000000  2068178225  Star Wars: The Force Awakens
         4    9.335014  190000000  1506249360                     Furious 7

                                                         cast          director  \
         0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...   Colin Trevorrow
         1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...     George Miller
         2  Shailene Woodley|Theo James|Kate Winslet|Ansel...  Robert Schwentke
         3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...       J.J. Abrams
         4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...         James Wan

            runtime                                    genres  \
         0      124  Action|Adventure|Science Fiction|Thriller
         1      120  Action|Adventure|Science Fiction|Thriller
         2      119         Adventure|Science Fiction|Thriller
         3      136   Action|Adventure|Science Fiction|Fantasy
         4      137                      Action|Crime|Thriller

                                         production_companies release_date  vote_count  \
         0  Universal Studios|Amblin Entertainment|Legenda...   2015-06-09        5562
         1  Village Roadshow Pictures|Kennedy Miller Produ...   2015-05-13        6185
         2  Summit Entertainment|Mandeville Films|Red Wago...   2015-03-18        2480
         3          Lucasfilm|Truenorth Productions|Bad Robot   2015-12-15        5292
         4  Universal Pictures|Original Film|Media Rights ...   2015-04-01        2947
```

```
   vote_average  release_year      profit
0           6.5          2015  1363528810
1           7.1          2015   228436354
2           6.3          2015   185238201
3           7.5          2015  1868178225
4           7.3          2015  1316249360
```

```
In [15]: # Movie with the highest profit

         df.loc[df['profit'].idxmax()]
```

```
Out[15]: popularity                                                    9.43277
         budget                                                      237000000
         revenue                                                    2781505847
         original_title                                                 Avatar
         cast                    Sam Worthington|Zoe Saldana|Sigourney Weaver|S...
         director                                               James Cameron
         runtime                                                           162
         genres                           Action|Adventure|Fantasy|Science Fiction
         production_companies    Ingenious Film Partners|Twentieth Century Fox ...
         release_date                                      2009-12-10 00:00:00
         vote_count                                                       8458
         vote_average                                                      7.1
         release_year                                                     2009
         profit                                                     2544505847
         Name: 1386, dtype: object
```

```
In [16]: # Movie with the lowest profit

         df.loc[df['profit'].idxmin()]
```

```
Out[16]: popularity                                                    0.25054
         budget                                                      425000000
         revenue                                                      11087569
         original_title                                         The Warrior's Way
         cast                    Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
         director                                                  Sngmoo Lee
         runtime                                                           100
         genres                          Adventure|Fantasy|Action|Western|Thriller
         production_companies                          Boram Entertainment Inc.
         release_date                                      2010-12-02 00:00:00
         vote_count                                                         74
         vote_average                                                      6.4
         release_year                                                     2010
         profit                                                     -413912431
         Name: 2244, dtype: object
```

Which movie had the highest and lowest profit?

As we can see that **Avatar** movie Directed by James Cameron earns the highest profit in all, making over 2.5B in profit whereas the highest loss incurred is **The Warrior's Way** with more than 400M directed by Sngmoo Lee.

1.0.3 Research Question 1.2 Which movie had the highest and lowest budget?

```
In [17]: # Movie with highest budget
         df.loc[df['budget'].idxmax()]
```

```
Out[17]: popularity                                                        0.25054
         budget                                                          425000000
         revenue                                                          11087569
         original_title                                          The Warrior's Way
         cast                    Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
         director                                                       Sngmoo Lee
         runtime                                                               100
         genres                          Adventure|Fantasy|Action|Western|Thriller
         production_companies                             Boram Entertainment Inc.
         release_date                                          2010-12-02 00:00:00
         vote_count                                                             74
         vote_average                                                          6.4
         release_year                                                         2010
         profit                                                         -413912431
         Name: 2244, dtype: object
```

```
In [18]: # Movie with lowest budget
         df.loc[df['budget'].idxmin()]
```

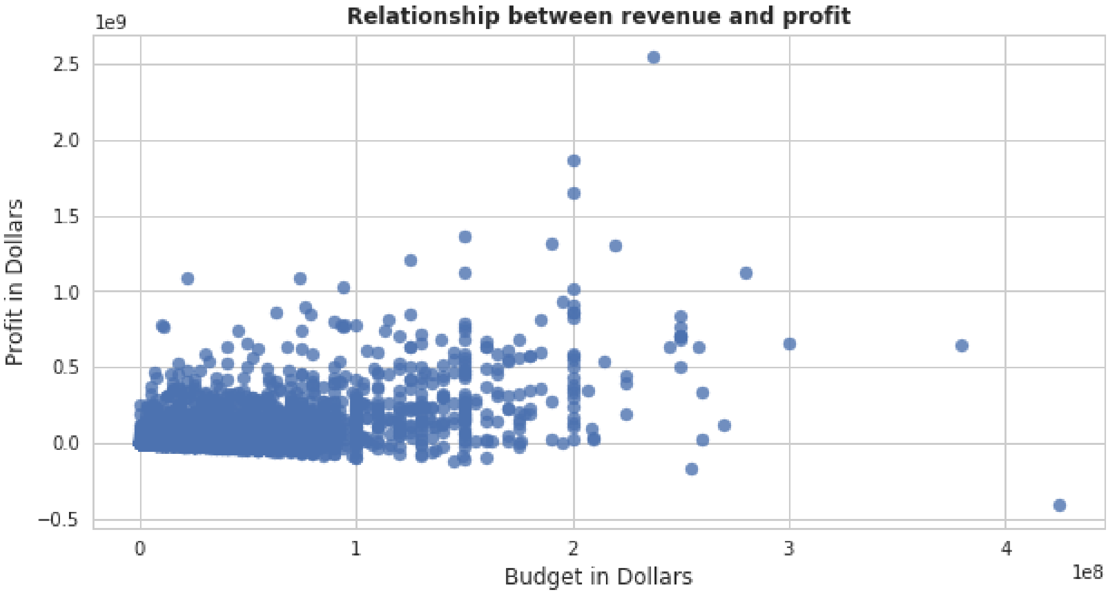
```
Out[18]: popularity                                                       0.090186
         budget                                                                  1
         revenue                                                               100
         original_title                                               Lost & Found
         cast                    David Spade|Sophie Marceau|Ever Carradine|Step...
         director                                                     Jeff Pollack
         runtime                                                                95
         genres                                                     Comedy|Romance
         production_companies             Alcon Entertainment|Dinamo Entertainment
         release_date                                          1999-04-23 00:00:00
         vote_count                                                             14
         vote_average                                                          4.8
         release_year                                                         1999
         profit                                                                 99
         Name: 2618, dtype: object
```

Which movie had the highest and lowest budget?ű

As we can see that, the movie with the highest budget was **The Warrior's Way** with budget of 425000000 dollars and the movie with the lowest was **Lost & Found** with budget of 1 dollar

Let's see if there's a relationship between the movie's budget and profit.

```
In [92]: #Plotting the Scatter plot
         # x-axis
         plt.xlabel('Budget in Dollars',fontsize=12)
         # y-axis
         plt.ylabel('Profit in Dollars',fontsize=12)
         # Title of the histogram
         plt.title('Relationship between revenue and profit', fontweight="bold", fontsize=12)
         plt.scatter(df['budget'], df['profit'], alpha=0.8)
         plt.show()
         #setup the figure size.
         sns.set(rc={'figure.figsize':(10,5)})
         sns.set_style("whitegrid")
```



We can see that there no as such relationship between budget and profits, But yes there are very less movies which didnt make profit when the budget was more then 20M Dollar.

1.0.4 Research Question 1.3 Which movie had the highest and lowest revenue?

```
In [33]: # Movie with highest revenue
         df.loc[df['revenue'].idxmax()]
```

```
Out[33]: popularity                                                 9.43277
         budget                                                   237000000
         revenue                                                        2781505847
         original_title                                                     Avatar
         cast                    Sam Worthington|Zoe Saldana|Sigourney Weaver|S...
         director                                                    James Cameron
         runtime                                                               162
         genres                           Action|Adventure|Fantasy|Science Fiction
         production_companies    Ingenious Film Partners|Twentieth Century Fox ...
         release_date                                          2009-12-10 00:00:00
         vote_count                                                           8458
         vote_average                                                          7.1
         release_year                                                         2009
         profit                                                         2544505847
         Name: 1386, dtype: object
```

```
In [34]: # Movie with lowest revenue
         df.loc[df['revenue'].idxmin()]
```

```
Out[34]: popularity                                                       0.462609
         budget                                                            6000000
         revenue                                                                 2
         original_title                                            Shattered Glass
         cast                    Hayden Christensen|Peter Sarsgaard|ChloÃń Sevi...
         director                                                        Billy Ray
         runtime                                                                94
         genres                                                      Drama|History
         production_companies    Lions Gate Films|Cruise/Wagner Productions|Bau...
         release_date                                          2003-11-14 00:00:00
         vote_count                                                             46
         vote_average                                                          6.4
         release_year                                                         2003
         profit                                                           -5999998
         Name: 5067, dtype: object
```

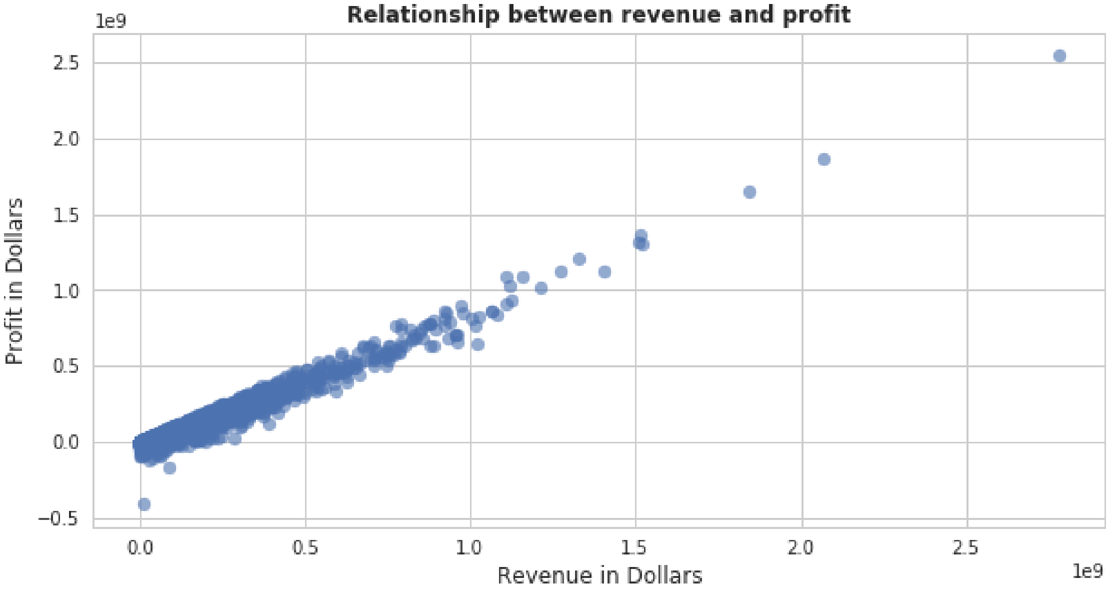
Which movie had the highest and lowest revenue?

> As we can see that, the movie **Avatar** had the highest revenue with revenue of 2781505847 dollars whereas the movie **Shattered Glass** had the lowest revenue with revenue of 2 dollars

Let us check if there a relation between the Revenue and Profit

```
In [91]: # x-axis
         plt.xlabel('Revenue in Dollars',fontsize=12)
         # y-axis
         plt.ylabel('Profit in Dollars',fontsize=12)
         # Title of the histogram
         plt.title('Relationship between revenue and profit',fontweight="bold", fontsize=12)
         plt.scatter(df['revenue'], df['profit'], alpha=0.6)
         plt.show()
```

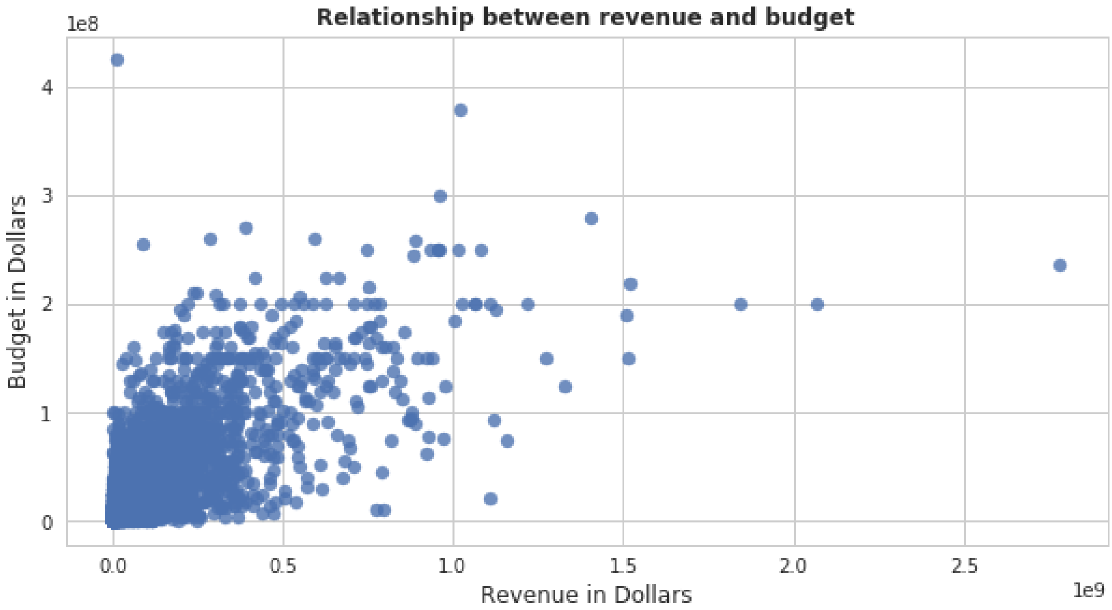
```
#setup the figure size.
sns.set(rc={'figure.figsize':(10,5)})
sns.set_style("whitegrid")
```



We can see that there is a strong relationship between profit and revenue, higher the revenue, higher the profit.

Let us check if there a relation between the Budget and Revenue

```
In [90]: # x-axis
         plt.xlabel('Revenue in Dollars',fontsize=12)
         # y-axis
         plt.ylabel('Budget in Dollars',fontsize=12)
         # Title of the histogram
         plt.title('Relationship between revenue and budget', fontweight="bold", fontsize=12)
         plt.scatter(df['revenue'], df['budget'], alpha=0.8)
         plt.show()
         #setup the figure size.
         sns.set(rc={'figure.figsize':(10,5)})
         sns.set_style("whitegrid")
```



From the above plot what we observe is, Most of the movies have a revenue upto 50M Dollars. ### Research Question 1.4 What is the average runtime of all movies?

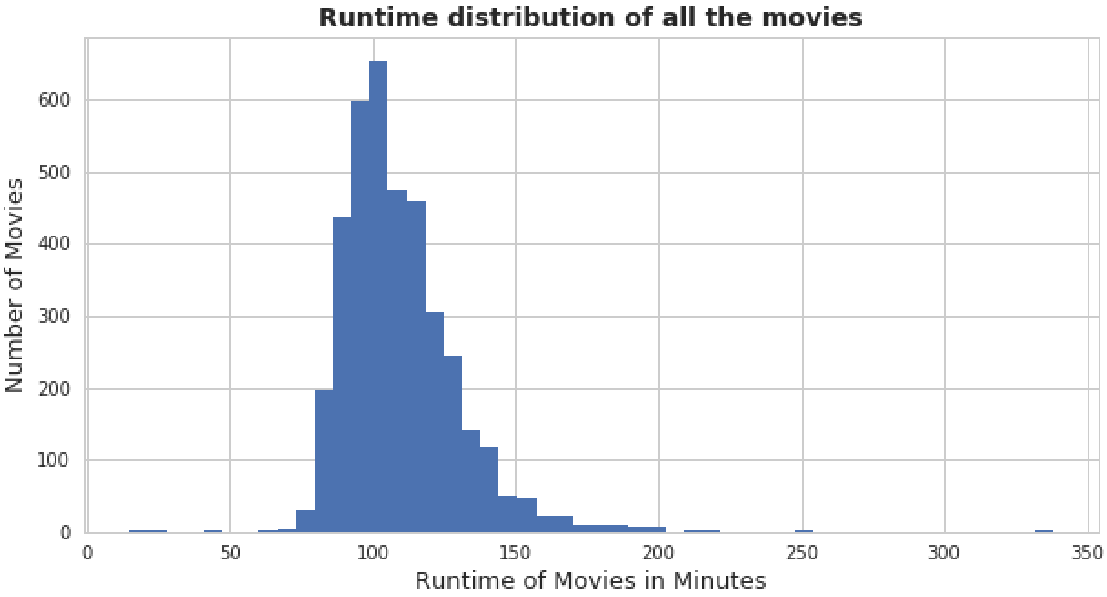
```
In [36]: # Average runtime of movies
         df['runtime'].mean()

Out[36]: 109.22029060716139
```

What is the average runtime of all movies?

So the average runtime of the movies is 109.22 minutes

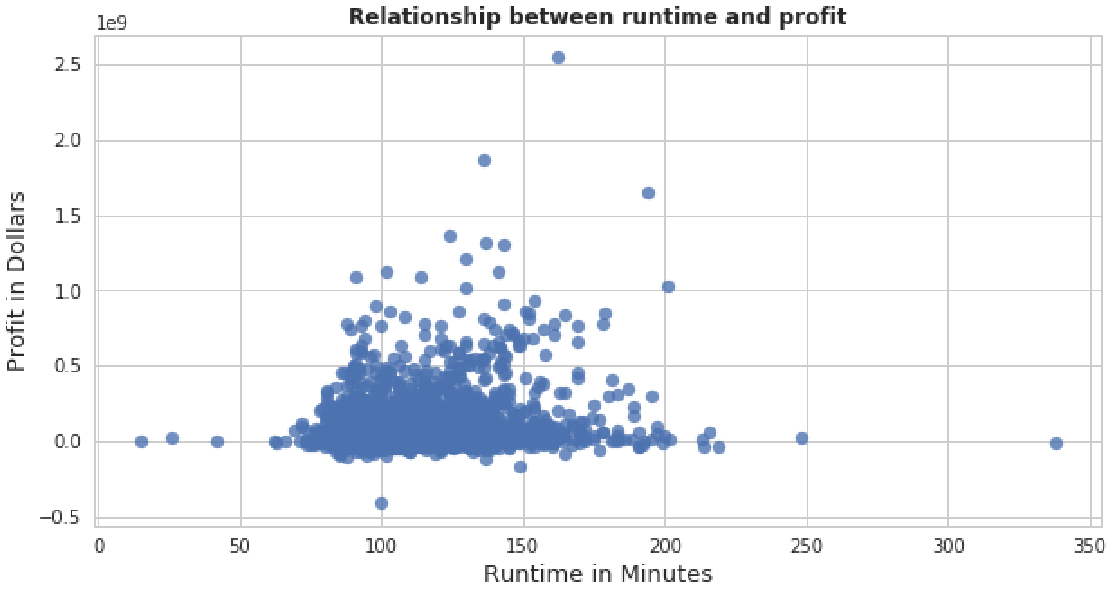
Let us plot a histogram to see the same.

```
In [88]: # x-axis
         plt.xlabel('Runtime of Movies in Minutes',fontsize=13)
         # y-axis
         plt.ylabel('Number of Movies', fontsize=13)
         # Title of the histogram
         plt.title('Runtime distribution of all the movies', fontsize=14, fontweight="bold")
         # Plot a histogram
         plt.hist(df['runtime'], bins = 50)
         #setup the figure size.
         sns.set(rc={'figure.figsize':(10,5)})
         sns.set_style("whitegrid")
```



We can see that most of the movie are in the range of 100 minutes to 120 minutes.
Let us check if there a relation between the Runtime and Profit

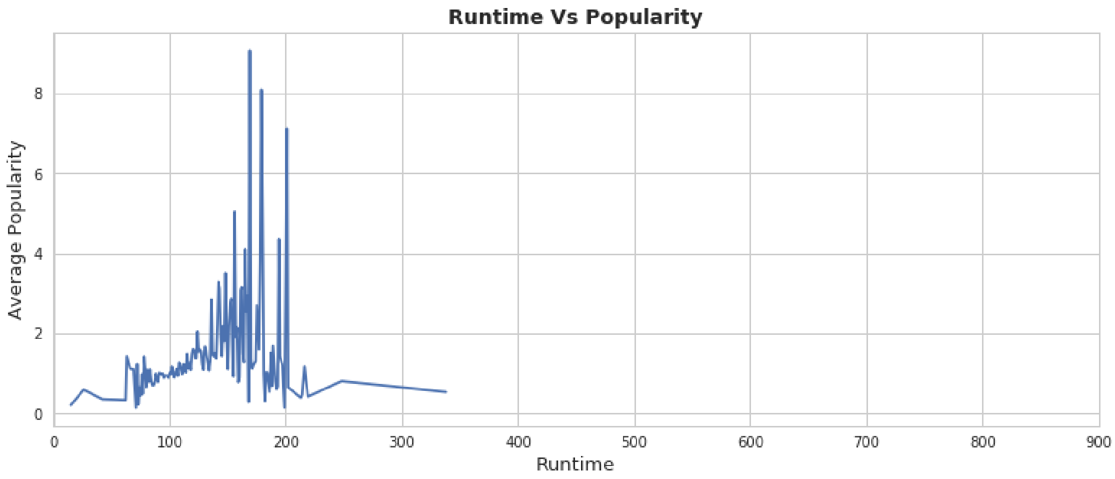
```
In [54]: # x-axis
         plt.xlabel('Runtime in Minutes',fontsize = 13)
         # y-axis
         plt.ylabel('Profit in Dollars',fontsize = 13)
         # Title of the histogram
         plt.title('Relationship between runtime and profit',fontsize=12, fontweight="bold")
         plt.scatter(df['runtime'], df['profit'], alpha=0.8)
         plt.show()
         #setup the figure size.
         sns.set(rc={'figure.figsize':(10,5)})
         sns.set_style("whitegrid")
```



Most of the movies have runtime in range of 85 to 120 Minutes.

1.0.5 Research Question 1.5 Which duration movies are most liked by the audiences according to their popularity?

```
In [49]: #use groupby function and group the data according to their runtime.
         #make a plot using their popularity and find which length movies are most popular.

         #make the group of the data according to their runtime and find the mean popularity rel
         df.groupby('runtime')['popularity'].mean().plot(figsize = (13,5),xticks=np.arange(0,100

         #setup the title of the figure
         plt.title("Runtime Vs Popularity",fontsize = 14, fontweight="bold")

         #setup the x-label and y-label of the plot
         plt.xlabel('Runtime',fontsize = 13)
         plt.ylabel('Average Popularity',fontsize = 13)

         #setup the figure size.
         sns.set(rc={'figure.figsize':(10,5)})
         sns.set_style("whitegrid")
```



Which duration movies are most liked by the audiences according to their popularity?

> From the above plot we can say that movies in the range of 100-200 runtime are more popular than other runtime movies.This is true as usually audiences prefer to watch less duration movies as there is a high tendency of them getting bored soon.

1.0.6 Research Question 2.1 What is the average budget of the movie?

Now since in all the remaining questions we are going to answer them with respect to profit, we will now clean our datset and only incudde data of movies who made profit of more then 25M Dollars.

```
In [55]: # Dataframe which has data of movies which made profit of more the 25M Dollars.
         tmdb_profit_data = df[df['profit'] >= 25000000]

         # Reindexing the dataframe
         tmdb_profit_data.index = range(len(tmdb_profit_data))

         #showing the dataset
         tmdb_profit_data.head()
```

```
Out[55]:    popularity     budget     revenue                original_title  \
         0   32.985763  150000000  1513528810                Jurassic World
         1   28.419936  150000000   378436354            Mad Max: Fury Road
         2   13.112507  110000000   295238201                     Insurgent
         3   11.173104  200000000  2068178225  Star Wars: The Force Awakens
         4    9.335014  190000000  1506249360                     Furious 7

                                                          cast          director  \
         0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...   Colin Trevorrow
         1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...     George Miller
         2  Shailene Woodley|Theo James|Kate Winslet|Ansel...  Robert Schwentke
         3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...       J.J. Abrams
         4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...         James Wan

            runtime                                     genres  \
         0      124  Action|Adventure|Science Fiction|Thriller
         1      120  Action|Adventure|Science Fiction|Thriller
         2      119         Adventure|Science Fiction|Thriller
         3      136   Action|Adventure|Science Fiction|Fantasy
         4      137                      Action|Crime|Thriller

                                         production_companies release_date  vote_count  \
         0  Universal Studios|Amblin Entertainment|Legenda...   2015-06-09        5562
         1  Village Roadshow Pictures|Kennedy Miller Produ...   2015-05-13        6185
         2  Summit Entertainment|Mandeville Films|Red Wago...   2015-03-18        2480
         3          Lucasfilm|Truenorth Productions|Bad Robot   2015-12-15        5292
         4  Universal Pictures|Original Film|Media Rights ...   2015-04-01        2947

            vote_average  release_year      profit
         0           6.5          2015  1363528810
         1           7.1          2015   228436354
         2           6.3          2015   185238201
         3           7.5          2015  1868178225
         4           7.3          2015  1316249360
```

```
In [56]: # Checking the new dataframe
         tmdb_profit_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1791 entries, 0 to 1790
Data columns (total 14 columns):
popularity              1791 non-null float64
budget                  1791 non-null int64
revenue                 1791 non-null int64
original_title          1791 non-null object
cast                    1790 non-null object
director                1791 non-null object
runtime                 1791 non-null int64
genres                  1791 non-null object
production_companies    1788 non-null object
release_date            1791 non-null datetime64[ns]
vote_count              1791 non-null int64
vote_average            1791 non-null float64
release_year            1791 non-null int64
profit                  1791 non-null int64
dtypes: datetime64[ns](1), float64(2), int64(6), object(5)
memory usage: 196.0+ KB
```

```
In [57]: # average budget of movies which made profit more than 25B dollars
         tmdb_profit_data['budget'].mean()

Out[57]: 51870307.757118925
```

What is the average budget of the movie w.r.t Profit of movies making more then 25M Dollars?ű

So the average budget of the movies is 51870307.75 Dollars

1.0.7 Research Question 2.2 What is the average revenue of the movie?

```
In [58]: # average revenue of movies which made profit more then 25M Dollars
         tmdb_profit_data['revenue'].mean()

Out[58]: 206359440.87269682
```

What is the average revenue of the movie w.r.t Profit of movies making more then 25M Dollars?

So the average revenue of the movies is 206359440.87 Dollars

1.0.8 Research Question 2.3 Which are the most frequent cast involved?

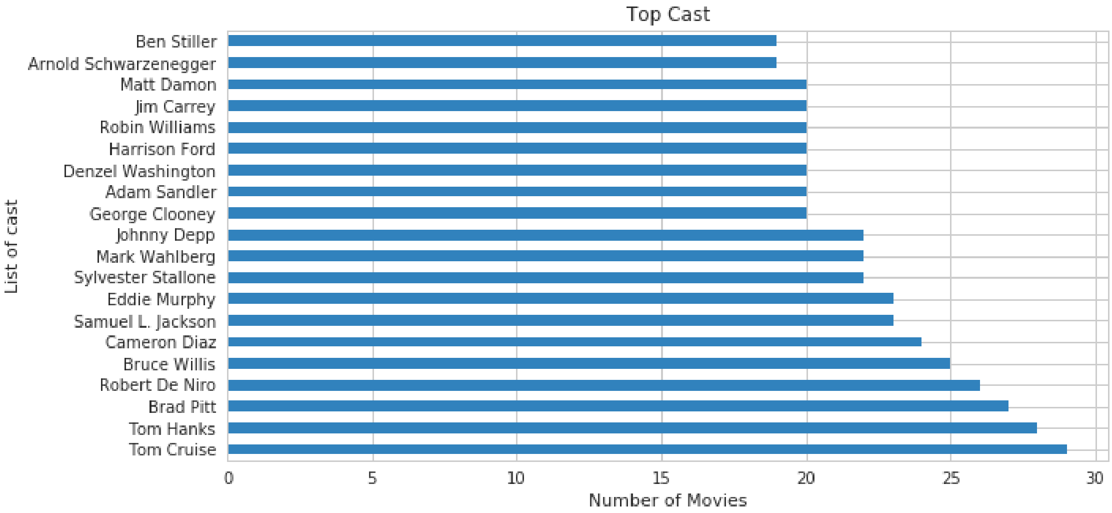
```
In [59]: # This will first concat all the data with | from the whole column and then split it us
         cast_count = pd.Series(tmdb_profit_data['cast'].str.cat(sep = '|').split('|')).value_co
         cast_count.head(20)

Out[59]: Tom Cruise                29
         Tom Hanks                 28
         Brad Pitt                 27
         Robert De Niro            26
         Bruce Willis              25
         Cameron Diaz              24
         Samuel L. Jackson         23
         Eddie Murphy              23
         Sylvester Stallone        22
         Mark Wahlberg             22
         Johnny Depp               22
         George Clooney            20
         Adam Sandler              20
         Denzel Washington         20
         Harrison Ford             20
         Robin Williams            20
         Jim Carrey                20
         Matt Damon                20
         Arnold Schwarzenegger     19
         Ben Stiller               19
         dtype: int64
```

Which are the most frequent cast involved w.r.t Profit of movies making more then 25M Dollars?

So the Top 5 cast are Tom Cruise, Tom Hanks, Brad Pitt, Robert De Niro, Bruce Willis

Lets visualize this with a plot

```
In [79]: # Initialize the plot
         figure = cast_count.head(20).plot.barh(fontsize = 10,colormap= 'tab20c')
         # Set a title
         figure.set(title = 'Top Cast')
         # x-label and y-label
         figure.set_xlabel('Number of Movies')
         figure.set_ylabel('List of cast')
         # Show the plot
         plt.show()
         #setup the figure size.
         sns.set(rc={'figure.figsize':(10,5)})
         sns.set_style("whitegrid")
```



We can clearly see in the visualization that most movies have Tom Cruise as a cast which has leaded to higher profit.

1.0.9 Research Question 2.4 Which are the successful genres?

```
In [68]: # This will first concat all the data with | from the whole column and then split it us
         genres_count = pd.Series(tmdb_profit_data['genres'].str.cat(sep = '|').split('|')).valu
         genres_count
```

```
Out[68]: Drama              688
         Comedy             645
         Action             566
```

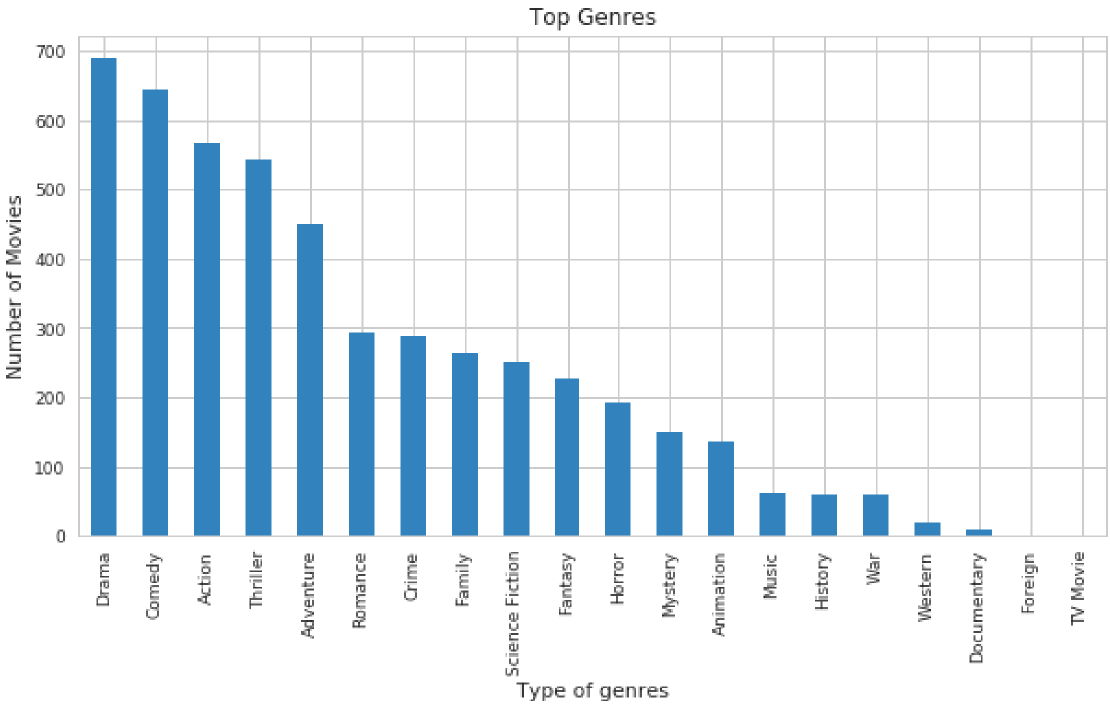
```
Thriller           542
Adventure          451
Romance            292
Crime              287
Family             265
Science Fiction    250
Fantasy            227
Horror             191
Mystery            150
Animation          136
Music               62
History             59
War                 58
Western             20
Documentary          9
Foreign              1
TV Movie             1
dtype: int64
```

Which are the successful genres w.r.t Profit of movies making more then 25M Dollars?

The Top 10 Genres are Drama, Comedy, Action, Thriller, Adventure, Romance, Crime, Family, Scince Fiction, Fantasy

Lets visualize this with a plot

```
In [76]: # Initialize the plot
         diagram = genres_count.plot.bar(fontsize = 9, colormap= 'tab20c')
         # Set a title
         diagram.set(title = 'Top Genres')
         # x-label and y-label
         diagram.set_xlabel('Type of genres')
         diagram.set_ylabel('Number of Movies')
         # Show the plot
         plt.show()
         #setup the figure size.
         sns.set(rc={'figure.figsize':(10,5)})
         sns.set_style("whitegrid")
```



We can clearly see in the visualization that most movies which have drame as a genre tend to have higher profit.

#### Conclusions

Based on the analysis for profit more than 25M dollars, we have found the following:

The average budget of the movies can be around 51870307.75 Dollars

The Top 10 Genres we should focus on should be Drama, Comedy, Action, Thriller, Adventure, Romance, Crime, Family, Scince Fiction, Fantasy

The average revenue of the movies will be around 206359440.87 Dollars

The Top 5 cast we should focus on should be Tom Cruise, Tom Hanks, Brad Pitt, Robert De Niro, Bruce Willis

Limitations : Findings are tentative and not verified by the principles of statistics and machine learning.The conclusion is not full proof that this formula is gonna work, but it shows us that we have high probability of making high profits if we had similar characteristics as such.This was just one example of an influential factor that would lead to different results, there are many that have to be taken care of.

```
In [96]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])

Out[96]: 0
```