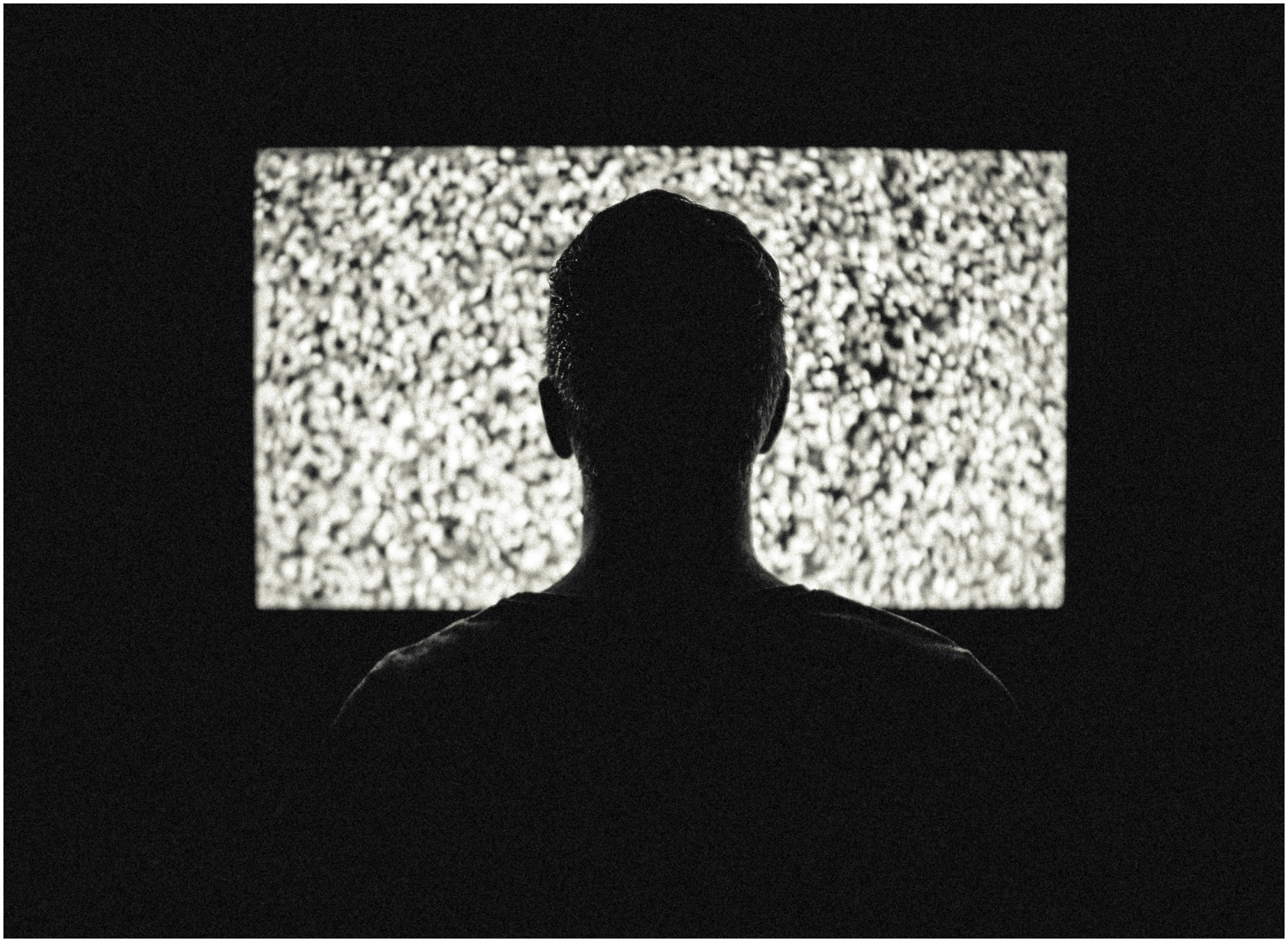
Project MovieRatings

FROM OBJECTS TO DATA

Marijn Koolen



Florian Blom
Fenna Levenbach
Martijn Muller

INTRODUCTION

In the starting phase of this project, in which we were asked to form groups based on presentations we just held, Project MovieRatings consisted of four persons: Florian, Maarten, Fenna and Martijn. Three of us had proposed research questions relating to movie ratings, which is a very useful common ground for a starting point. After short debate, we decided to look at how different things might influence viewers' appreciation of movies. Hence the research question:
Q: "Which external factors influence movie appreciation?"

Early on, we decided to divide our research into three more substantive sub-questions. One of us would focus on the effect of a film's duration on the rating (Q.a), another on the effect of a film's genre (Q.b), a third on the effect of an actor's passing away (Q.c) and the fourth would prepare the dataset and figure out how best to use it.
It came to be, just after we had handed in the first Portfolio version, that Maarten decided to leave our group. Due to this sudden change in manpower and subsequently in load balancing, our project is not as structured as we had hoped it to be.

All graphs, codes, and files we used can be found in the GitHub. We did not document every command we used, but the file in /Files/Q.a/Codes.txt gives an idea of the most important and/or most frequently used ones.

METHOD

To be able to answer research questions Q.a and Q.b, we had to gather data about a substantial amount of movies. This data should include the name, runtime, rating, genre and release year. Additionally, we had to be able to filter out every entry that did not meet these requirements—this would be the case when not all data was available, when the entry was not a movie, etc.
For Q.c we would need reviews from over a period of time, stretching from before the actor's passing away to after. The database we would use, then, should provide reviews which included a date and a rating.

OMDb
The first online movie database that we thought of was the conveniently named *Internet Movie Database*, or *IMDb*. This is by far the most extensive and up-to-date source of movie data, but unfortunately does not have its own API. We did find links to websites which held all kinds of files *IMDb* uses, among which a file called ratings.list. This turned to be a 45 Megabyte plain text file containing the titles of all movies, documentaries, series, videogames and other entries that are on *IMDb* and their number of votes, vote distribution and rating. Very useful indeed, but still only half the data we needed.
We soon discovered that there are multiple websites which do offer access to the rest of the *IMDb* data, via their own complicated APIs. The website we chose to use is the *Open Movie Database*, or *OMDb*. It is quite easy to use and offers enough options for us to effectively request the desired data. The *OMDb* API could provide us with the runtime, rating (ranging from 1 to 10), genre and release year of all entries on the ratings.list file. It did not, however, include reviews.
The exact link to the *OMDb* API is http://omdbapi.com.

Amazon
Martijn, responsible for Q.c, decided to use the *Amazon* database. This international online market place contains most movies in the world and, better yet, reviews with a date and a rating (ranging from 1 to 5). There is one disadvantage of these reviews, however; the reviews are written by people who frequently judge the bought product and not the movie itself. For instance, if someone had bought a DVD from *Amazon*, which was delivered a week late, they might write a negative review despite their enjoying the movie once it was delivered. It was decided to attach little weight to this problem, as the total number of reviews should still give us an acceptable estimate of a movie's appreciation.
The exact link to the *Amazon* data is https://snap.stanford.edu/data/web-Amazon.html.

The differences between the data which both databases offer, resulted in Florian and Fenna on the one hand and Martijn on the other working more independently from each other. The following section is therefore divided into subsections 'Q.a and Q.b', for which the *OMDb* API was used, and 'Q.c', for which the *Amazon* database was used.

PREPARATIONS AND ANALYSES

Q.a and Q.b

We struggled immensely to prepare the data for these research questions. The first, most tiresome task was to clean the ratings.list file as to only keep the names of movie entries and nothing else.

We used commands to delete all but the title column, to remove entries that were either pornography, part of a series or contained non-alphanumerical symbols, to create a sample list from the 300000+ total entries, to convert spaces between words into the symbol '+' as to make it useable for *OMDb*, and to loop it through the API. All this took more than a week to figure out, but finally we had a file in JSON format containing all *IMDb* data on over 2000 different movies or documentaries—at this point we decided that standalone documentaries were as good a movie as any other audiovisual production.

We wanted to use jq on the JSON file to filter out only the necessary information, but it did not work. It took us another half week and email correspondence with Mr Koolen to discover what the problem was. After using a command to insert a comma after every bracket '}' except the last, the file could finally be submitted to the jq commands. The few series entries still on our sample list could now be filtered out as well.

Next, we saved our data to a CSV file, which we thought would be most convenient to use. In the 'Runtime' column we removed the word "min", manually converted entries in the format of 1h33m into a single number (in this case: '93'), and deleted all entries which were short films (runtime < 40min) or longer than 180min. We used this, finally useable, data frame in *Rstudio* to analyse the data and plot graphs.

We used ggplot2 to create more sophisticated, better-looking graphs. The first graph we produced was a plot of 'imdbRating' against 'Runtime', coloured by genre (see **Figure 1**). A larger version can be found in the GitHub repository.

Figure 1 - 'Runtime','imdbRating'

As is clearly visible, there are too many different genre labels. This is due to the structure of the OMDb data; when a movie belongs in multiple genres, its genre value contains all those genres. Hence, a comedy about sport has a completely different genre value than a comedy about Sci-Fi. Because there was too little time left, we decided that restructuring the entire 'Genre' column wasn't feasible and to drop Q.b altogether.

With only Q.a to focus on, we compared runtime against both rating and release year. We used the cor.test command in Rstudio to test for correlation, and in both cases p was smaller than 0.05, which means that there is a statistically significant correlation between runtime and rating, and between runtime and release year (see **Figure 2** and **Figure 3**).

```
> cor.test(lijst$Runtime, lijst$imdbRating)

        Pearson's product-moment correlation

data:  lijst$Runtime and lijst$imdbRating
t = 7.5447, df = 3374, p-value = 5.795e-14
alternative hypothesis: true correlation is not equal to
95 percent confidence interval:
 0.09548637 0.16183737
sample estimates:
      cor
0.128806
```

Figure 2 - correlation test 'Runtime','imdbRating'

```
> cor.test(lijst$Year, lijst$Runtime)

        Pearson's product-moment correlation

data:  lijst$Year and lijst$Runtime
t = 7.0411, df = 3374, p-value = 2.301e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08695558 0.15344878
sample estimates:
      cor
0.1203371
```

Figure 3 - correlation test 'Year','Runtime'

However small the effect is, these numbers do predict that longer movies will generally be more appreciated than shorter ones, and that longer movies will be made as years go by. For a more readable graph, we used ggplot to add a line that shows the average increase, as can be seen in **Figure 4** and **Figure 5** (larger versions can be found in the GitHub).

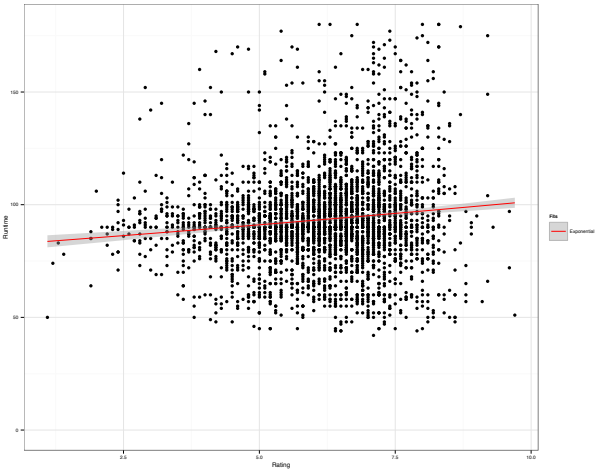


Figure 4 - 'Rating','Runtime' with correlation

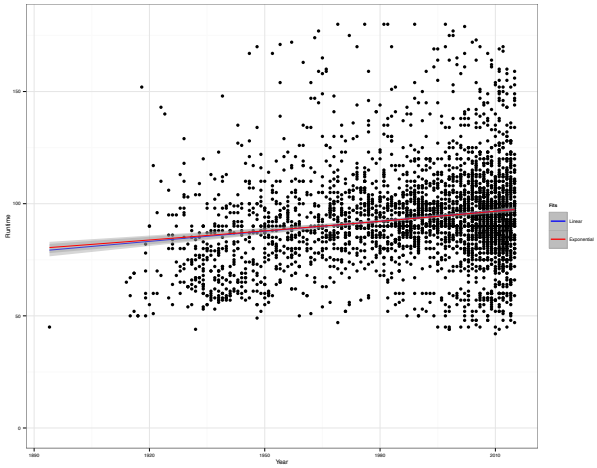


Figure 5 - 'Year','Runtime' with correlation

Q.c

The first thing to do, was to find out which deceased actors and actresses would likely have left an imprint on the viewers' minds. Luckily, the internet was keen to answer this question and by the aid of *IMDb* we soon had a list of useable people.
Not certain how to get the necessary reviews exactly we asked Mr Koolen for help, and he sent us all we needed. Using the program *Notepad++*, we created files containing all movies a certain actor/actress had starred in. From each files we selected two or more different movies, made not too far before the actor's passing away. We used *Rstudio* to analyse the data and plot graphs of movies with the actors Paul Walker († 30-11-2013) and Heath Ledger († 22-1-2008).
Figure 6 is an example of Paul Walker, who starred in *Fast and Furious*, *2 Fast 2 Furious*, and *Fast and Furious Tokyo Drift*. The upper graph shows the ratings per month from 2001 onwards; the lower graph is the amount of reviews per month from 2001 onwards. As you can see, there is not much to say about the ratings in the upper graph, as they don't really show a lot of consistency. Ratings themselves, then, do not seem to be affected by Walker's passing away. There is something to say about the popularity, however: we see an immense rise at the beginning and one at the end of

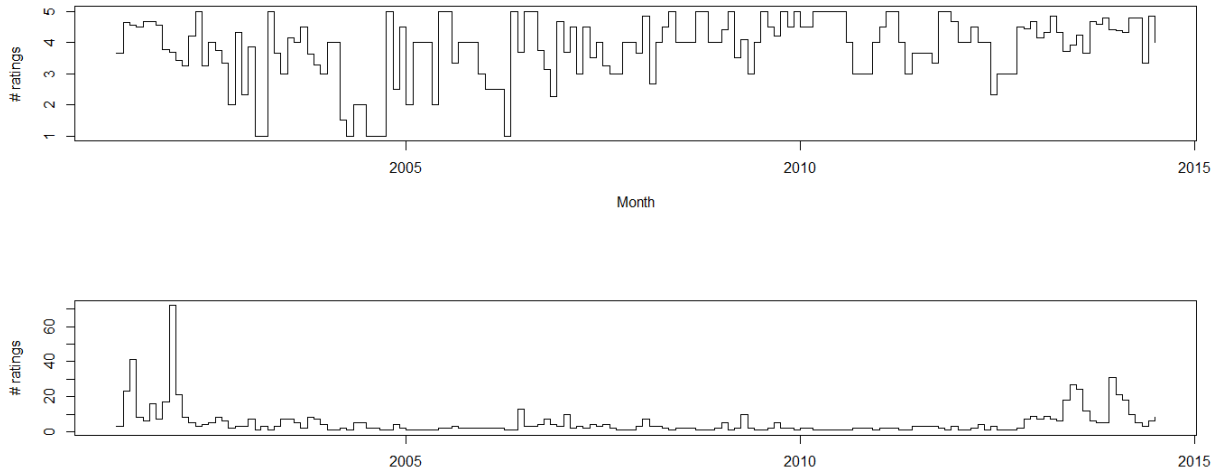


Figure 6 - Ratings of the movies *Fast and Furious*, *2 Fast 2 Furious*, and *Fast and Furious Tokyo Drift*, all movies with Paul Walker

the graph. **Figure 7** is zoomed in on the rise at the end, showing a surge in amount of ratings somewhere in summer 2013 and at the very end of that same year.

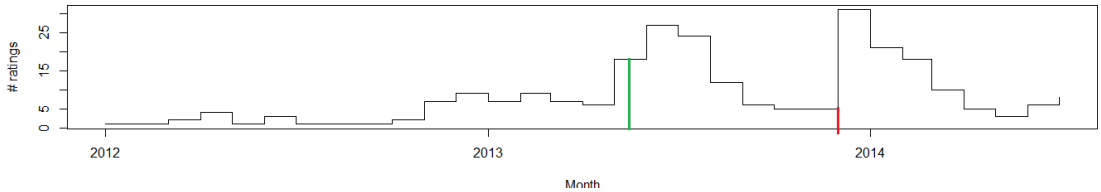


Figure 7 - Ratings of the movies *Fast and Furious*, *2 Fast 2 Furious*, and *Fast and Furious Tokyo Drift*, zoomed in

There are possible explanations for these rises: the green line indicates the release of the movie *Fast & Furious 6*, also starring Paul Walker, and the red line indicates his death. The ratings themselves do not change in any direction at either point in time, just the amount of reviews.

The example of Heath Ledger, in **Figure 8**, shows more or less the same. The ratings go up and down quite randomly, but the amount of reviews shows a significant abnormality. At the beginning of the bump in the line, Brokeback Mountain was released on DVD, which explains the rise in amount of reviews. There is no significant abnormality around the time of Ledger's death however.

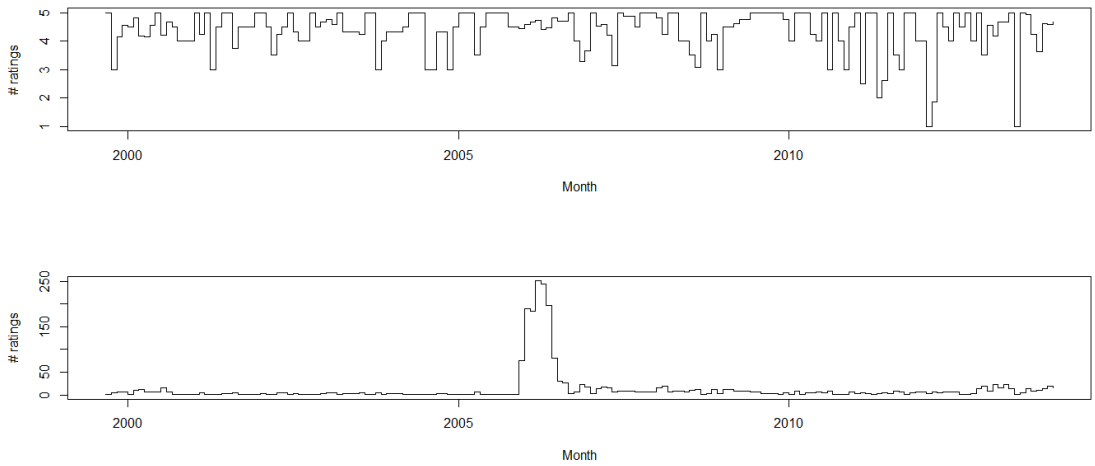


Figure 8 – Ratings of movies starring Heath Ledger

DISCUSSION AND EVALUATION

To answer Q.a, we would have to say that the Runtime of a movie does influence movie appreciation in a statistically significant way. We have discovered that the longer a movie is, the higher its rating on *IMDb*. Additionally, in the past years increasingly longer movies have been made. Taking both findings into account, we could predict that films in the future will generally be rated increasingly higher, since the runtime increases and there is a positive correlation between runtime and rating. This seems to be an odd statement to make and we would not put any money on this prediction.
To answer Q.b, we would have to say that an actor's death does not influence movie appreciation. Although the amount of reviews written does sometimes increase significantly upon the passing away of an actor, this is not always the case. We did find an second a reason for the amount of reviews to surge dramatically, which is the release of a film. No surprises there.
Taking everything into account, we could only answer our research question Q: "Which external factors influence movie appreciation?" with: runtime does, the death of an actor does not, and the influence of a movie's genre has still to be researched.

Apart from learning to work with a Command Line Interface, *Rstudio* and *GitHub* and how to gather data from APIs, we have learned that the single most important part of doing research is preparing the data. We had several issues with our datasets: for the *OMDb* part the dataset was too large but for the *Amazon* part the dataset was too small; the reviews we used were about the physical product instead of the movie itself; the *OMDb* movie list contained too many non-movie productions. In future projects we are likely to choose a smaller scope of the project, a ready-to-use and extensive dataset, and more concrete research questions.
We did, however, enjoy doing this project and would like to thank Mr Koolen for his help, patience and understanding.