Mohammed Almoneef

DAND

Project 2

Investigate a Dataset

P2: Investigate a Dataset

Data set name: TMDb movie data
Description: This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

Introduction:

this project to complete my DAND program from Udacity. i chose the TMDb dataset which has a collection of more than 10K movies details (budget, revenues, release date, ... etc). So let's start to exploring the dataset csv file.

loading libraries:

```
# Use this cell to set up import statements for all of the packages tha
t you
#   plan to use.

#loading libraries
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
```

Data Wrangling

General Properties

Load data & some dataset details:

```
# Load data
data = pd.read_csv('tmdb-movies.csv')
print(data.shape)
```

result: (10866, 21)

```
# number of rows
rows, col = data.shape
print('We have {} total entries of movies & {} columns.'.format(rows-1,
col))
```

result: We have 10865 total entries of movies & 21 columns.

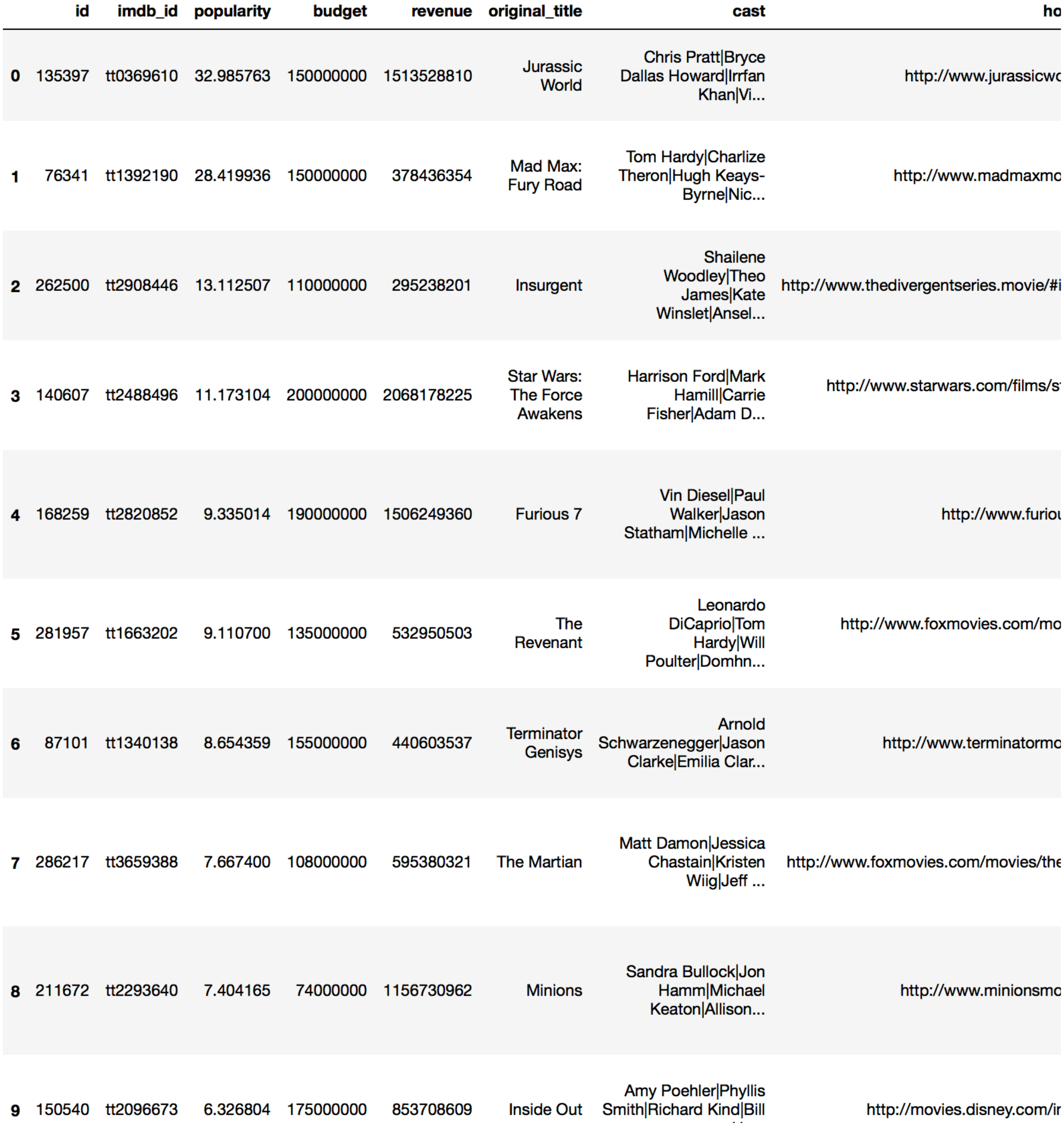
P2: Investigate a Dataset

```
# here we will prent all the columns name in dataset..
print(list(data.columns.values))
```

result:
```
['id', 'imdb_id', 'popularity', 'budget', 'revenue', 'original_tit
le', 'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview',
'runtime', 'genres', 'production_companies', 'release_date', 'vote_coun
t', 'vote_average', 'release_year', 'budget_adj', 'revenue_adj']
```

```
# here we will present the first 10 values in the dataset
data.head(10)
```

result:



	id	imdb_id	popularity	budget	revenue	original_title	cast	ho
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicwo
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmo
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#i
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/s
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furiou
5	281957	tt1663202	9.110700	135000000	532950503	The Revenant	Leonardo DiCaprio|Tom Hardy|Will Poulter|Domhn...	http://www.foxmovies.com/mo
6	87101	tt1340138	8.654359	155000000	440603537	Terminator Genisys	Arnold Schwarzenegger|Jason Clarke|Emilia Clar...	http://www.terminatormo
7	286217	tt3659388	7.667400	108000000	595380321	The Martian	Matt Damon|Jessica Chastain|Kristen Wiig|Jeff ...	http://www.foxmovies.com/movies/the
8	211672	tt2293640	7.404165	74000000	1156730962	Minions	Sandra Bullock|Jon Hamm|Michael Keaton|Allison...	http://www.minionsmo
9	150540	tt2096673	6.326804	175000000	853708609	Inside Out	Amy Poehler|Phyllis Smith|Richard Kind|Bill	http://movies.disney.com/ir

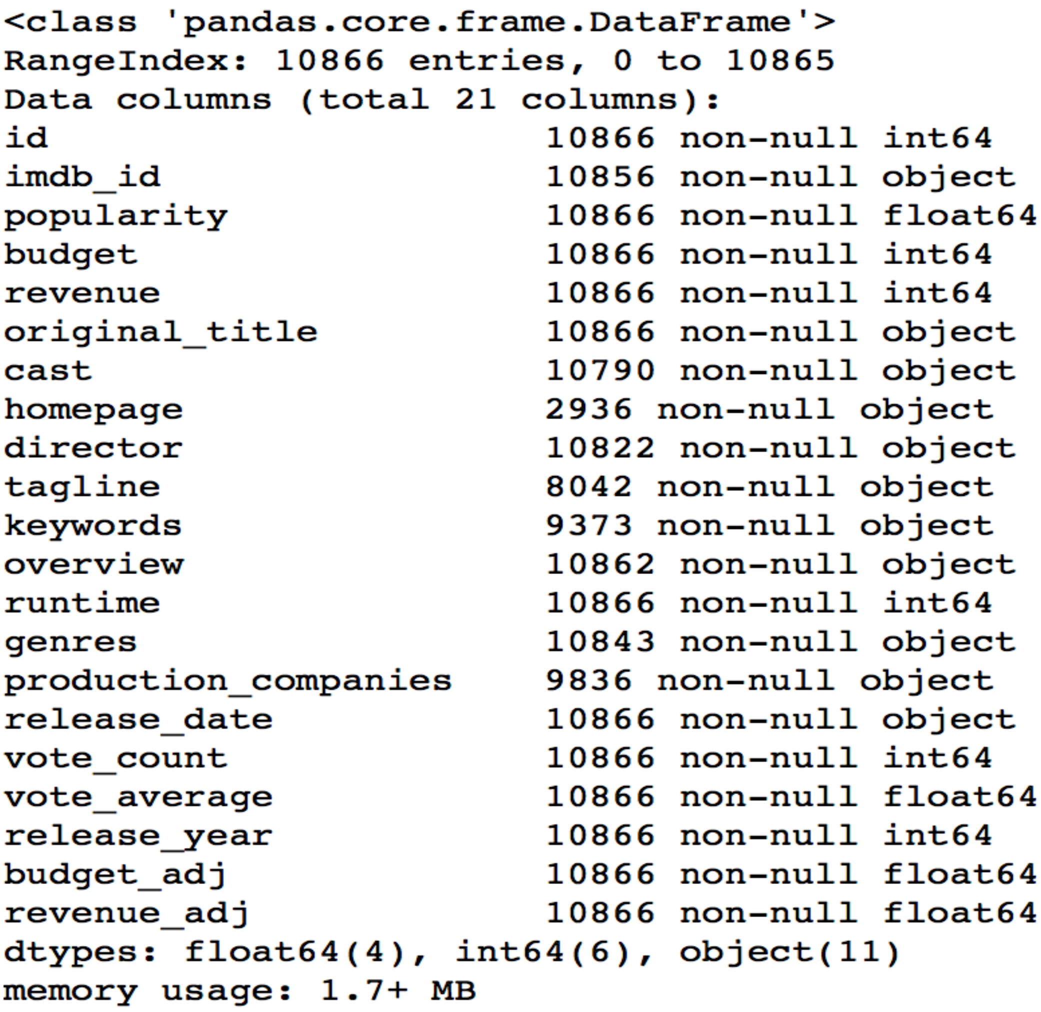
P2: Investigate a Dataset

comments on the dataset:
1) ID columns, is a unique value
2) There are no currency in the dataset, so I will assume it's in the dollar

Data Cleaning (remove null values & duplicate, changing date formate, ...)

```
# 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 se
ction.
data.info()
```

result:



```
<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
```

```
#remove nan values from cast column
#remove rows if revenue_adj & budget_adj = zero

data = data[data["cast"].isnull() == False]
data = data[data["genres"].isnull() == False]

data = data[data.budget_adj != 0]
data = data[data.revenue_adj != 0]

#remove unusefull column
un_necessary_columns=['id', 'imdb_id','overview','popularity','homepage
','tagline','keywords','overview']
data = data.drop(un_necessary_columns,axis = 1)
data.shape
```

result: `(3851, 14)`

duplicated rows

```
#duplicated rows
sum(data.duplicated())
```

result: 1

```
#remove duplicate row
data.drop_duplicates(inplace=True)

#check
print (sum(data.duplicated()))
```

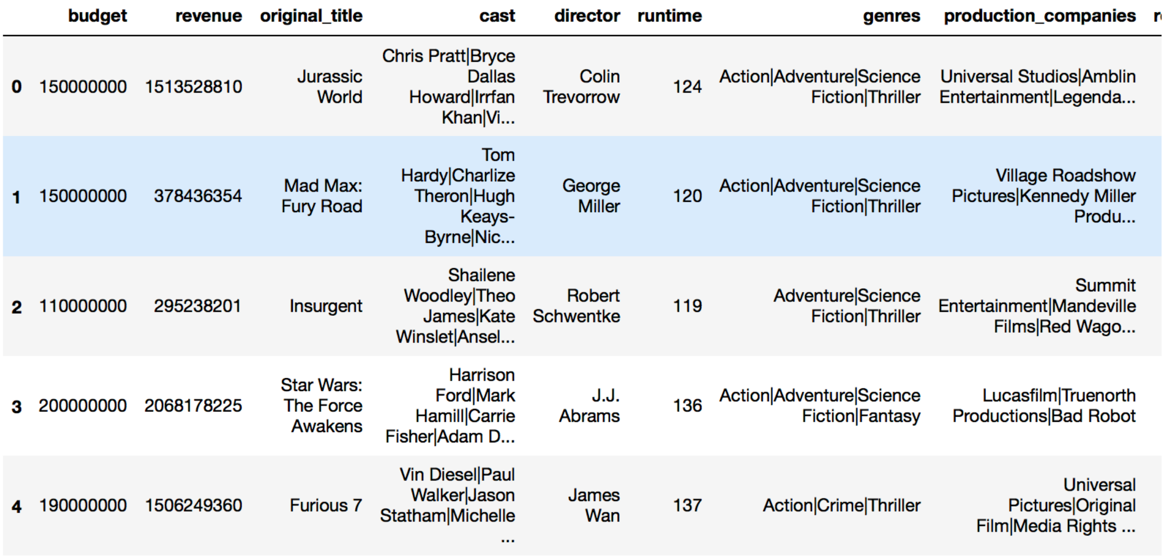
result: 0

Changing release date column into standard date formating

```
data.release_date = pd.to_datetime(data['release_date'])
data.head(5)
```

result:

P2: Investigate a Dataset



	budget	revenue	original_title	cast	director	runtime	genres	production_companies	r
0	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	124	Action|Adventure|Science Fiction|Thriller	Universal Studios|Amblin Entertainment|Legenda...	
1	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	120	Action|Adventure|Science Fiction|Thriller	Village Roadshow Pictures|Kennedy Miller Produ...	
2	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	119	Adventure|Science Fiction|Thriller	Summit Entertainment|Mandeville Films|Red Wago...	
3	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	136	Action|Adventure|Science Fiction|Fantasy	Lucasfilm|Truenorth Productions|Bad Robot	
4	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	137	Action|Crime|Thriller	Universal Pictures|Original Film|Media Rights ...	

```
data.info()
```

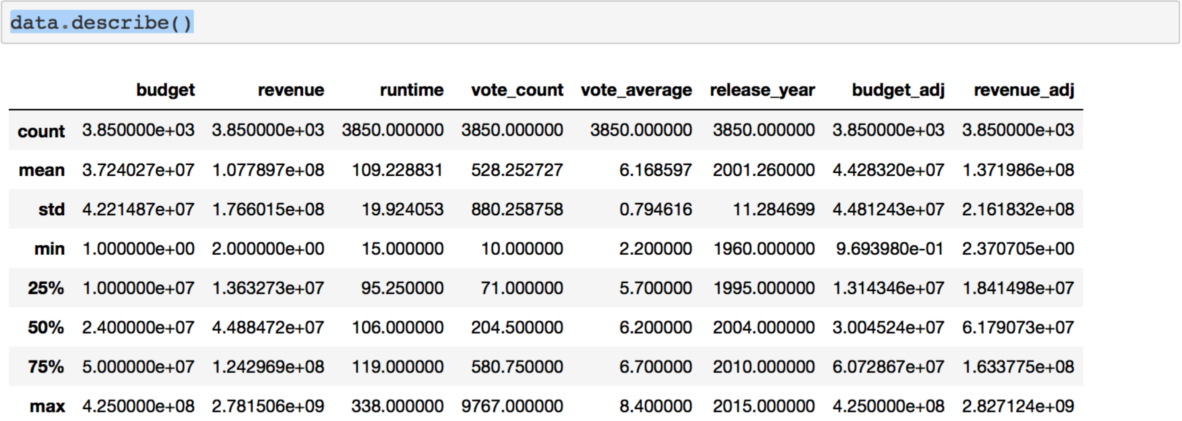
result:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3850 entries, 0 to 10848
Data columns (total 14 columns):
budget                  3850 non-null int64
revenue                 3850 non-null int64
original_title          3850 non-null object
cast                    3850 non-null object
director                3849 non-null object
runtime                 3850 non-null int64
genres                  3850 non-null object
production_companies    3806 non-null object
release_date            3850 non-null datetime64[ns]
vote_count              3850 non-null int64
vote_average            3850 non-null float64
release_year            3850 non-null int64
budget_adj              3850 non-null float64
revenue_adj             3850 non-null float64
dtypes: datetime64[ns](1), float64(3), int64(5), object(5)
memory usage: 451.2+ KB
```

```
data.describe()
```

result:

P2: Investigate a Dataset



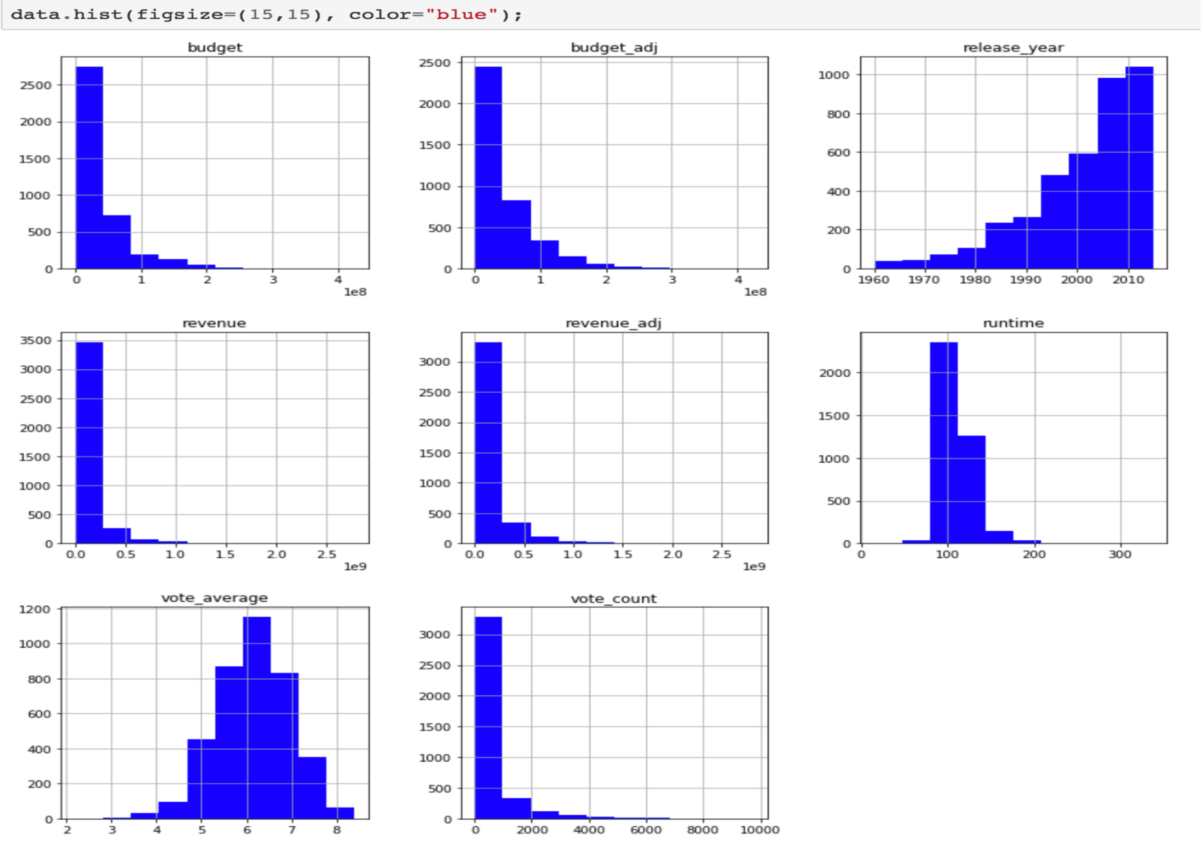
```
data.describe()
```

	budget	revenue	runtime	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	3.850000e+03	3.850000e+03	3850.000000	3850.000000	3850.000000	3850.000000	3.850000e+03	3.850000e+03
mean	3.724027e+07	1.077897e+08	109.228831	528.252727	6.168597	2001.260000	4.428320e+07	1.371986e+08
std	4.221487e+07	1.766015e+08	19.924053	880.258758	0.794616	11.284699	4.481243e+07	2.161832e+08
min	1.000000e+00	2.000000e+00	15.000000	10.000000	2.200000	1960.000000	9.693980e-01	2.370705e+00
25%	1.000000e+07	1.363273e+07	95.250000	71.000000	5.700000	1995.000000	1.314346e+07	1.841498e+07
50%	2.400000e+07	4.488472e+07	106.000000	204.500000	6.200000	2004.000000	3.004524e+07	6.179073e+07
75%	5.000000e+07	1.242969e+08	119.000000	580.750000	6.700000	2010.000000	6.072867e+07	1.633775e+08
max	4.250000e+08	2.781506e+09	338.000000	9767.000000	8.400000	2015.000000	4.250000e+08	2.827124e+09

Here we will present the Data Visualization

```
data.hist(figsize=(15,15), color="blue");
```

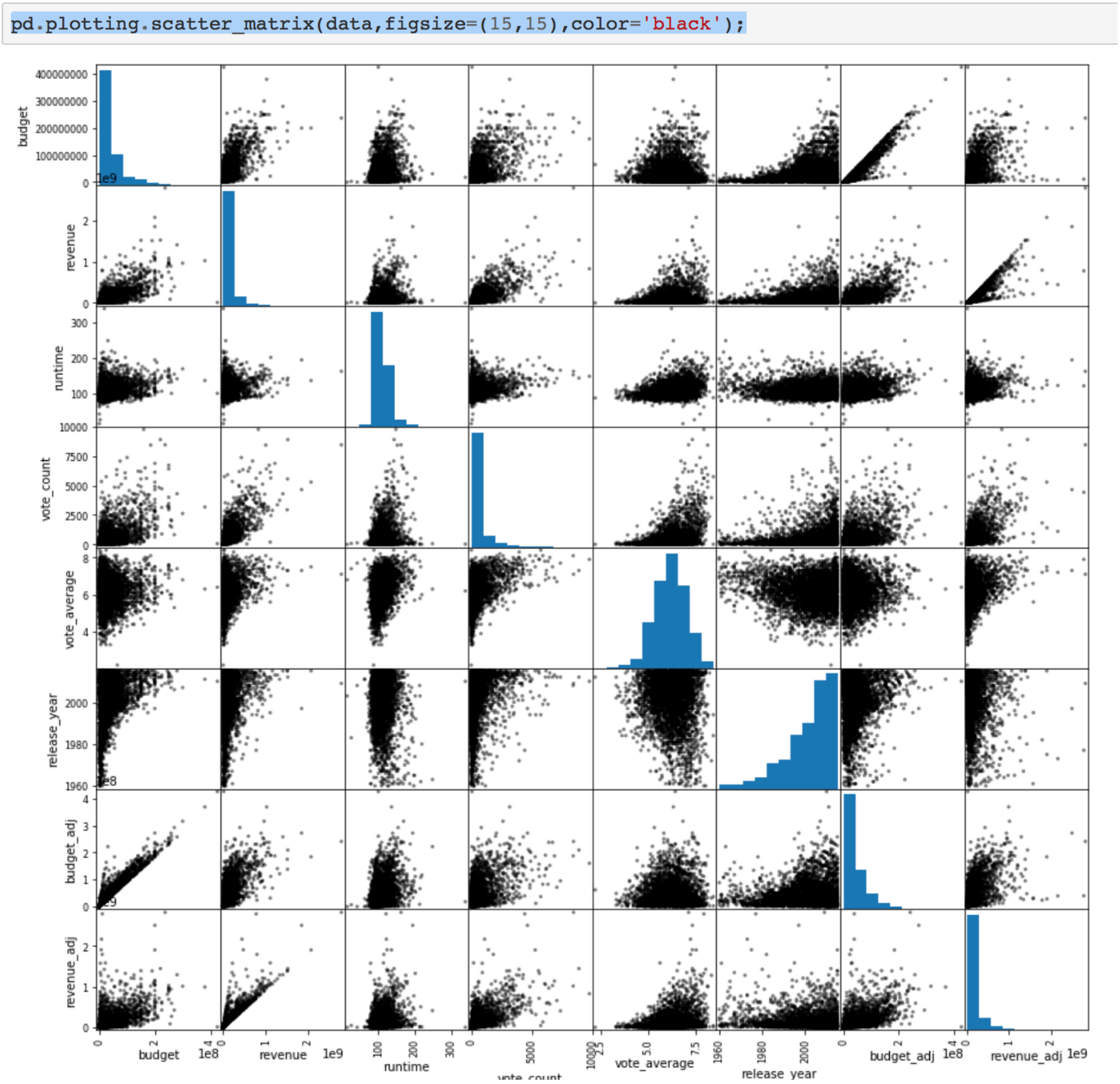
result:



```
data.hist(figsize=(15,15), color="blue");
```

```
pd.plotting.scatter_matrix(data,figsize=(15,15),color='black');
```

result:



```
pd.plotting.scatter_matrix(data,figsize=(15,15),color='black');
```

Exploratory Data Analysis

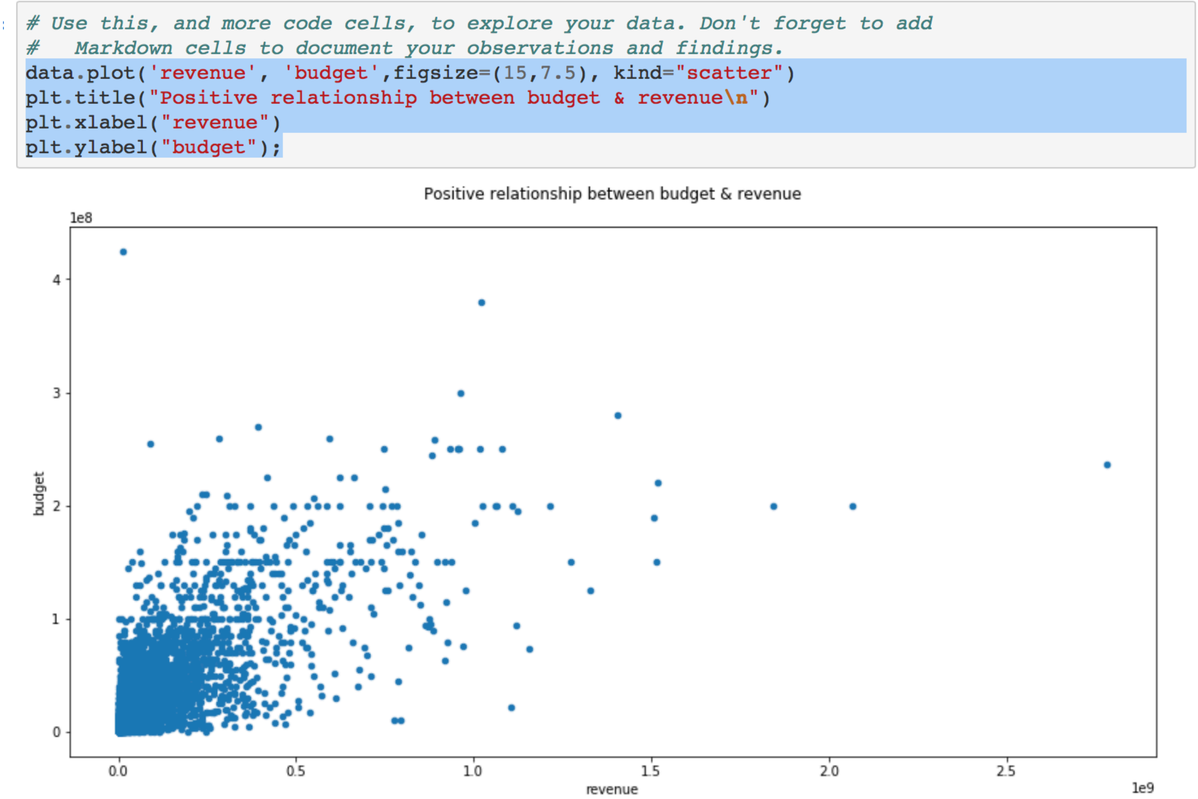
Q1: what is the kind of relationship between budget and revenue?

Answer: Positive relationship, the below graph show that.

Prove:

```
data.plot('revenue', 'budget',figsize=(15,7.5), kind="scatter")
plt.title("Positive relationship between budget & revenue\n")
plt.xlabel("revenue")
plt.ylabel("budget");
```

result:

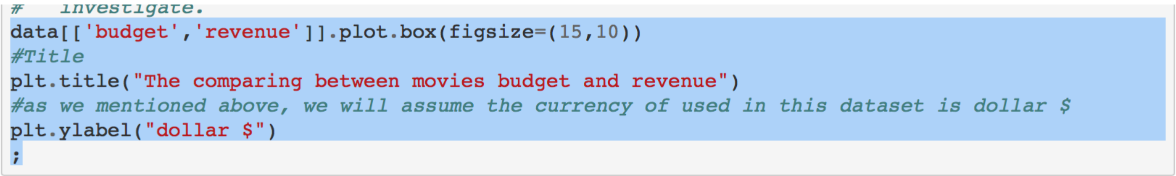


```
# Use this, and more code cells, to explore your data. Don't forget to add
#   Markdown cells to document your observations and findings.
data.plot('revenue', 'budget',figsize=(15,7.5), kind="scatter")
plt.title("Positive relationship between budget & revenue\n")
plt.xlabel("revenue")
plt.ylabel("budget");
```

Q 2: Comparing between budgets of movies and revenues.

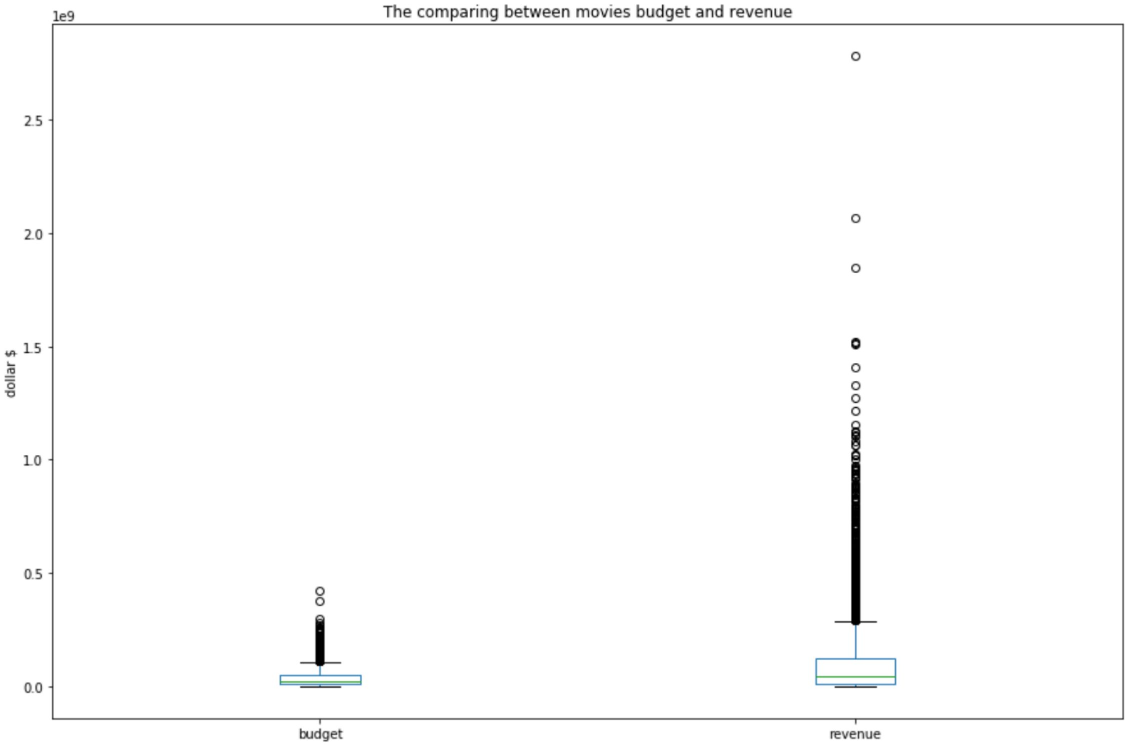
```
data[['budget','revenue']].plot.box(figsize=(15,10))
#Title
plt.title("The comparing between movies budget and revenue")
#as we mentioned above, we will assume the currency of used in this dat
aset is dollar $
plt.ylabel("dollar $")
;
```

result:



```
#   investigate.
data[['budget','revenue']].plot.box(figsize=(15,10))
#Title
plt.title("The comparing between movies budget and revenue")
#as we mentioned above, we will assume the currency of used in this dataset is dollar $
plt.ylabel("dollar $")
;
```

```
: ''
```

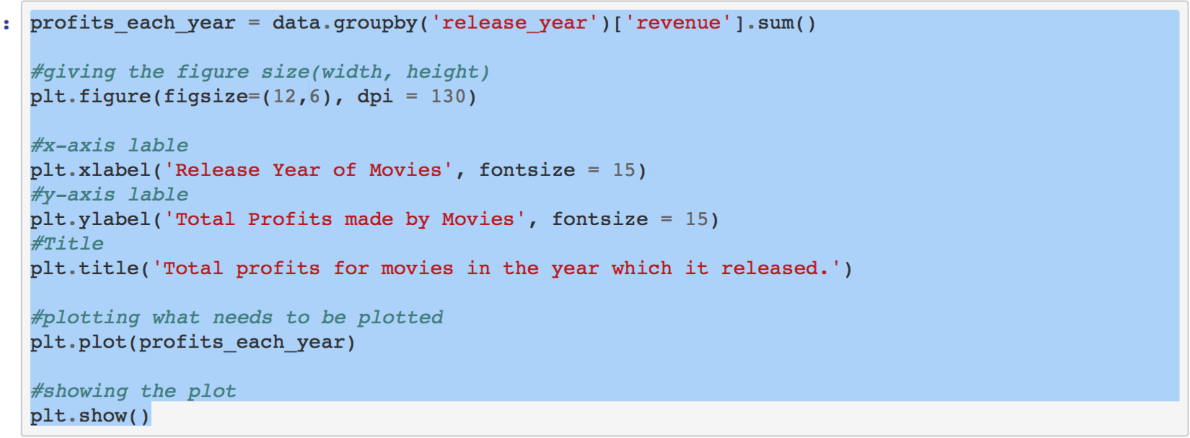


Q3: Calculate the total profits made by all movies in year which it released

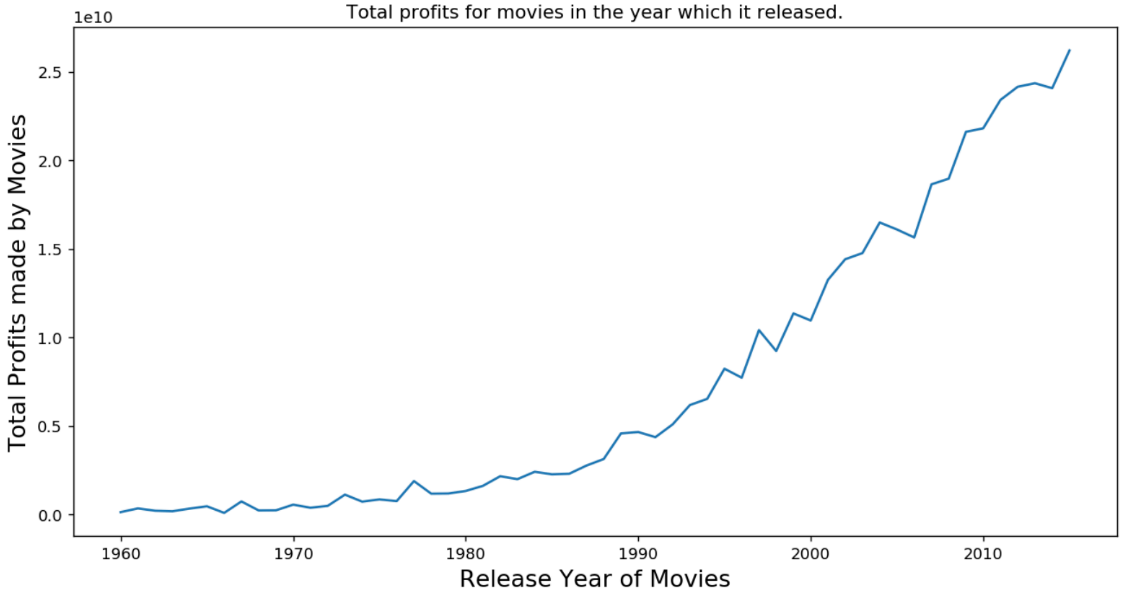
```
profits_each_year = data.groupby('release_year')['revenue'].sum()

#giving the figure size(width, height)
plt.figure(figsize=(12,6), dpi = 130)

#x-axis lable
plt.xlabel('Release Year of Movies', fontsize = 15)
#y-axis lable
plt.ylabel('Total Profits made by Movies', fontsize = 15)
#Title
plt.title('Total profits for movies in the year which it released.')

#plotting what needs to be plotted
plt.plot(profits_each_year)

#showing the plot
plt.show()
```

result:

Q3: Calculate the total profits made by all movies in year which it released.



```
profits_each_year = data.groupby('release_year')['revenue'].sum()

#giving the figure size(width, height)
plt.figure(figsize=(12,6), dpi = 130)

#x-axis lable
plt.xlabel('Release Year of Movies', fontsize = 15)
#y-axis lable
plt.ylabel('Total Profits made by Movies', fontsize = 15)
#Title
plt.title('Total profits for movies in the year which it released.')

#plotting what needs to be plotted
plt.plot(profits_each_year)

#showing the plot
plt.show()
```



Conclusions

The dataset is very rich with information, and very interesting to analysis. there are some facts we can summry in:

Average budget to be in successful criteria around 60$

Average duration for the movies should be around 113 min

Movies business is profitable

If the budget is high the profit will be high

Some movies get unbelavable revenue