Movie Rating Comparison based on the genre Comedy vs Action and Adventure

Vinoth Kaladeepan

April 20, 2015

1 Introduction

Rotten Tomatoes is a website launched in 1998 devoted to film reviews and news; it is widely known as a film review aggregatr.Coverage now includes TV content as well. The name derives from the practice of audiences throwing rotten tomatoes when disapproving of a poor stage performance.The Rotten Tomaoes API (Application Program Interface) provides limited access to critic and audience ratings and reviews, allowing developers to incoporate Rotten Tomatoes data on other websites. The free service is intended for use in the US only; permission is required for use elsewhere.

This paper will focus on the comparison of movies rating based on the genre which are Comedy and Action and Adventure type repsectively.So in order to compare it we are considereing the top 100 movies in each category,finding out the mean of both the types,obtaining the final result and plotting the graph for it.

Data will be obtained from mentioning the url using rvest and table number and the data is stored into a data frame. Data will be explored in greater detail after it is obtained.Tables and graphs will be used as a visual description of the results gathered from analyzing the data.

2 Data

This section will conduct an in-depth exploration of the data, and will display detailed information about the dataset. Data will need to be obtained before being analyzed.We try and load all the data into a data frame in R for our further analysis. The data extracted will contain all the rows and columns without anyrestrictions. We need to perform data cleaning operations to get the data that we want.Data will be obtained from mentioning the url using rvest and table number and the data is stored into a data frame. Not all the data is needed.

2.1 Extracting data

In this process we are extracting different datas from the same resource. Data for comedey type will be obtained from mentioning the url and it is stored in a dataframe called address using rvest and table number command and the data is stored into a data frame called comedydata.In the dataframe comedydata we can see the columns Rank,RatingTomatometer,title and no.of reviews. Not all the data is needed.

```
#Represents the table number containing data
tablenumber <- 3;

#The source from which the data is to fetched
address <- rvest::html("http://www.rottentomatoes.com/top/bestofrt/?category=6")
comedydata <- rvest::html_table(rvest::html_nodes(address, "table")[[tablenumber]])

#Displays the first few rows of the dataframe comedydata
head(comedydata)
```

```
##   Rank RatingTomatometer                         Title No. of Reviews
## 1    1               100%          All About Eve (1950)            63
## 2    2                99%     A Hard Day's Night (1964)           101
## 3    3               100%           Modern Times (1936)            53
## 4    4               100%    Singin' in the Rain (1952)            48
## 5    5                98%  It Happened One Night (1934)            49
## 6    6               100% The Philadelphia Story (1940)            54
```

Code chunk 2: In this we are feteching the data for action and adventure type movies from the url provided and the table number is provided to get that particular data from the table and it is stored into a dataframe called actiondata.After which it is displayed using a head command.In the dataframe actiondata we can see the columns Rank,RatingTomatometer,title and no.of reviews in which we do not need all the data from it.So we proceed for the next step called Scrubbing of data.

```
#Represents the table number containing data
tablenumber <- 3;

#The source from which the data is to fetched
address <- rvest::html("http://www.rottentomatoes.com/top/bestofrt/?category=1")
actiondata <- rvest::html_table(rvest::html_nodes(address, "table")[[tablenumber]])

#Displays the first few rows of the data frame actiondata
head(actiondata)
```

```
##   Rank RatingTomatometer                                  Title
## 1    1               99%                     Metropolis (1927)
```

```
## 2    2              100%         The Adventures of Robin Hood (1938)
## 3    3               98%                             King Kong (1933)
## 4    4              100%     The Treasure of the Sierra Madre (1948)
## 5    5              100% Seven Samurai (Shichinin no Samurai) (1954)
## 6    6               98%                                   Up (2009)
##   No. of Reviews
## 1            114
## 2             44
## 3             54
## 4             44
## 5             57
## 6            280
```

2.2 Scrubbing Data

In a table not all data are needed for the analysis but we do need some of them.So this can be done through the process of cleansing.In this process we remove the unwanted rows and columns and we also rename them.So while considering our data we need only one column which is RatingTomatometer which has ratings percentage in it.So using the keeps command we are keeping only the particular column and storing it in a new data frame called mydata.Since the coulmn RatingTomatometer has an atomic vector in it,it is removed by using the percentvec command or using gsub command.Since the values of the datatype in the data frame you need are factors so we need to convert them to either text or numbers so we use the function as.character or as.numeric.

```
#Scrubbing of comedy data
comedydata.new<-comedydata[,1:2]

#Keeps command for selecting only particular column
keeps <-c("RatingTomatometer")
comedydata.new1<-comedydata.new[keeps]

#Displays the first few rows of the data frame comedydata.new1
head(comedydata.new1)

##   RatingTomatometer
## 1              100%
## 2               99%
## 3              100%
## 4              100%
## 5               98%
## 6              100%

#Renaming the column
```

```r
colnames(comedydata.new1)[colnames(comedydata.new1)=="RatingTomatometer"] <- "Rating"

#Replacing the percentage sign using gsub
comedydata.new1$Rating  <- gsub(pattern = "\\%",
                        replacement = "",
                        x = comedydata.new1$Rating,
                        ignore.case = TRUE,
                        perl = FALSE,
                        fixed = FALSE,
                        useBytes = FALSE)

#Displays the first few rows of the data frame comedydata.new1
head(comedydata.new1)

##   Rating
## 1    100
## 2     99
## 3    100
## 4    100
## 5     98
## 6    100

#Converts factor into numbers
comedydata.new1$Rating <- as.numeric(gsub(",","",comedydata.new1$Rating))
```

Code chunk 4: Find the mean of the particular column using the mean command and display the result.

```r
# To find mean of the paticular column
mycdata<-mean(comedydata.new1$Rating)

#Displays the result
mycdata

## [1] 96.63
```

Code chunk 5: So while considering our data we need only one column which is RatingTomatometer which has ratings percentage in it.So using the keeps command we are keeping only the particular column and storing it in a new data frame called myaadata.Since the coulmn RatingTomatometer has an atomic vector in it,it is removed by using the percentvec command or using gsub command.Since the values of the datatype in the data frame you need are factors so we need to convert them to either text or numbers so we use the function as.character or as.numeric.

```r
#Cleansing of Action and Adventure data
actiondata.new<-actiondata[,1:2]

#Displays the first few rows of the data frame actiondata.new
head(actiondata.new)

##   Rank RatingTomatometer
## 1    1                99%
## 2    2               100%
## 3    3                98%
## 4    4               100%
## 5    5               100%
## 6    6                98%

#Keeps command for selecting only particular column
keeps <-c("RatingTomatometer")
actiondata.new1<-actiondata.new[keeps]

#Renaming the column
colnames(actiondata.new1)[colnames(actiondata.new1)=="RatingTomatometer"] <- "Rating"


#Replacing the percentage sign using gsub
actiondata.new1$Rating  <- gsub(pattern = "\\%",
                      replacement = "",
                      x = actiondata.new1$Rating,
                      ignore.case = TRUE,
                      perl = FALSE,
                      fixed = FALSE,
                      useBytes = FALSE)

#Displays the first few rows of the data frame actiondata.new1
head(actiondata.new1)

##   Rating
## 1     99
## 2    100
## 3     98
## 4    100
## 5    100
## 6     98

#Converts factor into numbers
actiondata.new1$Rating <- as.numeric(gsub(",","",actiondata.new1$Rating))
```

Code chunk 6: In this we find the mean of the particular column using the

mean command and display the result.

```
# To find mean of the paticular column
myaadata<-mean(actiondata.new1$Rating)

#Displays the result
myaadata

## [1] 95.46
```

2.3 Exploring data

Exploring the data that we have gathered is as important as analasing the data.We need to know the data we have and the format of the data and the datatypes and the amount of data we have.

```
# Display the data type of comedydata.new1
class(comedydata.new1)

## [1] "data.frame"

# Display the data type of actiondata.new1
class(actiondata.new1)

## [1] "data.frame"
```

The above code display the data type of dataset.whether it is data.frame,arrays,lists,vetcor etc.

```
# Display the summary of the whole data set
summary(comedydata.new1)

##      Rating
##  Min.   : 91.00
##  1st Qu.: 95.00
##  Median : 97.00
##  Mean   : 96.63
##  3rd Qu.: 98.00
##  Max.   :100.00

# Display the summary of the whole data set
summary(actiondata.new1)

##      Rating
##  Min.   : 88.00
##  1st Qu.: 93.75
```

```
##  Median : 96.00
##  Mean   : 95.46
##  3rd Qu.: 98.00
##  Max.   :100.00
```

The above code tells us the summary of the data. It tells us the max,min,mean values,etc. for the dataset and also other properties of the data.

```
# Display the structure of the whole data set
str(comedydata.new1)

## 'data.frame': 100 obs. of  1 variable:
##  $ Rating: num  100 99 100 100 98 100 100 99 98 99 ...

# Display the structure of the whole data set
str(actiondata.new1)

## 'data.frame': 100 obs. of  1 variable:
##  $ Rating: num  99 100 98 100 100 98 99 99 94 98 ...
```

The above code tells us the data types for different columns and the form in which the data is present. Here we see that the rating column is in number datatype.

3 Results

This section will all display the result

```
V1=(c("Comedy","Action and Aventure"))
V2=(c(mycdata,myaadata))
result<-cbind(V1,V2)

# Display all the data
result

##      V1                    V2
## [1,] "Comedy"              "96.63"
## [2,] "Action and Aventure" "95.46"
```

From the result we can see that mean of comedy and action and adventure types of movies have not much of a difference between them.And the end of anlaysis,extraction and cleansing of data,the result tend to be almost same for both the movie type which seems to be intresting which gives a new perspective on both the movie types.We can clearly make out from it that people do watch good movies irrespective of their type or genre and also give good rating and feedback about the movies.

4 Plotting of graphs

Just getting tabular data is never enough for analysis. Analysing requires proper understanding of the data, which we can usually get from visualisations.There are many ways to plat a graph but the most common way to plot is using bar graph.

```
# Use ggplot command from ggplot2 package to draw a bar graph
## that shows the total number of rating based on their genre
# data is obtained from movies data frame
# x axis shows the genre, y axis shows the number of ratings
# display a bar graph

library(ggplot2)
Movies <- data.frame(
  Genre = factor(c("Comedy","Action and adventure"),
                 levels=c("Comedy","Action and adventure")),
  Ratings = c(mycdata,myaadata))

ggplot(data=Movies, aes(x=Genre, y=Ratings, fill=Genre)) +
  geom_bar(stat="identity")+
  ggtitle("Movie Rating Comaparison")
```

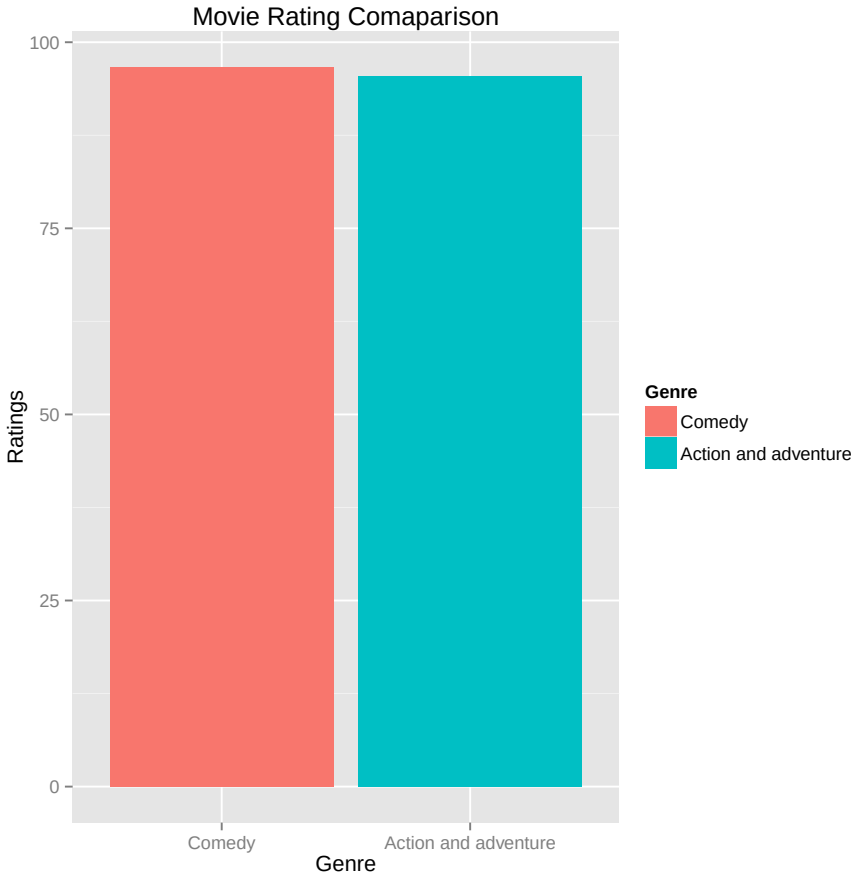

Figure 1: Figure 1.1 shows the Genre along the bottom and the counts of ratings along the side (y-axis)

The above graph Figure 1.1 shows the movie rating comparison between comedy and action and adventure type for top 100 movies in each category.We see that the mean of both comedy and action and adventure seems to have difference in their rating but not much of it.

```
# Use ggplot command from ggplot2 package to draw a line graph
## that shows the total number of rating based on their genre
# data is obtained from school.new data frame
# x axis shows the genre, y axis shows the ratings
# all data points in this graph need to be connected, so group=1
# display a line graph with points


library(ggplot2)
```

```
Movies <- data.frame(
  Genre = factor(c("Comedy","Action and adventure"),
                 levels=c("Comedy","Action and adventure")),
  Ratings = c(mycdata,myaadata))

ggplot(data=Movies, aes(x=Genre, y=Ratings, group=1)) +
  geom_point(stat="identity")+geom_line()+
  ggtitle("Movie Rating Comaparison")
```

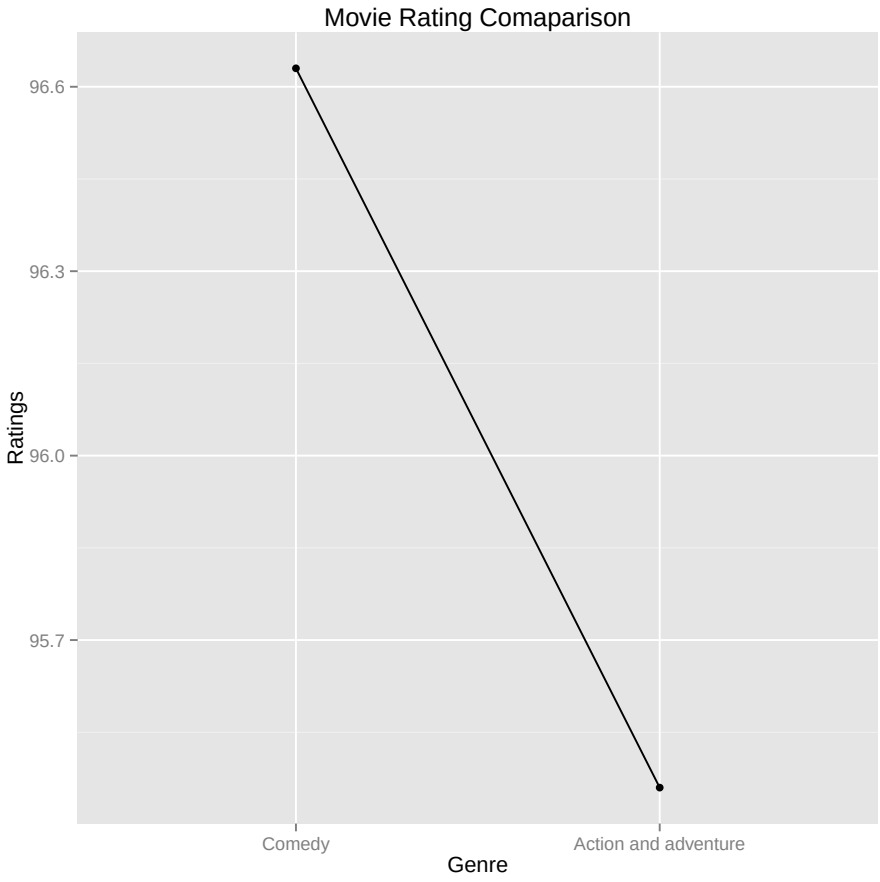

Figure 2: Figure 1.2 shows the Genre along the bottom and the counts of ratings along the side (y-axis)

The above graph Figure 1.2 shows the movie rating comparison between comedy and action and adventure type for top 100 movies in each category.We see that the mean of both comedy and action and adventure types have different values but not much of a difference though.