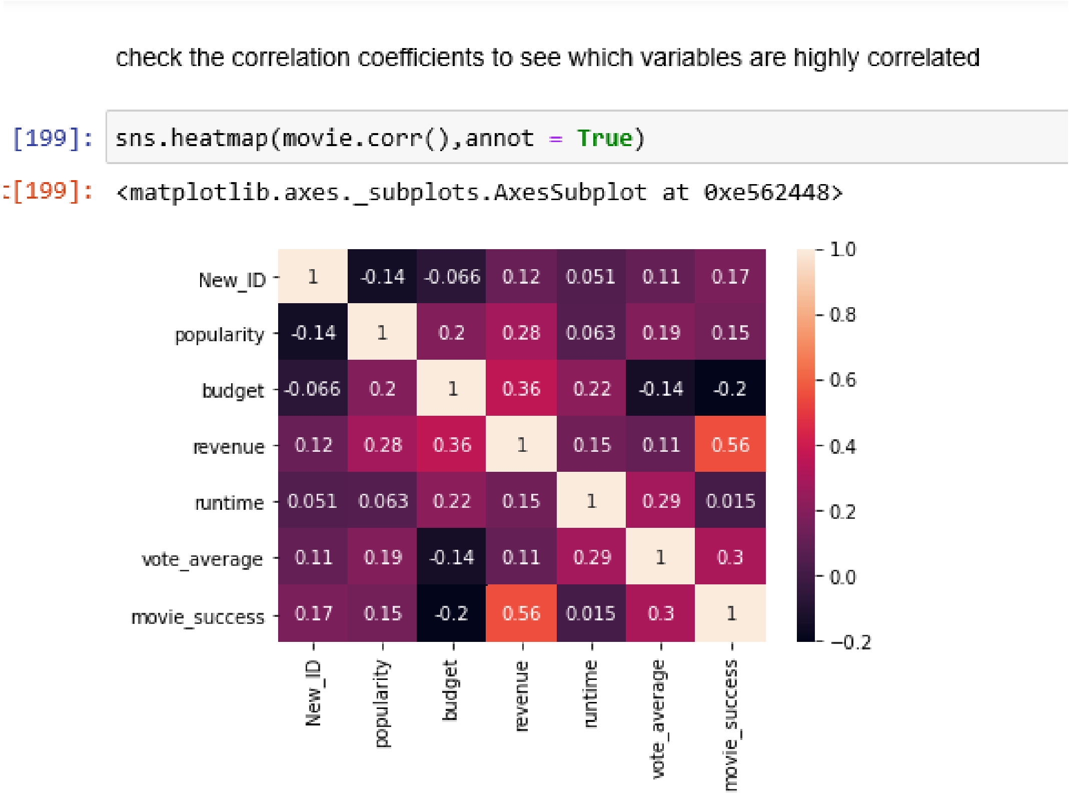
Data visualization

-After cleaning our dataset we need to draw some attractive and informative statistical graphics and plots to increase the data correlation to achieve a deep understanding of it…😊

-First : we import two important data visualization and plotting libraries:

1- Seaborn 2-Matplotlib

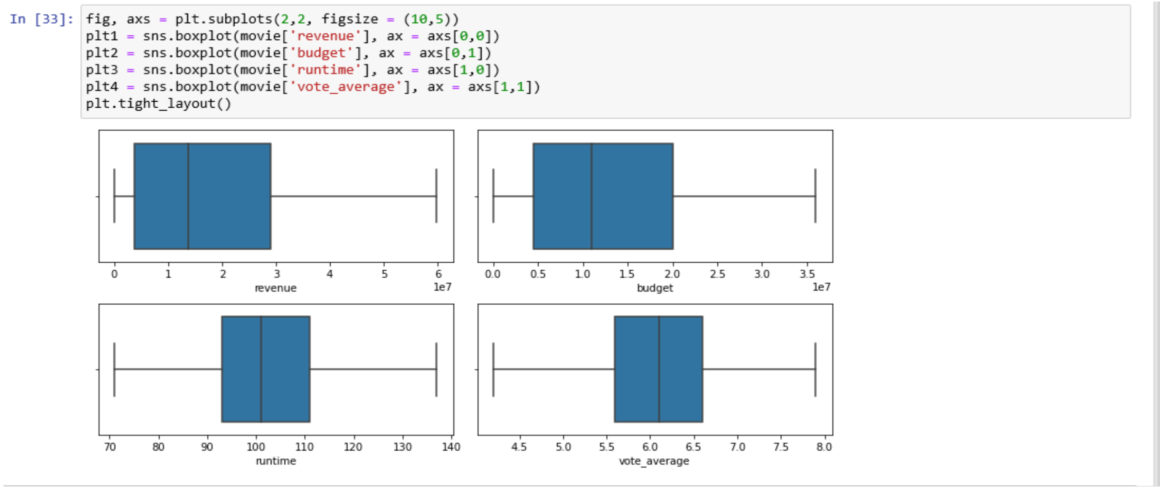


check the correlation coefficients to see which variables are highly correlated

```
[199]: sns.heatmap(movie.corr(),annot = True)
```

```
:[199]: <matplotlib.axes._subplots.AxesSubplot at 0xe562448>
```

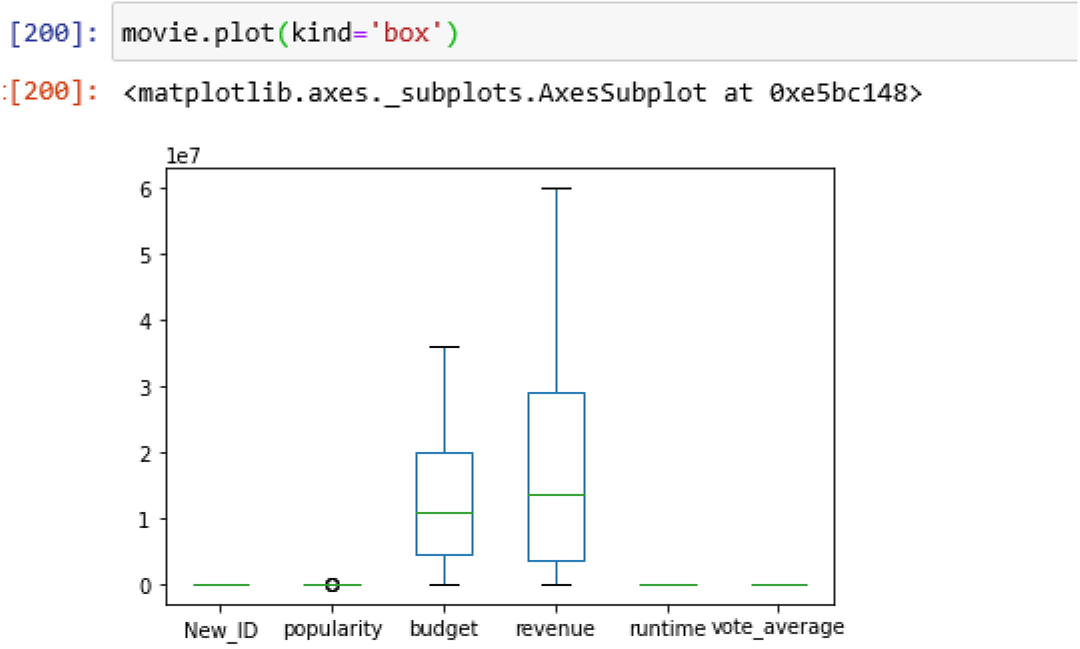
1) In this case we use "heatmap" to know the correlation between variables…and find out that there is a 56% strong correlation between "revenue" and "movie_success".



```
In [33]: fig, axs = plt.subplots(2,2, figsize = (10,5))
         plt1 = sns.boxplot(movie['revenue'], ax = axs[0,0])
         plt2 = sns.boxplot(movie['budget'], ax = axs[0,1])
         plt3 = sns.boxplot(movie['runtime'], ax = axs[1,0])
         plt4 = sns.boxplot(movie['vote_average'], ax = axs[1,1])
         plt.tight_layout()
```

2) In this case we use “boxplot” to express the data and to know if there are an outliers in the columns [“revenue” , “budget” , “runtime” , “vote_average”] and find out that:

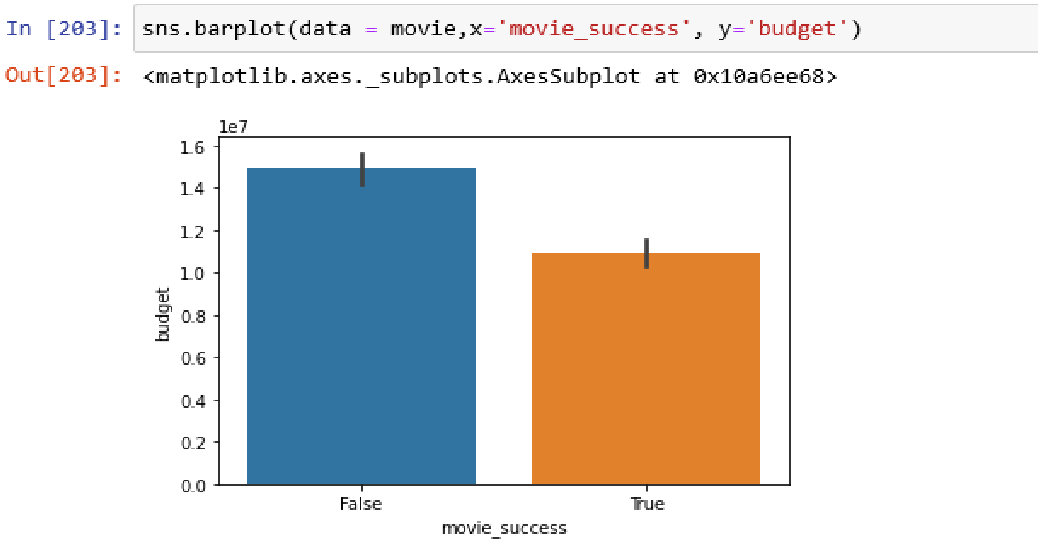
- After using the “IQR” method there are almost no outliers in the cleaned data which makes the model more accurate .



```
[200]: movie.plot(kind='box')
```

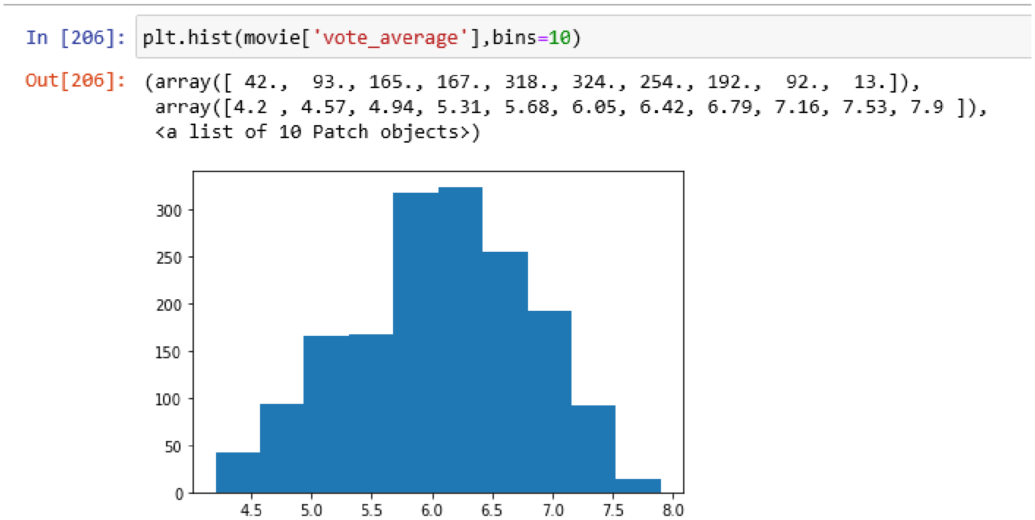
```
[200]: <matplotlib.axes._subplots.AxesSubplot at 0xe5bc148>
```

Hint: after removing the outliers, the number of rows decrease therefore, the outliers in “popularity“ don’t have to be removed.



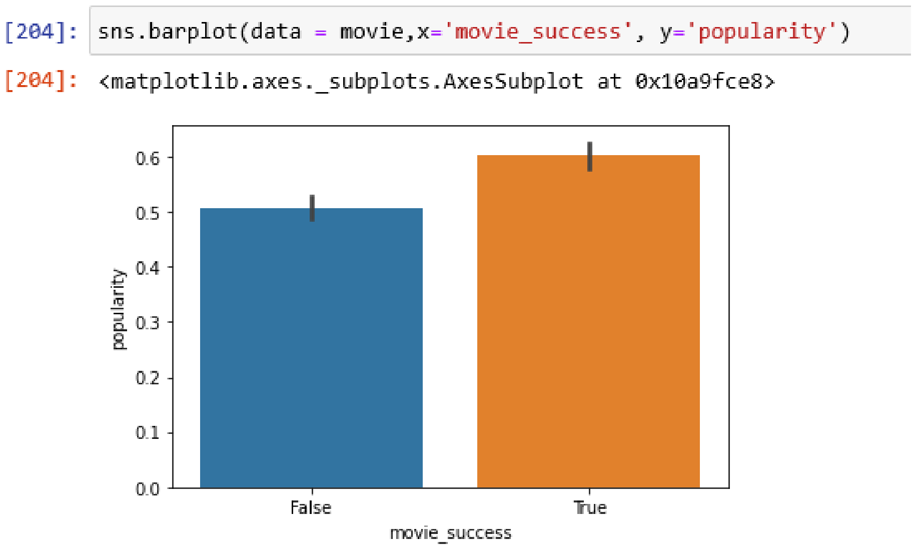
```
In [203]: sns.barplot(data = movie,x='movie_success', y='budget')
Out[203]: <matplotlib.axes._subplots.AxesSubplot at 0x10a6ee68>
```

3) In this case we used the “barplot” to visualize the relation between “budget” and “movie_success” We find out that it’s not necessary that the success movies need high budget, as shown most of the unsuccessful movies need a high budget.



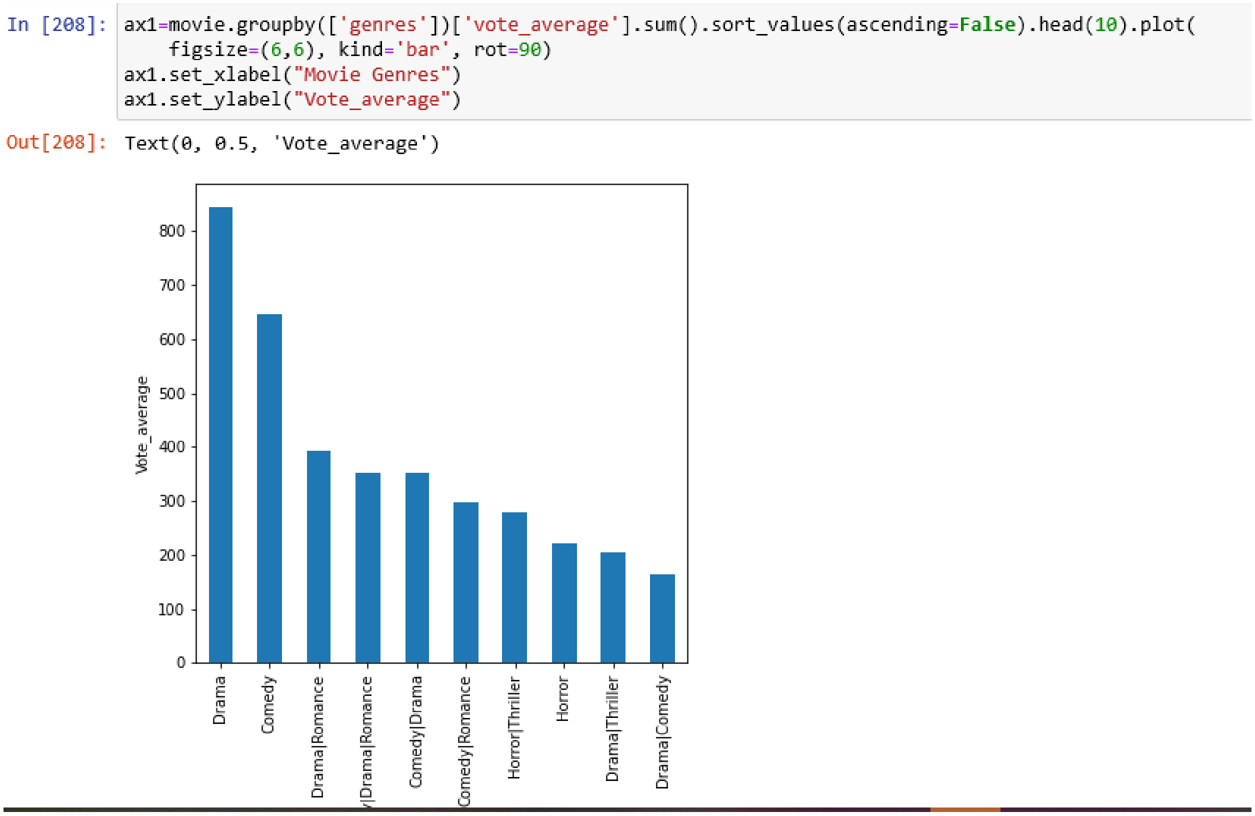
```
In [206]: plt.hist(movie['vote_average'],bins=10)
Out[206]: (array([ 42.,  93., 165., 167., 318., 324., 254., 192.,  92.,  13.]),
           array([4.2 , 4.57, 4.94, 5.31, 5.68, 6.05, 6.42, 6.79, 7.16, 7.53, 7.9 ]),
           <a list of 10 Patch objects>)
```

4) In this case we used “histogram” to represent the distribution of vote average by splitting the range of the data of the column which called “vote_average” into 10 equal sized bins(classes),as shown that (approximately from 6.25 to 6.5) the highest percentage of votes.



```
[204]: sns.barplot(data = movie,x='movie_success', y='popularity')
[204]: <matplotlib.axes._subplots.AxesSubplot at 0x10a9fce8>
```

5)we used “barplot” to show that the most successful movies are which have the most popularity .



```
In [208]: ax1=movie.groupby(['genres'])['vote_average'].sum().sort_values(ascending=False).head(10).plot(
              figsize=(6,6), kind='bar', rot=90)
          ax1.set_xlabel("Movie Genres")
          ax1.set_ylabel("Vote_average")
Out[208]: Text(0, 0.5, 'Vote_average')
```

6) In this case we tried to know the first 10 genres which have the highest vote average and find that “Drama movies have the most votes”.