CA682 Data management and visualisation

Name	Aishwarya Gupta
Student Number	18210298
Programme	MCM
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	16, Dec 2018
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: Aishwarya Gupta Date: 16-12-2018

Introduction:

For this data visualisation assignment, I have used the IMDB data set available on IMDB website to perform exploratory analysis of the movie data. I have explored three questions:

1. The movies released across 5 genres (Drama, Horror, Action, Comedy and Romance) released year wise with rating by number of votes.
2. Genre wise number of movies released from the year 1950 to 2018, this graph shows the trend in the number of movies released every year for these genres.
3. The top 15 actors, actresses and both from 2010 to 2018 in terms of score generated by movie ratings and number of votes.

By performing exploratory analysis using these visualisation, we can find various information like, the highest rated movies of various genres in each year, the contrast in the number of movies released each year of these 5 genres and the trend that can be seen across the years, the actors and actresses who have the highest scoring movies in the current decennium.

The purpose of this visualisation is to explore the IMDB data set and find trends and interesting information from it.

I have used Jupyter Notebooks for data cleaning and visualisation. I have created the visualisations using the Bokeh library. As the bokeh library does not support interactivity directly within the notebook, I have exported my code into a .py file and run that file on the bokeh server.

Dataset:

I have acquired my data set from the official IMDB website using the link – https://datasets.imdbws.com/ . The information regarding each of these files can be found from the link - http://www.imdb.com/interfaces/ . I have used the below files for my visualisations –

1. name.basics.tsv.gz – consists of information of name, knownfortitles array, primary profession, etc. The number of rows present in the file is 8998554.

nconst	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0053137,tt0050419,tt0043044,tt0072308

2. title.ratings.tsv.gz – consists of movie names, rating and number of votes. The number of rows in the file is 896805

tconst	averageRating	numVotes
tt0000001	5.8	1441

3. title.principals.tsv.gz – contains main cast and crew for the titles. The number of records in this file is 31153380

tconst	ordering	nconst	category	job	characters
tt0000001	1	nm1588970	self	\N	["Herself"]

4. title.basics.tsv.gz – contains genre, primarytitle, startyear, etc. The number of records in this file is 5456841.

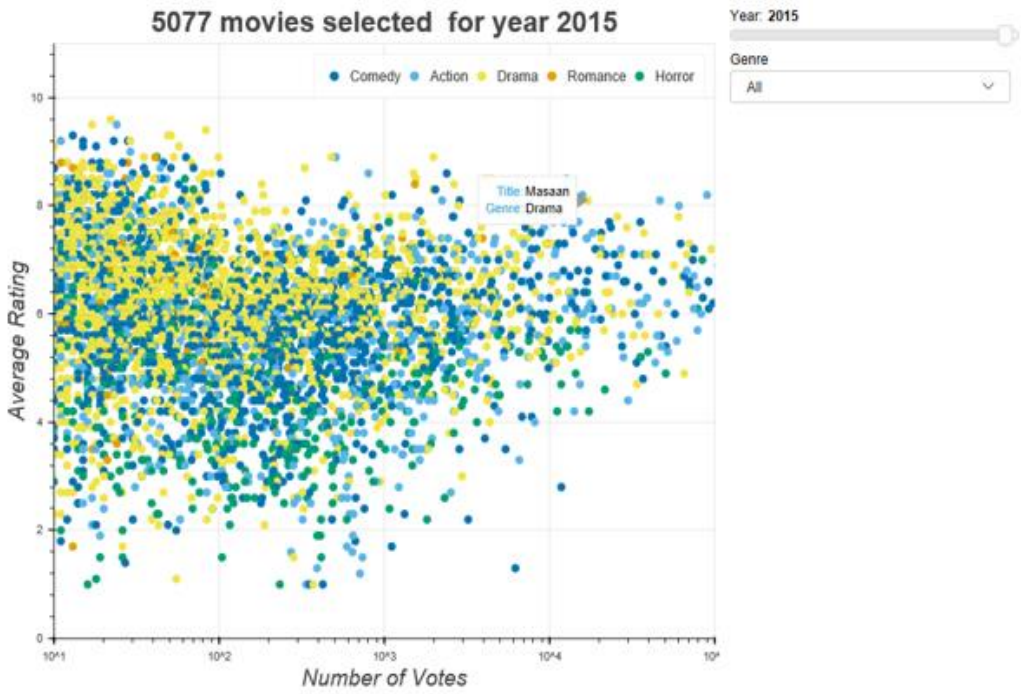
tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary, Short

This site is updated every day, I had downloaded the data on 7th, Dec 2018 for my visualisations.

Process:

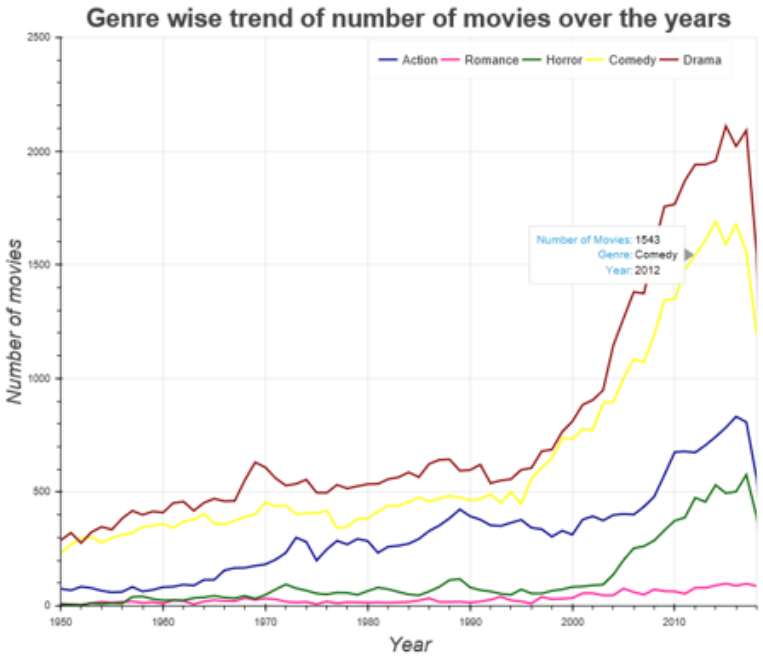
I followed the below steps for creating my visualisations.

1. Data cleaning: Since all my files had many records the first thing I did was to clean the data in such a way that it was useful to me.
   a. For the comma separated fields present in files, I first converted them into columns and removed extra columns which I was not interested in e.g. the genre column contained comma separated values which I partitioned into genre1 and genre2, I only kept genre1 in my dataset as I did not want to consider sub genres.
   b. I wanted to consider only actor and actresses for my analysis, for that I first converted comma separated values in primaryProfession into columns and removed data where the 1st profession was not either actor or actress.
   c. I dropped columns which were of no interest to me like birth year, death year of the actors and actresses, end year of movie or series.
   d. I removed records having NA values for start year considering that most of these records were of very old movies before 1943.
   e. I converted the data type of start year to numeric for easy processing of information.
   f. I have merged the title ratings and title basics into a single data frame as these values will be used frequently together.
   g. Finally, I exported this data into new files for easy access to clean data for visualisation.
2. Data visualisation 1 – For creating my first visualisation, I wanted to create a graph which gave year and genre wise information of all the movies released in that year based on the number of votes and rating.



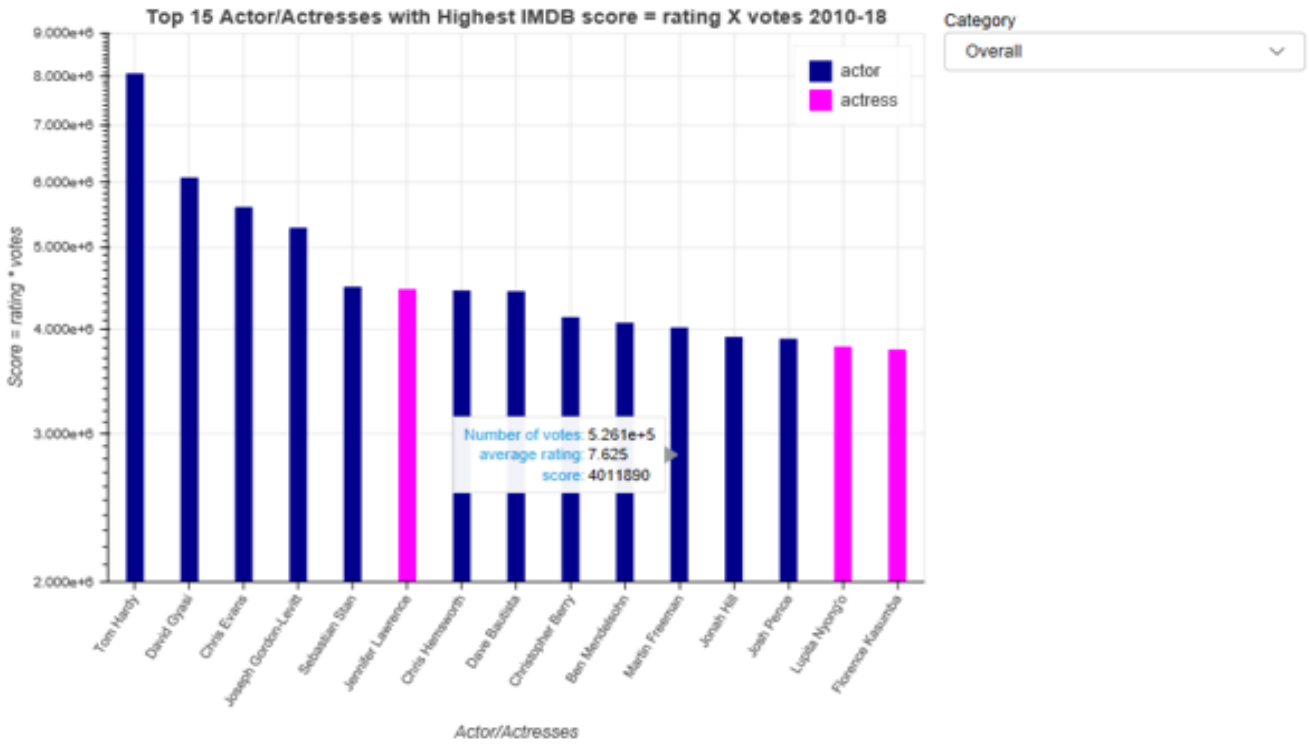
a. I decided to make a scatter plot graph, which would plot movies based on the ratings and number of votes.
b. I decided to use five genres for my analysis as it would be easy to visually comprehend the information and be meaningful.
c. The title of the visualisation showed the total number of movies released that year for these 5 genres.
d. The slider on the side can be used to change the value of the year.
e. The drop down allows you to select the genre that you are interested in and see the movies of that genre in the graph.
f. Hovering over the point in the scatter plot provides the name of the movie and the genre that it belongs to.
g. I have used log scale on x axis as the number of votes ranged from 10 to lakhs of values.
h. For the colour scheme, I have used the colour-blind palette from bokeh so that my graph is understandable to colour blind people as well.
i. I have added legend on the top right corner where I will not have any of my values, such that none of my values are covered by the legend.
j. The legend tells which colour represents which genre. I have used a distinct colour palette as none of these values are related and are distinct categorical values.
k. When a value is selected from the drop down for genre, the number of movies change as well to represent the number of movies of that genre in that year

3. Data visualisation 2 – For my second visualisation, I wanted to explore the trend of the number of movies made across years for the above 5 genres.



a. For making this visualisation, I had to perform some more data manipulation on the previously cleaned data.
b. Since, I was interested in movies only of only these 5 genres (Action, Romance, Horror, Comedy, Drama) I decided to drop rest of the data from the data frame.
c. I decided to drop the data other than movies since, I wanted to concentrate only on movies for my exploration.
d. I wanted my timeline to be from 1950 to 2018, hence I removed data for movies before 1950.
e. I grouped the data according to genre and counted the number of movies in each genre for a year and stored this information into a new data frame.
f. Since I was working with time on my x axis, I decided to use a line graph as they are preferable when working with time.
g. The Y axis represents the number of movies, and the coloured lines represent the genre.
h. On hovering over a line at a point, information about the number of movies, year and genre is shown.
i. On clicking the legend for a genre, the line representing that genre is hidden so that comparison can be made on genres as the viewer desires.
j. The colours used for the genres are according to colours that are generally associated with that genre e.g. yellow for comedy, Blue for action, pink for romance, etc.
k. A clear trend can be seen from the graph, depicting that the number of movies made for drama are way higher than that of romance.

4. Data visualisation 3 – For the third graph, I wanted to show the top 15 actors/actresses whose movies have the highest ratings and votes.



a. The major issue in this graph was identifying actors and actresses and creating an overall score based on the known titles of the actors and actresses. For this I first reduced the size of my data as I was interested in only movie data from the year 2010.
b. I merged the names data with the titles_prin data, which had information about major actor and actresses of a title. By doing this I was able to filter only the people who had done a major role in at least one movie.
c. Merging with names data gave me the 4 known movie titles of these actors and actresses. I combined this data with the rating data 4 times, each time with my join condition being on the 4 different known titles fields.
d. After getting the rating and number of votes for each title, I added it to the overall rating and number of votes.
e. I divided the final value by 4, as this was the number of movies per person that I had.
f. I combined the value of ratings and number of votes to generate a score. This score was to remove bias such as having high rating with less no of votes vs. having a medium rating but with high number of votes.
g. The actors, actresses and overall list of 15 people was maintained in separate data frames.
h. I used a bar chart as it clearly reflected the score of each person on the x axis. It can be seen how high the score of each individual is, and how it decreases as the ranking of the person decreases.
i. I added a drop down which makes the graph interactive, based on the value selected the graph can be viewed for top 15 actresses, top 15 actors or overall top 15.

j. Each time the value is changed in the drop down, a call back function is used to change the source data and point to the correct data frame.
k. The labels, titles and values also change as per the value selected in drop down.
l. I used the colours dark blue and magenta, as they can be easily related to actors and actresses. The colours are also colour blind friendly making it perceivable by all.
m. Hovering of the bars gives information regarding the overall score, number of votes and average rating of the individual.
n. I slanted the values on x axis to 45 degrees to make them clearly visible.

Results:

While creating the visualisations, I paid close attention to the type of graph suitable for the kind of information that I wanted to show. All my graphs are exploratory rather than being explanatory, however we can see that each graph presents a lot of information.
I paid close attention to the following design principles while making my graphs:

1. Keeping the titles short and to the point.
2. Clearly labelling the x and y axis to provide maximum information about the graphs.
3. Data on the axis is always ordered, in terms of numerical values or while representing categorical values.
4. The colours for categorical values in line graph are distinct representing each genre to colour closely associated with it.
5. The colours chosen for all the graphs are colour blind friendly.
6. Graph is two dimensional, especially the bar chart to show all the values clearly.
7. Unnecessary labels and clip arts are not present on the graphs.
8. Kept the Y axis as 0, wherever possible.
9. Used only 5 things at a time to show information, making it easy for the viewer to comprehend the information.
10. Made graphs where it is easy to view the information for a single entity as well as whole, in terms of the availability to select a genre for scatter plot and seeing its movies on the graph, and the ability to reduce the number of lines on the timeline, to closely see and compare the values of number of movies over the years.

Possible shortcomings for my graphs could be the following:

1. At first seeing the scatter plot graph can be a bit overwhelming, since it contains a lot of values. However, carefully viewing the graph we can see that it is very informative and can be helpful in selecting which movie to see next.
2. In the bar graph, I couldn't start the Y axis from 0 as the numbers were too large, this can make it a bit difficult to understand the values relative to 0.
3. The graphs are exploratory and not explanatory, they do not present a single answer but can answer many questions.