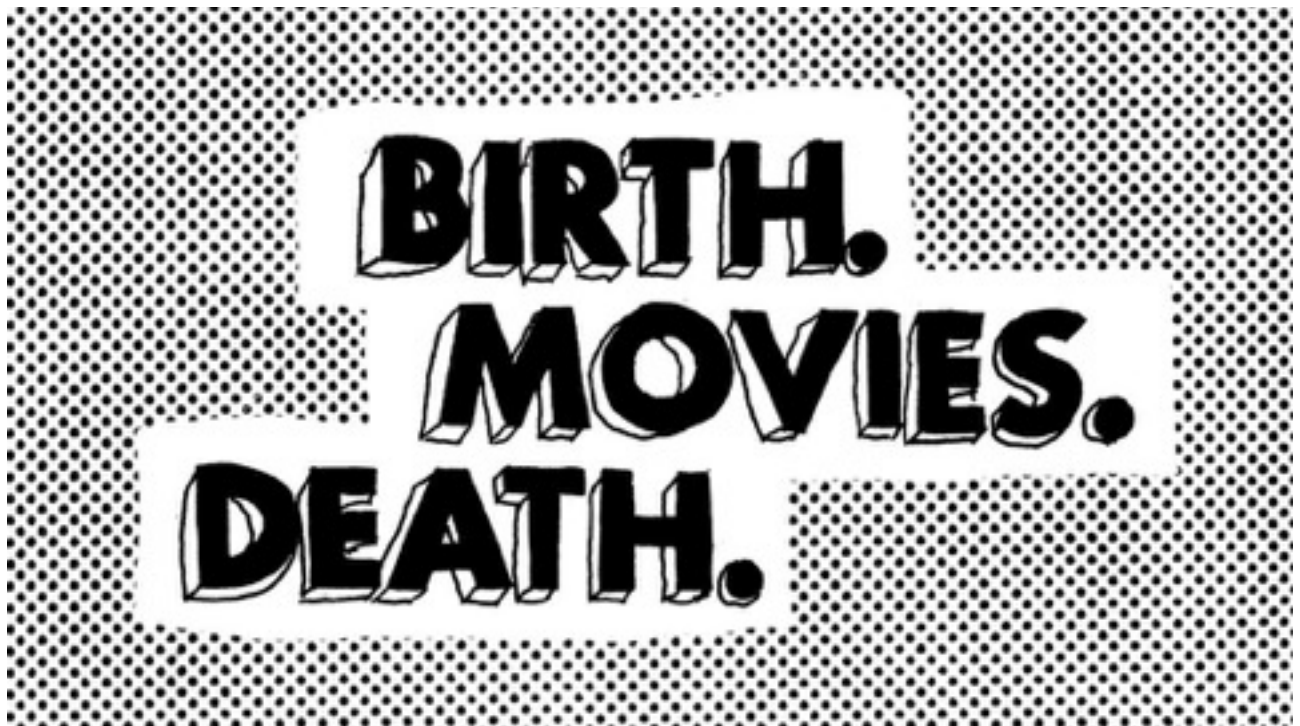Data Analytics

Project report

IMDB 5000 Movie Dataset

Daria Sankova
MoTIS 8

4/05/2017

For my project I chose «IMDB 5000 Movie Dataset» that includes 5000+ movie data scraped from IMDB website.

The original dataset comprised a lot of different data but we didn't need all of it. So, preparing data for a work we left only some columns in dataset in order to answer on questions: Does gross (that, actually, box office of movies) depend on IMDb score or Facebook likes?

Also, we removed directors who has less that 5 movies even if this films brought to the author a lot of money because it made data not so objective.


We started to explore data. Firstly, we wanted to know does director name affect on gross or not? And who is the most paid director. So, we did a chart where you can see that movies of Christopher Nolan, James Cameron, Peter Jackson and George Lucas are the most profitable (what a «surprise») for the cinemas. As well all of them have a quite high score on IMDb that also could be a connection.
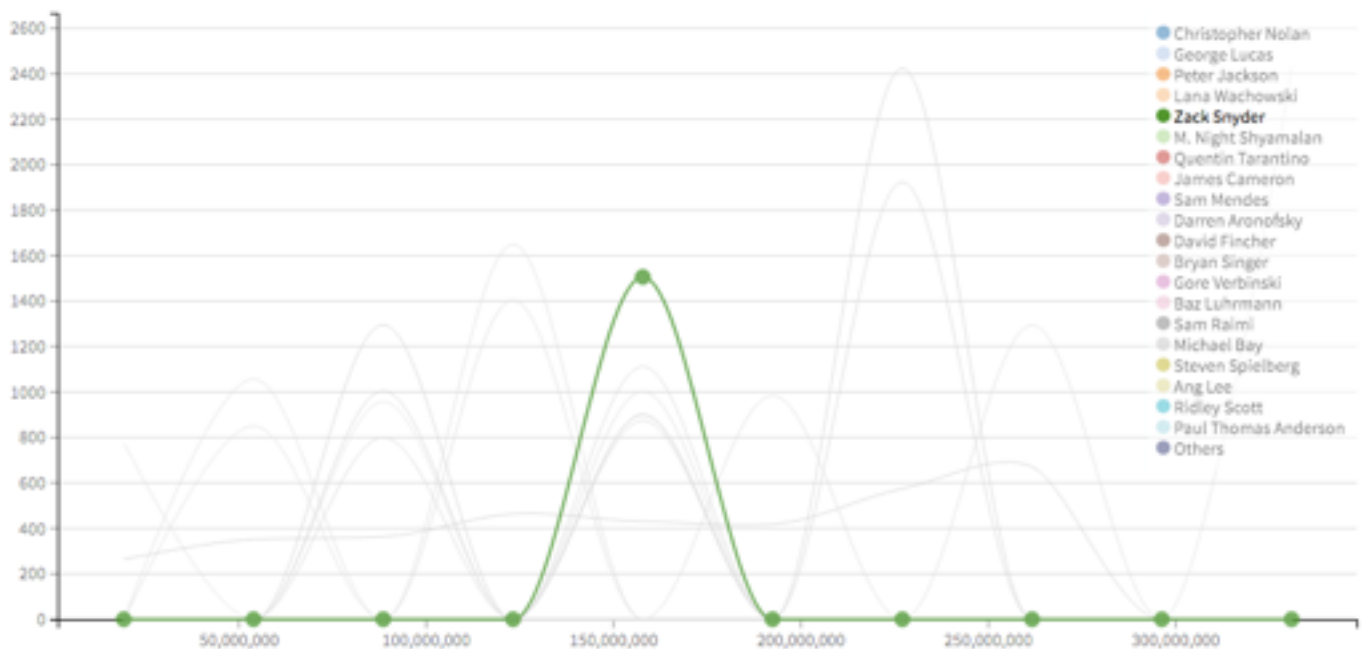
The last conclusion proved a prediction that we did with a data. According to a result characteristic that affects most on a gross of the movies is average number of users for reviews.
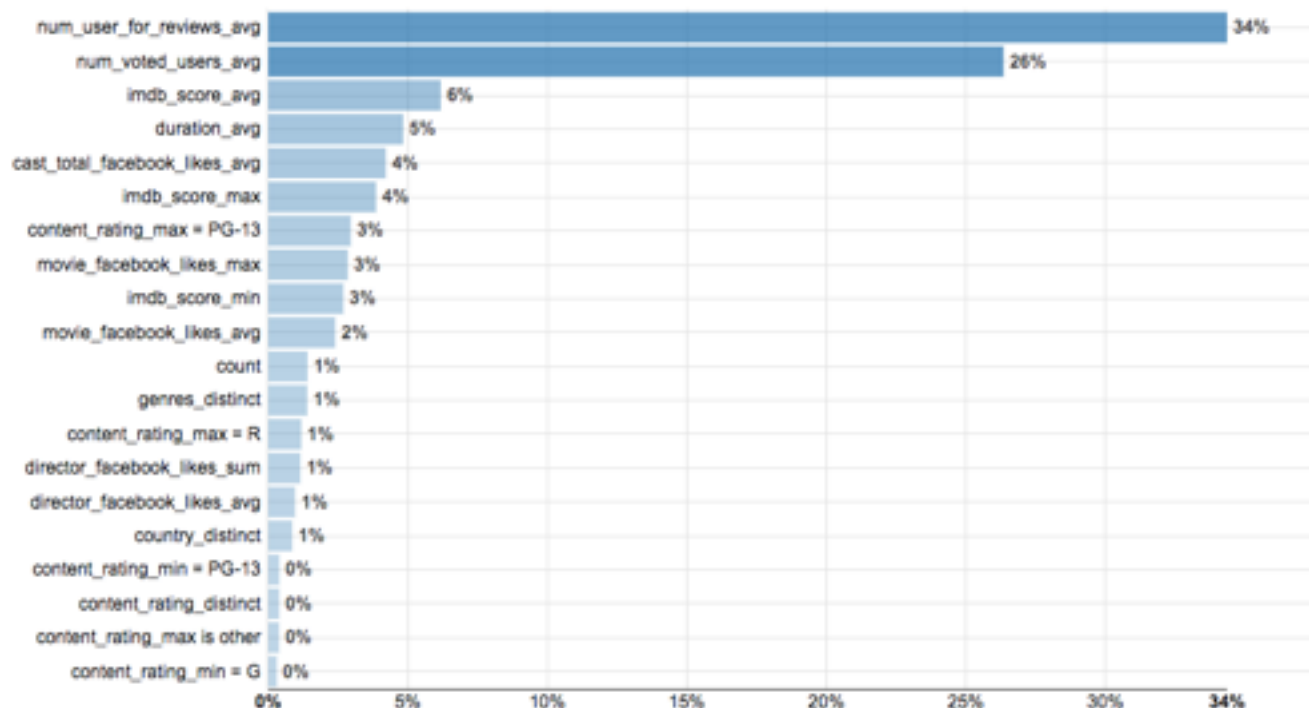
So, after the prediction we made a chart to see how is it look.
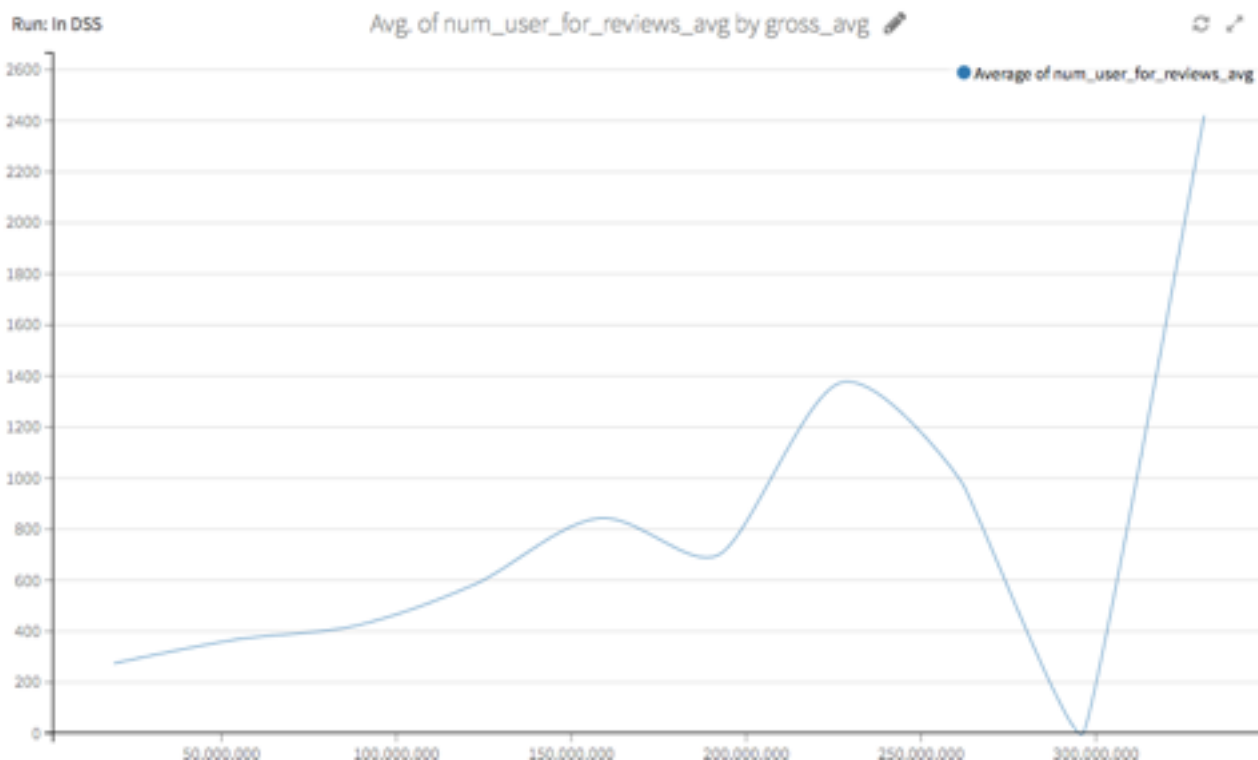
Avg. of num_user_for_reviews_avg by gross_avg on movie_metadata_prepared_by_director_name_prepared


As we could see big number of users reviews not always means a good box office.

And one more thing that was intersection for us – does gross depend on Facebook likes? We couldn't ignore this data in the age of social media. So, the result you can see on the charts.

Avg. of movie_facebook_likes_avg by gross_avg on movie_metadata_prepared_by_director_name_prepared

Avg. of cast_total_facebook_likes_avg by gross_avg ✎
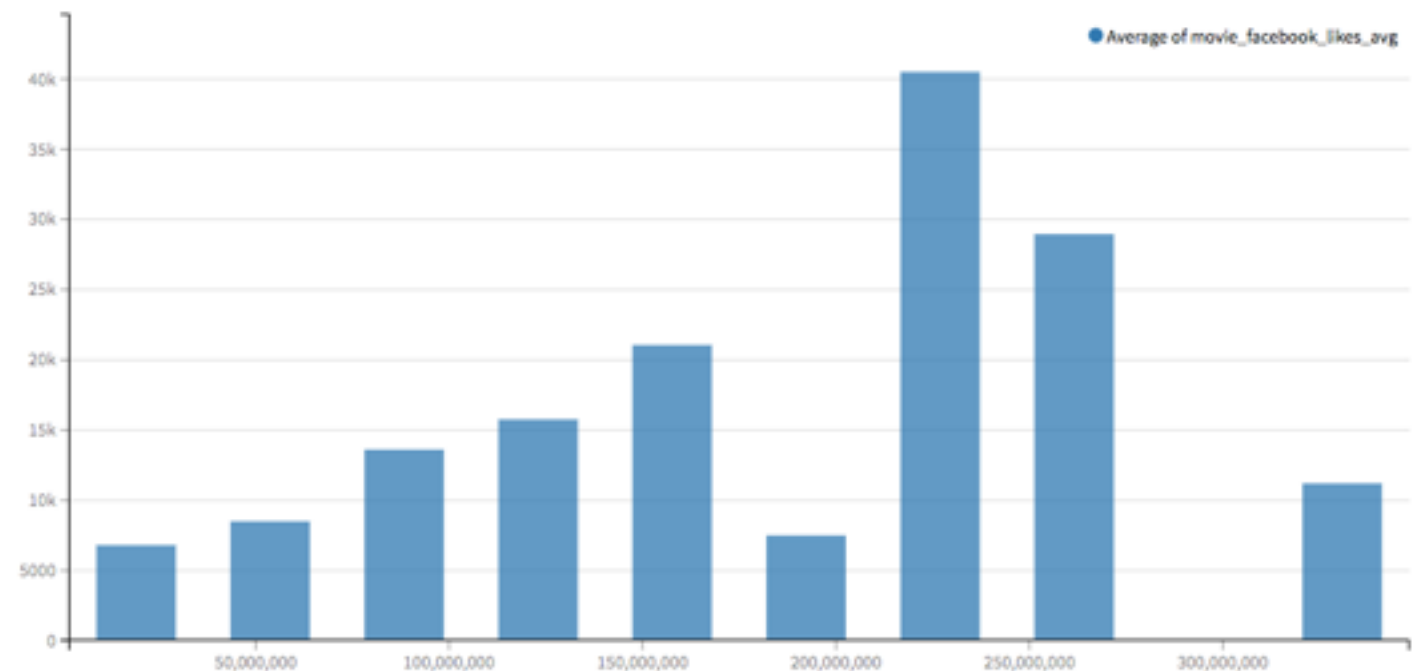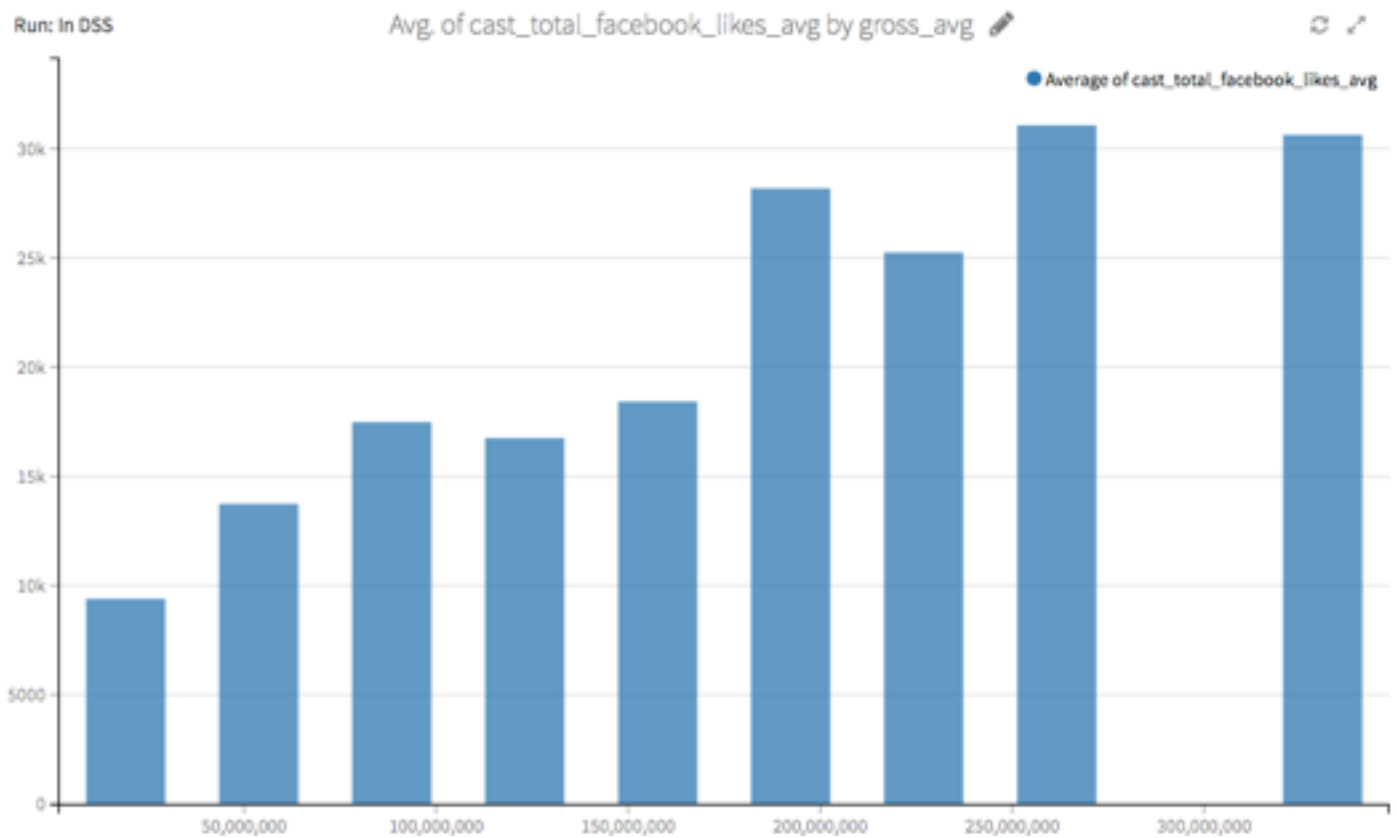
The results are not so surprising just another proof how deep social media got in our life and affect on choices.

Conclusion

To sum up I would like to say that this results and tools of Data Analytics could help cinema theatres to make more profitable choice according to analysis. As well as they can predict which movie gather more people to a cinema. The data that we studied is only a small part from data array.

P.S. :)