An Illustrative Data Analysis

Carlos Paniagua

9/8/2021

Types of Movies Your Favorite Actor Makes

Inspired by the Hollywood Taxonomy by Walt Hickey at fivethirtyeight.com (https://fivethirtyeight.com/tag/hollywood-taxonomy/).

Idris Elba

- We will do a similar analysis for Idris Elba (https://www.rottentomatoes.com/celebrity/idris_elba)
- We will use a clustering algorith to classify the different types of movies this actor makes
- Goal: Write an application that perform similarly for any other actor

Step 1: Get the data!

- We will use movie ratings data from Rotten Tomatoes (https://www.rottentomatoes.com/celebrity/idris_elba).
- Getting data from a webpage is called ***scrapping***.

Let’s get the ratings first!

```
##   TomatometerÂ.                                 Title Year
## 1          91%                     The Suicide Squad 2021
## 2          19%                                  Cats 2019
## 3          67% Fast & Furious Presents: Hobbs & Shaw 2019
## 4          85%                Avengers: Infinity War 2018
## 5          54%                                Yardie 2018
## 6          16%                        The Dark Tower 2017
```

The original dataset includes US domestic gross information but we will get this from another source.

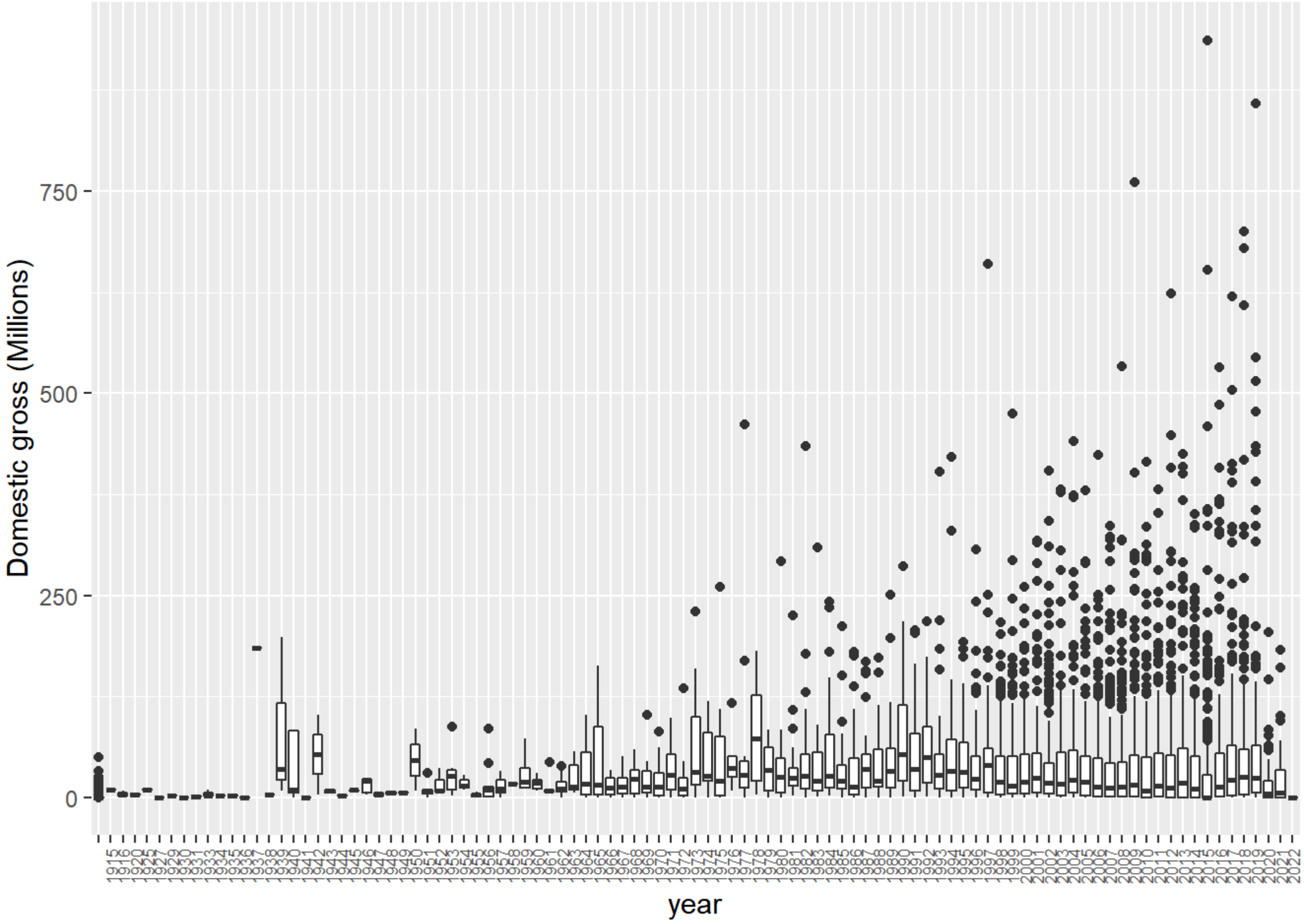
Let’s get the movie budgets and revenue!

We will scrape data from the-numbers.com (https://www.the-numbers.com/movie/budgets/all/101).

```
##    ReleaseDate                                        Movie ProductionBudget
## 1 Apr 23, 2019                            Avengers: Endgame     $400,000,000
## 2 May 20, 2011 Pirates of the Caribbean: On Stranger Tides     $379,000,000
## 3 Apr 22, 2015                      Avengers: Age of Ultron     $365,000,000
## 4 Dec 16, 2015         Star Wars Ep. VII: The Force Awakens     $306,000,000
## 5 Apr 25, 2018                       Avengers: Infinity War     $300,000,000
## 6 May 24, 2007  Pirates of the Caribbean: At Worldâ\200\231s End     $300,000,000
##   DomesticGross WorldwideGross
## 1  $858,373,000 $2,797,800,564
## 2  $241,071,802 $1,045,713,802
## 3  $459,005,868 $1,395,316,979
## 4  $936,662,225 $2,064,615,817
## 5  $678,815,482 $2,044,540,523
## 6  $309,420,425   $960,996,492
```

Question: Are the `DomesticGross` and `WorldwideGross` columns adjusted for inflation?

```
## Warning in gsub("[\\$,]", "", x) %>% as.integer(): NAs introduced by coercion to
## integer range
```



Probably not. Can you see why?

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

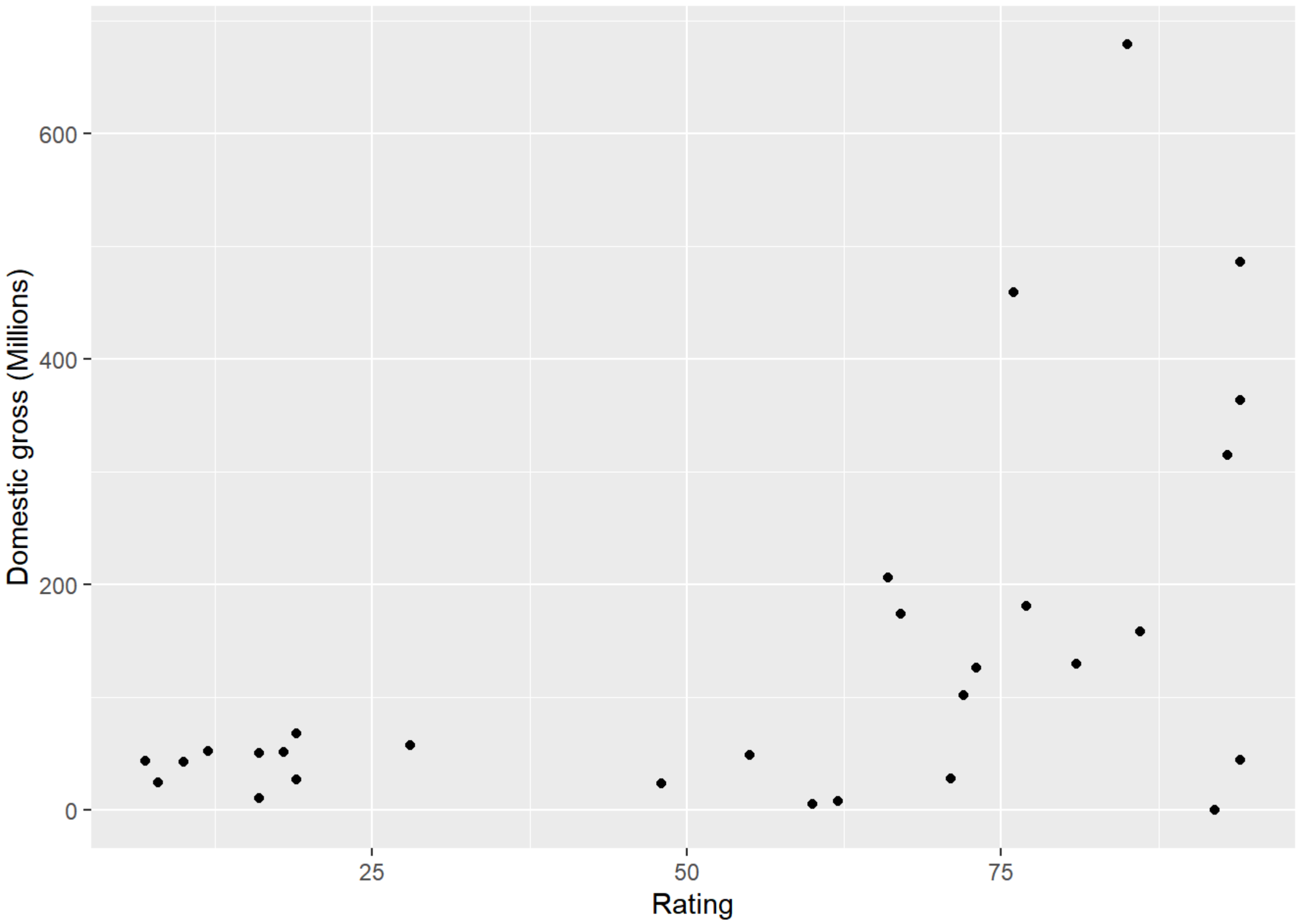
Step 2: Data Wrangling

We have to combine these two datasets into one so we can analyse it. This is called *Data Wrangling* or *Data Munging*.

```
## # A tibble: 29 x 3
##    Rating Title                                    ProductionBudget
##    <chr>  <chr>                                               <int>
##  1 19%    Cats                                            100000000
##  2 67%    Fast & Furious Presents: Hobbs & Shaw           200000000
##  3 85%    Avengers: Infinity War                          300000000
##  4 16%    The Dark Tower                                   60000000
##  5 93%    Thor: Ragnarok                                  180000000
##  6 94%    The Jungle Book                                 175000000
##  7 94%    The Jungle Book                                  27000000
##  8 86%    Star Trek Beyond                                185000000
##  9 94%    Finding Dory                                    200000000
## 10 16%    The Gunman                                       40000000
## # ... with 19 more rows
```

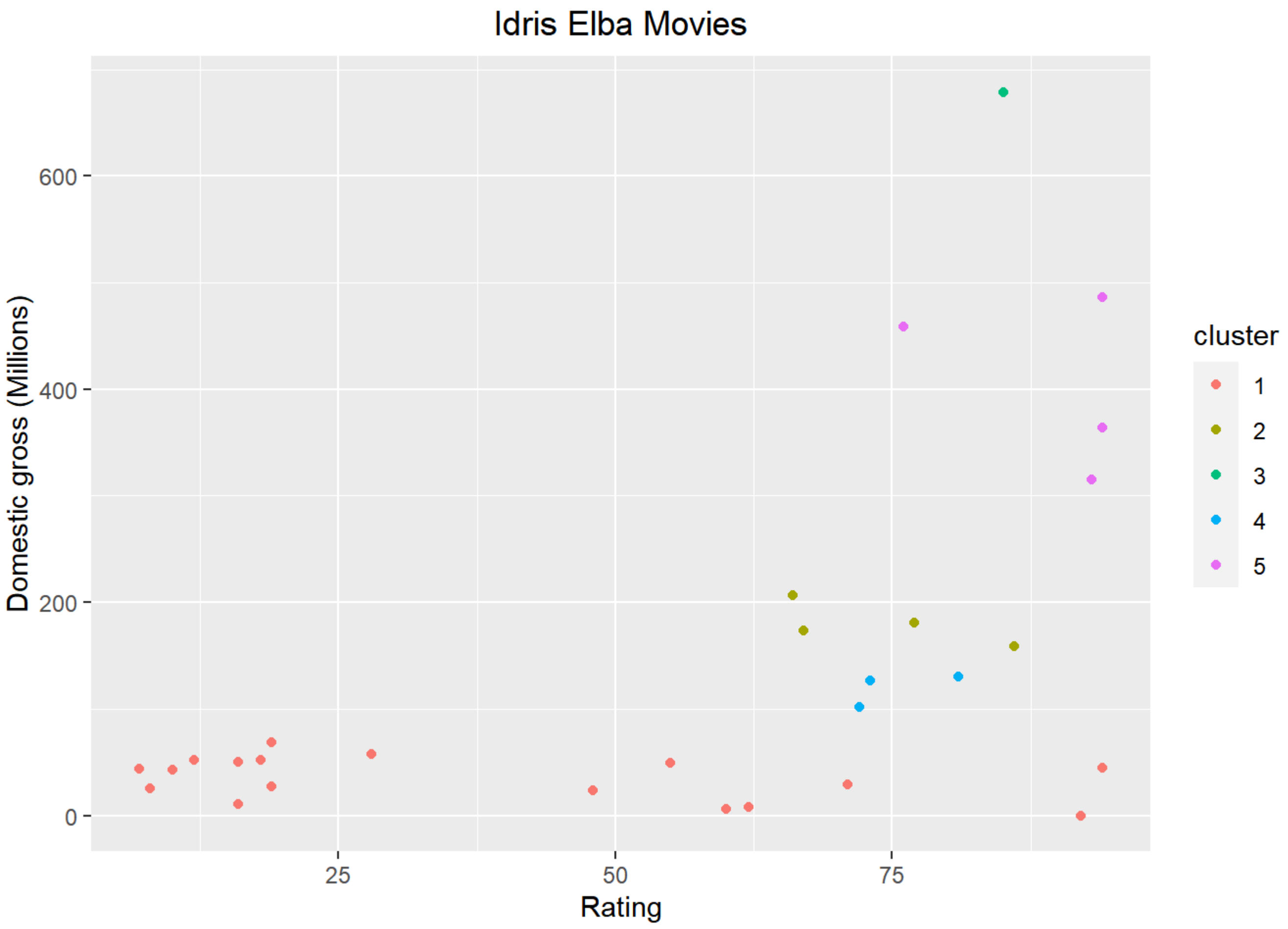
Step 3: Visualize the data

Let us plot our data!



Step 4: Modeling the data

We will use a clustering algorithm (https://en.wikipedia.org/wiki/Cluster_analysis) called *k*-means clustering (https://en.wikipedia.org/wiki/K-means_clustering) to group Idris Elba’s movies. To do this we must choose the number of groups (clusters). Five clusters seems a good choice. Here are the results:



Here is the finished dataset including the clusters.

```
##                                     Title Rating DomesticGross cluster
## 2                                   Cats     19      27166770       1
## 3  Fast & Furious Presents: Hobbs & Shaw     67     173956935       2
## 4                  Avengers: Infinity War     85     678815482       3
## 6                          The Dark Tower     16      50701325       1
## 9                          Thor: Ragnarok     93     315058289       5
## 10                        The Jungle Book     94     364001123       5
## 11                        The Jungle Book     94      44342956       1
## 12                       Star Trek Beyond     86     158848340       2
## 13                           Finding Dory     94     486295561       5
## 15                             The Gunman     16      10664749       1
## 16                Avengers: Age of Ultron     76     459005868       5
## 17                    Beasts of No Nation     92         90777       1
## 18                           No Good Deed     12      52543632       1
## 19          Mandela: Long Walk to Freedom     62       8323085       1
## 20                   Thor: The Dark World     66     206362140       2
## 21                            Pacific Rim     72     101802906       4
## 22                             Prometheus     73     126477084       4
## 23                                   Thor     77     181030624       2
## 24       Ghost Rider: Spirit of Vengeance     18      51774002       1
## 25                                 Takers     28      57744720       1
## 26                             The Losers     48      23591432       1
## 27                             The Unborn     10      42670410       1
## 28                               Obsessed     19      68261644       1
## 29                             RocknRolla     60       5700626       1
## 30                             Prom Night      7      43869350       1
## 31                      American Gangster     81     130164645       4
## 32                         This Christmas     55      49121934       1
## 34                            The Reaping      8      25126214       1
## 35                         28 Weeks Later     71      28638916       1
```