IOANNA PAPAGIANNI 2790

QUESTION 4

FROM RAW DATA TO FIXED

Taking a careful look at the dataset of movies.csv file, there are some issues that need to be fixed before moving on with the data analysis.

At first, there are columns that are unnecessary and is better to be dropped off.

These are:
'Unnamed: 0', 'Unnamed: 0.1', 'US Gross', 'US DVD Sales', 'MPAA Rating', 'Running Time (min)', 'Distributor', 'Source', 'Creative Type'

The rest of them (9 columns), are the attributes we need for the data analysis. Note that, all movies are 3201 but they are indices start from 0 to 3200.

- **Title** attribute is fine.
  3201 movies (after pruning Worldwide Gross and Release Date pruning 3187)
  Type string

- **Worldwide Gross** attribute needs some changes. Although it has all 3201 movies there are some marked as "unknown" (7 movies). These movies are pruned. After pruning, data are converted to float. Renamed to: **Gross**
  3201 (after pruning 3187)
  Type string (converting to float)

- **Release Date** attribute format is month/day/year for the most, but there are movies with year only and others cointain alpharithmetics. For our analysis, year only fits the criteria and we keep only that. Firstly, we prune spaces (if any) on the right side of every value for this attribute and then we keep only the last two characters if and only if they are numbers, in two digits format. Anything else is pruned. After pruning, depending on the century they belong (e.g: if the number is 01 → 2000 century, 97 → 1900 century, the logic is [0,19]*→ 2000 century else → 1900 century) two more digits are added in front. Renamed to: **Date**
  3201 movies (after pruning Worldwide Gross and Release Date pruning 3187)
  Type string (converting to int)

  * 19 stands for 2019 (the current year). Notice that, there is a movie "Over the Hill to the Poorhouse" that was released in 1920. Fortunately, this dataset contains movies till 2016.

- **Mayor Genre** attribute has a lot of different genres. Where there is <space> we consider the category as a whole, but where there is </> means that this movie belongs in two categories. We made a dataframe called genres_df, with 'First' and 'Second' as columns to store the different genres a movie can be. We do not need titles as long as we have their indices.
  Renamed to: **Genre**
  2926 movies
  Type string

- **Rotten Tomato Rating** attribute has ratings on the [0,100] climax. We divide its of them by 10 so as to fit IMDB Rating climax [0,10]. Renamed to **RTRating**

2321 movies
Type float

- **IMDB Rating** attribute is fine. Renamed to **IMDBRating**
  2988 movies
  Type float

- **IMDB Votes** attribute is fine. Renamed to **IMDBVotes**
  2988 movies
  Type float

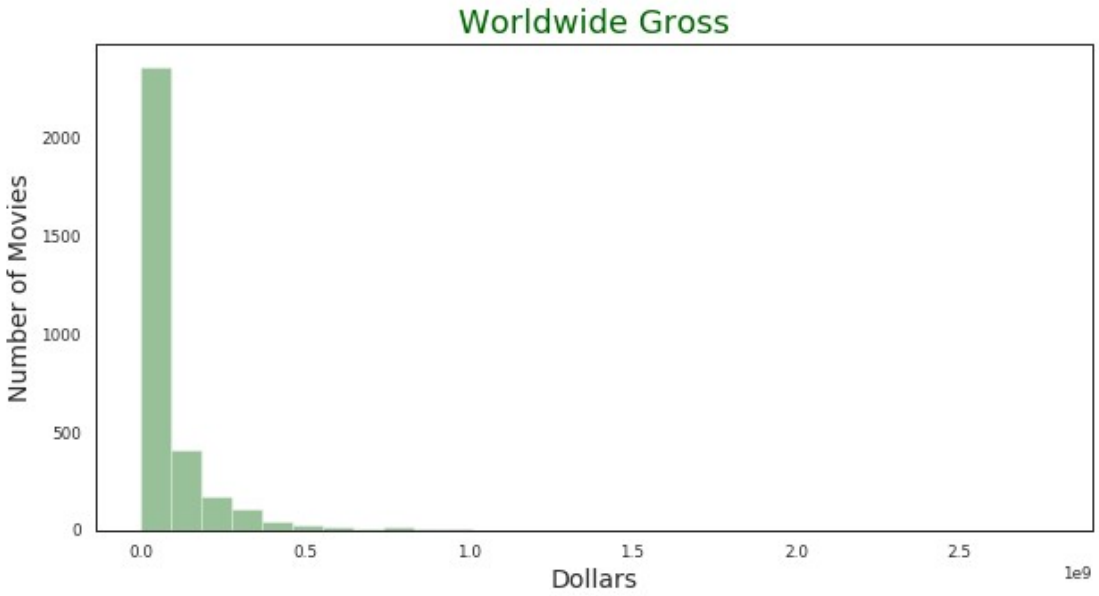
- **Production Budget** attribute is fine. (This attribute is needed for the analysis in part 4.).
  Renamed to: **Budget**
  3200 movies (after pruning Worldwide Gross and Release Date pruning 3187)
  Type float

- **Director** attribute is fine. (This attribute is needed for the analysis in part 4.).
  1870 movies
  Type float

After Worldwide Gross and Release Date pruning, indices will be reset and renamed as “ID” starting from 1 instead of 0.

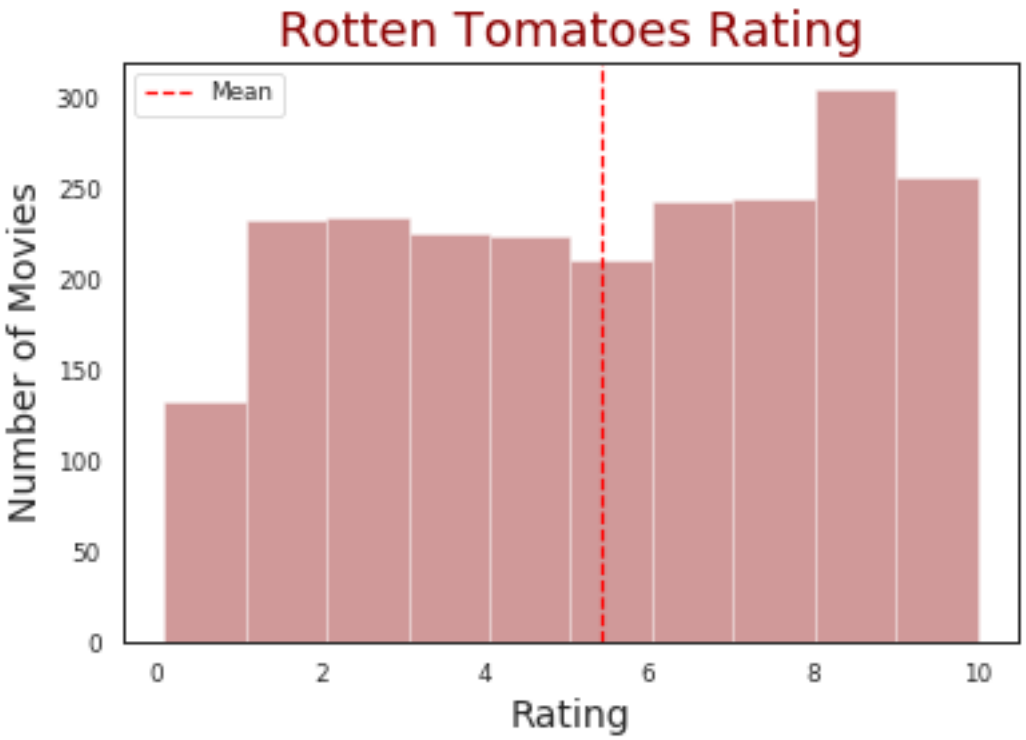
1. HISTOGRAMS AND DISTRIBUTIONS

WORLDWIDE GROSS: Skewed Non Symmetric (Non-Normal) Right-long-tailed Histogram Geometric Distribution:



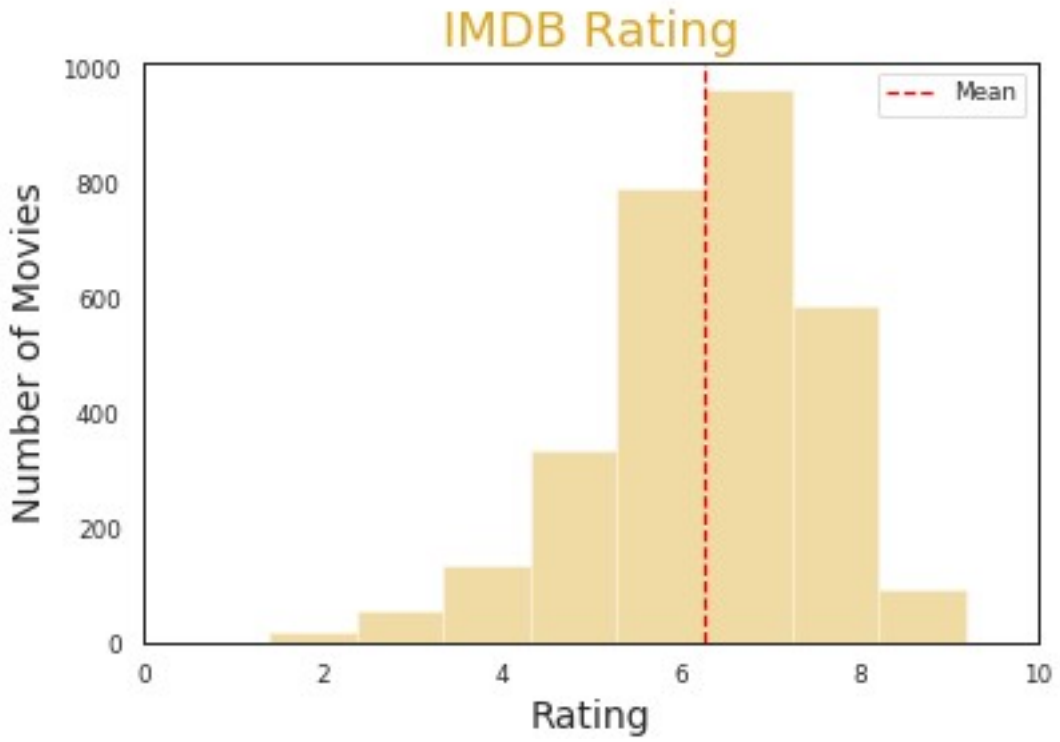
- Data are located in the left side.
- Scale of data: interval [0.0 – 2767891499.0] dollars.
  The spread of data is not really wide. Note that our data points, range from $0 up to $2767891499, only one reaches 2 billions and a few 1 million.
- The histogram is “skewed” , there is no-mirroring of the data.
- This distribution is non-symmetric. It has a long tail (on the right) relative to the other tail (left side has no tail at all). This phenomenon occurs, beacause the lower bounds of the data are significantly more frequent than the upper bounds.

ROTTEN TOMATOES RATING: Symmetric and Unimodal Histogram
1 peak Unimodal short–tailed Distribution



- Data, cluster around a single mode (1 peak unimodal). Tails here, approach zero very fast. We could say that Rotten Tomatoes ratings are well-distributed, ratings vary a lot and overspread through whole data scale .
- Scale of data : interval [0.1,10.0] points**.
- No skewness.

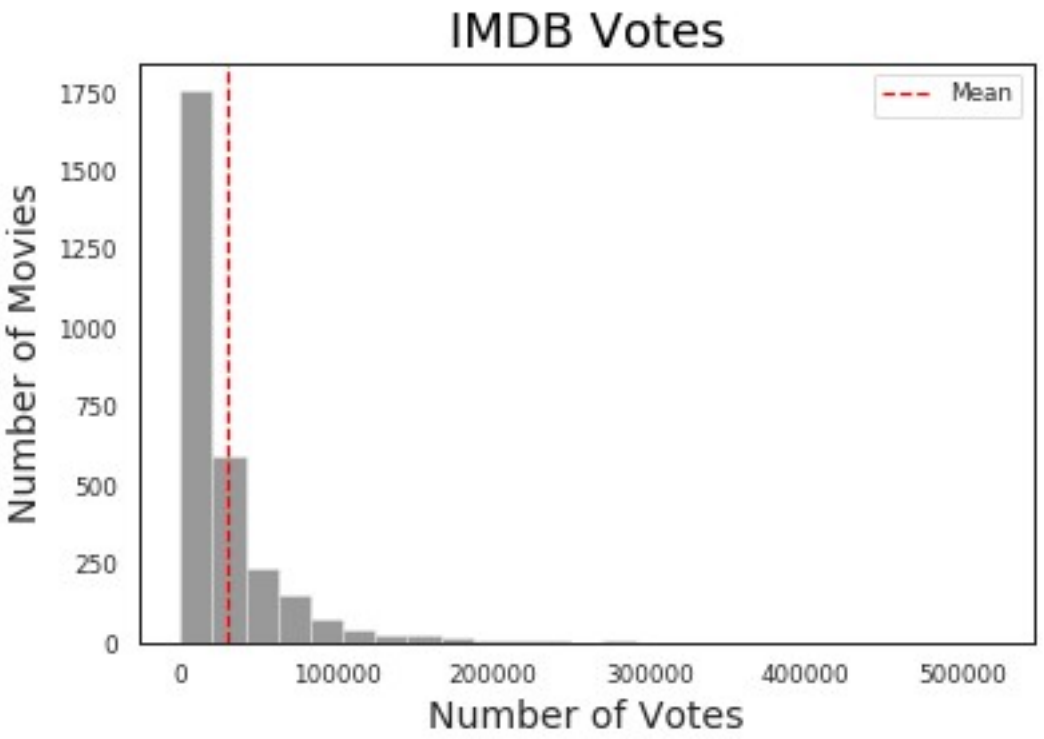
IMDB RATING: Skewed Non-Symmetric Normal moderate-tailed Histogram
Normal Distribution



- The data are normally distributed with a litle skewness to the right. IMDB Ratings follow the normal distribution, centrality is around 6 and 7 points and as we can see there is a tendency for higher voting than lower.
- Scale of data : interval [1.4, 9.2] points**. We managed to contain the whole scale.
- There is a litle right skewness and we conclude that most ratings tend to be higher than the middle rating point of the scale (that is, number 5).

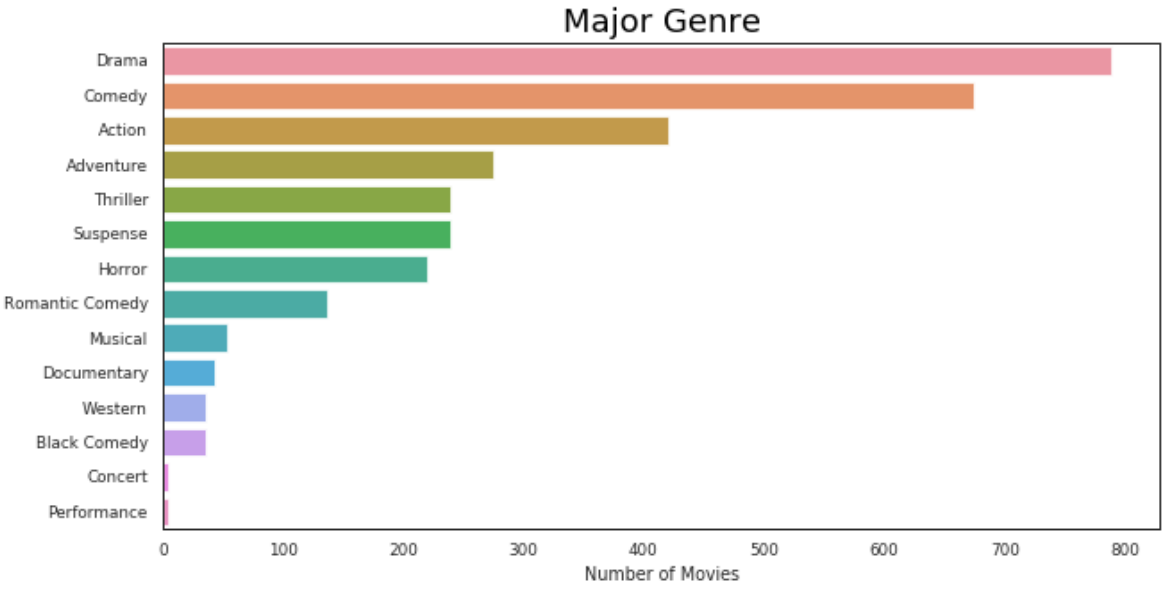
**For all ratings points are between 0 and 10.

IMDB VOTES: Skewed Non Symmetric (Non-Normal) Right-long-tailed Histogram Geometric Distribution



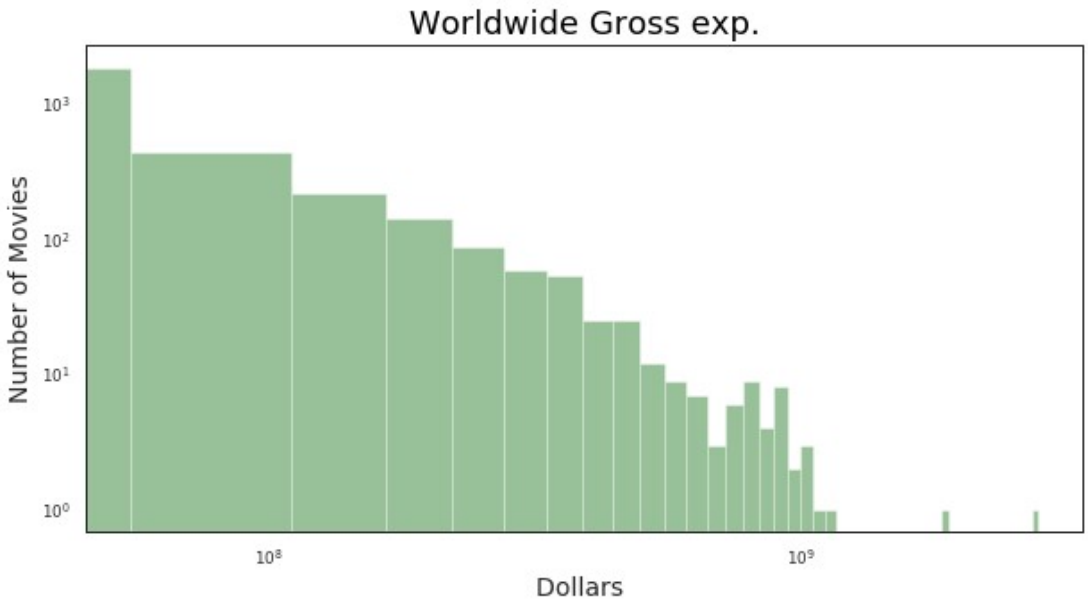
- Most of the data are located in the left side and with a first glance we could say that the centrality of the number of votes is between 0-30000.
- Scale of data : interval [18, 519541] votes. We managed to contain the whole scale. Data are not normally spread all along the interval.
- Again, data are skewed to the left and this distribution is non-symmetric. It has a long tail (on the right) relative to the other tail (left side has no tail at all). This phenomenon occurs, due to huge dispersion of our data points, huge difference between lower and upper bounds (as we explained above).

MAJOR GENRE: Number of Movies per Genre, Bar Plot *"Oh! Too much DRAMA"*



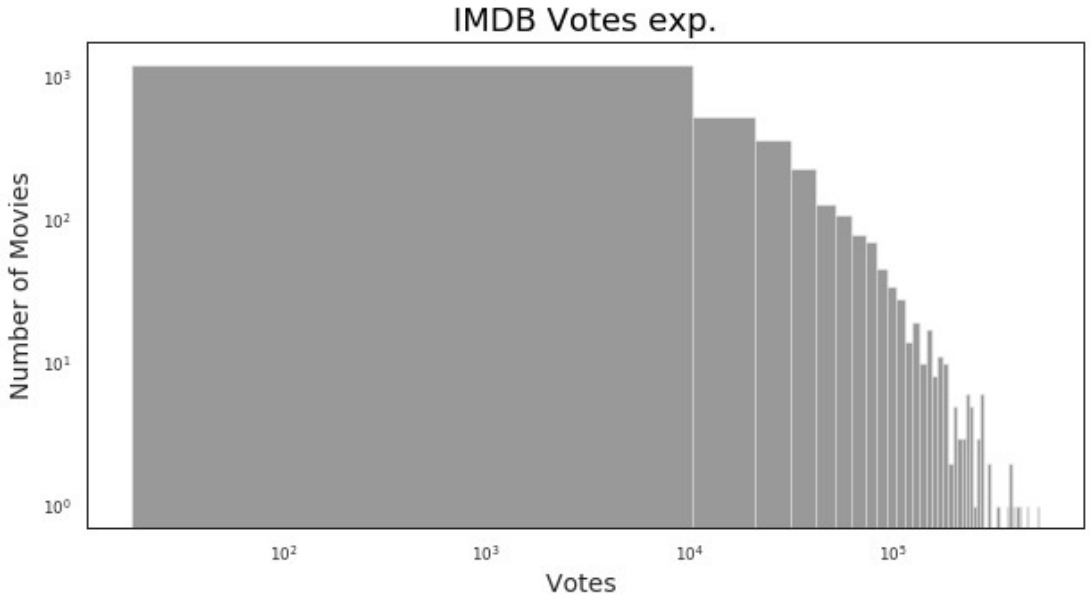
The barplot shows some interesting analytics for what kind of movie, Production Companies, prefer to release. There is a tendency to fund and therefore, produce more "Drama" and "Comedy" movies. The bronze metal goes to "Action" films and "Performance" and "Concert" genres come last.

WORLDWIDE GROSS (exp.) HISTOGRAM



It seems that Worldwide Gross follow exponential regression with two outliers in the right side that managed to form small bins. Outliers should be seriously considered. They show us the great dispersion that movies can have on their revenue. Big money is for the minority!

IMDB VOTES (exp.) HISTOGRAM

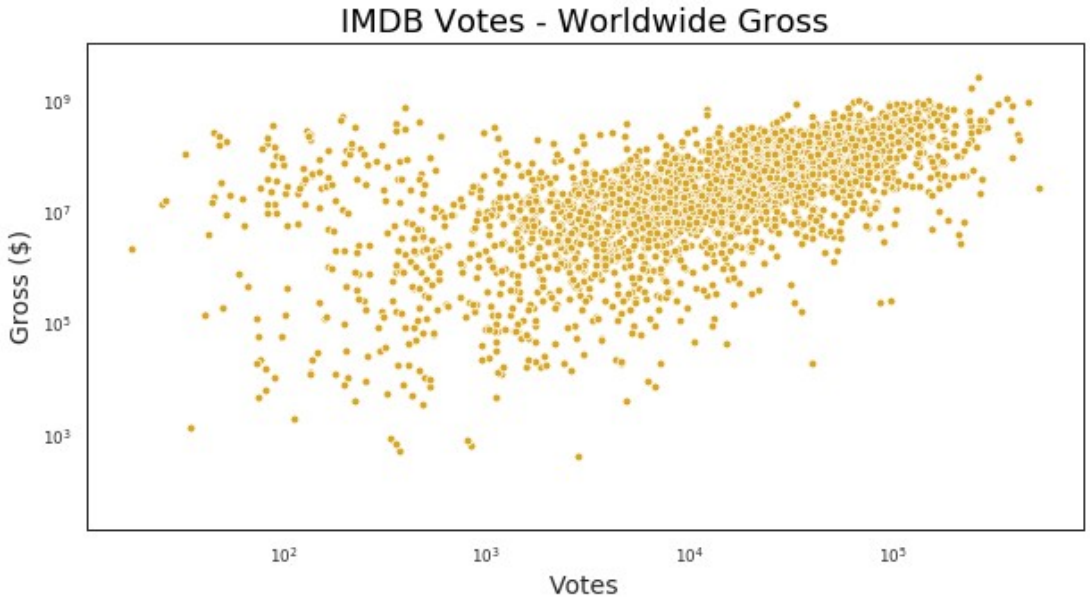


IMDB votes follow exponential regression. There are a few movies that reach up to 10^5 votes . As the votes increase, bins contain less and less quantity of data.

Summing up, in both cases, data are gathered in the left side and we cannot even see the end of the tail, the tail drops significantly and the bins get “thinner”. There are in both scatterplots outliers on the right side, that can strongly affect the result of our analysis.
In other words, there are huge deviations from the mean and the climax of data is large enough (high upper bounds) to increase the mean number a lot.

IMDB VOTES – WORLDWIDE GROSS (exp.) SCATTERPLOT

Scatterplots are useful tools to compare data sets against each other to see if there is a relationship.



In this scatterplot, it is easy to realize that there is a **high positive linear correlation** between the revenues and IMDB votes. Most data points are clustered together.

2. SCATTERPLOT, Z-TEST, P-VALUE, CORRELATION COEFFICIENTS

P-value: For a given statistical model, the probability that, when the null hypothesis is true, the absolute value of the sample mean difference between two compared groups would be greater than or equal to the actual observed results. It measures how compatible our data are with the null hypothesis.

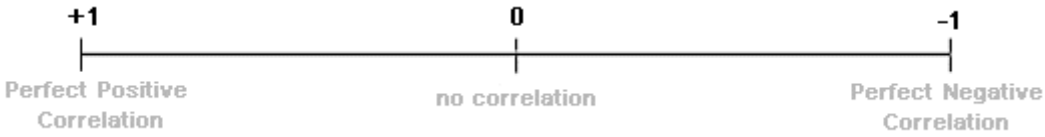
We run two sample **z-test** to find p-value. The value in a two sample **z-test,** is the difference between mean of sample1 and mean of sample2 under the Null Hypothesis.

There is a formula that we follow to interprete Pearson's and Spearman's Correlation Coefficient:

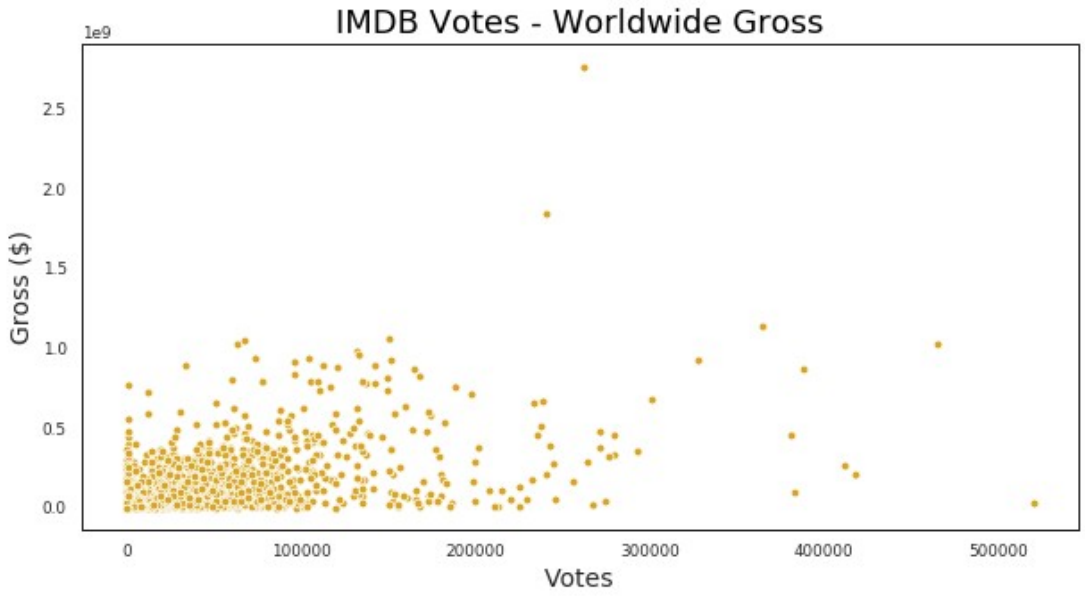
PEARSON:

Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

SPEARMAN:



IMDB VOTES – WORLDWIDE GROSS SCATTERPLOT :



As we can see from the IMDB Votes – Worldwide Gross Scatterplot, there is a **strong clustering** in the left corner of the axis. As vote axis increases, gross axis looks to stay around the same values. Therefore, we could make a hypothesis that says: " Movies are more popular online ".

2-Sample Z-Test (Worldwide Gross and IMDB Votes):

Our code stests.ztest has:

x1: IMDB Votes
x2: Worldwide Gross
value : 0 so, mean(IMDB Votes) - mean(Worldwide Gross) – 0
alternatives: 'larger' thus, mean(IMDB Votes) - mean(Worldwide Gross) > 0

H0: Movies are more popular online, than in the cinema.

If the calculated p-value is less than 0.05 our Null Hypothesis (H0) does not hold good, so we need to reject it and propose (H1) that movies are equal or more popular at the cinema, than online . Below, are the results of our z-test:

```
H0: Movies are more popular online, than in the cinema.

p-value:  1.0

Accept Null Hypothesis (H0)
```

Pearson Correlation Coefficient results:

	Gross	IMDBVotes
Gross	1.000000	0.551452
IMDBVotes	0.551452	1.000000

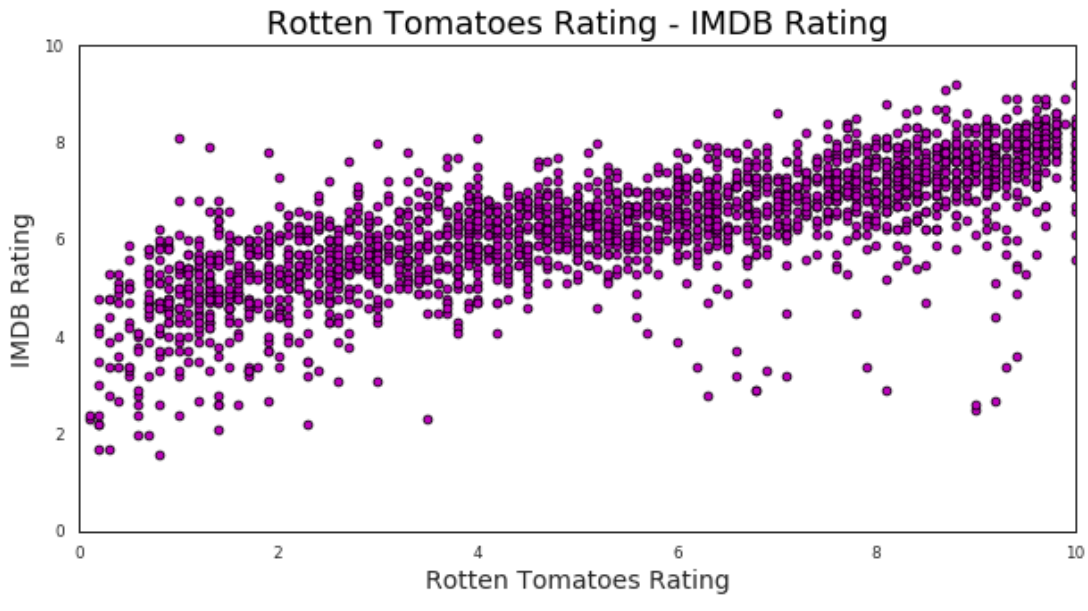
Spearman Correlation Coefficient results:

	Gross	IMDBVotes
Gross	1.000000	0.655319
IMDBVotes	0.655319	1.000000

That means, **our data have large strength association**.

Taking into account, p-value, Pearson's and Spearman's Correlation Coefficients we could say that our assumptions (H0) are absolutely right and Worldwide Gross is strongly correlated with IMDB Votes.

ROTTEN TOMATOES – IMDB RATING SCATTERPLOT:



As we can see from the Rotten Tomatoes – IMDB Rating Scatterplot, there is a **strong linear positive correlation** between them.

Small values of RTRating correspond to small values of IMDBRating – Large values of RTRating correspond to large values of IMDBRating.

If we take a careful look, we see that when people vote movies under 5 in Rotten Tomatoes, people in IMDB vote (for the same movies) a litle higher. However, when people vote above 6 in Rotten Tomatoes, people in IMDB vote lower than those in Rotten, still above 6.

2-Sample Z-Test (Rotten Tomatoes Rating and IMDB Rating):

1st test:

Our code stests.ztest has:

x1: Rotten Tomatoes Rating
x2: IMDB Rating
value : 0 so, mean(Rotten Tomatoes Rating) - mean(IMDB Rating)) – 0
alternatives: 'two-sided' thus, mean(Rotten Tomatoes Rating) - mean(IMDB Rating)) != 0

H0: People in Rotten Tomatoes and IMDB vote similarly.

```
H0: People in Rotten Tomatoes and  IMDB vote similarly.

p-value:  7.21302915029273e-43

Reject Null Hypothesis (H0)
```

Our assumption was wrong, although the scatterplot seems to follow linear correlation.
We have to assumpt now that:

H1: People in Rotten Tomatoes vote higher than in IMDB.
or
H2: People in Rotten Tomatoes vote lower than in IMDB.

2nd test:

Our code stests.ztest has:

x1: Rotten Tomatoes Rating
x2: IMDB Rating
value : 0 so, mean(Rotten Tomatoes Rating) - mean(IMDB Rating)) – 0
alternatives: 'larger' thus, mean(Rotten Tomatoes Rating) - mean(IMDB Rating)) > 0

H1: People in Rotten Tomatoes vote higher than in IMDB.

```
H1: People in Rotten Tomatoes vote higher than in IMDB.

p-value:  1.0

Accept Null Hypothesis (H1)
```

Our assumptions are true, people vote higher in Rotten Tomatoes than in IMDB. We do not need to proceed with H2 hypothesis.

Pearson Correlation Coefficient results:

	RTRating	IMDBRating
RTRating	1.000000	0.742951
IMDBRating	0.742951	1.000000

Spearman Correlation Coefficient results:

	RTRating	IMDBRating
RTRating	1.000000	0.776429
IMDBRating	0.776429	1.000000

That means, **our data have indeed large strength association** based on the formulas. Nonetheless, people vote higher in Rotten Tomatoes.

NOTE: If we concatenate in a column "Ratings" all the ratings, and make a new column "Websites", that it keeps the websites the votes come from, we could **have a better visualization of this correlation,** making a scatterplot that it has different colors for each of these websites, x-axis and for the "Ratings" and y-axis for the amount of movies.

2. BARPLOT WITH CONFIDENCE INTERVALS, T-TEST

For this analysis we group movies by genre and for every group we calculate the mean gross and the standard deviation.

In order to find the confidence intervals, we use :

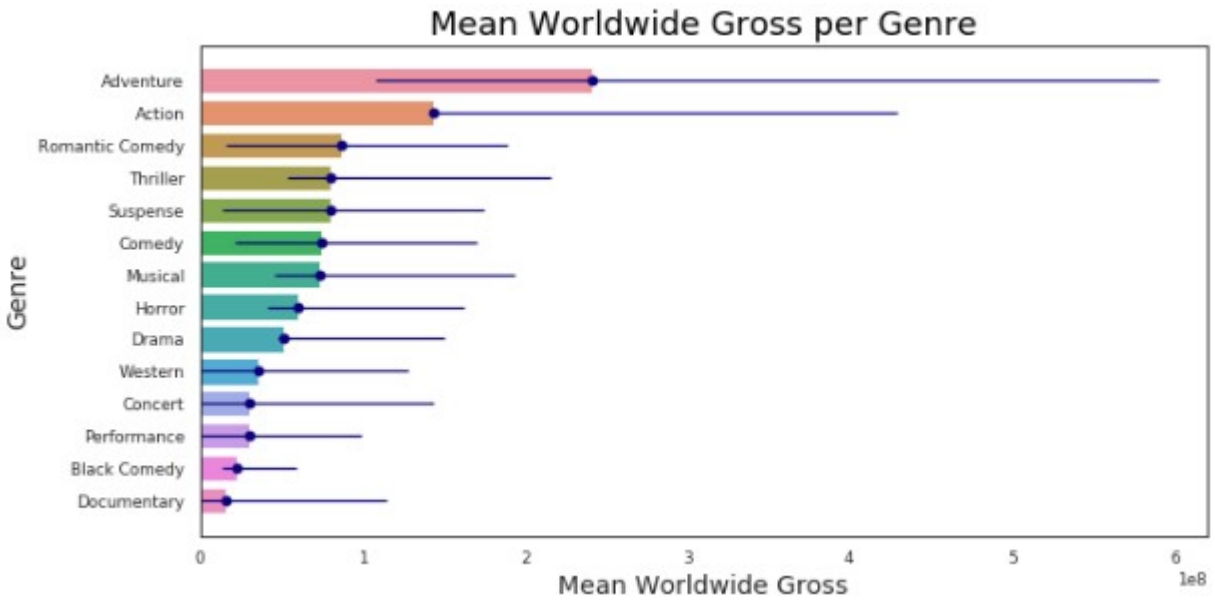
*stats.norm.interval(**0.95**, **loc** = m_gross_genre1_df['Mean Gross'],*
__scale__ = std_gross_genre1_df['Std Gross'] / np.sqrt(std_gross_genre1_df['Std Gross'].count()))

- **95% confidence interval** means that there is a 95% chance that the confidence interval we calculated contains the true population mean. The confidence interval covers the true value in 95 of 100 studies performed.
- **loc** is the mean gross we calculated
- **scale** is: (std gross)/sqrt(number of std grosses)

And as such, it returns the bottom and upper boundaries of the confidence interval for every genre.
For bottom errors that contain negative values, we substitute them with the gross mean in order to have only positive axis.

NOTE: In spite of the fact that this dataset has not many movies in some genres , we did **not** pull them out. The size of the confidence interval depends on the sample size and the standard deviation of the study groups. In general, the higher the probability to cover the true value (95% in our case) the wider the confidence interval.
If the confidence interval is wide, either it means that the sample is small or the dispersion is high.

MAJOR GENRE – MEAN WORLDWIDE GROSS BARPLOT:



Firstly, it is clear that adventure movies are far more profitable than the rest. Interesting is, that drama movies, the most preferable among production companies, are not that prosperous at all and that is also for comedies. Secondly, the confidence interval for adventure movies is enormous! As it comes 4rth in the production ranking (checking back on the previous barplot about popularity of genres), the reason seems to be that it has wide dispersion.

In order to find out which of these differences are statistical significant we run **t-tests.**
T-test, compares two means and tells us if the average difference between these two, is really significant or if it is due instead to random chance.

The Null Hypothesis in T-Tests is assuming that they are equal.

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

Mx : sample1 mean
My: sample2 mean
Sx: sample1 std
Sy: sample2 std

n_x: number of sample1
n_y: number of sample2

Thanks to scipy.stats we do not need to do that via hard coding. For obvious reasons we set sample1 as the one with the higher mean.

We use *equal_var = False* to specify that they do not have equal variances.

2-Sample T-Test (Adventure - Action Genre):

We examine the first two genres : adventure movies and action movies. They both have huge ci's and t-test will testify if the average difference between two groups is really significant.

H0: There is no significant difference between adventure movies mean gross and action movies mean gross.

```
H0: There is no significant difference between adventure movies mean gross and action movies mean gross.

p-value:  5.520054519574185e-07

Reject Null Hypothesis (H0)
```

Our hypothesis is incorrect, that means the difference between those two is significant.

2-Sample T-Test (Comedy- Drama):

We examine the first two genres in the "number of movies produced" race, these are drama movies and comedy movies. They both have similar ci's and t-test will testify if the average difference between two groups is really significant.

H0: There is no significant difference between comedy movies mean gross and drama movies mean gross.

```
H0: There is no significant difference between comedy movies mean gross and drama movies mean gross.

p-value:  4.012664537132461e-06

Reject Null Hypothesis (H0)
```

Our hypothesis is incorrect, that means the difference between those two is significant.

2-Sample T-Test (Drama - Western):

Now, we test if drama movies and western movies have signifant levels in their mean difference. They both have similar ci's BUT t**he sample of western movies is really small whereas drama movies come first in the sample size** from this dataset and t-test uses std and number size to calculate the result.

H0: There is no significant difference between drama movies mean gross and western movies mean gross.

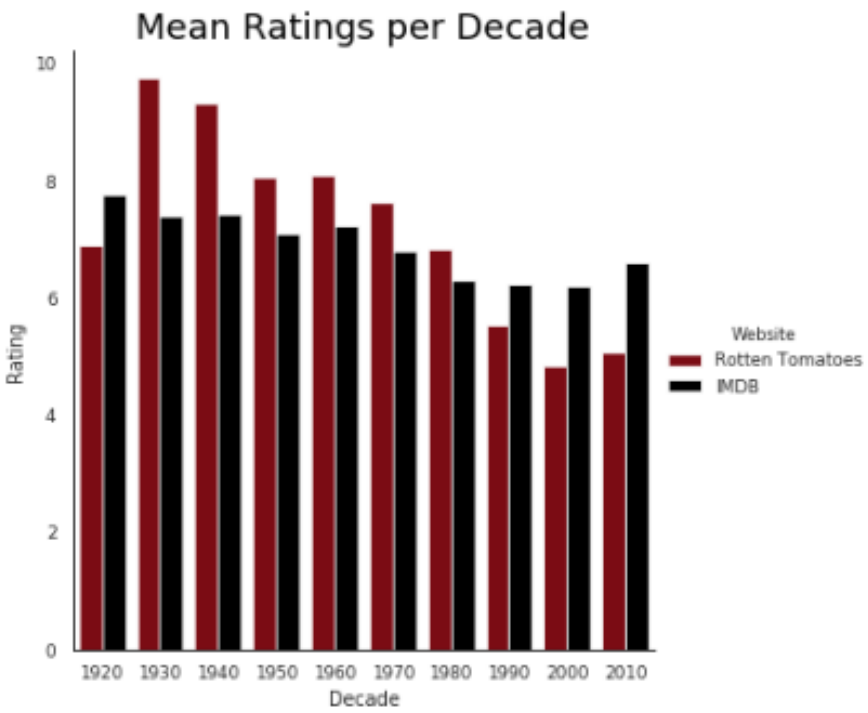
```
H0: There is no significant difference between drama movies mean gross and western movies mean gross.

p-value:  0.23096572640997634

Accept Null Hypothesis (H0)
```

Incredible result that is! There is no signifant difference between them, nontheless sample size is 789 in the former and 36 in the latter.

3. “MOVIES ARE NOT AS GOOD AS THEY USED TO BE”

Another assumption is that quality of movies has dropped significantly laters.
For this analysis we group decades and for every group we calculate the mean rating.

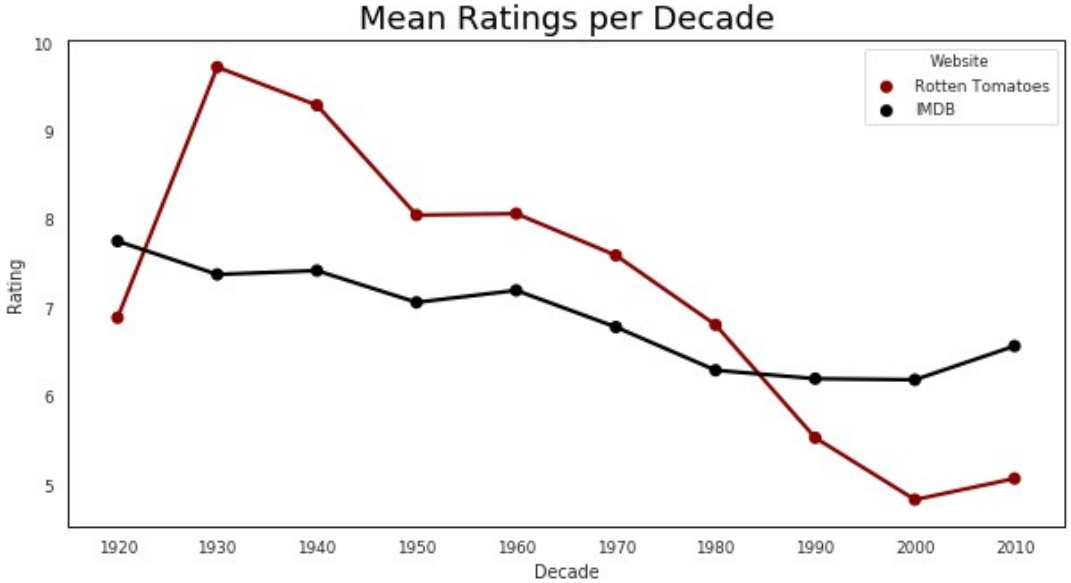
ROTTEN TOMATOES - IMDB MEAN RATINGS PER DECADE BARPLOT:



From this plot, we realise that ratings are higher back in the years that what they are today.

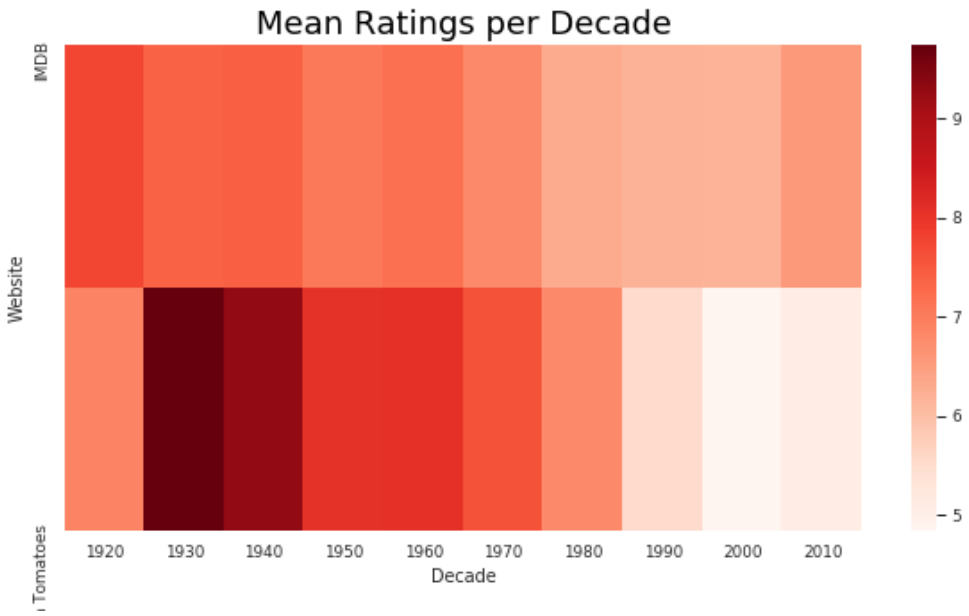
This catplot gives us more info about the votes as it categorizes them with different colors. Another inference is that IMDB Votes are more symmetric than Rotten Tomatoes votes, which indicates that users think movies are not that bad nowadays, whereas, critics in Rotten Tomatoes seem to be dissapointed by recent movies. So yes, ratings are worse than what they used to be.

ROTTEN TOMATOES - IMDB MEAN RATINGS PER DECADE POINTPLOT



Pointplot shows exactly the “line drop” we mentioned above. The red line used to be really high, even higher than the black, and nearby 1985 dropped rappidly under the black one.

ROTTEN TOMATOES - IMDB MEAN RATING PER DECADE HEATMAP:



The color-coding here helps us understand when users voted higher. Rotten Tomatoes critics voted higher from 1930 to 1960 movies and color turns whitey the last two decades. IMDB users stay in orange-beige palette throughout most of the decades, with a more whiter tone from 1980 to 2000.

To sum up, recent movies are indeed low rated, the decline is evident.

NOTE: This analysis is for **all** decades from the dataset, even though 2010 decade has not finished yet. The sample is not normally distributed, there are decades with up to 3 movies only, thus our plots might be a litle biased. If these movies, for example, are high-rated, then the decade gets easily high-rated.

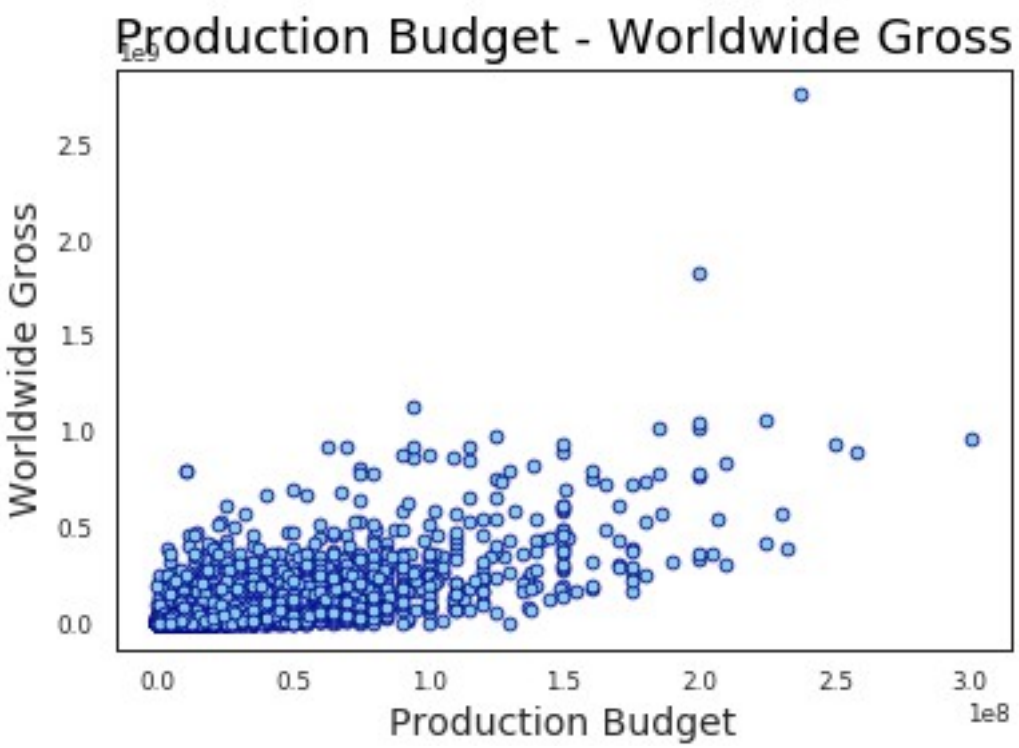
4. ANOTHER DATA MINING ANALYSIS IDEA , "MONEY MONEY"

They say that ***"The more you give, the more you get".***

Is that true for Production Companies, though? Is the invested budget for a movie worthwhile, and is it correlated with high revenues when the movie is out for the big screen?

The following analysis, is going to help us understand how things work, and maybe save some money of any new Production Company in the territory, that plans to invest a big budget for a movie, or it might be a message to those who are not in the top, how to do better.

PRODUCTION BUDGET – WORLDWIDE GROSS SCATTERPLOT :



Scatterplot gives us a first thought that the bigger the budget the better the gross (le8 vs le9).
It follows a weak linear positive correlation, data points are less as budget goes up.

However, most of the data points are gathered in the left down corner.

2-Sample Z-Test (Production Budget and Worldwide Gross):

1st test:

Our code stests.ztest has:

x1: Worldwide Gross
x2: Production Budget
value : 0 so, mean(Worldwide Gross) - mean(Production Budget) – 0
alternatives: 'larger' thus, mean(Worldwide Gross) - mean(Production Budget) > 0

H0: Grosses are larger than budgets invested.

```
H0: Grosses are larger than budgets invested.

p-value:  1.8468088059599744e-88

Reject Null Hypothesis (H0)
```

Oh that is dissapointing! Our assumption was wrong, we now test:

H1: Grosses are smaller than budgets invested.

2st test:

x1: Worldwide Gross
x2: Production Budget
value : 0 so, mean(Worldwide Gross) - mean(Production Budget) – 0
alternatives: 'two-sided' thus, mean(Worldwide Gross) - mean(Production Budget) < 0

H1: Grosses are smaller than budgets invested.

```
H1: Grosses are smaller than budgets invested.

p-value:  1.0

Accept Null Hypothesis (H1)
```

Sad news! This should be considered as a beware notice for any newcomer…

Apparently, budgets often are extreme and movies do not pay off these investments.

Pearson Correlation Coefficient results:

	Gross	Budget
Gross	1.000000	0.665871
Budget	0.665871	1.000000

Spearman Correlation Coefficient results:
Coefficient shows a good strength association based on the formulas.

	Gross	Budget
Gross	1.000000	0.677304
Budget	0.677304	1.000000

Eventually … "The more you give … "