CA682 Data management and visualisation

Name	Vikas Chhillar
Student Number	16212887
Programme	M.SC Computing
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	09/12/16
Module coordinator	Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name:Vikas Chhillar
Date: 09/12/16

Introduction

Dataset: The dataset consist of IMDB movies and their ratings. The dataset gives information about IMDB rating, Movie title, Director name etc. Two files of dataset are merged using excel. The second file of dataset has got facebook likes of the movies,actor 1, actor 2, actor 3 and director. The movie title is common between both the files of Dataset. All

the column names are filled with the appropriate names. All the blank values are filled with the appropriate information.

Data Refining and cleaning:

Two tools are used:
Microsoft Excel:
Some basic changes are done like arranging the data in order for my better understanding and then merging the 2 files of Dataset namely, IMDB movie lists and rating and the other one is facebook likes of actors and directors.

Google OpenRefine :

- Column names are changed.
- Remove special character A in column movie title using excel by opening finder and then replace the special character with null value because the character exists at the end of every movie title. The actors with facebook likes more than 10K are selected to work on the data of only the famous actors.
- Some high IMDB scores and low IMDB are selected using the open refine by clicking on facet and then clicking text facet.

Languages used:

- R language and Tableau are used in this project.
    - Different packages of R programming are used in this project. The package can be installed using install.packages("name of package").
    - These are the packages used in this project.
        - library(dplyr)
        - library(ggplot2)
        - library(plotly)
        - library(data.table)
        - library(formattable)

Some high IMDB scores and low IMDB are selected using the open refine.
Reading the data:
moviedata<-read.csv('/Users/vikaschhillar/Downloads/movie_metadata.csv'),

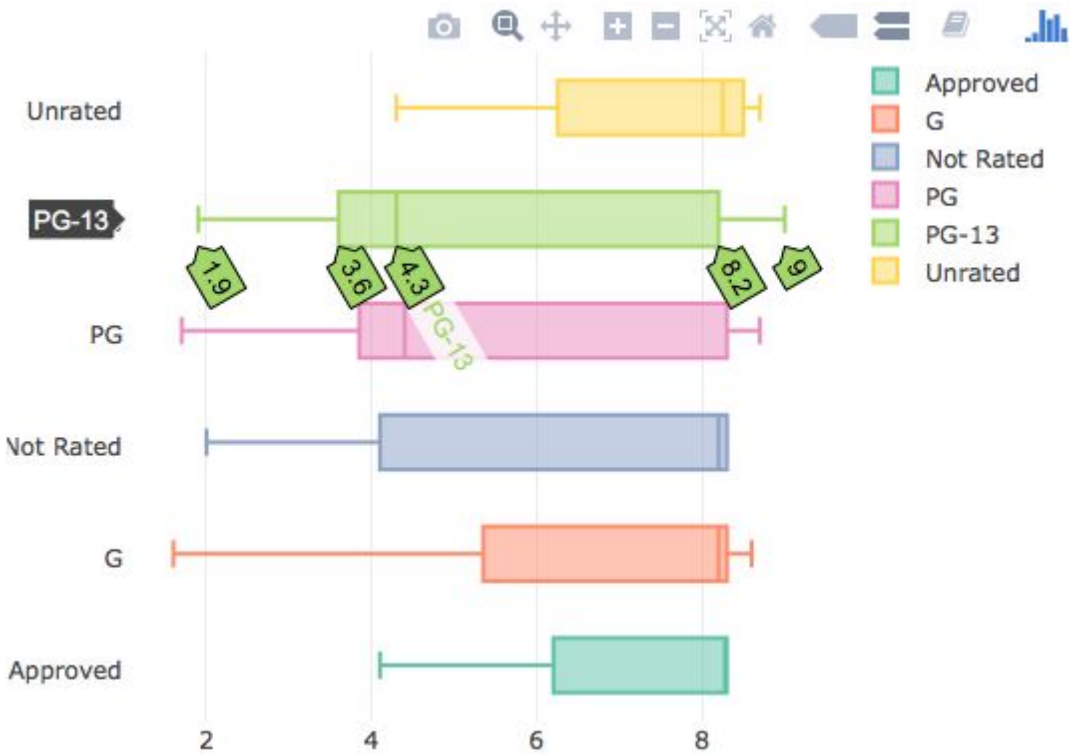
Graph 1: The data is cleaned in the openrefine by clicking on different variables and then including or excluding them. The IMDB score are filtered. Some of the low scores and high IMDB scores are chosen. The same is done with the content rating. Once we have filtered the data, We go into R studio and download different libraries. The different colors in the graph gives the content rating and the cursor on the screen gives the IMDB rating.
When the cursor is moved on the graph, it gives the values of IMDB score of each category that is selected from the data.

```
move<-read.csv("/Users/vikaschhillar/Downloads/movie_metadata-csv (1).csv,header=T,stringsAsFactors = FALSE")

movie_dataset <- read.csv('/Users/vikaschhillar/Downloads/movie_metadata-csv (1).csv',header=T,stringsAsFactors = F)

New <- movie_dataset %>% select(imdb_score,content_rating)
New <- na.omit(New)
plot_ly(New, x = New$imdb_score, color = New$content_rating, type = "box")
```

- The information on the right gives the content of the movies.



- The x-axis gives the IMDB Score. The y-axis gives the detail about rating content. In the second graph, all the movies with facebook likes more than 10K are taken. This is done by openrefine. click on the facet and then click on text facet and choose the values with more than 10K Facebook likes.

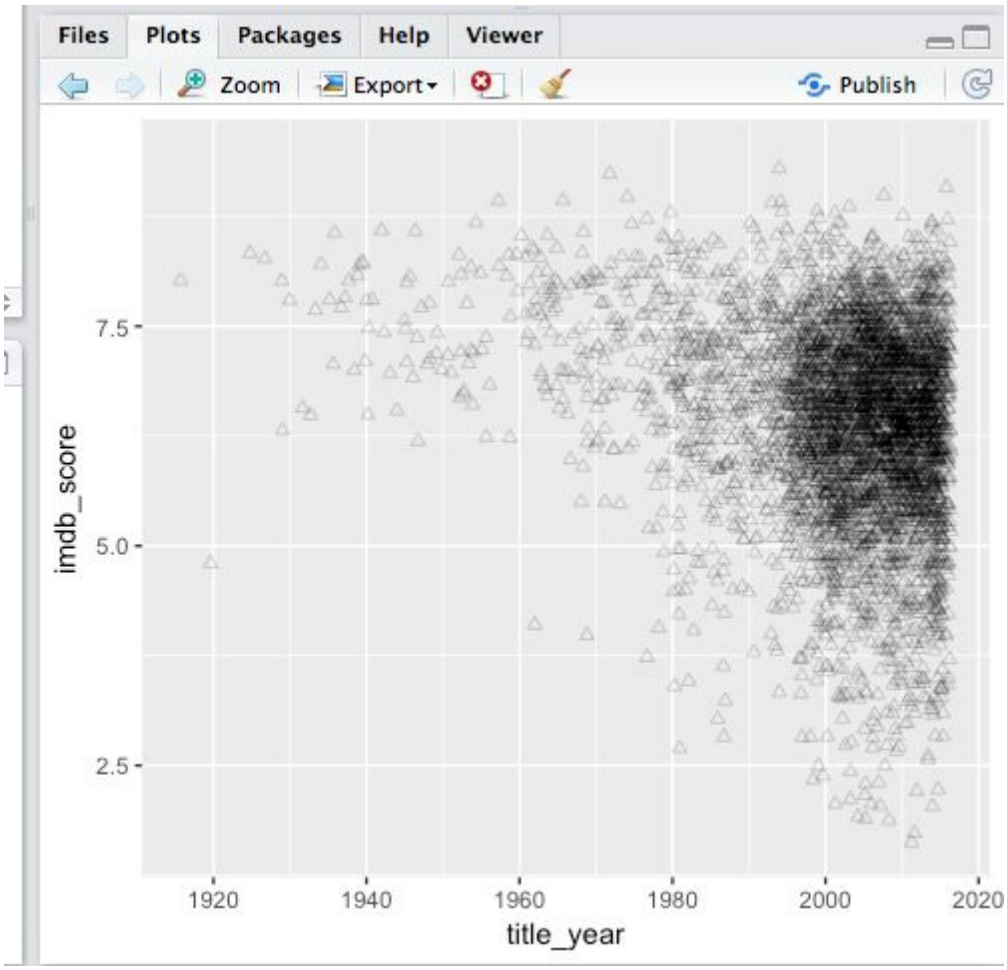
Graph 2**:Is the no. of likes for the director related to the IMDB score?**

```
movie_dataset<-read.csv("/Users/vikaschhillar/Downloads/movie_metadata-csv (2).csv")
New <- movie_dataset %>% select(imdb_score,movie_facebook_likes)
New <- na.omit(New)
plot_ly(New, x = New$imdb_score, color = New$movie_facebook_likes, type = "box")
```

- The x-axis of the graph tells about the number of facebook likes of the director and y-axis tells us about the the IMDB rating.
- The graph is interactive. By taking the cursor on the graph, it gives the value of x-axis and y-axis.

Graph 3: The relationship between IMDB score and year.

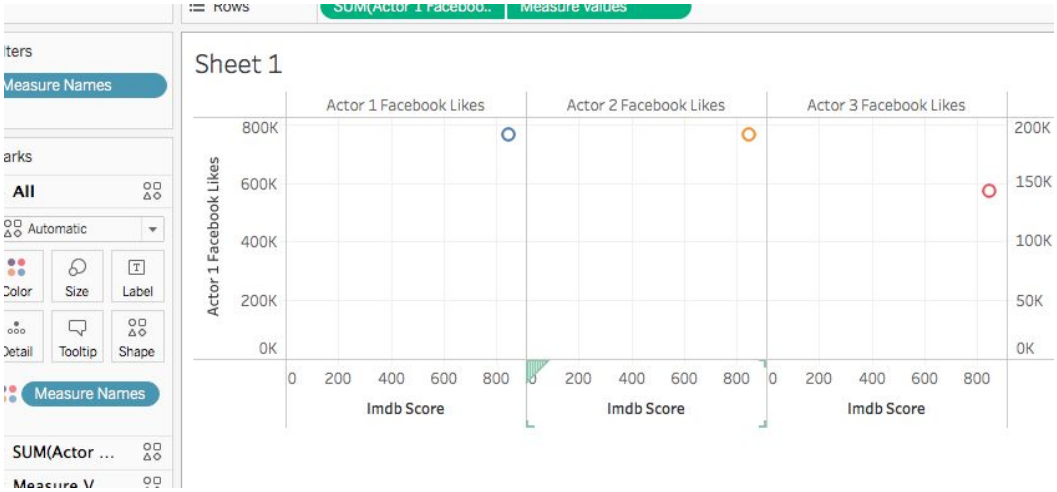
```
read.csv('/Users/vikaschhillar/Downloads/movie_metadata-csv (2).csv')
library(ggplot2)
ggplot(data=moviesdata,aes(x=title_year,y=imdb_score))+geom_jitter(alpha=0.2,shape=2,color="black")
```



- This is clear from the above graph that IMDB score in early 1900's was always high for the movies. But with increase in number of year, more movies can be seen in the given year and IMDB rating for some movies are less than 1.

Tableau

- The csv files are loaded in the tableau.
- The actor names for actor1, actor2, actor3 are selected and dragged into Rows section.
- The IMDB scores are selected and dragged into the column section.
- The graph clearly shows the total of facebook likes and their dependency with IMDB score.



The above graph is made with the tableau. The IMDB score is compared with the facebook likes for each actor. The 3 dot in the graph gives the total of facebook likes for all actors in each graph.

Conclusion

- The relationship between between IMDB score and rating content can be clearly seen in the first interactive graph.
- The data is analysed between number of facebook likes for director and actors and their dependency on the IMDB score.
- The numbers of movies between 2003 and 2012 are almost same.
- There is only one movie in 1920 which was given the rating of 7.8 .
- It is easier to work in the tableau as no programming skill is needed.

References

https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset