# Movie Data Analysis

# Does movie’s budget, director’s name and runtime influence or effect the success of a movie?

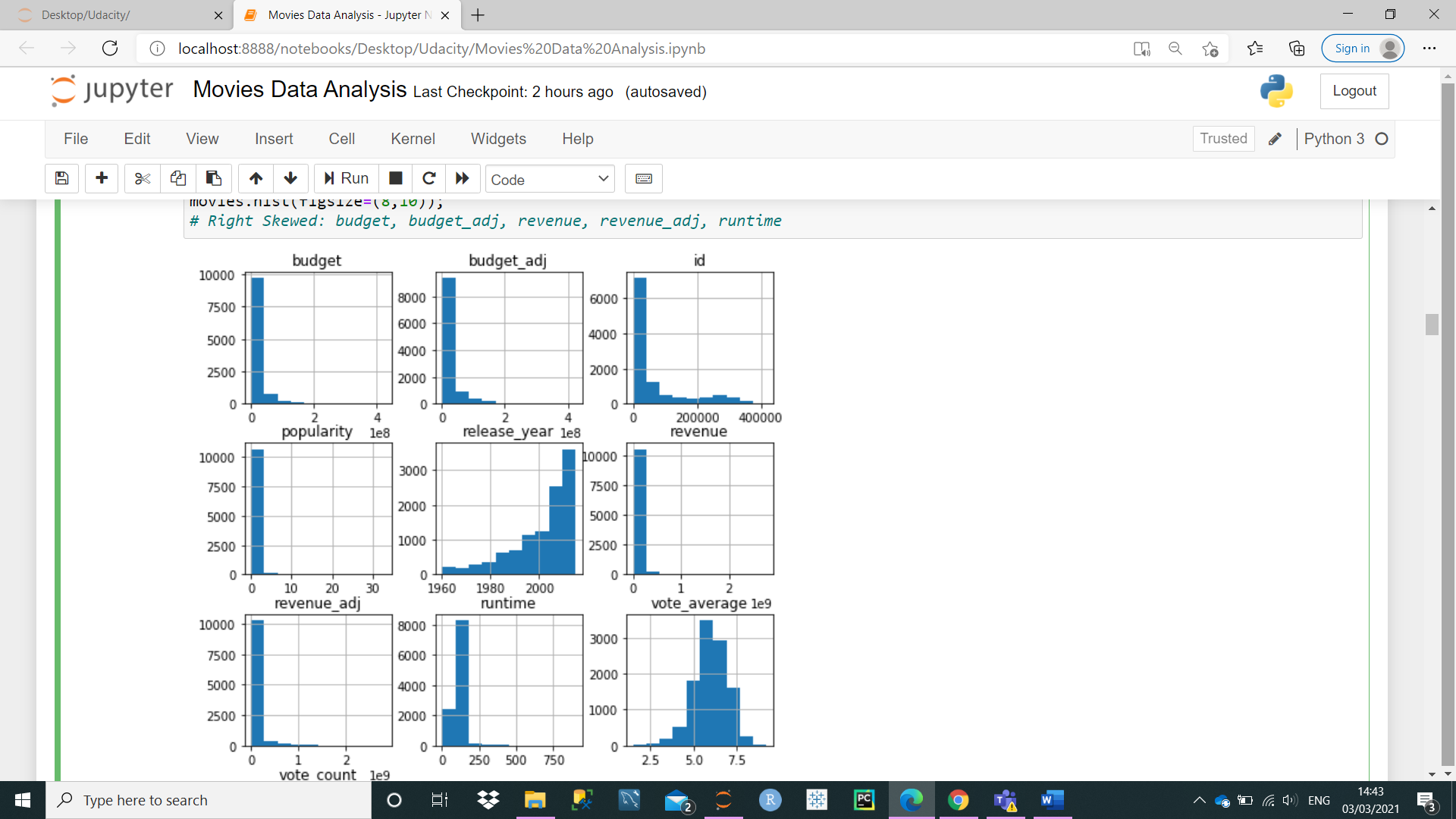
For this project movies dataset was used to analyse whether movie's budget, director and runtime had an impact on revenue. Thus, from quantitative dataset, revenue, runtime and budget figures were used and for qualitative dataset, director dataset was used.

Before using the data, it is good practise to understand and visual the data. The csv was first downloaded, and after downloading file the data frame was named “movies”.

I used .shape() to to acknowledge how many rows and columns there are in the dataset. Then I used .info() to acknowledge the data types. The majority were object types followed by integer and float.

### Key facts (of original dataset)

To visualise the data, **.hist()** was used and as there were 10 columns in numeric form it showed 10 histograms. Amongst, these graphs the most relevant is budget, budget\_adj, revenue, revenue\_adj, runtime and these graphs are right skewed which means mean value is greater than median and median is greater than mode. Most of the values are clustered around left hand size of the graph so there is possibility there are outliers which causes mean value is be greater.

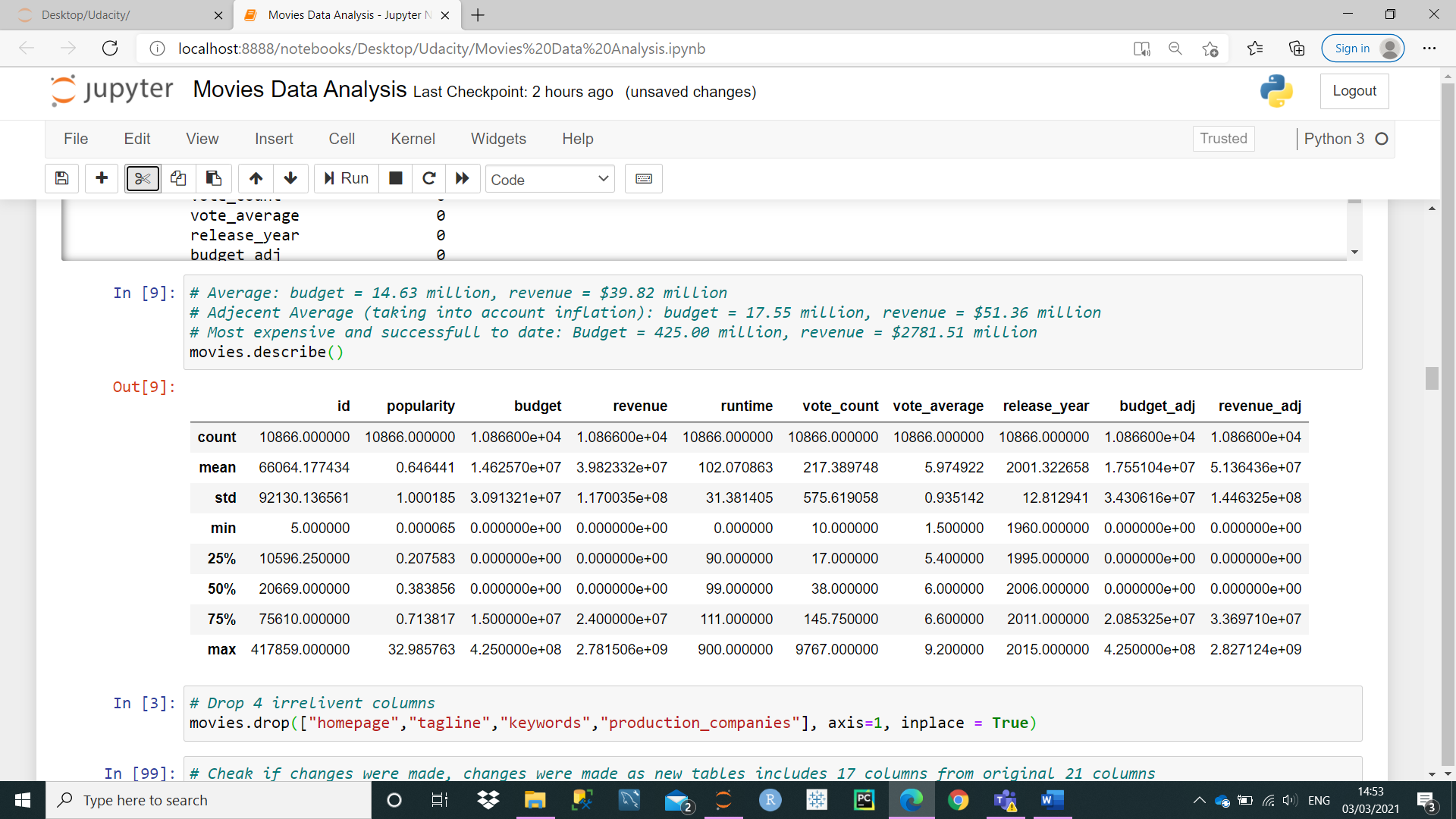


As a good practise, it is always a good idea to find out if there are any missing/null data.

So, I used **.isnull().sum()** function to acknowledge missing or null data. The following columns had null values: homepage, tagline, keywords, production companies, cast, director, genres, imbd\_id, overview.

Then, **.describe()** function was used to find out mean, minimum, maximum and quartiles.

The average budget, revenue and runtime was 14.6 million, 39.8 million, 102 minutes respectively. The minimum budget, revenue and runtime was 0 for all. *Most likely this is an error* and needs to be checked because a released movie cannot be made with 0 budget or gained 0 revenue or made with 0 runtime. The maximum budget, revenue and runtime was 425 million, 2782million (or 2.78 billion), 900 minutes (which is 15 hours, seems very long but could be possible).



## Data Wrangling

Data cleaning and preparation was the most important part of the project because if appropriate and right data is not available then it leads to flawed analysis. So, majority of the time and effort spend on the project was done on data wrangling.

1. The 4 columns, homepage, tagline, keywords, production companies were deleted with the following function because it has the most null values and it was irrelevant to our analysis. So, now we have 17 columns.
2. **movies.drop(["homepage","tagline","keywords","production\_companies"], axis=1, inplace = True)**

1. Then, we had to make sure same movie was not repeated. I initially, thought of dropping duplicate values in “original title” column but then there could be movies with the same name. One column which we could use (which acts like a primary key) is “id” column. There should be one if for each movie.
   1. Another test I did to ensure same movies did not repeat is do the same process for imdb\_id. The imdb\_id is unique should be only 1 for each movie. The following functions were used.

**2) movies.drop\_duplicates(subset = 'id',keep = 'first')**

**3) movies.drop\_duplicates(subset = 'imdb\_id',keep = 'first')**

No duplicates found so. Although there were movies with the same title.

1. There were lot of 0 value fields in revenue, budget and runtime columns which looks wrong because it is not possible to make a movie with 0 budget or movie which makes 0 revenue or movie which has 0 runtime.
   1. So, following formula was used to delete 0 rows for these columns and make a new data frame.

**4) new\_movies = movies[movies.loc[:]!= 0]**

* 1. But, then I observed another problem there were lot of null values (Nan). So, I used the following function to delete null fields.

**5) new\_movies = new\_movies[pd.notnull(new\_movies['budget'])]**

**new\_movies = new\_movies[pd.notnull(new\_movies[‘revenue’])]**

**new\_movies = new\_movies[pd.notnull(new\_movies[‘runtime’])]**

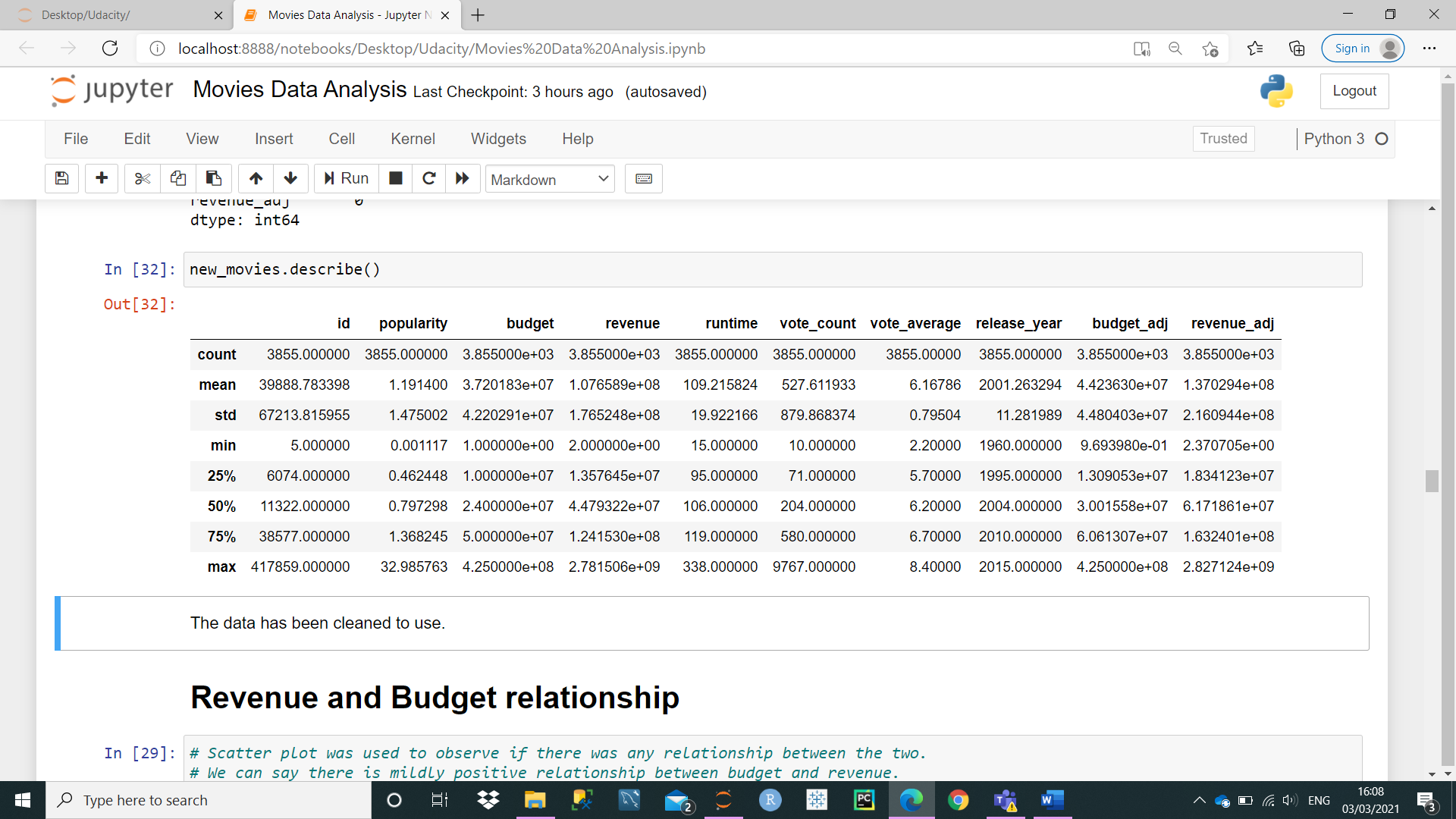
After taking away 0 and null values, the row total is 3854.

Now after when data was checked again .isna().sum() and isnull().sum(), there 4 null values for cast and 1 null value for director. It is possible to have missing cast and director data for these as record may not have been available at the time of writing.

Now, it acknowledge if data cleaning had any effect, I applied the .describe() function again.

Then, **.describe()** function was used to find out mean, minimum, maximum and quartiles.

The average budget, revenue and runtime increased from 14.6 million to 37.3 million, increased from 39.8 million to 107.7 million and increased from 102 to 109 minutes respectively. The minimum budget, revenue and runtime was 0 for all previously. Now, the minimum is 1, 2 and 15 minutes respectively. It is still arguably very low but could be possible. The maximum budget and revenue is the same at 425 million, 2782million (or 2.78 billion) but the maximum runtime decreased from 900 minutes to 338 and this figure looks more reasonable.



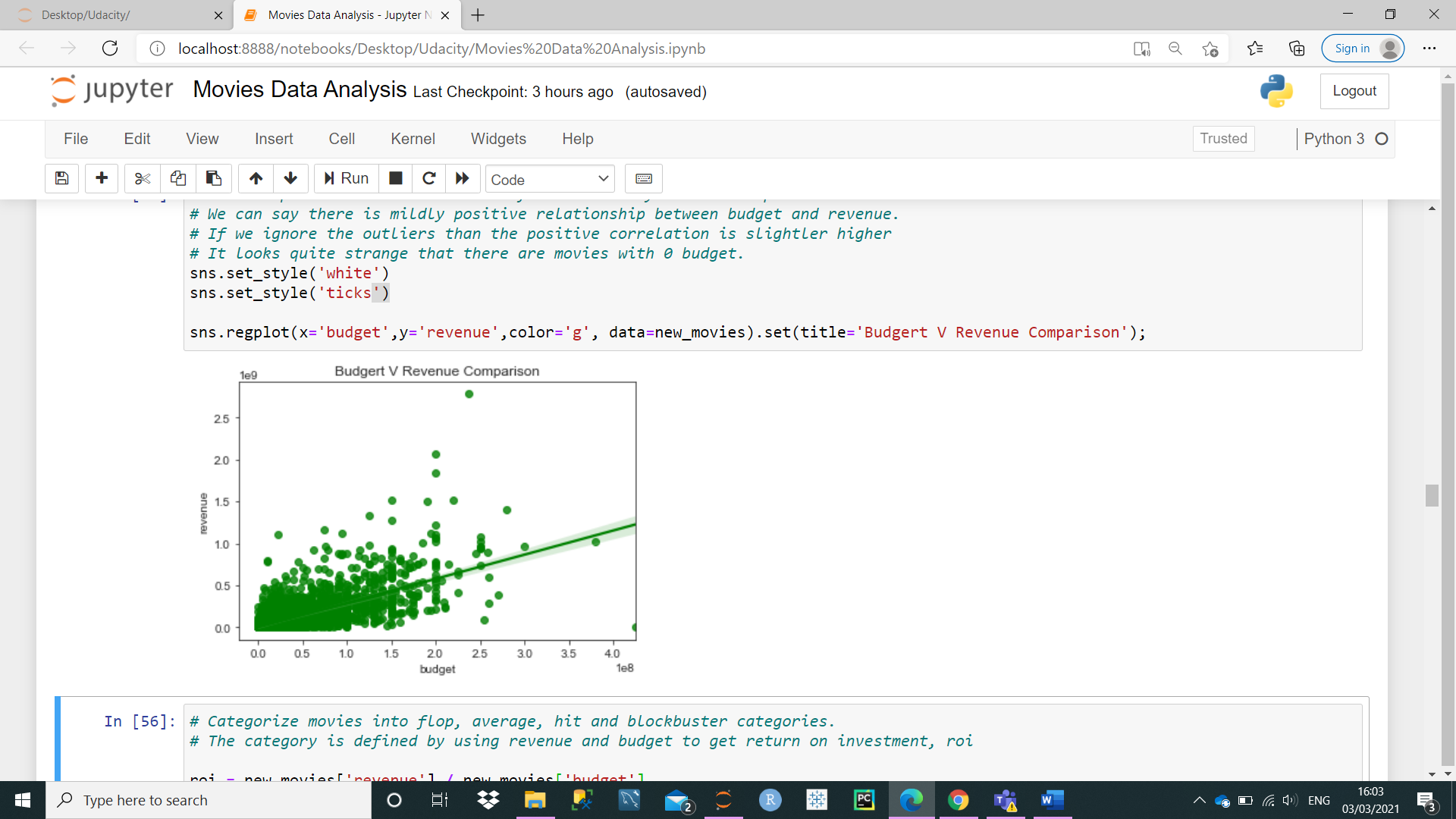
## Budget and Revenue relationship

After cleaning the data, to answer the first question if a movie’s success is dependent upon its budget, scatter plot was made with the following functions.

**6) sns.set\_style('white')**

**sns.set\_style('ticks')**

**sns.regplot(x='budget',y='revenue',color='g', data=movies).set(title='Budgert V Revenue Comparison');**



Budget was labelled on the x-asis as it is an independent variable. There seems to be a positive trend with trendline appearing positive and data points also increasing in horizontally. So, vaguely we can say that the budget is one of the factors which can determine/influence revenue figures.

However, it is not always that easy to suggest this conclusion from just one scatter plot and especially when there have been instances where scatter plot has shown inaccurately positive relationships. If car sales had gone up during the summer and likewise ice cream sales has gone up then scatter plot may make it seem like by making and selling ice creams during the summer makes people buy cars as well.

So, another variable, return on investment (Ad-hoc) was made due to the following reason. A movie’s success is dependent up on many factors and there are many ways to measure movie’s success. But the most obvious method to measure a movie’s success is by finding out how many much money it had earned and divide it by the cost (budget) of the movie. If the number is positive, then it made some/reasonable profit.

There is also another aspect which measures a movie’s success, comparative success. If a movie has made 40 million and also recovered money (budget of 30 million) then it is a success but if there were range of movies which has earned more than 100 million then the movie which earned 40 million would not really be considered successful by critics or fans or at least memorable movie. In some countries/markets, the blockbuster status is only given to movies which has made money much more than its counterparts along with recovering its costs.

### Return on Investment (ROI)

To classify if a movie is successful or not the revenue figure is greater than the budget figure. A formula named return on investment (ROI) was used to find out how much revenue was generated in comparison to money being invested. The following formula was used to find out ROI for each movie.

1. **roi = new\_movies['revenue'] / new\_movies['budget']**

Then, ROI column was created within the movies table with the following formula.

1. **new\_movies["roi"] = roi**

Using this ROI, quartile ranges were calculated with the following formula to find out which movies with certain ROI are in the bottom or top quartile.

1. **new\_movies.roi.quantile([0.25,0.5,0.75])**

So, movies which were in the lower quartile (25%) had ROI under 0.87, movies which were in 2nd quartile (over 25% and under 50%) had ROI between 0.87 and 2.21. Movies which were in 3rd quartile had ROI between 2.21 and 4.21 and movie in the top 25% had ROI over 4.21.

Then, ROI was used to categorise movies into four groups, flop, average, hit, blockbusters based on these quarters and then box office verdict column was made with the following formulas.

1. **roi\_conditions = [**

**(new\_movies['roi'] < 0.87),**

**(new\_movies['roi'] >= 0.87) & (new\_movies['roi'] < 2.12),**

**(new\_movies['roi'] >= 2.12) & (new\_movies['roi'] < 4.21),**

**(new\_movies['roi'] >= 4.21)**

**]**

1. **verdict = ['flop','average','hit','blockbuster']**

**new\_movies['box\_office\_verdict'] = np.select(roi\_conditions,verdict)**

Then, I used the following python function to group box office verdict and budget class to find out if movies made at certain budget were more likely to get better box office verdict by making a new data frame named group. Then used this data frame to count in these column pairs.

1. **group = new\_movies.groupby(['budget\_class', 'box\_office\_verdict',])**

**group.count()**

### Budget

Then, same thing was applied to budget column. Quartile ranges were implemented for budget column and based on quartiles, 4 columns were made, where movies in the bottom 25% were classified of having “Very Low Budget”, 2nd quartile having “Low Budget”, 3rd quartile having “Average Budget” and top 25% having “High Budget”.

So, if a movie was made under the budget of 10m then it would be considered to be of “very low budget”, and if the budget is between 10m and 24m and it has “low budget”. If a movie has budget between 24m and 50m it has average budget and if it has budget over 50m then it is a “high budget” movie. Budget class column was made for respective movies.

1. **budget\_conditions = [**

**(new\_movies['budget'] < 10000000),**

**(new\_movies['budget'] >= 10000000) & (new\_movies['budget'] < 24000000),**

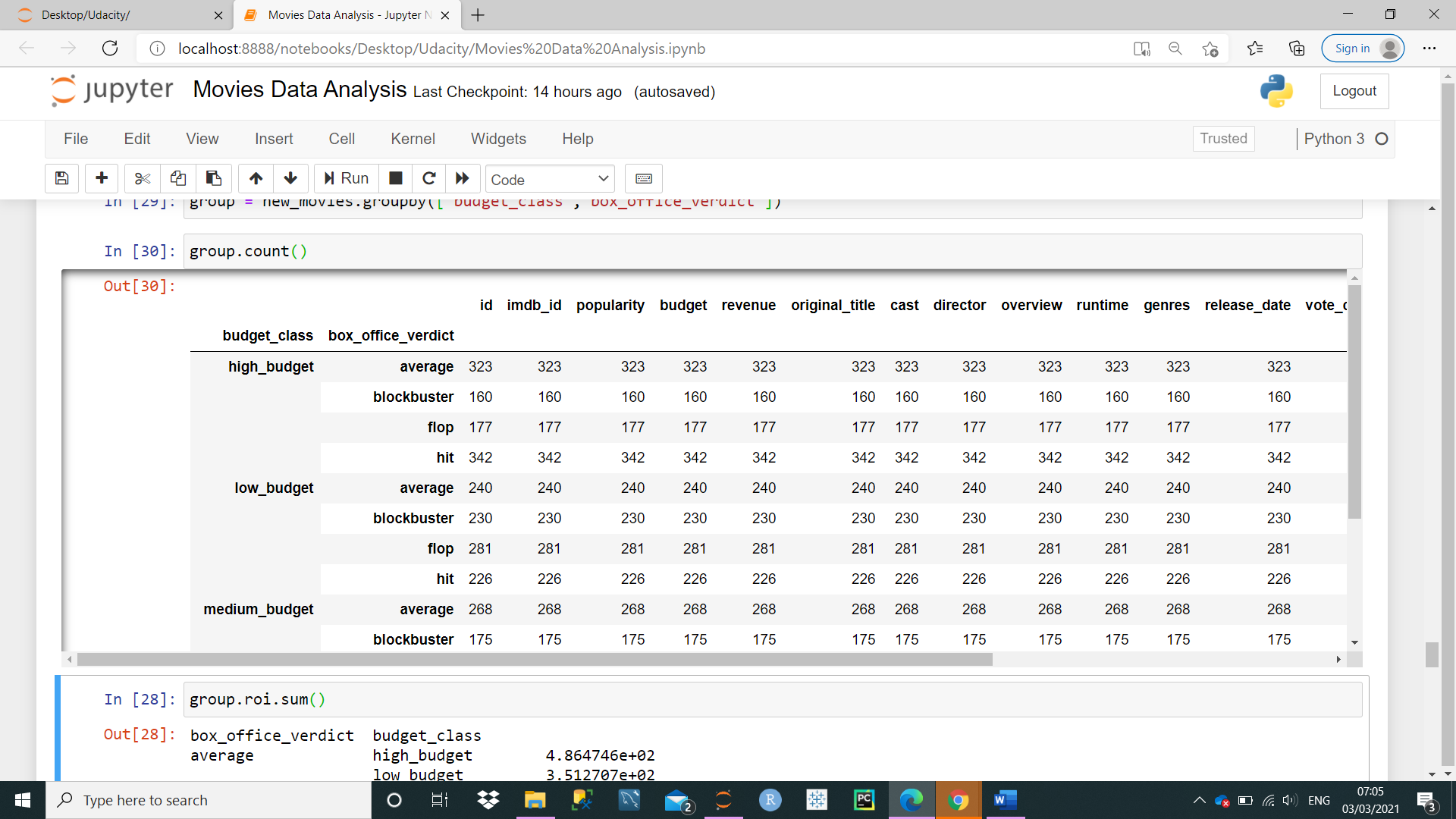
**(new\_movies['budget'] >= 24000000) & (new\_movies['budget'] < 50000000),**

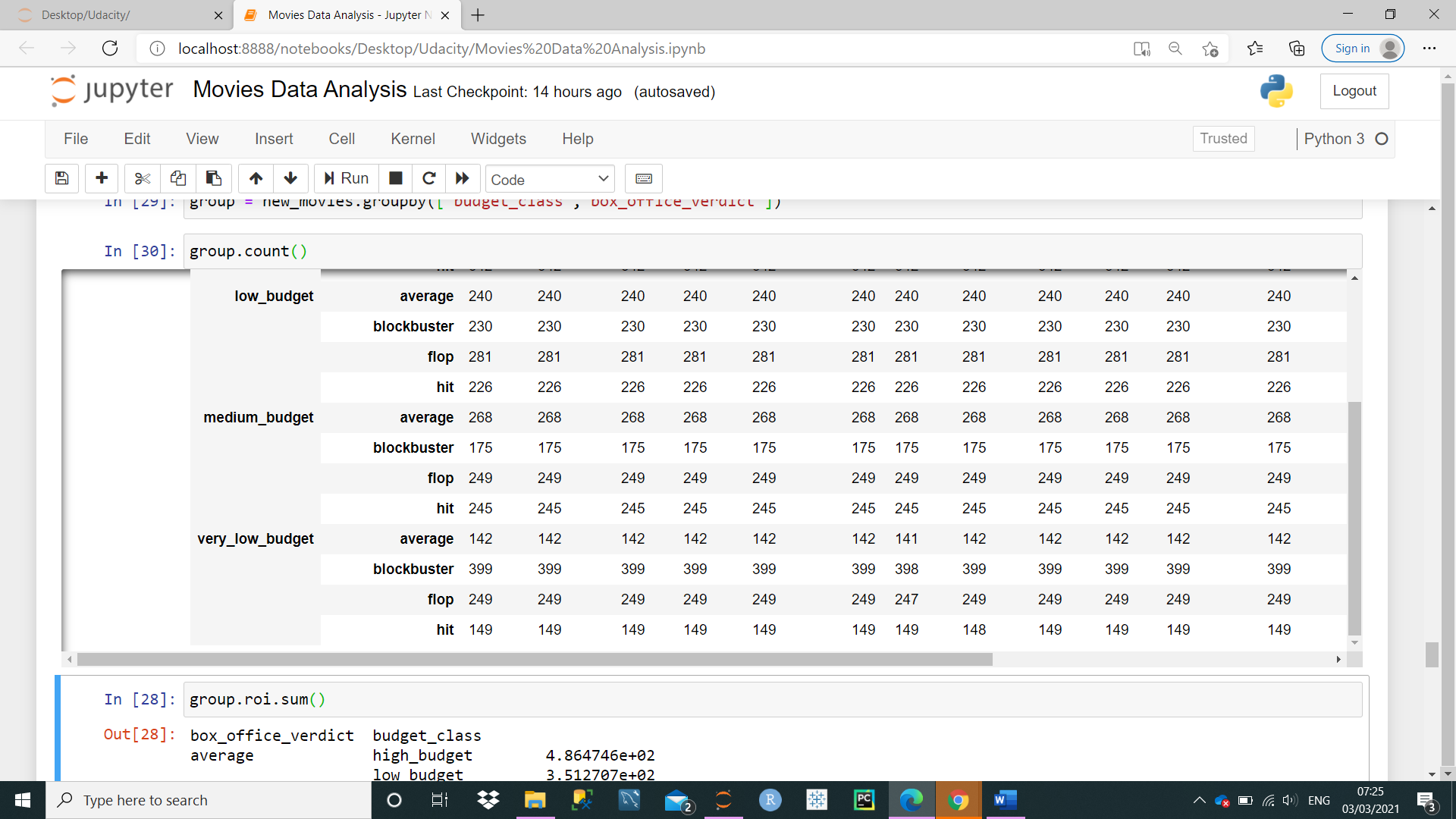
**(new\_movies['budget'] >= 50000000)**

**]**

1. **budget\_class = ['very\_low\_budget','low\_budget','medium\_budget','high\_budget']**
2. **new\_movies['budget\_class'] = np.select(budget\_conditions,budget\_class)**

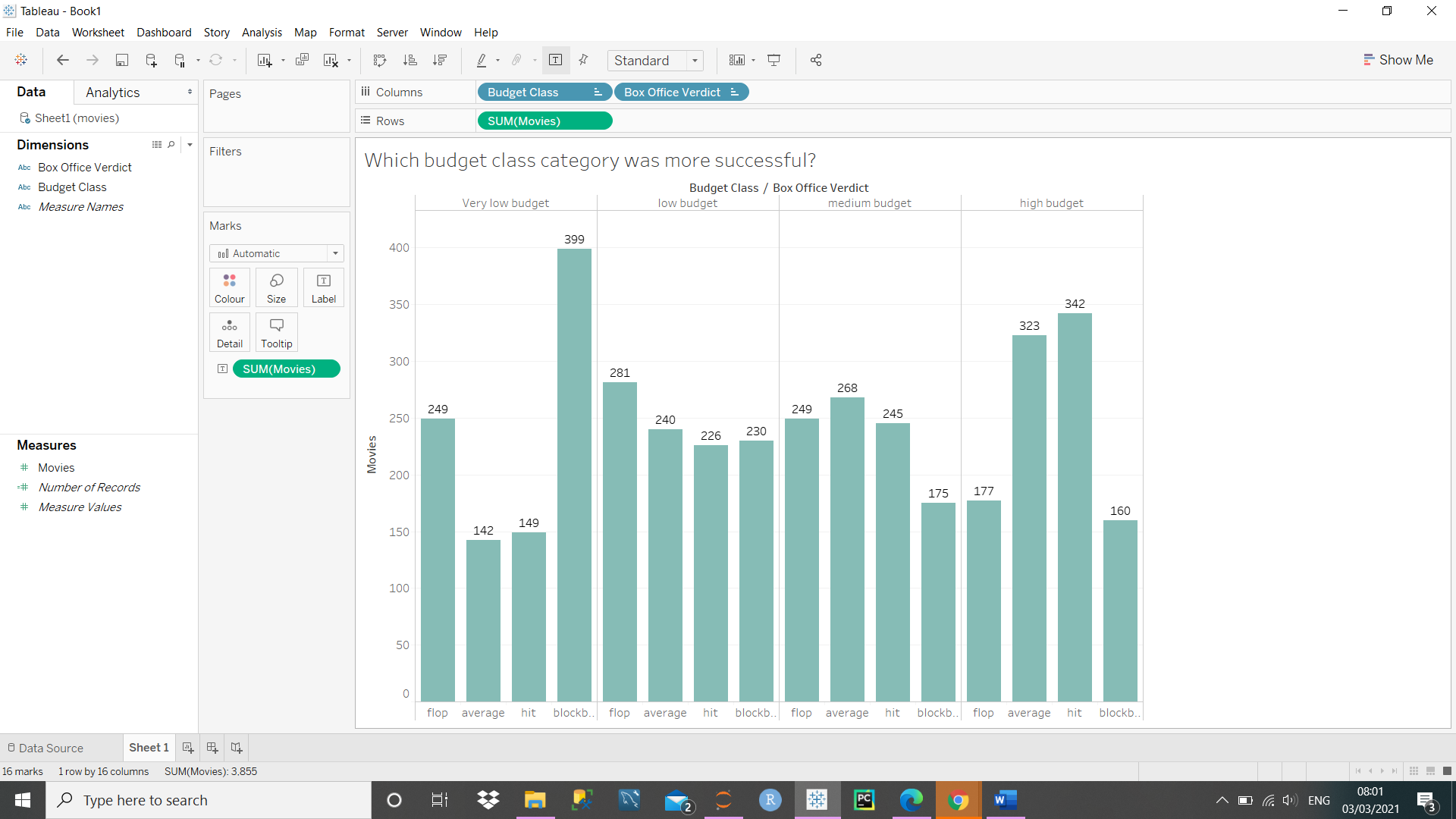
This group was based on budget class to find out which budget class was more successful. Then, these figures were used to make a visualisation graph on Tableau via excel.





### Tableau

Two graphs were made using Tableau. First is bar graph showing actual figure based on budget class and box office verdict and the 2nd graph is also a bar graph that shows for each box office verdict which budget class was most popular.

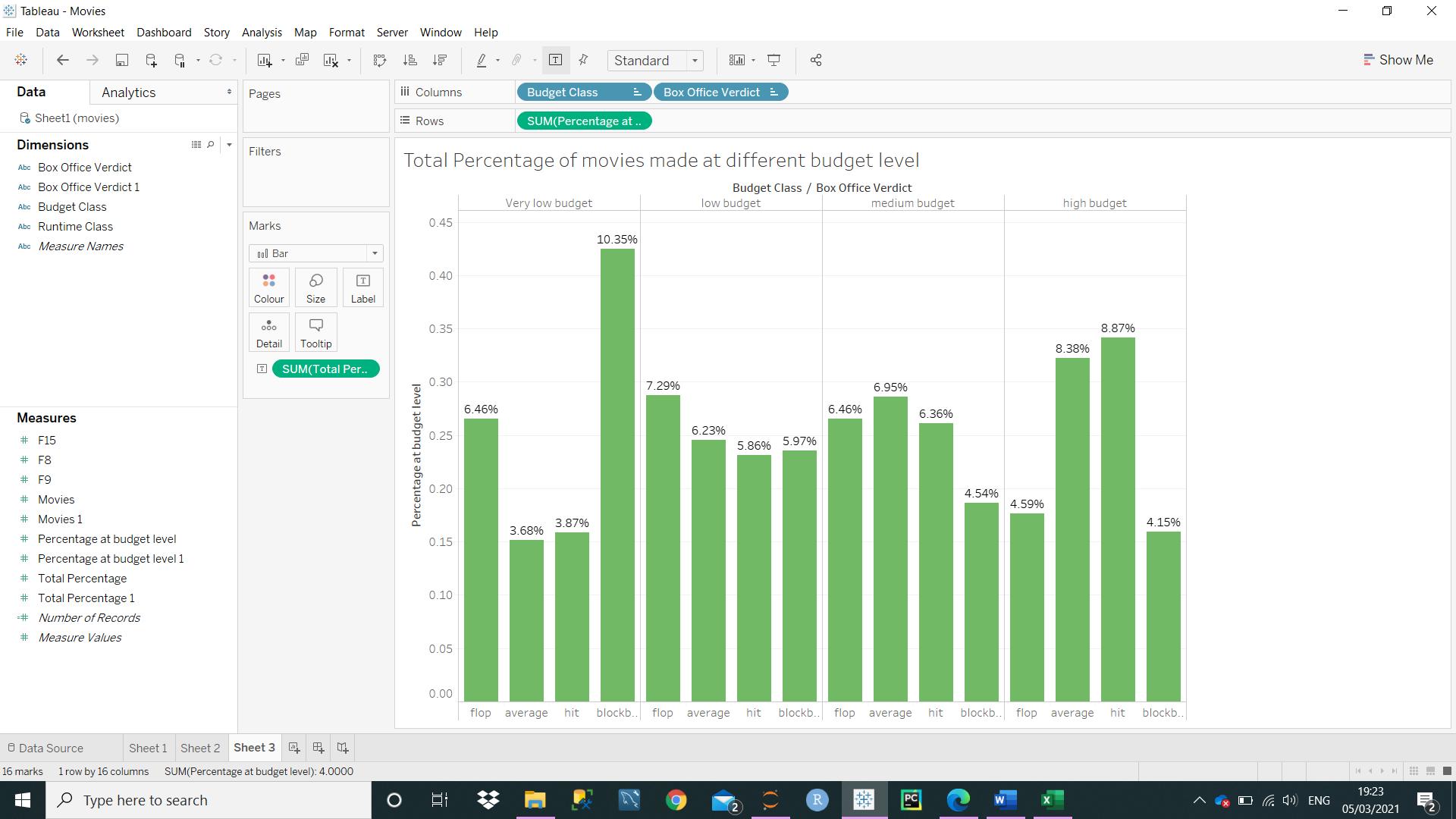


Amongst movies made with very low budget, the most popular box office verdict was blockbuster, followed by flop. Amongst movies made with low budget, the most popular box office verdict was flop, followed by average. Amongst movies made with medium budget, the most popular box office verdict was average, followed by flop. Amongst movies made with high budget, the most popular box office verdict was hit, followed by average.

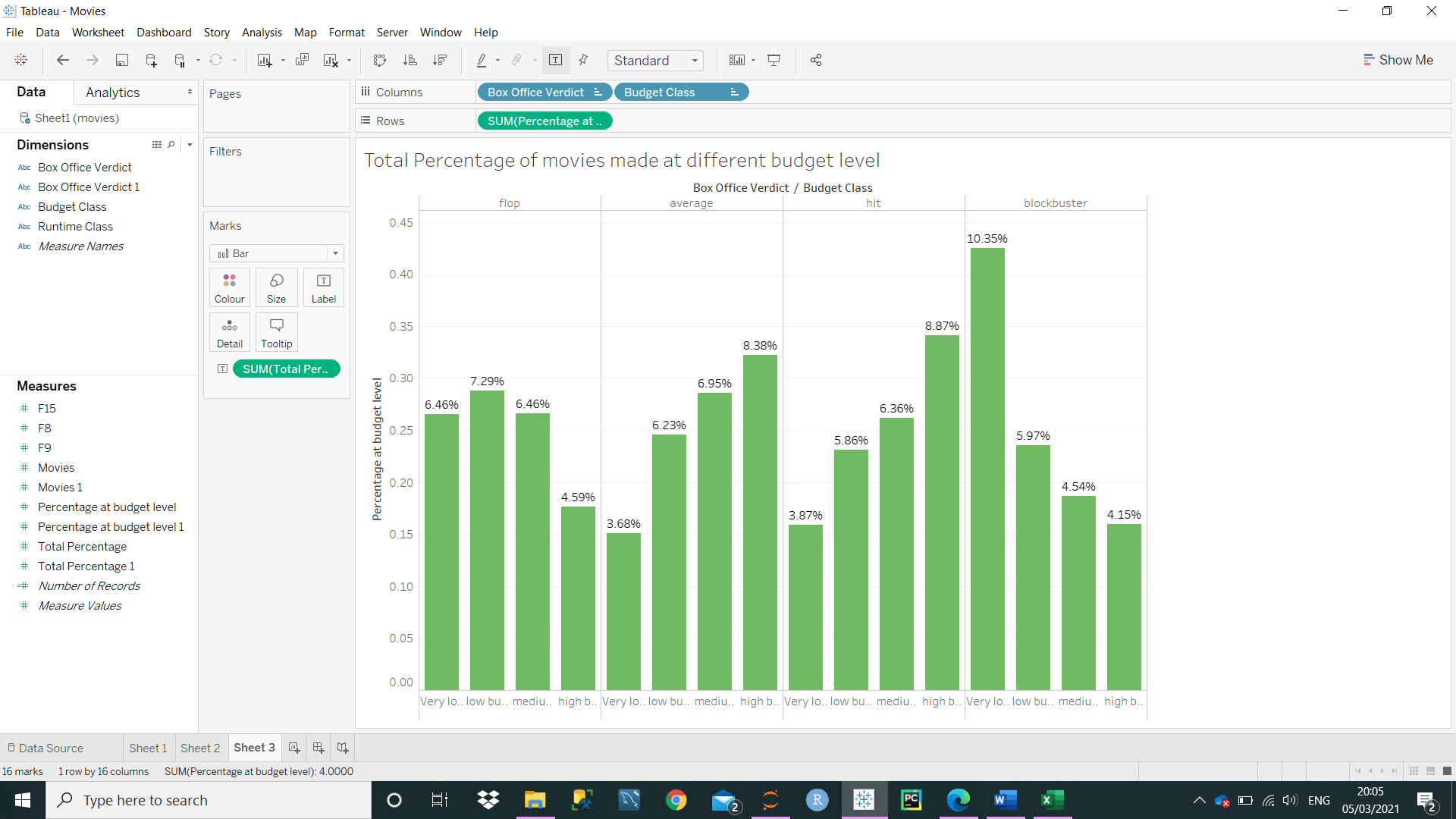
Amongst movies which were flops, the lowest number of movies were made with *“high budget”* (over 50 million budget). So, we can at least suggest that high budget movies were more likely to at least recover costs (break-even). So, in terms of safe investment high budget movies were safe bet.

Amongst movies which were blockbusters, the highest number of movies were made with “*very-low budget*” (less than 10 million budget). The cost of budget is very low so there is more chance to get better ROI than movies made with higher budget considering there is also good content within the movie. So, if investing in a movie under 10 million returned highest ROI (as expected).

The following graph shows total percentage ratio of movies made at different budget level. So, amongst the movies which were made at very low budget, movies which achieved blockbuster status got the highest figure at 10.35%. Then, movies which were made at low budget achieved the second best figure at 5.97% that achieved blockbuster status.



Most of the movies which are made at very low budget likely to achieve blockbuster status. For movies achieving average to hit status there is linear growth e.g., medium budget movie fared better than low budget movie.



If the information pointed out in the previous paragraph was taken into account, then it would have answered our first question that success of a movie is dependent on budget. But, then the last box office verdict completely is vice versa from others.

Amongst movies which had highest return on investment there was inverse positive trend. So, medium budget movies had better chance of getting ROI great than 4.21 than high budget movies. Low-cost movies performed better than medium budget and very low budget movies fared better than low budget movies.

So, in conclusion when we do deeper analysis then it is noticed that there is no relationship between budget and revenue but it can be implied that content plays a bigger role.

## Director Analysis

Another aim of this project was to analyse whether if a movie made by a specific director influenced the revenue. However, there are many directors in the data who made only 1 movie so it would not be right comparable to compare directors who made 10 plus movies with those who only made 1 or 2 movies. So, I removed directors who made less than 6 movies by using the following formulas. In the last formula I updated the table by not including rows with count of rows of directors less than 6.

1. **value\_count = new\_movies['director'].value\_counts()**
2. **to\_remove = value\_count[value\_count <=5].index**
3. **new\_movies\_director = new\_movies[~new\_movies.director.isin(to\_remove)]**

So, to find out how each director performed, I grouped the directors with the following functions. To find out directors from top 10, I changed the ascending to equal to false and when finding out directors in the bottom 10 I changed it to true. The values were sported based on ROI.

1. **director = new\_movies\_director.groupby(by='director').agg('count').roi.sort\_values(ascending = False)**
2. **director.head(10)**

According to the data, Steven Spielberg made most number of movies at 27 followed by Clint Eastwood at 24, then Ridley Scott at 21 and then Woody Allen at 18. There 136 directors in total who made more than 5 movies.

Then, to find out more information about specific director, the following function was used by changing director names.

1. **Director\_history = new\_movies\_director[new\_movies['director'] == ‘Steven Spielberg’]**

There are 136 directors so it would not be convenient and efficient to analyse all these directors so to analysis director performance, directors were analysed which were in the top 10 and bottom 10 based on number of movies made and ROI.

So, directors in the top 10 were as follows.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Total | Blockbuster (Ratio) | Hit | Average | Flop | Success (Blockbuster + Hit) Ratio |
| Steven Spielberg | 27 | 16 | 8 | 3 | 0 | 88.9% |
| Client Eastwood | 24 | 10 | 5 | 6 | 3 | 62.5% |
| Ridley Scott | 21 | 3 | 5 | 9 | 4 | 38.1% |
| Woody Allen | 18 | 10 | 1 | 3 | 4 | 61.1% |
| Steven Soderbergh | 17 | 4 | 3 | 5 | 5 | 41.2% |
| Martin Scorsese | 17 | 3 | 6 | 7 | 1 | 52.9% |
| Tim Burton | 16 | 4 | 6 | 5 | 1 | 62.5% |
| Oliver Stone | 15 | 3 | 2 | 7 | 3 | 33.3% |
| Renny Harlin | 15 | 1 | 2 | 5 | 7 | 26.7% |
| Robert Zemeckis | 15 | 8 | 1 | 6 | 0 | 60.0% |
|  |  |  |  |  |  | 50.0% |

In the bottom 10 as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Total | Blockbuster | Hit | Average | Flop | Success (Blockbuster + Hit) Ratio |
| Adam  McKay | 6 | 1 | 3 | 2 | 0 | 66.7% |
| Dwight H. Little | 6 | 0 | 3 | 1 | 2 | 50% |
| Francis Lawrence | 6 | 2 | 4 | 0 | 0 | 100% |
| Frank Coraci | 6 | 2 | 2 | 0 | 1 | 66.7% |
| Gary Fleder | 6 | 2 | 3 | 0 | 1 | 83.3% |
| George Lucas | 6 | 5 | 1 | 0 | 0 | 100% |
| George Miller | 6 | 2 | 2 | 1 | 1 | 66.7% |
| Guillermo del Toro | 6 | 0 | 2 | 3 | 1 | 33.3% |
| James Mangold | 6 | 1 | 4 | 1 | 0 | 83.3% |
| James Wan | 6 | 4 | 0 | 1 | 1 | 67.7% |
|  |  |  |  |  |  | 72.2% |

The average of top 10 director’s success ratio is 50.0% and the average of director’s success ratio in the bottom 10 is 72.2%. There is a phenomenon (in statistics) that the more movies one director makes the lower the success ratio, because the variance increases with size. This phenomenon can be seen by the fact that the average success ratio of top 100 is 50.0% but success ratio of bottom 10 is 71.2%.

The independent variable is qualitative and other function which is being used as quantitative measure is count function. So, to the previous example it is not that straightforward. The initial question is “does movie’s director’s name influence or effect the success of a movie.” To that question even if there are few examples with success ratio over 80% then we can say yes to the answer. That is because there are few well known directors and amongst them few are very successful.

The average success ratio of bottom 10 is 71.2% so if a director in the top 10 has success ratio over 71.2% then they are proportionally successful and perceived a linear growth throughout their career. From the list the most successful director appears to be Steven Spielberg who has made most number of movies at 27, had given most number of blockbusters at 16 and given most number of hits at 8 and success ratio was the highest amongst directors who made more than 15 movies at 89.9%. Then, there is Francis Lawrence who got an impressive success ratio of 100%, so all of his movies were successful but it also worthy to note that he only made 6 movies.

In the top 10, only Steven Spielberg’s success ratio is over 72% and it is impressive to note that he has made most number of movies, according to the data. So, if a movie involves Steven Spielberg’s name then there is 89.9% chance that it would be a success. Francis Lawrence and George Lucas had an impressvie success ratio of 100%. So, in a nutshell the answer would be yes, if movies are made by certain directors then it is more likely to be successful. Obviously other factors play a big role (budget, cast, genre).

## Runtime Analysis

The third question was whether movie’s runtime had any effect on the success of a movie.

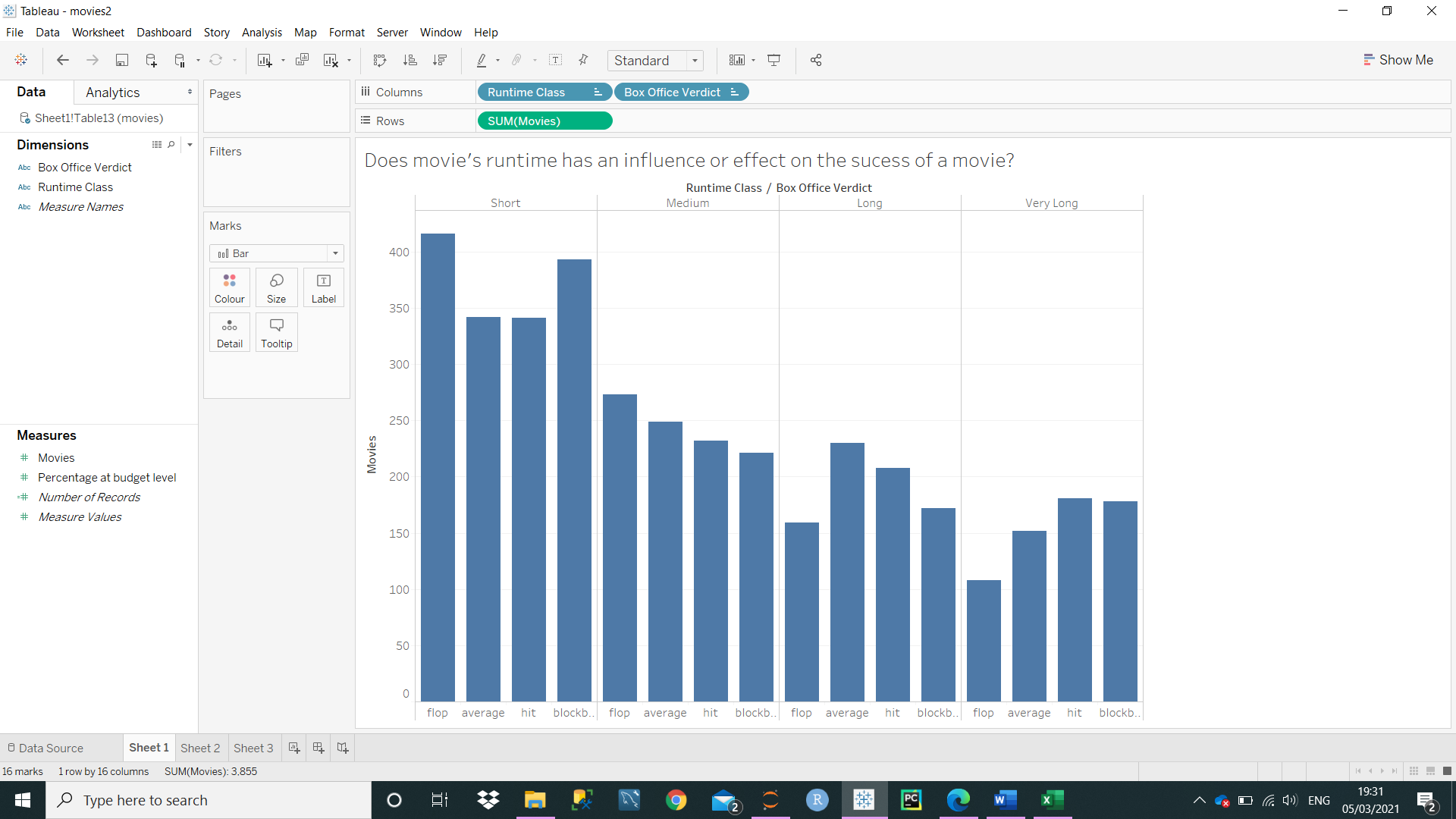
So, similar to budget and box office verdict class, runtime\_class was made based on quartiles.

First, mean of runtime was calculated with the following function, which was 102. So, the quartile which 102 appeared it that was mean medium length quartile.

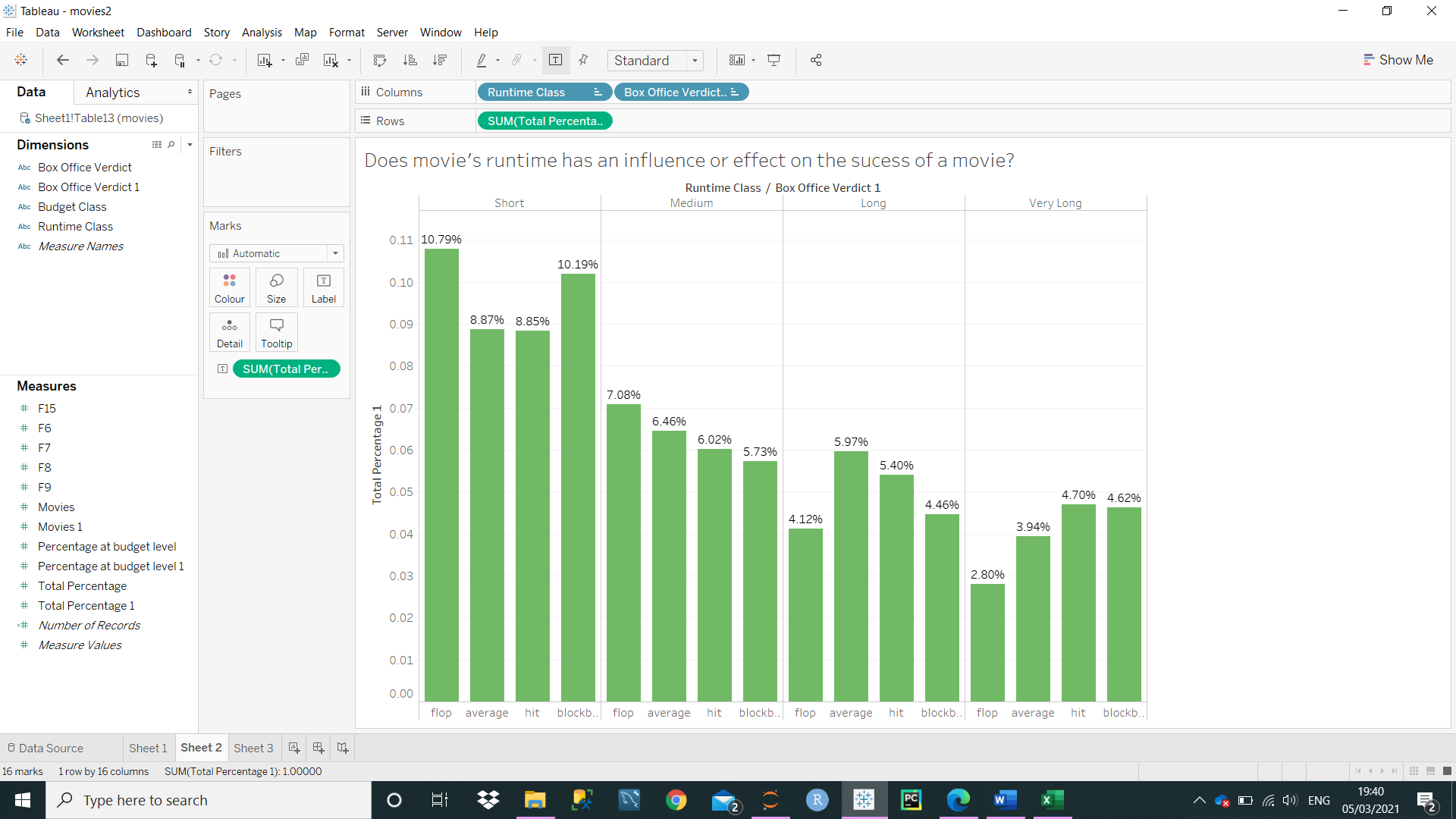
1. New\_movies.runtime.mean()

The quartiles were split into, short, medium, long and very long and a new column was made runtime class. Then, these quartiles were used with box office verdict classes to make visual graphs on Tableau.

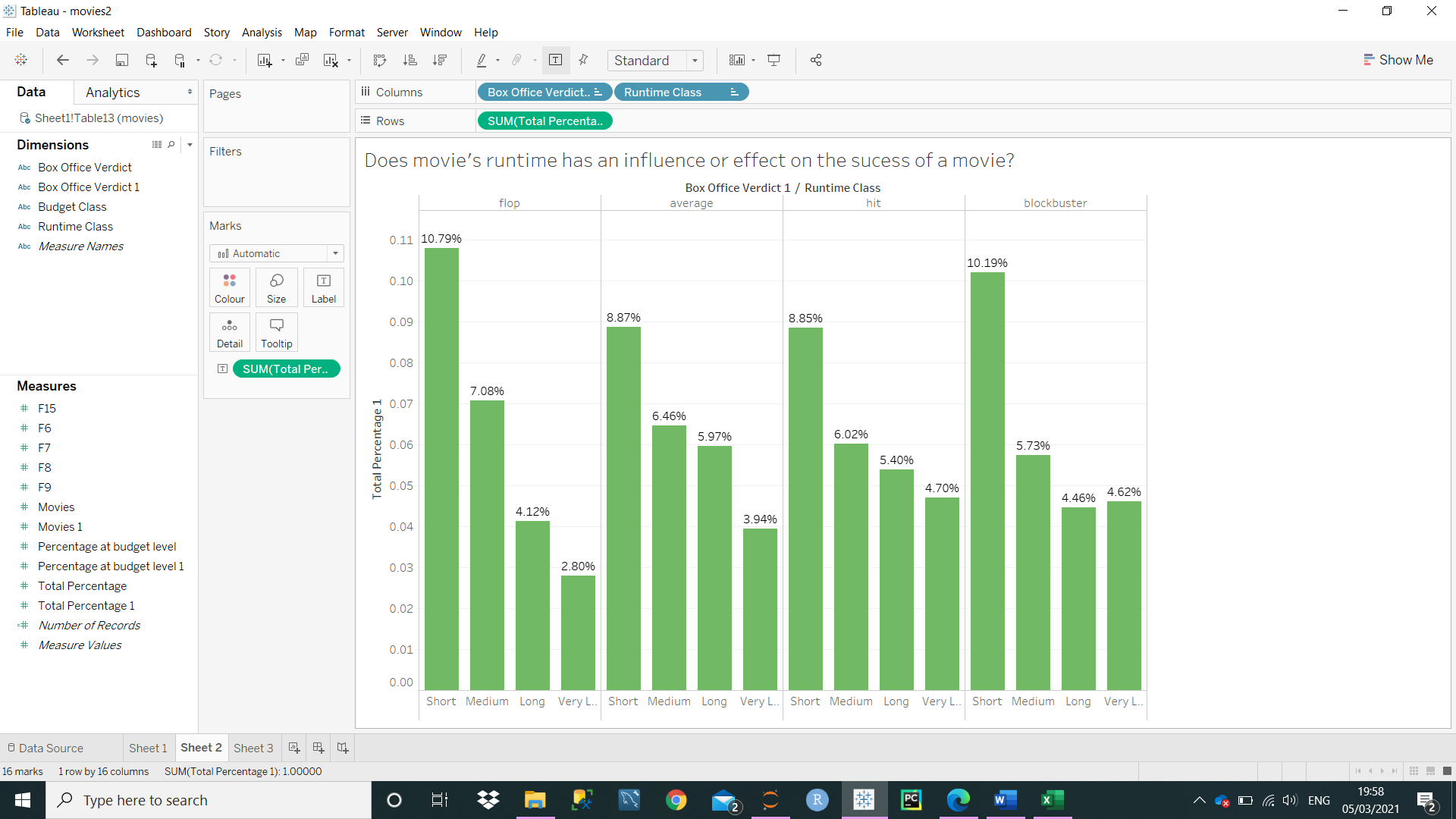
Bar graph is made by comparing movies released with specific length runtime and its box office verdict. First thing we can notice is that movies released with runtime of less 95 minutes was the most popular and then other runtime classes. Flop, average, hit and blockbuster figures were all highest in short runtime class. There is a downward trend in the graph which suggests that more movies are made with less runtime.



About 40% of total percentage of movies were made with runtime less than 95 minutes, the following graphs implies this. About 16% of total percentage of movies were made with runtime over 119 minutes.



The following graph shows at box office verdict pane level. Movies made at short runtime were more likely to achieve blockbuster status but it also apparent that it was also more likely that a movie will achieve flop status. This just applies that majority of movies are made with short runtime.



## Conclusion

In conclusion, the revenue to budget analysis shows that it is not always apparent that the higher the budget the more successful the movie will be because very low-cost budgets fared much better than high-cost movies. However, it will also true that high-cost movies were more secure investment than other classes but very low cost movies got highest return on investment.

When comparing director’s influence, if we had to answer the question if a movie is made by certain director then to that question we can say yes. We have acknowledged that movies made by Steven Spielberg attracted lot of revenues. Obviously, there would be other factors which would affect revenue of movie, but it was a fact that for someone who had made more than 10 movies he had the best success ratio. Francis Lawrence and George Lucas gained most impressive success ratio of 100%. So, there is good possibility that if certain directors are involved then it is more likely to be successful. Runtime analysis indicates that it there is no such indication that imply that runtime is related to success of a movie. It could be possible that movie makers acknowledge that people like to watch movies with short runtime, so most movies are made at shorter runtime. As more movies are short runtime there is more likely to be larger variance as size, k, is larger.