IDS 2017
Assignment 1

Bogdan Petre (s3480941)
Low Daniel (s3120155)
Xu Teng Andrea (s3548120)
Group 7

September 10, 2017

1

1.1 Identify Data Types (10P)

- Brightness as measured by a light meter: continuous because each measurement obtains a distinct score[1], quantitative (ratio) because this device could have an absolute zero (i.e., absence of light).
- Brightness as measured by people's judgments: if you use a Lickert scale to measure people's judgment, then brightness would be discrete and qualitative (ordinal) becuase the measurements have a logical order but do not reflect numerical true values.
- Time in terms of AM or PM: binary, qualitative (nominal if one considers there is not a logical order between AM and PM or ordinal if one views PM coming after AM).
- Coat check number (certain places offer you to leave your coat to someone who, in turn, gives you a number tag that you need to claim it back when you leave): discrete, qualitative (ordinal or perhaps nominal if the coats aren't placed in the order of the integers).

1.2 Collect It... Link it! (50P)

In this exercise we created enrich_script.R in order to collect additional data from the API of OMDB. A general description of the code is provided in the *ReadMe.txt* file, while more detailed explanations are given as comments.

1.3 Think About Types (20P)

- Title: discrete, qualitative (nominal) Discrete because there is a finite number of titles and qualitative nominal because it's like an ID
- ReleaseDate: discrete, quantitative (interval)

  Discrete for the same reason, and quantitative (interval) because they are calendar dates.
- Popularity: continuous, quantitative (ratio)

  It's continuos because there infinite numbers of popularity since they are real numbers and quantitative (ratio) because there is absolute zero
- Budget: continuous, quantitative (ratio)

  same reason as above
- Revenue: continuous, quantitative (ratio)

  same reason as above

- Genre: discrete, categorical (nominal)

  discrete because there are finite number of genres, qualitative (nominal) you cannot declare an order

- imdbRating: discrete, numeric (ratio)

  discrete because it has a precision to the first decimal number , it exists the absolute zero

- imdbVotes: continuos, quantitative (ratio)

  it's continuos because there is no finite number of imdbVotes, and ratio because we can have 0 votes.

- Director: discrete, categorical(nominal)

  discrete because there finite number of names, categorical (nominal) because they are names

- Country: discrete, qualitative (categorical)

  same reason as above

- PG rating: discrete, qualitative (ordinal)

  finite number of pg ratings, qualitative (ordinal) because you can order the ratings by the age the child can watch that movie

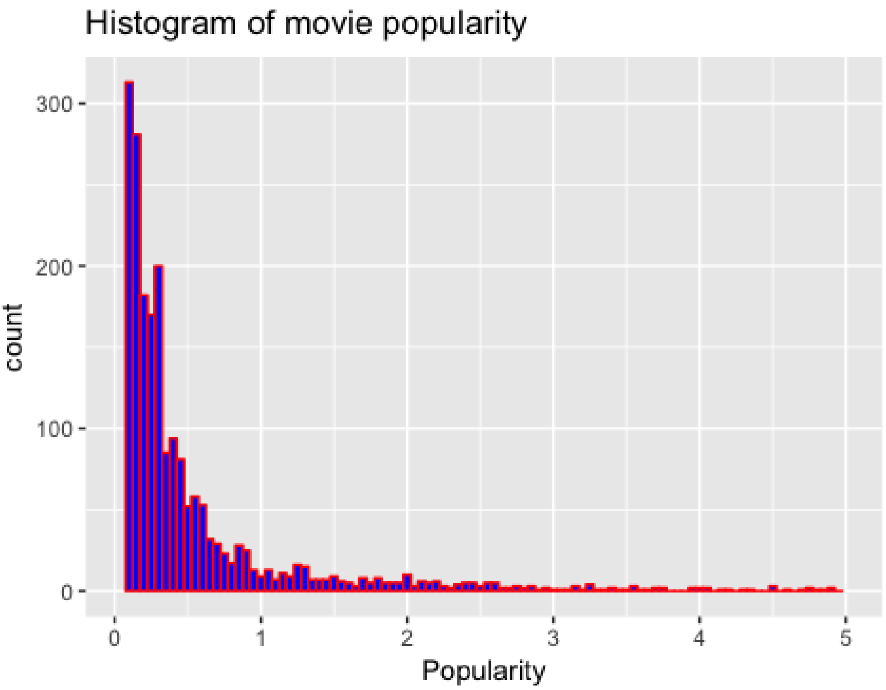
general comments: you should describe your result also in the exploratory analysis
there are problems also in the visualization of the plots
you should develop more your arguments (also in the descriptive questions)

10/20

1.4 And Finally... Analyze it! 20P

1.4.1 a

- **Figure 1** Popularity follows a power law distribution with most movies having a popularity of 0 and 0.5.



the red contours are disturbing, you don't need them

Figure 1: Histogram of popularity

- **Figure 2** Continuous variables like popularity are better visualized with a density plot.
- **Figure 3** Budget follows a power law distribution with most movies having spent less than 25 million dollars.

  you can't see the power law distribution from this plot, anyway it's true that most movies cost less than 25 milion dollars

- **Figure 4** Boxplot of most frequent genres shows that Comedy is the most frequent genre.

  not really interesting information, you just count how many movies for a specific genre you have in the dataset, figure 4 is not a boxplot

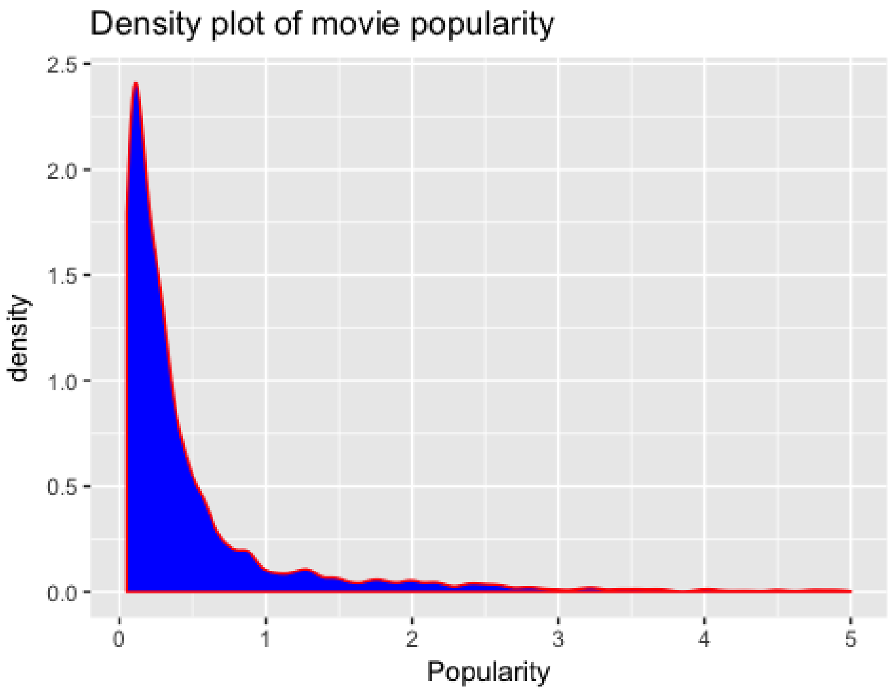


Figure 2: Density plot of popularity.

not needed, give the same information of the histogram in figure 1

1.4.2 b

- **Figure 5**
  Does a higher budget produce a higher rating movie? Interestingly, there seems to be a negative correlation with the movies spending more having a lower rating. But inferential statistics are needed to determine this. Moreover, very few movies had budgets reported, so more data is needed. the data are enough, looks there is no correlation

- **Figure 6**
  Does a higher budget produce a more popular movie? Popularity weighs the rating on the amount of votes, so it can be considered as a better measure of general audience appeal than IMDB rating. There does not seem to be a trend; however, over 10 million dollars in budget, there seems to be a slight positive correlation (i.e., producing higher popularity). But inferential statistics are needed to determine this. Moreover, very few movies had budgets reported, so more data is needed.

- **Figure 7 Are certain genres more popular than others?** Popularity of most frequent genres shows a power law distribution. Short films have a bimodal distribution, with a most of the short films having a low score but also certain short films having a high rating over 5. It is difficult to answer this question with this figure alone, but for instance we can conclude that action movies seem to have higher popularity than others. ok

1.5 Bonus (+10P)

1
We used a scatter plot in order to visualize the data that we have and compare properly the IMDB ratings and Rotten Tomatoes ratings. We can easily see that for some movie the ratings are pretty close for other there is some difference. That's because IMDB uses a weigthed mean [2] with all the users votes, in the other hand Rotten Tomatoes collects all the ratings from critical professionist from all over the world. There is also a Rotten Tomatoes rating from the users see [3].

3
For this exercise we obtained also the Production House Information, then obtained the first Genre for every Movie in Order to delete noise with the other subGenres.

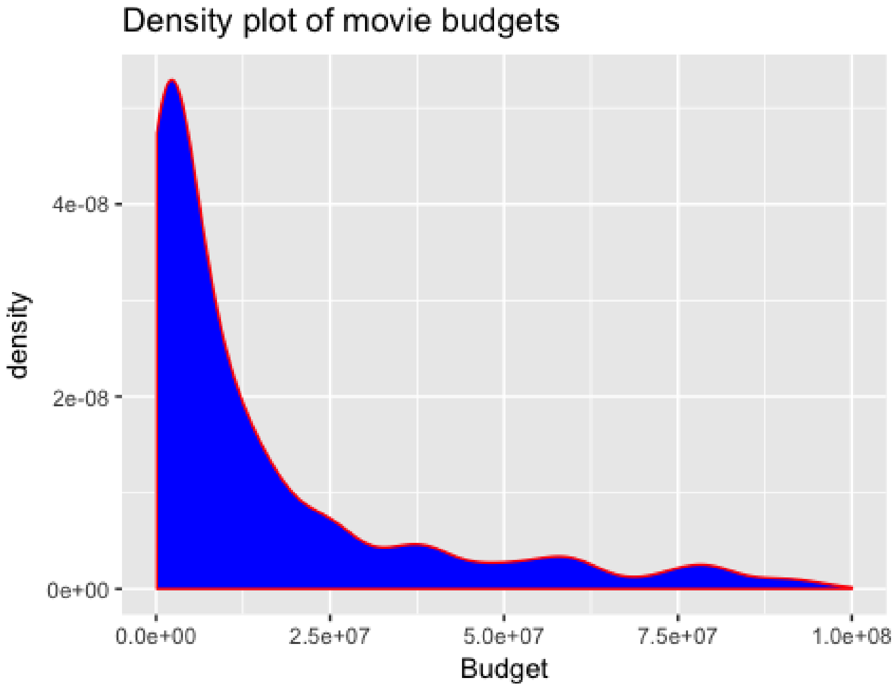


Figure 3: Density plot of budget.

Furthermore we obtained top10 Production Houses and then grouped each Production House with their genres counting per genre.
So with a simple group by function we can easily see in the groupedGenrePerProduction.csv that usually each production house is used to one genre and maybe one or two outliers.

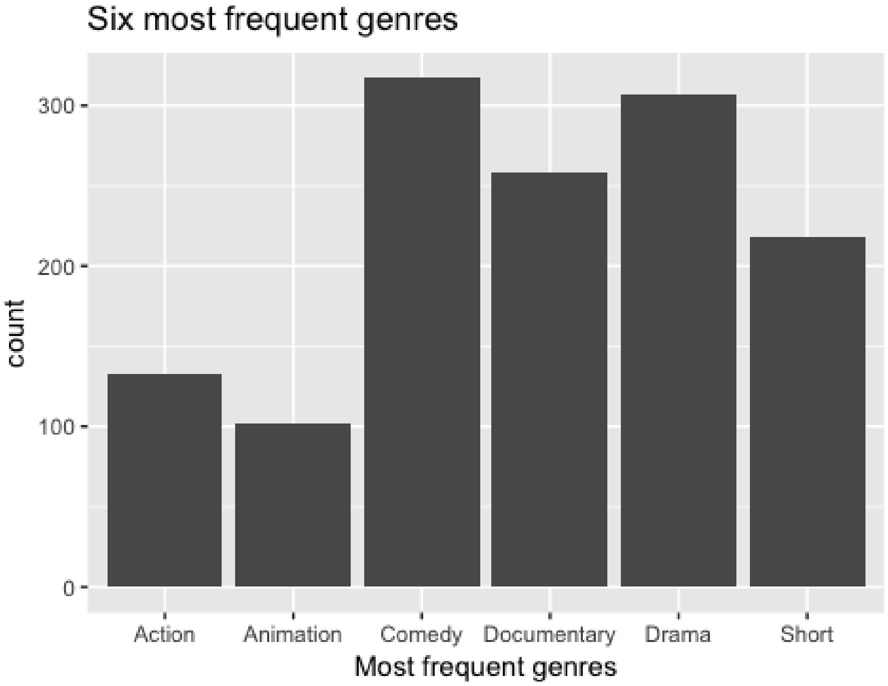


Figure 4: Most frequent genres

References

[1] Field, A. (2009). Discovering statistics using SPSS. Sage publications.

[2] https://math.stackexchange.com/questions/169032/understanding-the-imdb-weighted-rating-function-for-usage-on-my-own-website

[3] https://www.rottentomatoes.com/about/

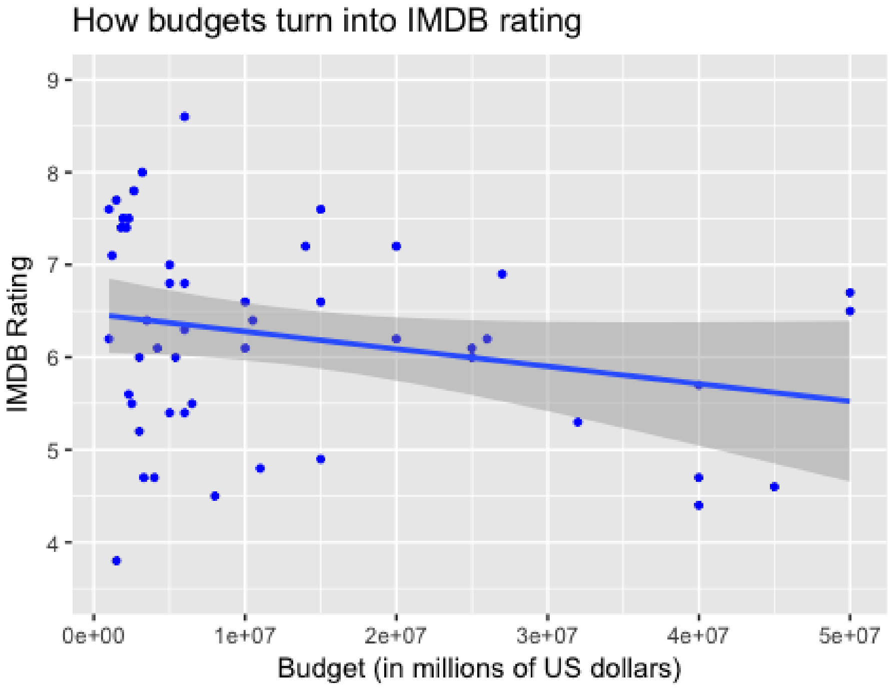


Figure 5: Scatter plot of budget and IMDB rating

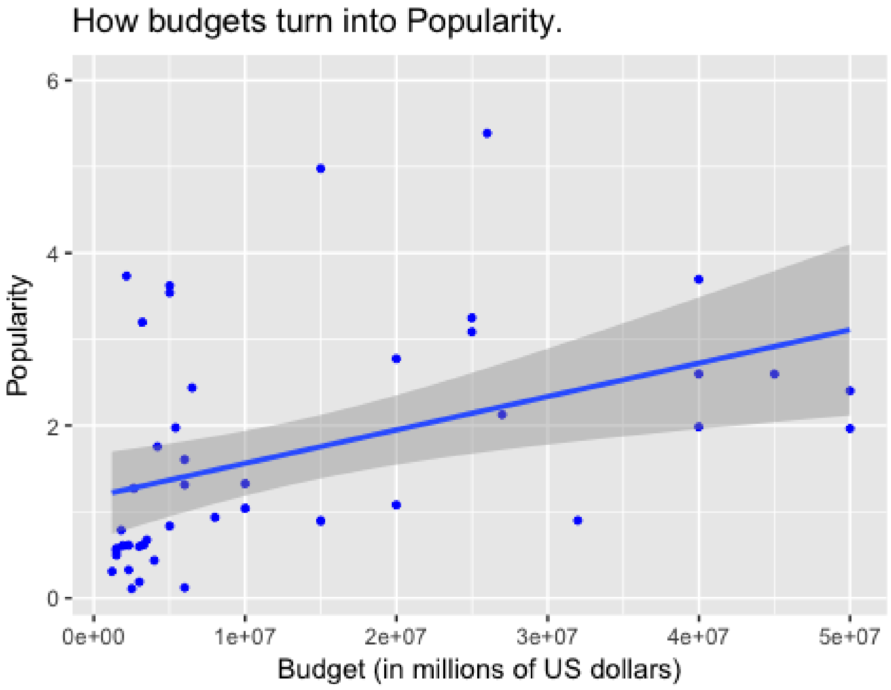


Figure 6: Scatter plot of budget and popularity

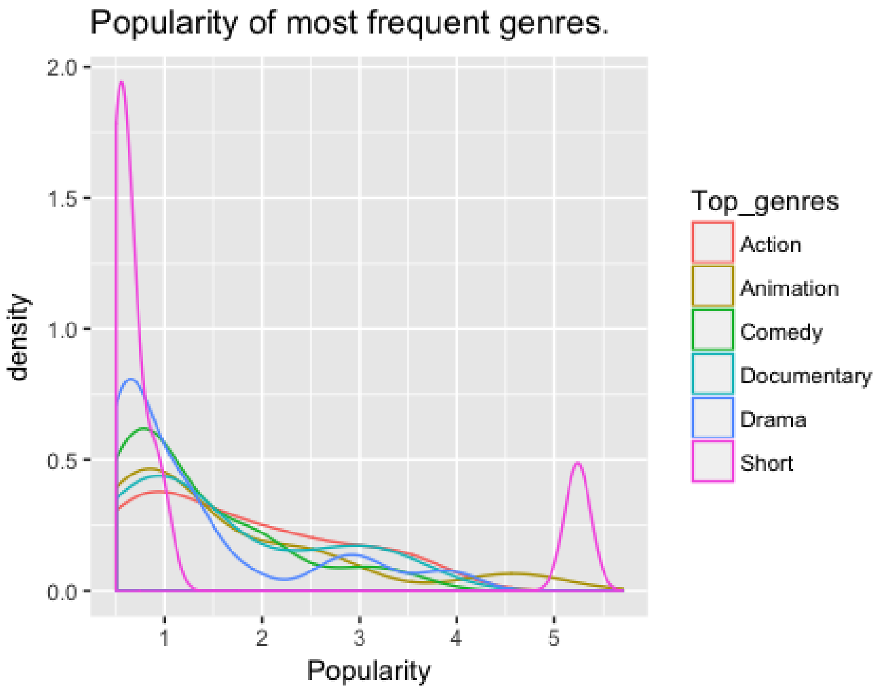


Figure 7: Popularity of most frequent genres

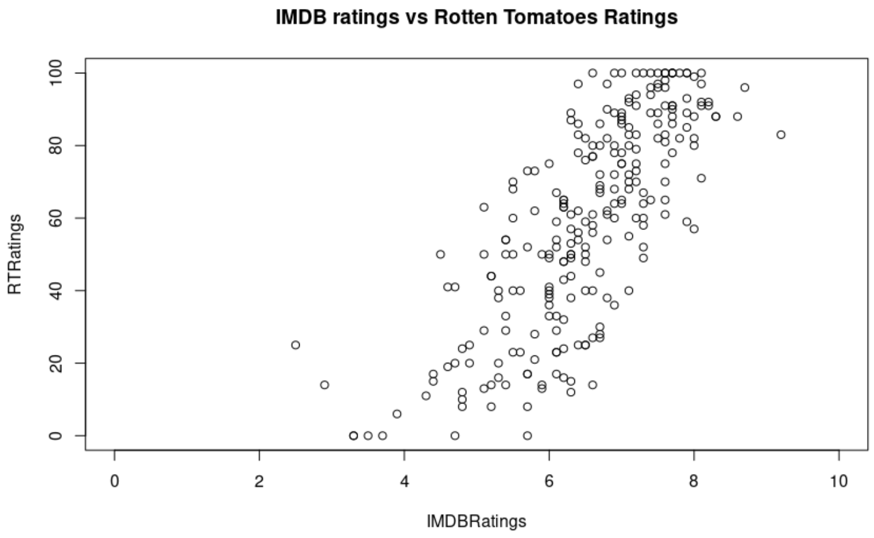


Figure 8: IMDB vs. Rotten Tomatoes ratings