Assignment 4 Data Analytics Project

Sandeep Tom

12/12/2021

1. Discuss the business problem/goal :The goal of this analysis is to conduct a sentiment analysis of movie ratings from IMDB website. 2.identify where the dataset was retrieved from: For this we will be using 50,000 ratings that were available from their website to conduct the visualizations.
2. The code that imported and saved dataset in R is below

## Rows: 1193894 Columns: 3

## -- Column specification --------------------------------------------------------  
## Delimiter: "\t"  
## chr (1): tconst  
## dbl (2): averageRating, numVotes

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## # A tibble: 6 x 3  
## tconst averageRating numVotes  
## <chr> <dbl> <dbl>  
## 1 tt0000001 5.7 1845  
## 2 tt0000002 6 236  
## 3 tt0000003 6.5 1603  
## 4 tt0000004 6 153  
## 5 tt0000005 6.2 2424  
## 6 tt0000006 5.2 158

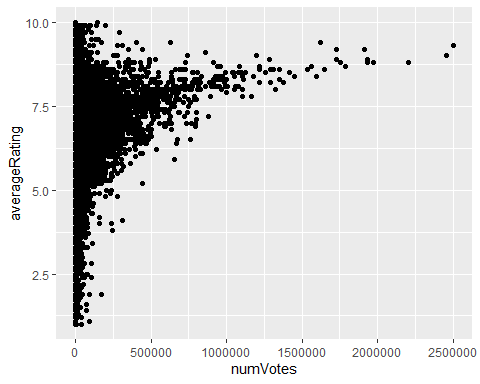
1. describe your data set (using the common attributes such as #rows, #columns, variable names, types, means, SD, min/max, NAs, etc…)

summary(imdb\_ratings)

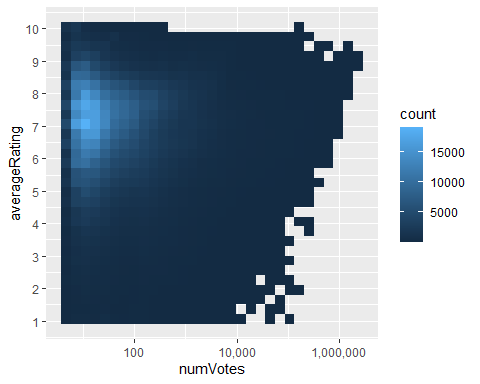
## tconst averageRating numVotes   
## Length:1193894 Min. : 1.000 Min. : 5.0   
## Class :character 1st Qu.: 6.200 1st Qu.: 11.0   
## Mode :character Median : 7.100 Median : 24.0   
## Mean : 6.926 Mean : 993.2   
## 3rd Qu.: 7.900 3rd Qu.: 93.0   
## Max. :10.000 Max. :2502655.0

1. The data is available on the IMDB website that contains rating information for title and the number of votes each of those titles received. From review of the dataset it looks like all missing values were removed from the dataset. For datasets of these sizes on of the key issues would be missing values / missing ratings. The dataset we are working with have had all missing values removed from them.
2. Modelling - Below we will be using visualizations techniques to analysis and gather insights from the data.
3. Below is a graph of the ratings of all the movies by the number of votes they received in IMDB. The plot below gives you an idea of how each of those titles were voted but does not give you a very clear idea due to the volume of the data points

ggplot(imdb\_ratings, aes(x = numVotes, y = averageRating)) +  
 geom\_point()



So we use a heat map to plot the same information.The graph below gives you a slightly better picture of votes by title rating



Now we will use the movies to tie them to their ratings. For that we will first use another dataset that is avialble on website.

We introduce the data and then now join this data to the existing dataset.

## Rows: 1193894 Columns: 3

## -- Column specification --------------------------------------------------------  
## Delimiter: "\t"  
## chr (1): tconst  
## dbl (2): averageRating, numVotes

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Rows: 8506043 Columns: 9

## -- Column specification --------------------------------------------------------  
## Delimiter: "\t"  
## chr (5): tconst, titleType, primaryTitle, originalTitle, genres  
## dbl (4): isAdult, startYear, endYear, runtimeMinutes

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Joining, by = "tconst"

## # A tibble: 6 x 11  
## tconst averageRating numVotes titleType primaryTitle originalTitle isAdult  
## <chr> <dbl> <dbl> <chr> <chr> <chr> <dbl>  
## 1 tt0000001 5.7 1845 short Carmencita Carmencita 0  
## 2 tt0000002 6 236 short Le clown et ~ Le clown et ~ 0  
## 3 tt0000003 6.5 1603 short Pauvre Pierr~ Pauvre Pierr~ 0  
## 4 tt0000004 6 153 short Un bon bock Un bon bock 0  
## 5 tt0000005 6.2 2424 short Blacksmith S~ Blacksmith S~ 0  
## 6 tt0000006 5.2 158 short Chinese Opiu~ Chinese Opiu~ 0  
## # ... with 4 more variables: startYear <dbl>, endYear <dbl>,  
## # runtimeMinutes <dbl>, genres <chr>

## Warning in grid.Call(C\_stringMetric, as.graphicsAnnot(x$label)): font family not  
## found in Windows font database  
  
## Warning in grid.Call(C\_stringMetric, as.graphicsAnnot(x$label)): font family not  
## found in Windows font database

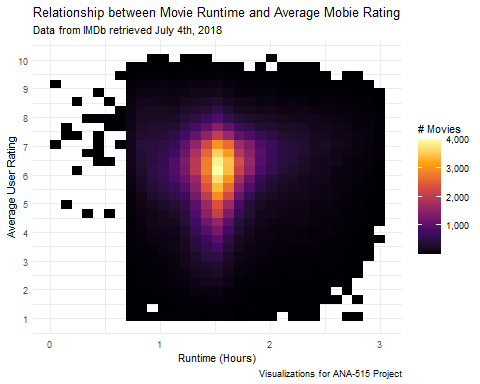
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database

## Warning in grid.Call(C\_stringMetric, as.graphicsAnnot(x$label)): font family not  
## found in Windows font database

## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database

## Warning in grid.Call.graphics(C\_text, as.graphicsAnnot(x$label), x$x, x$y, :  
## font family not found in Windows font database

## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database



## Warning: Removed 16 rows containing non-finite values (stat\_bin2d).

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 16 rows containing non-finite values (stat\_smooth).

