Project of WebScraping

Robert Carulla Solsona

13 de mayo de 2018

WebScraping from Metacritic

Facultat de Matemàtiques i Estadística
Màster en Estadística i Investigació Operativa

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Contents

1 Goal

The aim of this project is to scrab data from the www.metacritic.com and to carry out a descriptive analysis of this data. Metacritic is a web where there are ratings of movies and videogames. In this case we are going to analyse the game rates.

Metacritic is a reference web to check the ratings because it has the mean for each videogame of a lot of recognized magazines and webs from around the world that rate that videogame. Apart from this, the web also contains the users mean, which is the mean of the grades that the users give to a determined game.

The purpose is to analyse if there is difference between the rates of the magazines and the rates of the users. If there is evident difference between users and magazines could be for different reasons such as the magazines tend to rate better the games that come from a strong company, because of the company pays money to the magazines; or because the users sometimes are "trolls" that rate very low a game simply because they don't like it.

Another interesting thing to analyse is the way that the rates have evolved through the time, then we can see which was the best period of time for videogames. We can also see which is the platform with more games in the top 100, and then we will see which could have been the best console of all times.

2 The data source and approach used for downloading

As I explained before, the data source is www.metacritic.com and the web technology is HTML. In order to scrap the data I used the gadget selector and the source code of the web. I did a tutorial of CSS selectors to learn how to get an specific element from an specific class, descendant selector, descendant selector with ID, etc. Here the link of this tutorial http://flukeout.github.io/. I scrapped the data of the top 100 videogames of all times (whatever the platform), and the top 100 videogames for the following platforms, PS4, Xbox One, Nintendo Switch, PC, WiiU, Nintendo 3DS, PS vita and ios. For all that I scrapped the data of the top 100 videogames regarding the metascore criteria and regarding the userscore criteria, obtaining then 18 datasets. In this project we are going to explore only the data related to "all" games of all consoles of all times, and we are not going to analyse the datasets of the specific consoles, since it would be too much long. Once I had scrapped the data, I used functions from the package `stringr()` and regular expressions in order to clean and to prepare the data to analyse.

3 The results of web scraping

Here I am going to show the "head" of the dataset of the top 100 games of all times and all consoles by metascore, once the data is well prepared. The other datasets look similar.

```
##                                   game console metacritic.ratings
## 1 The Legend of Zelda: Ocarina of Time     N64                 99
## 2             Tony Hawk's Pro Skater 2      PS                 98
## 3                  Grand Theft Auto IV     PS3                 98
## 4                           SoulCalibur      DC                 98
## 5                  Grand Theft Auto IV    X360                 98
## 6                   Super Mario Galaxy     WII                 97
##         date
## 1 1998-11-23
## 2 2000-09-20
## 3 2008-04-29
## 4 1999-09-08
## 5 2008-04-29
## 6 2007-11-12
```

4 Analysis

Here I am going to carry out a descriptive analysis of the data set of the top 100 games of all times. I use the data set by metascore and the data set by userscore, in order to compare. I don't carry out the analysis of each data set (ps4, xboxone, switch...), the analysis would be similar.

First of all we can start comparing the mean of the top 100 grades by metascore and userscore. We have to take into acount that the top 100 games for users might be different from the top 100 games for the metacritic.

The mean of top 100 metascore:

```
## [1] 95.29
```

The mean of the top 100 userscore:

```
## [1] 91.04
```

We can see that the mean of the grades of the users is lower than the mean of the metascore, it seems that the magazines inflate the rates, may be to keep the videogame business alive and to keep the people buying videogames.

Next we can see the histograms of the grades, for both cases.

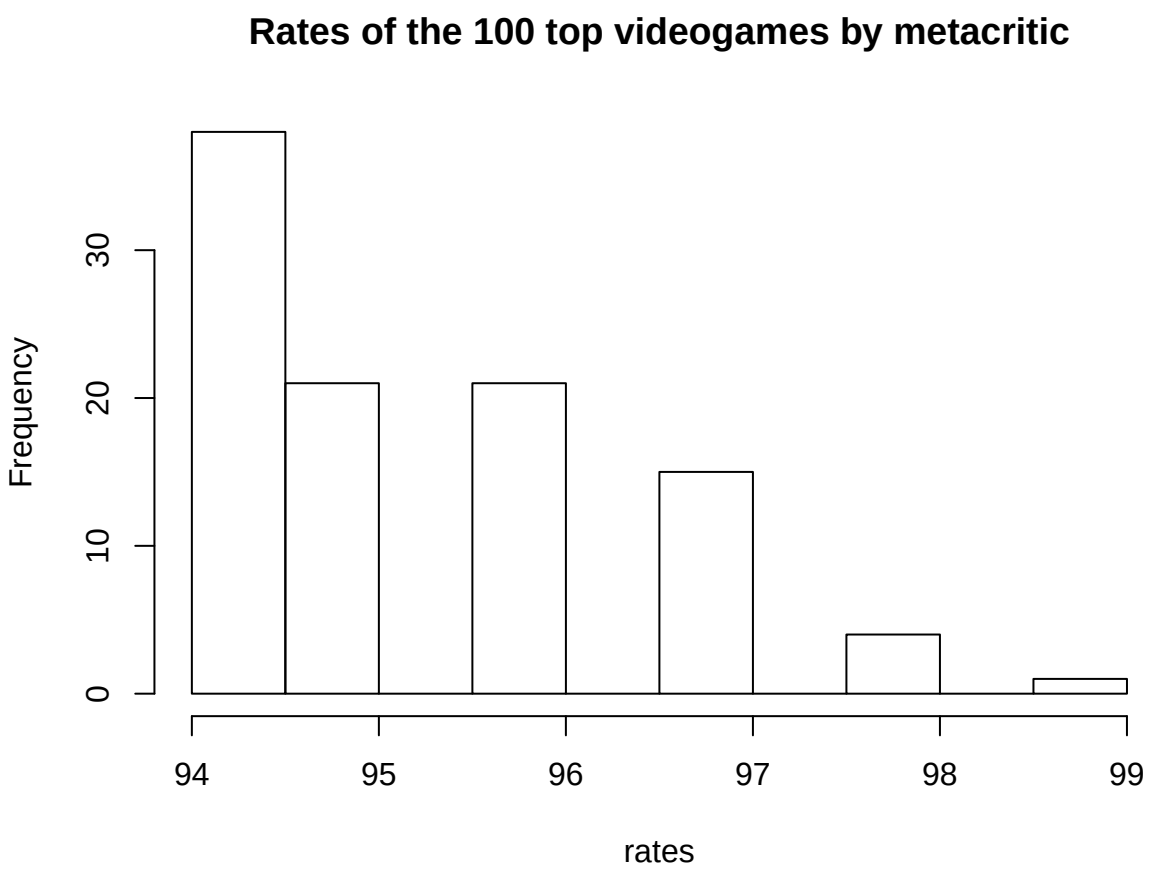

Figure 1: Histogram of the rates for the top 100 videogames by metascore

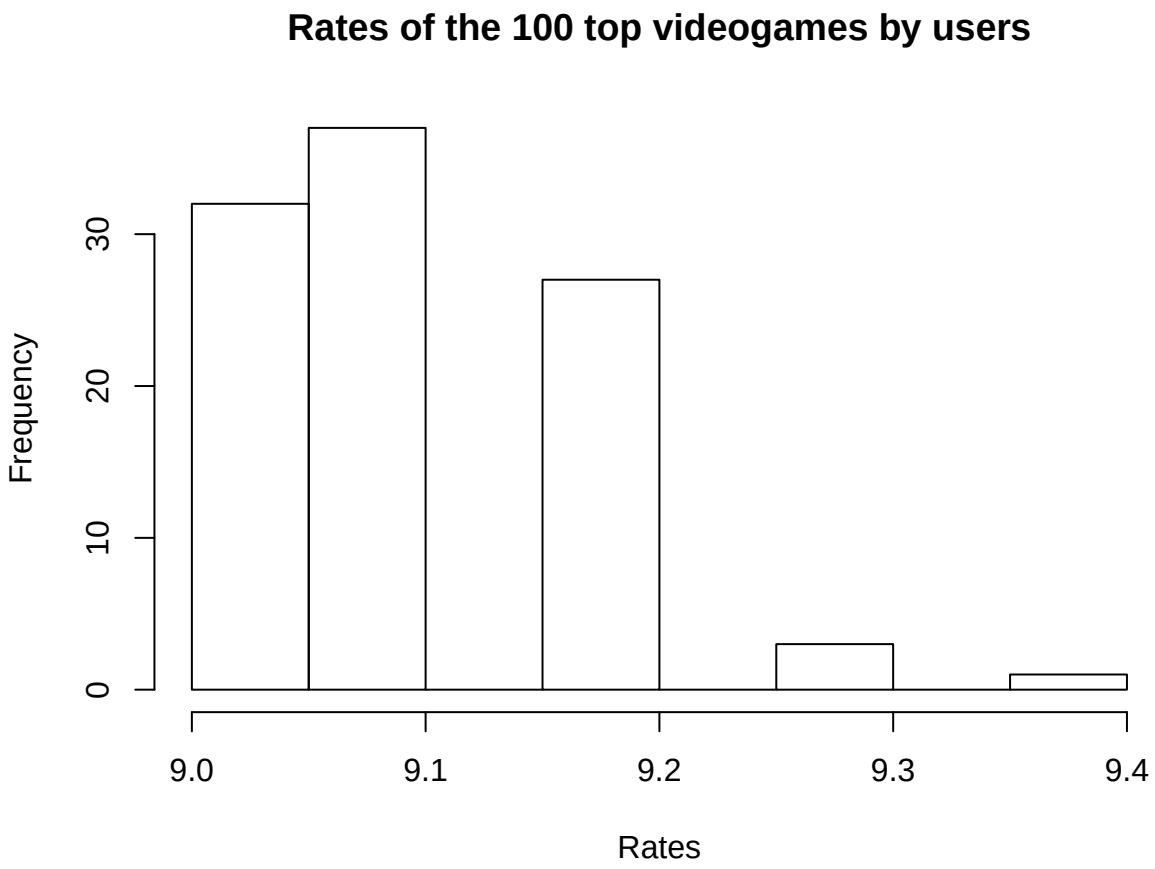

Figure 2: Histogram of the rates for the top 100 videogames by userscore