Movie Dataset Classification

Team Members

Kanagala Aashitha -19BCD7159

U. Hrushikesh Chowdary – 19BCI7073

Tadisetty Sai Ravi Teja -19BCE7150

Nittoor Vishnu Bharadwaj – 19BCE7478

VIT-AP UNIVERSITY

Guided by

Dr.Gopi Krishna

CSE 4027 Data Analytics

Contents

Introduction:

Movie data set classification is a topic of interest both to academics and industry. Most of the classification schemes are focused on users' preferences in selecting future movies. But a classification scheme targeted for the future popularity of movies enables producers, financiers, academics, or even viewers to understand the contributing factors that lead to movies' success. This is because too many parameters of different degrees are related and finding a suitable way to represent all the information related to a movie in a single instance is a cumbersome task. Even if a way is found out to represent a movie the final choice of classifiers to generate the model requires considerable research. Again in the case of the post-release movie the point of interest centers on the financial return. The problem of data representation and classification exists in this case also. So it is required to clean, preprocess and visualize the dataset after analyzing it in the form of graphs to predict the popularity of the pre-release and post-release movie.

Dataset Description:

Here, we are using an IMDB dataset having more than 50K movie reviews. We use this IMDB movie data set as it contains more precise and accurate data than any other datasets available online. We clean and classify this data set for identifying the genre with the highest success rate. We have information on the Movie title, original title, year, director, votes, average vote, reviews from the critics as well as from the users, language, actor, genres, episodes, production company, and release date. Missing Values, Outside of key fields, missing values are common. Sometimes the data seems to be unavailable, sometimes it hasn't been entered. Some information is inherently incomplete. Censored Data is ignored.

System requirements:

1. **Software Requirements:**

   Linux or Windows Operating System
   IDE: R.4.2.1
   R language: R.4.2.1

2. **Hardware Requirements:**

   Processor: Intel Core 13 and above
   RAM: 4 GB+

Key Steps:

Step 1: Collecting dataset.
Step 2: Loading dataset into R-studio.
Step 3: Cleaning data.
Step 4: Preprocess the data.
Step 5: Visualising the data in form of graphs to get the results.

Methods And Analysis

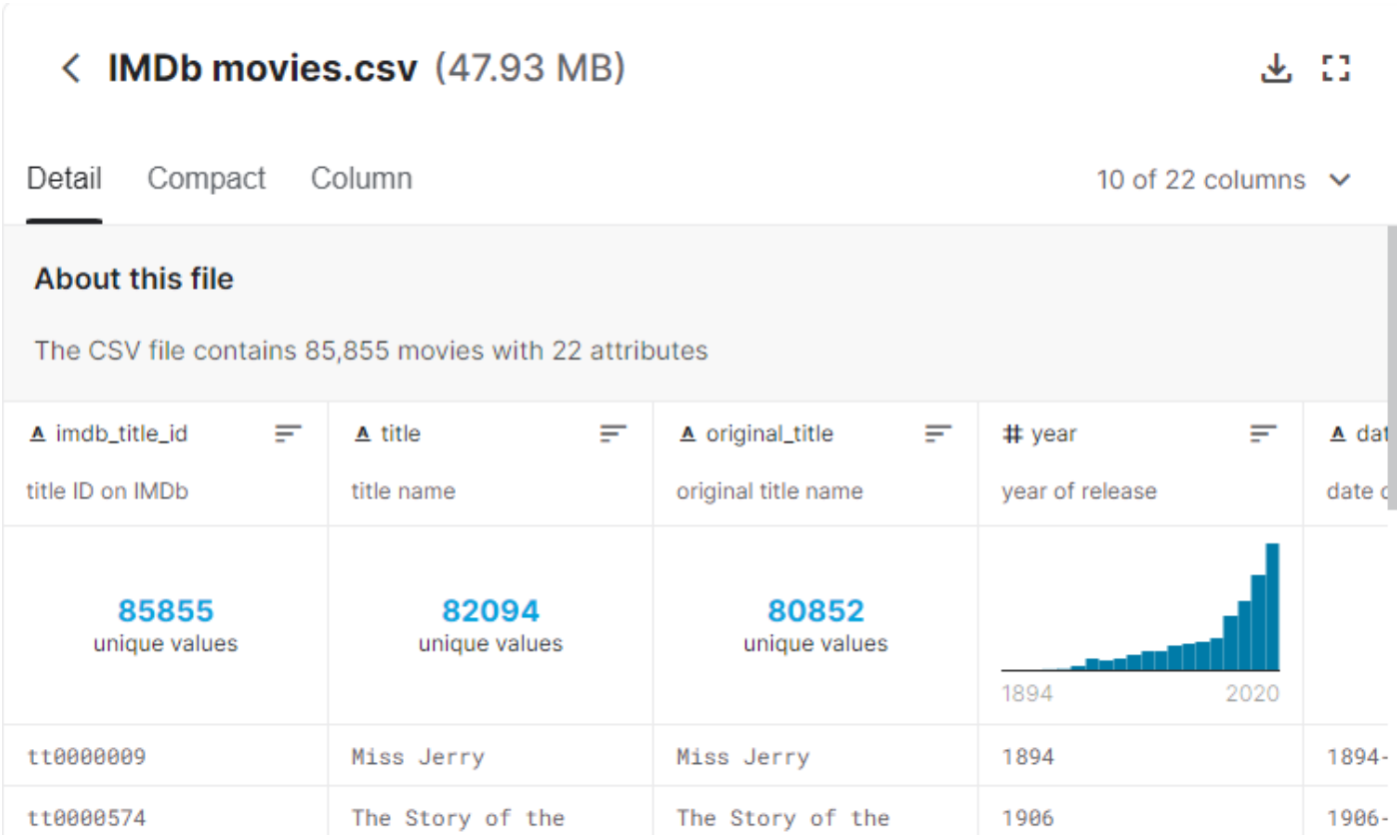
Data Preparation:

1. Gather data

The data preparation process begins with finding the right data. We downloaded the dataset from Kaggle after searching from different websites.

2. Discover and assess data

We thought of doing our project on data of covid-19 cases but as we decided that it is the most generalized topic which is already done by many websites so we shifted to the prediction of movie success. After finalizing the topic data scientist download a large dataset from Kaggle as mentioned above. After searching different websites for the dataset we found Kaggle where there are two datasets on movies but picked the most unique data with large numbers of test cases. There are imdb_title_id, title, year and so on, there are 22 columns as such and 85855 rows. We should process the data such that we should predict the movie's success.



3. Cleaning and validating data

Cleaning up the data is traditionally the most time-consuming part of the data preparation process, but it's crucial for removing faulty data and filling in gaps. Important tasks here include:

- Removing extraneous data and outliers.
- Filling in missing values.
- Conforming data to a standardized pattern.
- Masking private or sensitive data entries.

Once data has been cleansed, it must be validated by testing for errors in the data preparation process up to this point. Oftentimes, an error in the system will become apparent during this step and will need to be resolved before moving forward.

4. Transform and enrich data

Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience. *Enriching* data refers to adding and cleaning data with other related information to provide deeper insights.

Codes and their outputs:-

```
> getwd()
[1] "C:/Users/ravit/Documents"
> library(dplyr)
> library(ggplot2)
> data=read.csv("IMDb movies.csv")
> summary(data)
 imdb_title_id         title          original_title         year              genre
 Length:85855       Length:85855       Length:85855       Length:85855       Length:85855
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



    duration       country            language           director           writer
 Min.   : 41.0   Length:85855       Length:85855       Length:85855       Length:85855
 1st Qu.: 88.0   Class :character   Class :character   Class :character   Class :character
 Median : 96.0   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :100.4
 3rd Qu.:108.0
 Max.   :808.0

 production_company    actors             avg_vote        votes          reviews_from_users
 Length:85855       Length:85855       Min.   :1.000   Min.   :     99   Min.   :    1.00
 Class :character   Class :character   1st Qu.:5.200   1st Qu.:    205   1st Qu.:    4.00
 Mode  :character   Mode  :character   Median :6.100   Median :    484   Median :    9.00
                                       Mean   :5.899   Mean   :   9494   Mean   :   46.04
                                       3rd Qu.:6.800   3rd Qu.:   1766   3rd Qu.:   27.00
                                       Max.   :9.900   Max.   :2278845   Max.   :10472.00
                                                                         NA's   :7597

 reviews_from_critics
 Min.   :  1.00
 1st Qu.:  3.00
 Median :  8.00
 Mean   : 27.48
 3rd Qu.: 23.00
 Max.   :999.00
 NA's   :11797
> |
```

```
> unique_genre=unique(newgenre)
> unique_genre
 [1] "Romance"     "Biography"  "Drama"       "Adventure"   "History"    "Crime"       "Western"     "Fantasy"    "Comedy"      "Horror"       "Family"      "Action"      "Mystery"
[14] "Sci-Fi"      "Animation"  "Thriller"    "Musical"     "Music"      "War"         "Film-Noir"   "Sport"      "Adult"       "Documentary"
```

```
> head(data,5)
  imdb_title_id                      title             original_title year                      genre duration          country language                               director
1     tt0000009                 Miss Jerry                 Miss Jerry 1894                    Romance       45              USA     None                        Alexander Black
2     tt0000574 The Story of the Kelly Gang The Story of the Kelly Gang 1906    Biography, Crime, Drama       70        Australia     None                           Charles Tait
3     tt0001892             Den sorte drÃ¸m             Den sorte drÃ¸m 1911                      Drama       53 Germany, Denmark                                     Urban Gad
4     tt0002101                  Cleopatra                  Cleopatra 1912             Drama, History      100              USA  English                     Charles L. Gaskill
5     tt0002130                  L'Inferno                  L'Inferno 1911 Adventure, Drama, Fantasy       68            Italy  Italian Francesco Bertolini, Adolfo Padovan
                        writer        production_company
1               Alexander Black Alexander Black Photoplays
2                  Charles Tait          J. and N. Tait
3 Urban Gad, Gebhard SchÃ¤tzler-Perasini        Fotorama
4             Victorien Sardou Helen Gardner Picture Players
5               Dante Alighieri               Milano Film
                                                                                                                                                         actors
1                                                                                                              Blanche Bayliss, William Courtenay, Chauncey Depew
2           Elizabeth Tait, John Tait, Norman Campbell, Bella Cola, Will Coyne, Sam Crewes, Jack Ennis, John Forde, Vera Linden, Mr. Marshall, Mr. McKenzie, Frank Mills, Ollie Wilson
3                                                                                  Asta Nielsen, Valdemar Psilander, Gunnar Helsengreen, Emil Albes, Hugo Flink, Mary Hagen
4 Helen Gardner, Pearl Sindelar, Miss Fielding, Miss Robson, Helene Costello, Charles Sindelar, Mr. Howard, James R. Waite, Mr. Osborne, Harry Knowles, Mr. Paul, Mr. Brady, Mr. Corker
5                                                         Salvatore Papa, Arturo Pirovano, Giuseppe de Liguoro, Pier Delle Vigne, Augusto Milla, Attilio Motta, Emilise Beretta
  avg_vote votes reviews_from_users reviews_from_critics
1      5.9   154                  1                    2
2      6.1   589                  7                    7
3      5.8   188                  5                    2
4      5.2   446                 25                    3
5      7.0  2237                 31                   14
```

```
> newgenre=data$genre
> size=length(newgenre)
> for (i in 1:size)
+ {
+   str1=""
+   str2=""
+   str2=newgenre[i]
+   s=nchar(str2)
+   string_split = strsplit(str2, "")[[1]]
+   for (x in string_split)
+   {
+     if(x ==',')
+     {
+       break;
+     }
+     str1=paste(str1, x, sep ="")
+   }
+   newgenre[i]=str1
+ }
> newgenre
  [1] "Romance"   "Biography" "Drama"     "Drama"     "Adventure" "Biography" "Biography" "Drama"     "History"   "Drama"     "Drama"     "Crime"     "Drama"     "Crime"     "Drama"
 [16] "Drama"     "Crime"     "Drama"     "Drama"     "Crime"     "Adventure" "Drama"     "Crime"     "Western"   "Adventure" "Fantasy"   "Crime"     "Crime"     "Comedy"    "Horror"
 [31] "Drama"     "Family"    "Drama"     "Drama"     "Drama"     "Adventure" "Action"    "Action"    "Drama"     "Comedy"    "Drama"     "Drama"     "Western"   "Comedy"    "Drama"
 [46] "Crime"     "Adventure" "Drama"     "Drama"     "Drama"     "Drama"     "Drama"     "Drama"     "Crime"     "Drama"     "Adventure" "Drama"     "Drama"     "Drama"     "Drama"
 [61] "Biography" "Action"    "Comedy"    "Action"    "Drama"     "Drama"     "Drama"     "Comedy"    "Adventure" "Crime"     "Romance"   "Crime"     "Western"   "Romance"   "Comedy"
 [76] "Comedy"    "Drama"     "Adventure" "Comedy"    "Drama"     "Action"    "Action"    "Comedy"    "Mystery"   "Drama"     "Fantasy"   "Comedy"    "Drama"     "Drama"     "Drama"
 [91] "Western"   "Biography" "Horror"    "Comedy"    "Adventure" "Adventure" "Drama"     "Drama"     "Drama"     "Adventure" "Biography" "Comedy"    "Adventure" "Comedy"    "Adventure"
[106] "Drama"     "Western"   "Drama"     "Drama"     "Comedy"    "Comedy"    "Adventure" "Drama"     "Comedy"    "Comedy"    "Drama"     "Drama"     "Drama"     "Family"    "Romance"
[121] "Drama"     "Drama"     "Comedy"    "Drama"     "Drama"     "Western"   "Comedy"    "Comedy"    "Drama"     "Comedy"    "Drama"     "Comedy"    "Drama"     "Drama"     "Comedy"
[136] "Drama"     "Action"    "Adventure" "Drama"     "Drama"     "Crime"     "Drama"     "Comedy"    "Action"    "Adventure" "Drama"     "Drama"     "Drama"     "Drama"     "Comedy"
[151] "Comedy"    "Comedy"    "Drama"     "Drama"     "Comedy"    "Drama"     "Adventure" "Drama"     "Adventure" "Drama"     "Drama"     "Comedy"    "Comedy"    "Drama"     "Drama"
[166] "Fantasy"   "Adventure" "Comedy"    "Drama"     "Drama"     "Drama"     "Drama"     "Comedy"    "Comedy"    "Adventure" "Drama"     "Adventure" "Drama"     "Comedy"    "Fantasy"
[181] "Drama"     "Western"   "Action"    "Drama"     "Comedy"    "Fantasy"   "Biography" "Drama"     "Adventure" "Drama"     "Drama"     "Comedy"    "Crime"     "Comedy"    "Drama"
[196] "Horror"    "Fantasy"   "Biography" "Drama"     "Drama"     "Drama"     "Drama"     "Comedy"    "Drama"     "Drama"     "Comedy"    "Action"    "Comedy"    "Drama"     "Adventure"
[211] "Drama"     "Adventure" "Drama"     "Family"    "Comedy"    "Drama"     "Comedy"    "Comedy"    "Comedy"    "Drama"     "Drama"     "Adventure" "Comedy"    "Adventure" "Mystery"
[226] "Adventure" "Drama"     "Romance"   "Drama"     "Comedy"    "Drama"     "Crime"     "Action"    "Comedy"    "Comedy"    "Comedy"    "Drama"     "Adventure" "Drama"     "Drama"
[241] "Comedy"    "Comedy"    "Drama"     "Comedy"    "Drama"     "Drama"     "Drama"     "Drama"     "Adventure" "Adventure" "Drama"     "Comedy"    "Drama"     "Comedy"    "Drama"
[256] "Drama"     "Comedy"    "Comedy"    "Drama"     "Action"    "Drama"     "Crime"     "Drama"     "Comedy"    "Drama"     "Drama"     "Crime"     "Adventure" "Drama"     "Drama"
[271] "Action"    "Comedy"    "Drama"     "Drama"     "Drama"     "Drama"     "Drama"     "Drama"     "Drama"     "Romance"   "Drama"     "Drama"     "Comedy"    "Adventure" "Drama"
[286] "Drama"     "Drama"     "Drama"     "Comedy"    "Comedy"    "Drama"     "Mystery"   "Drama"     "Drama"     "Drama"     "Adventure" "Drama"     "Fantasy"   "Drama"     "Drama"
[301] "Drama"     "Adventure" "Adventure" "Biography" "Comedy"    "Drama"     "Drama"     "Drama"     "Drama"     "Action"    "Drama"     "Drama"     "Action"    "Drama"     "Drama"
[316] "Drama"     "Drama"     "Romance"   "Romance"   "Drama"     "Drama"     "Drama"     "Adventure" "Comedy"    "Drama"     "Drama"     "Adventure" "Drama"     "Comedy"    "Comedy"
[331] "Comedy"    "Drama"     "Romance"   "Drama"     "Action"    "Drama"     "Drama"     "Drama"     "Comedy"    "Drama"     "Drama"     "Biography" "Comedy"    "Fantasy"   "Drama"
[346] "Drama"     "Crime"     "Adventure" "Drama"     "Romance"   "Adventure" "Drama"     "Drama"     "Drama"     "Drama"     "Adventure" "Drama"     "Drama"     "Drama"     "Drama"
[361] "Drama"     "Comedy"    "Comedy"    "Comedy"    "Drama"     "Comedy"    "Comedy"    "Adventure" "History"   "Drama"     "Drama"     "Drama"     "Drama"     "Drama"     "Comedy"
[376] "Drama"     "Drama"     "Action"    "Comedy"    "Adventure" "Adventure" "Comedy"    "Crime"     "Comedy"    "Sci-Fi"    "Adventure" "Drama"     "Drama"     "Adventure" "Drama"
[391] "Action"    "Drama"     "Adventure" "Drama"     "Drama"     "Comedy"    "Romance"   "Animation" "Drama"     "Drama"     "Western"   "Drama"     "Drama"     "Adventure" "Drama"
[406] "Adventure" "Drama"     "Action"    "Drama"     "Drama"     "Comedy"    "Drama"     "Drama"     "Drama"     "Comedy"    "Adventure" "Drama"     "Drama"     "Drama"     "Comedy"
[421] "Biography" "Comedy"    "Drama"     "Drama"     "Action"    "Drama"     "Drama"     "Comedy"    "Adventure" "Drama"     "Comedy"    "Drama"     "Action"    "Crime"     "Comedy"
[436] "Horror"    "Comedy"    "Drama"     "Drama"     "Drama"     "Comedy"    "Drama"     "Biography" "Action"    "Drama"     "Comedy"    "Comedy"    "Drama"     "Comedy"    "Adventure"
[451] "Drama"     "Drama"     "Comedy"    "Drama"     "Western"   "Crime"     "Western"   "Crime"     "Drama"     "Comedy"    "Comedy"    "Adventure" "Comedy"    "Comedy"    "Action"
[466] "Drama"     "Drama"     "Mystery"   "Action"    "Action"    "Crime"     "Comedy"    "Comedy"    "Adventure" "Crime"     "Action"    "Drama"     "Comedy"    "Action"    "Drama"
[481] "Drama"     "Adventure" "Comedy"    "Comedy"    "Comedy"    "Drama"     "Comedy"    "Comedy"    "Drama"     "Action"    "Drama"     "Western"   "Action"    "Drama"     "Comedy"
[496] "Adventure" "Comedy"    "Comedy"    "Drama"     "Crime"     "Comedy"    "Action"    "Drama"     "Comedy"    "Drama"     "Drama"     "Drama"     "Adventure" "Comedy"    "Drama"
[511] "Drama"     "Biography" "Drama"     "Drama"     "Comedy"    "Drama"     "Comedy"    "Drama"     "Action"    "Romance"   "Drama"     "Comedy"    "Comedy"    "Comedy"    "Adventure"
[526] "Drama"     "Drama"     "Comedy"    "Drama"     "Drama"     "Drama"     "Romance"   "Drama"     "Adventure" "Drama"     "Drama"     "Comedy"    "Drama"     "Drama"     "Drama"
[541] "Drama"     "Drama"     "Drama"     "Adventure" "Drama"     "Drama"     "Biography" "Comedy"    "Comedy"    "Drama"     "Comedy"    "Comedy"    "Adventure" "Drama"     "Romance"
```

```
> unique_country=unique(newcountry)
> unique_country
  [1] "USA"                      "Australia"                      "Germany"                  "Italy"                              "Romania"
  [6] "France"                   "Denmark"                        "Sweden"                   "Belgium"                            "Hungary"
 [11] "Russia"                   "Mexico"                         "Canada"                   "Norway"                             "Japan"
 [16] "UK"                       "Soviet Union"                   "Austria"                  "Chile"                              "India"
 [21] "Switzerland"              "China"                          "Spain"                    "Czechoslovakia"                     "Brazil"
 [26] "Portugal"                 "Turkey"                         "Poland"                   "Netherlands"                        "Finland"
 [31] "Argentina"                "Yugoslavia"                     "Greece"                   "East Germany"                       "Egypt"
 [36] "West Germany"             "Albania"                        "Israel"                   "Cuba"                               "Ireland"
 [41] "Philippines"              "South Korea"                    "Hong Kong"                "Puerto Rico"                        "Bulgaria"
 [46] "Algeria"                  "Lebanon"                        "Sri Lanka"                "South Africa"                       "Taiwan"
 [51] "Senegal"                  "Bolivia"                        "Liechtenstein"            "Iran"                               "Croatia"
 [56] "Peru"                     "Syria"                          "Angola"                   "Jamaica"                            "Ethiopia"
 [61] "Indonesia"                "CÃ´te d'Ivoire"                 "Suriname"                 "Venezuela"                          "New Zealand"
 [66] "Mali"                     "Vietnam"                        "Libya"                    "Iceland"                            "Nicaragua"
 [71] "Zambia"                   "Burkina Faso"                   "Bahamas"                  "Colombia"                           "Iraq"
 [76] "Tunisia"                  "North Korea"                    "Gibraltar"                "Ukraine"                            "Armenia"
 [81] "Latvia"                   "Federal Republic of Yugoslavia" "Slovenia"                 "Kazakhstan"                         "Estonia"
 [86] "Slovakia"                 "Czech Republic"                 "Bangladesh"               "Tajikistan"                         "Cambodia"
 [91] "Georgia"                  "Republic of North Macedonia"    "Guatemala"                "Singapore"                          "Dominican Republic"
 [96] "Bosnia and Herzegovina"   "Lithuania"                      "Guinea"                   "Thailand"                           "Aruba"
[101] "Kuwait"                   "Greenland"                      "Kyrgyzstan"               "Luxembourg"                         "Bhutan"
[106] "Ecuador"                  "Pakistan"                       "Mozambique"               "Morocco"                            "Serbia"
[111] "Uruguay"                  "Moldova"                        "Malta"                    "Nepal"                              "Palestine"
[116] "Malaysia"                 "Costa Rica"                     "North Vietnam"            "Afghanistan"                        "Isle Of Man"
[121] "Serbia and Montenegro"    "Nigeria"                        "Kosovo"                   "Saudi Arabia"                       "Chad"
[126] "Jordan"                   "Azerbaijan"                     "Rwanda"                   "Uzbekistan"                         "Uganda"
[131] ""                         "Lesotho"                        "Mongolia"                 "United Arab Emirates"               "Montenegro"
[136] "Mauritius"                "Panama"                         "Belarus"                  "Cyprus"                             "Papua New Guinea"
[141] "Honduras"                 "Kenya"                          "Qatar"                    "The Democratic Republic Of Congo"   "Korea"
[146] "Trinidad and Tobago"      "Paraguay"                       "Netherlands Antilles"     "Yemen"                              "Tanzania"
[151] "Haiti"                    "Andorra"                        "Brunei"                   "Mauritania"                         "Ghana"
[156] "Myanmar"                  "Laos"                           "Bermuda"                  "Somalia"                            "Oman"
[161] "Zimbabwe"                 "Sudan"
> data$country=newcountry
```

```
> newcountry=data$country
> size=length(newcountry)
> for (i in 1:size)
+ {
+   str1=""
+   str2=""
+   str2=newcountry[i]
+   s=nchar(str2)
+   string_split = strsplit(str2, "")[[1]]
+   for (x in string_split)
+   {
+     if(x ==',')
+     {
+       break;
+     }
+     str1=paste(str1, x, sep ="")
+   }
+   newcountry[i]=str1
+ }
> newcountry
```

```
  [1] "USA"       "Australia" "Germany"   "USA"       "Italy"     "USA"       "Germany"   "Italy"     "Romania"   "France"    "Denmark"
 [12] "France"    "Sweden"    "France"    "Italy"     "Belgium"   "France"    "USA"       "Germany"   "USA"       "Italy"     "Italy"
 [23] "USA"       "USA"       "Italy"     "USA"       "France"    "France"    "USA"       "Germany"   "Denmark"   "USA"       "USA"
 [34] "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "Hungary"   "USA"       "USA"       "USA"
 [45] "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "Denmark"   "USA"
 [56] "Italy"     "USA"       "USA"       "Russia"    "Italy"     "USA"       "France"    "USA"       "USA"       "USA"       "USA"
 [67] "USA"       "Sweden"    "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"
 [78] "France"    "USA"       "Russia"    "USA"       "USA"       "Germany"   "USA"       "USA"       "USA"       "USA"       "Sweden"
 [89] "USA"       "USA"       "USA"       "USA"       "Germany"   "USA"       "Denmark"   "USA"       "USA"       "USA"       "France"
[100] "USA"       "Russia"    "USA"       "USA"       "USA"       "USA"       "Russia"    "USA"       "USA"       "Sweden"    "Sweden"
[111] "USA"       "USA"       "Sweden"    "USA"       "USA"       "Hungary"   "Germany"   "Sweden"    "USA"       "USA"       "Germany"
[122] "France"    "Germany"   "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"
[133] "USA"       "USA"       "USA"       "USA"       "USA"       "France"    "USA"       "USA"       "Germany"   "USA"       "Germany"
[144] "Mexico"    "Canada"    "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"
[155] "USA"       "USA"       "USA"       "Germany"   "USA"       "USA"       "Sweden"    "USA"       "Germany"   "Sweden"    "France"
[166] "Germany"   "USA"       "Germany"   "USA"       "USA"       "Germany"   "Denmark"   "Germany"   "USA"       "USA"       "Australia"
[177] "Germany"   "Sweden"    "USA"       "Germany"   "USA"       "USA"       "USA"       "USA"       "USA"       "Germany"   "Germany"
[188] "USA"       "France"    "Denmark"   "USA"       "Sweden"    "Norway"    "USA"       "Germany"   "Germany"   "Germany"   "USA"
[199] "France"    "France"    "USA"       "USA"       "USA"       "Germany"   "Sweden"    "Germany"   "USA"       "USA"       "USA"
[210] "USA"       "Norway"    "USA"       "USA"       "USA"       "Sweden"    "USA"       "Germany"   "USA"       "USA"       "USA"
[221] "USA"       "Germany"   "USA"       "Germany"   "USA"       "USA"       "Germany"   "Germany"   "USA"       "USA"       "USA"
[232] "USA"       "USA"       "USA"       "USA"       "Germany"   "USA"       "USA"       "USA"       "USA"       "USA"       "USA"
[243] "France"    "USA"       "USA"       "USA"       "Germany"   "Germany"   "Germany"   "Germany"   "Sweden"    "USA"       "Sweden"
[254] "USA"       "USA"       "USA"       "USA"       "USA"       "Germany"   "USA"       "USA"       "USA"       "Japan"     "USA"
[265] "Germany"   "Germany"   "Germany"   "USA"       "USA"       "France"    "USA"       "USA"       "USA"       "USA"       "Germany"
[276] "USA"       "USA"       "USA"       "Germany"   "USA"       "France"    "Denmark"   "USA"       "USA"       "France"    "USA"
[287] "USA"       "Germany"   "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "Germany"
[298] "Germany"   "USA"       "Germany"   "Germany"   "USA"       "USA"       "USA"       "USA"       "Germany"   "USA"       "USA"
[309] "USA"       "USA"       "Germany"   "USA"       "USA"       "USA"       "USA"       "UK"        "Germany"   "USA"       "USA"
[320] "Germany"   "USA"       "France"    "USA"       "USA"       "Sweden"    "Sweden"    "USA"       "USA"       "USA"       "USA"
[331] "USA"       "Germany"   "USA"       "France"    "USA"       "USA"       "Germany"   "USA"       "USA"       "Germany"   "Germany"
[342] "USA"       "USA"       "Germany"   "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "Soviet Union" "USA"
[353] "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "France"    "Germany"   "USA"
[364] "USA"       "USA"       "USA"       "USA"       "France"    "USA"       "USA"       "USA"       "France"    "Germany"   "USA"
[375] "USA"       "Germany"   "USA"       "USA"       "Soviet Union" "Germany" "Germany"   "USA"       "Austria"   "Soviet Union" "France"
[386] "USA"       "USA"       "USA"       "USA"       "USA"       "USA"       "Soviet Union" "USA"    "USA"       "UK"        "USA"
[397] "USA"       "Germany"   "USA"       "USA"       "USA"       "Soviet Union" "USA"    "USA"       "USA"       "USA"       "Denmark"
```

Data Exploration :

```
> unique_country=unique(newcountry)
> unique_country
```

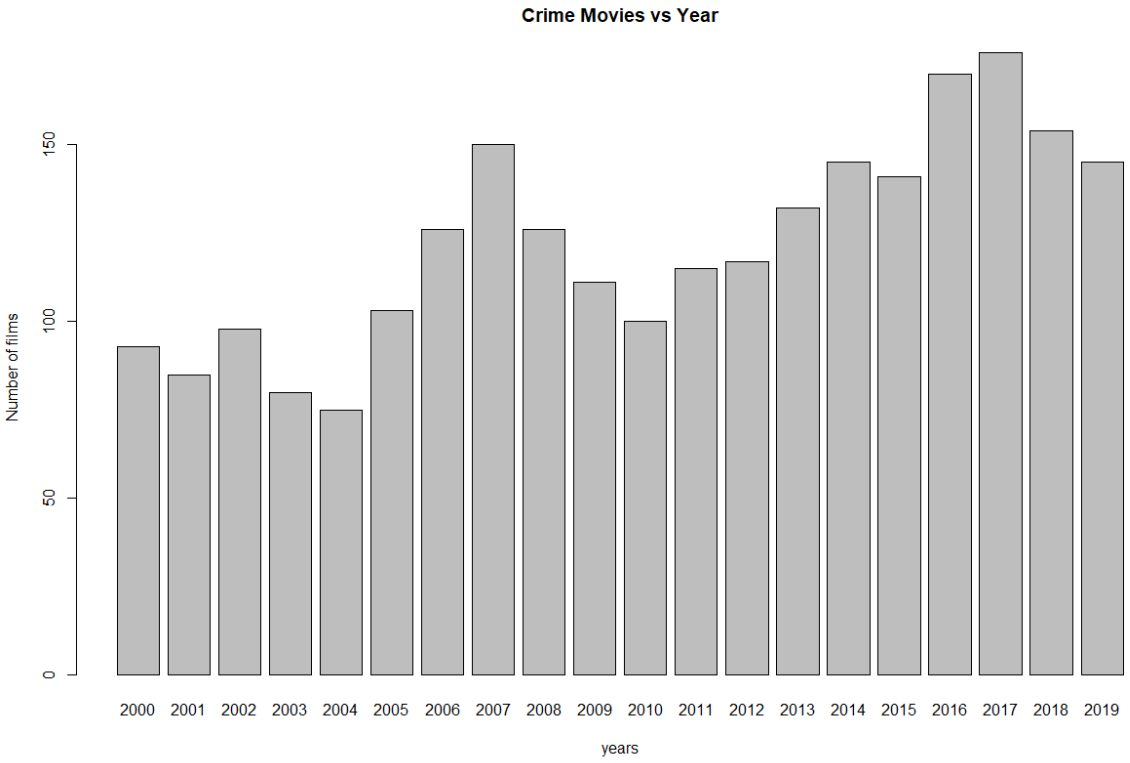
```
  [1] "USA"                       "Australia"                 "Germany"                   "Italy"                     "Romania"
  [6] "France"                    "Denmark"                   "Sweden"                    "Belgium"                   "Hungary"
 [11] "Russia"                    "Mexico"                    "Canada"                    "Norway"                    "Japan"
 [16] "UK"                        "Soviet Union"              "Austria"                   "Chile"                     "India"
 [21] "Switzerland"               "China"                     "Spain"                     "Czechoslovakia"            "Brazil"
 [26] "Portugal"                  "Turkey"                    "Poland"                    "Netherlands"               "Finland"
 [31] "Argentina"                 "Yugoslavia"                "Greece"                    "East Germany"              "Egypt"
 [36] "West Germany"              "Albania"                   "Israel"                    "Cuba"                      "Ireland"
 [41] "Philippines"               "South Korea"               "Hong Kong"                 "Puerto Rico"               "Bulgaria"
 [46] "Algeria"                   "Lebanon"                   "Sri Lanka"                 "South Africa"              "Taiwan"
 [51] "Senegal"                   "Bolivia"                   "Liechtenstein"             "Iran"                      "Croatia"
 [56] "Peru"                      "Syria"                     "Angola"                    "Jamaica"                   "Ethiopia"
 [61] "Indonesia"                 "CÃ´te d'Ivoire"            "Suriname"                  "Venezuela"                 "New Zealand"
 [66] "Mali"                      "Vietnam"                   "Libya"                     "Iceland"                   "Nicaragua"
 [71] "Zambia"                    "Burkina Faso"              "Bahamas"                   "Colombia"                  "Iraq"
 [76] "Tunisia"                   "North Korea"               "Gibraltar"                 "Ukraine"                   "Armenia"
 [81] "Latvia"                    "Federal Republic of Yugoslavia" "Slovenia"             "Kazakhstan"                "Estonia"
 [86] "Slovakia"                  "Czech Republic"            "Bangladesh"                "Tajikistan"                "Cambodia"
 [91] "Georgia"                   "Republic of North Macedonia" "Guatemala"               "Singapore"                 "Dominican Republic"
 [96] "Bosnia and Herzegovina"    "Lithuania"                 "Guinea"                    "Thailand"                  "Aruba"
[101] "Kuwait"                    "Greenland"                 "Kyrgyzstan"                "Luxembourg"                "Bhutan"
[106] "Ecuador"                   "Pakistan"                  "Mozambique"                "Morocco"                   "Serbia"
[111] "Uruguay"                   "Moldova"                   "Malta"                     "Nepal"                     "Palestine"
[116] "Malaysia"                  "Costa Rica"                "North Vietnam"             "Afghanistan"               "Isle of Man"
[121] "Serbia and Montenegro"     "Nigeria"                   "Kosovo"                    "Saudi Arabia"              "Chad"
[126] "Jordan"                    "Azerbaijan"                "Rwanda"                    "Uzbekistan"                "Uganda"
[131] ""                          "Lesotho"                   "Mongolia"                  "United Arab Emirates"      "Montenegro"
[136] "Mauritius"                 "Panama"                    "Belarus"                   "Cyprus"                    "Papua New Guinea"
[141] "Honduras"                  "Kenya"                     "Qatar"                     "The Democratic Republic of Congo" "Korea"
[146] "Trinidad and Tobago"       "Paraguay"                  "Netherlands Antilles"      "Yemen"                     "Tanzania"
[151] "Haiti"                     "Andorra"                   "Brunei"                    "Mauritania"                "Ghana"
[156] "Myanmar"                   "Laos"                      "Bermuda"                   "Somalia"                   "Oman"
[161] "Zimbabwe"                  "Sudan"
```

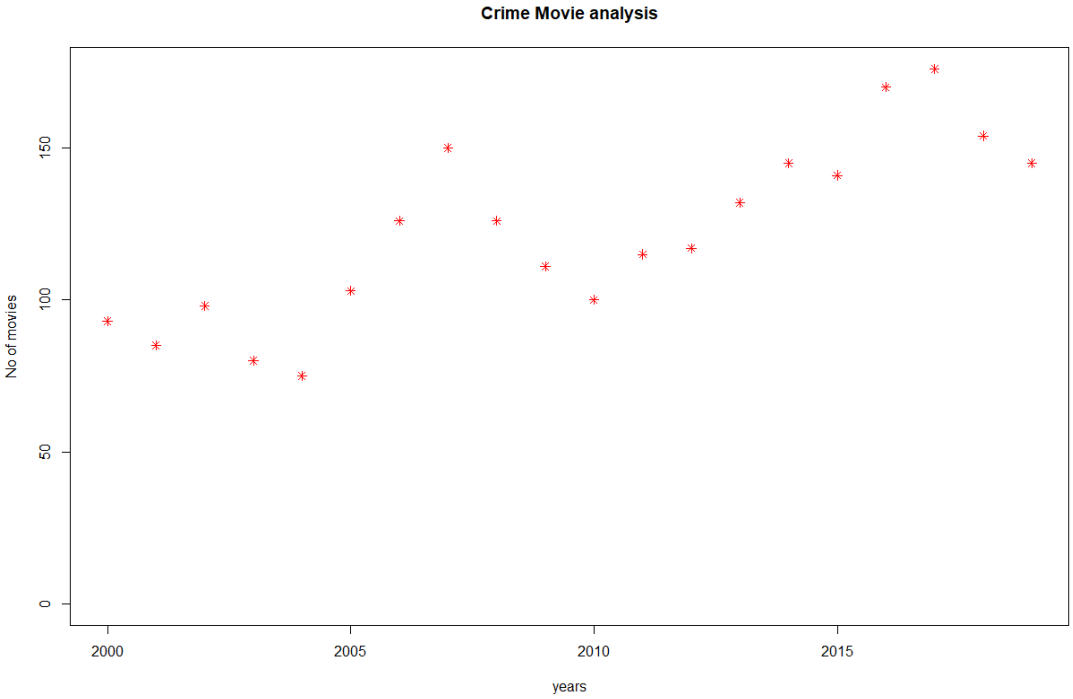
```
> data$country=newcountry
> data=select(data, year, genre, country, avg_vote)
> summary(data)
     year               genre              country            avg_vote
 Length:85855       Length:85855       Length:85855       Min.   :1.000
 Class :character   Class :character   Class :character   1st Qu.:5.200
 Mode  :character   Mode  :character   Mode  :character   Median :6.100
                                                          Mean   :5.899
                                                          3rd Qu.:6.800
                                                          Max.   :9.900

> sum(is.na(data))
[1] 0
```

Data Visualisation :

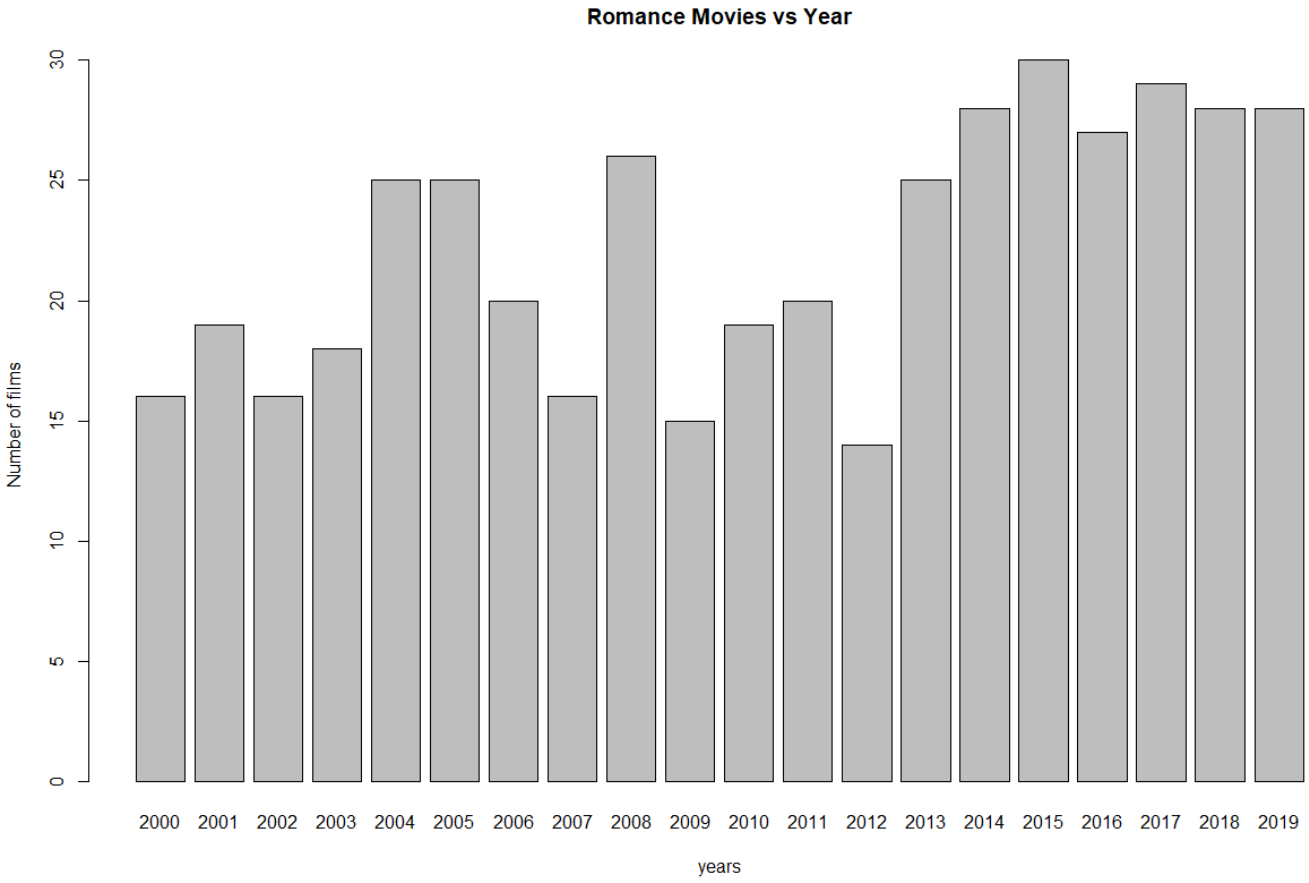
Data visualization itself means representing data in a process of translating large datasets and metrics into charts, graphs, and other visuals. From the general cases, we have taken the analysis of the genre based on the number of movies released per year and by this, we can easily come to a conclusion that this particular genre has been growing instantly and they have a higher success rate compared to the other genres.

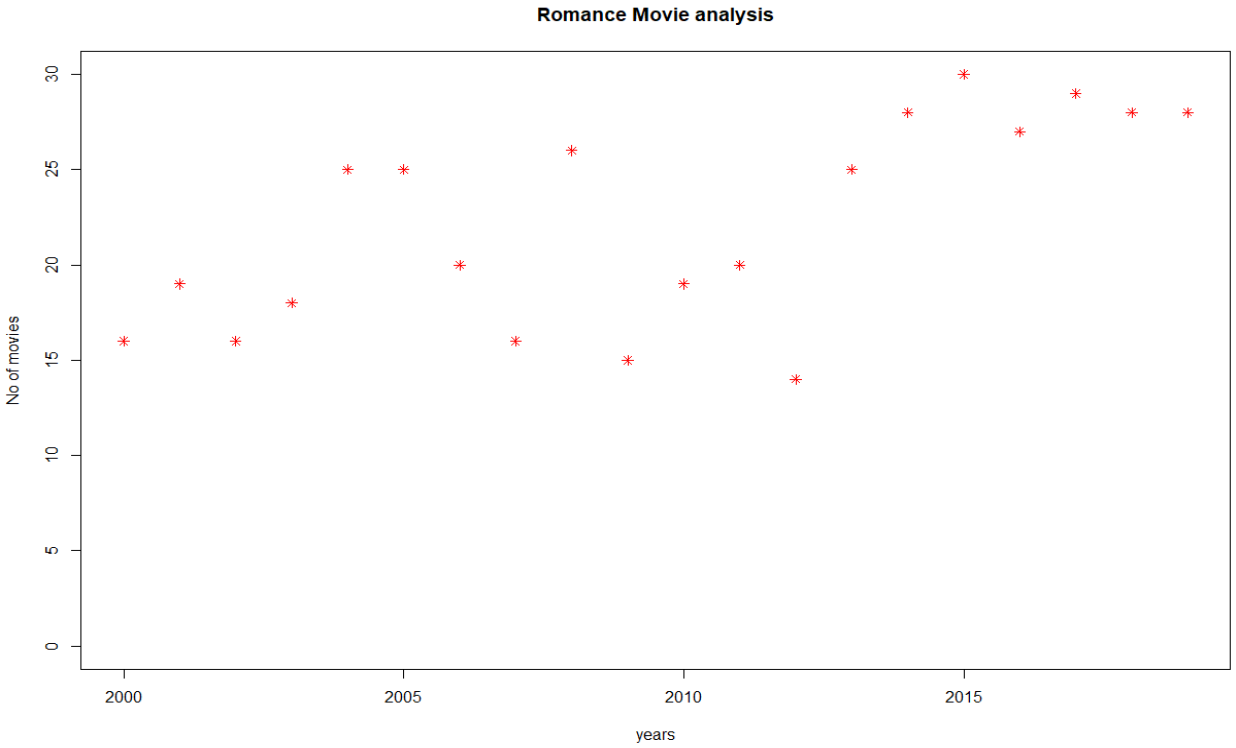
Case 1: Crime



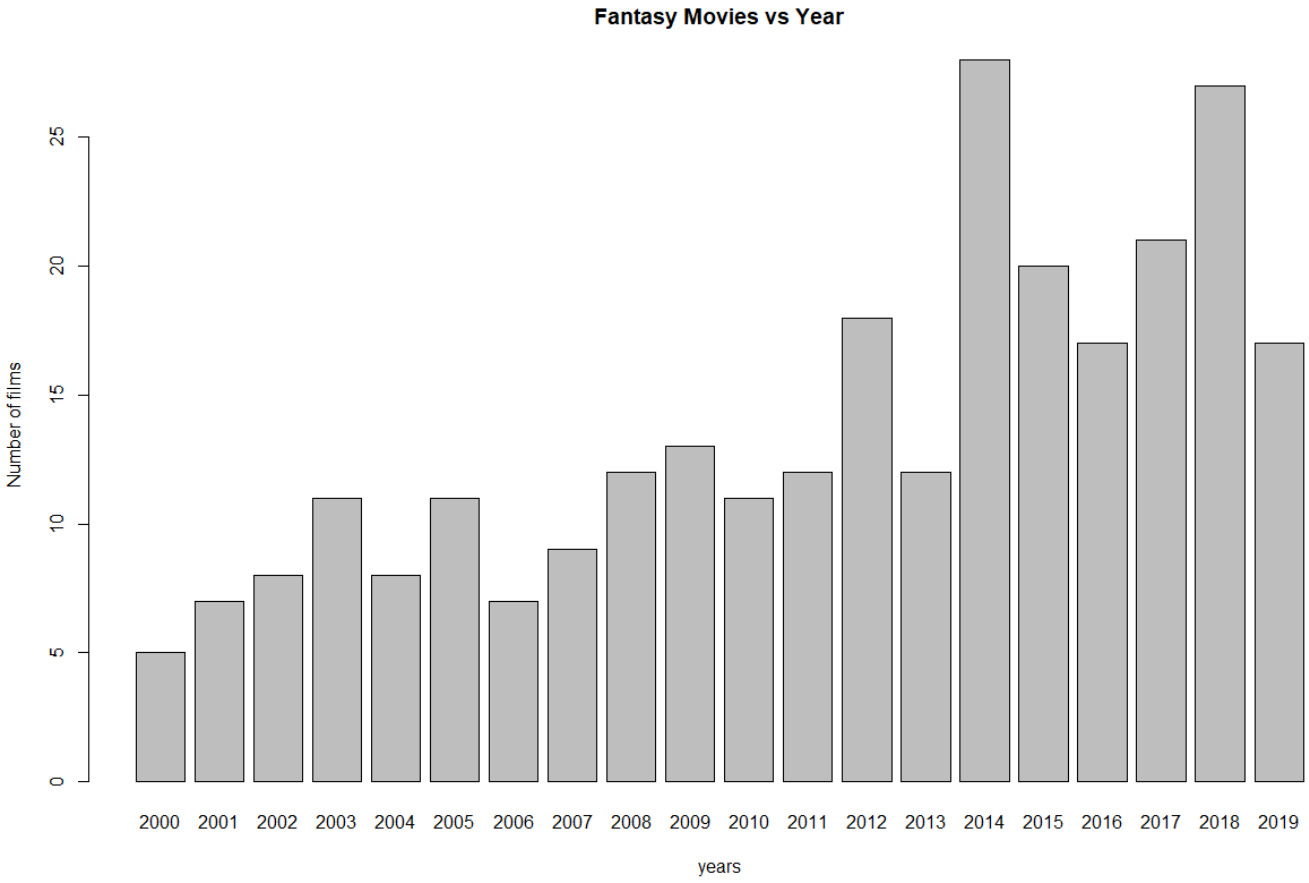


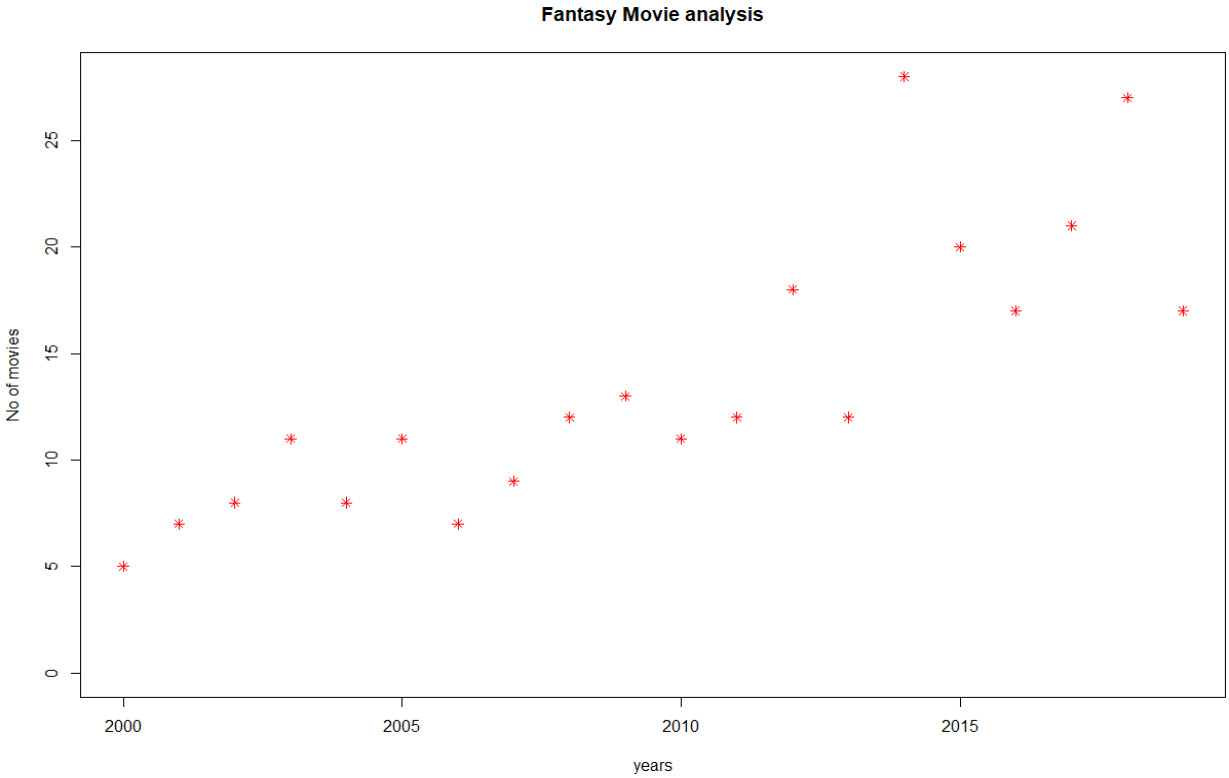
Case 2: Romance



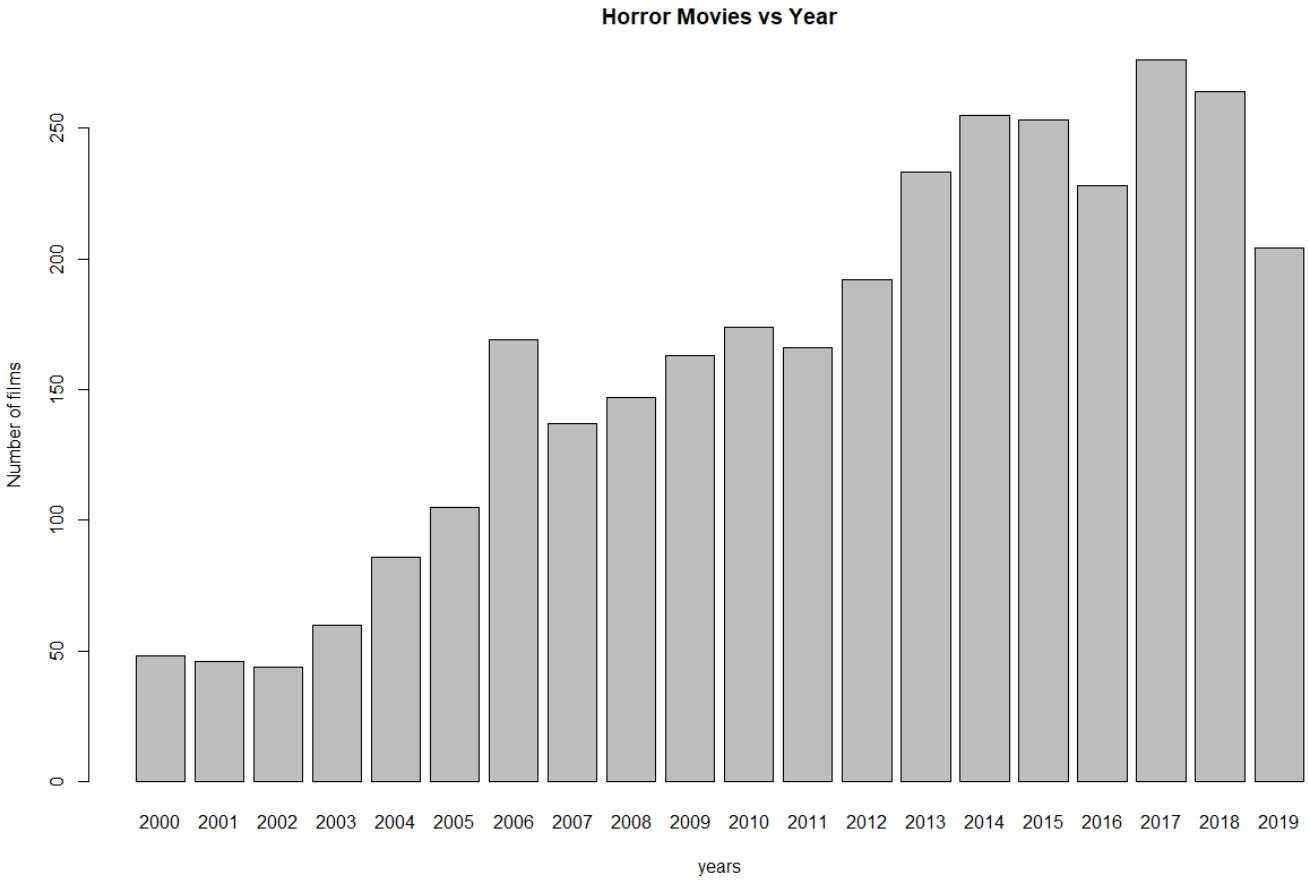


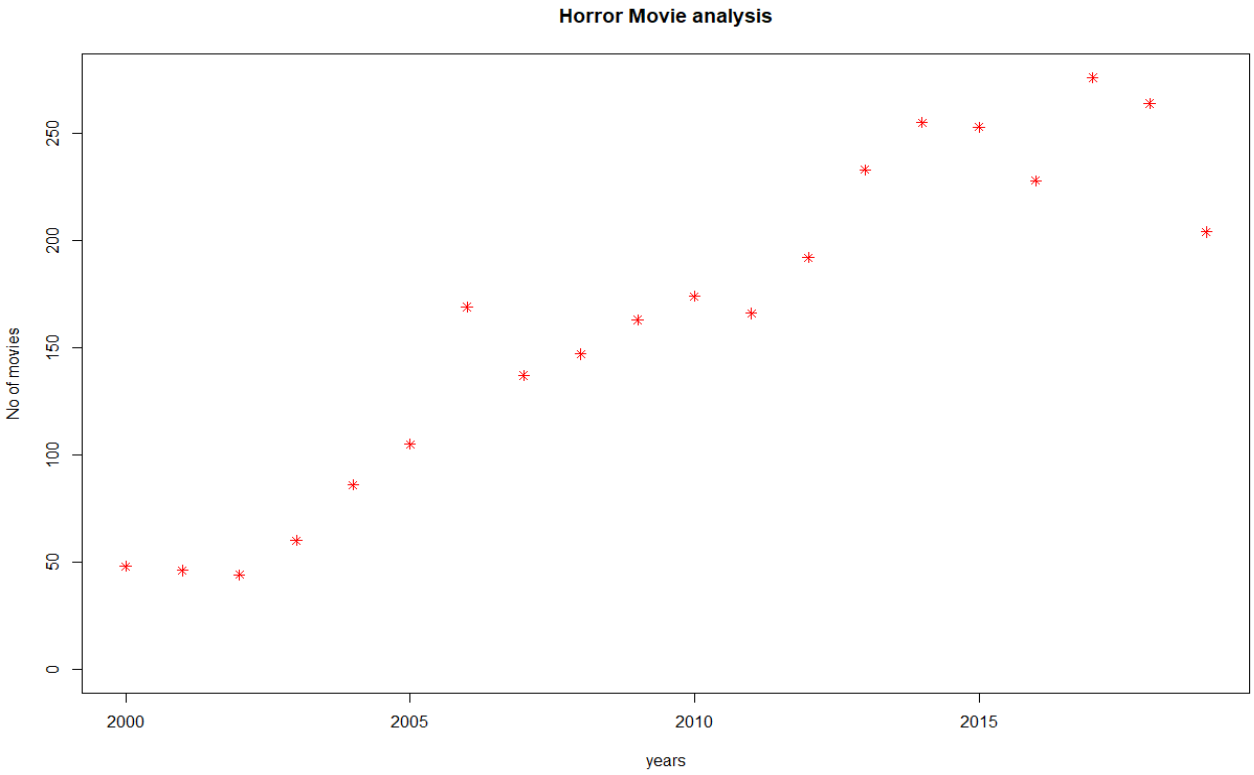
Case 3: Fantasy



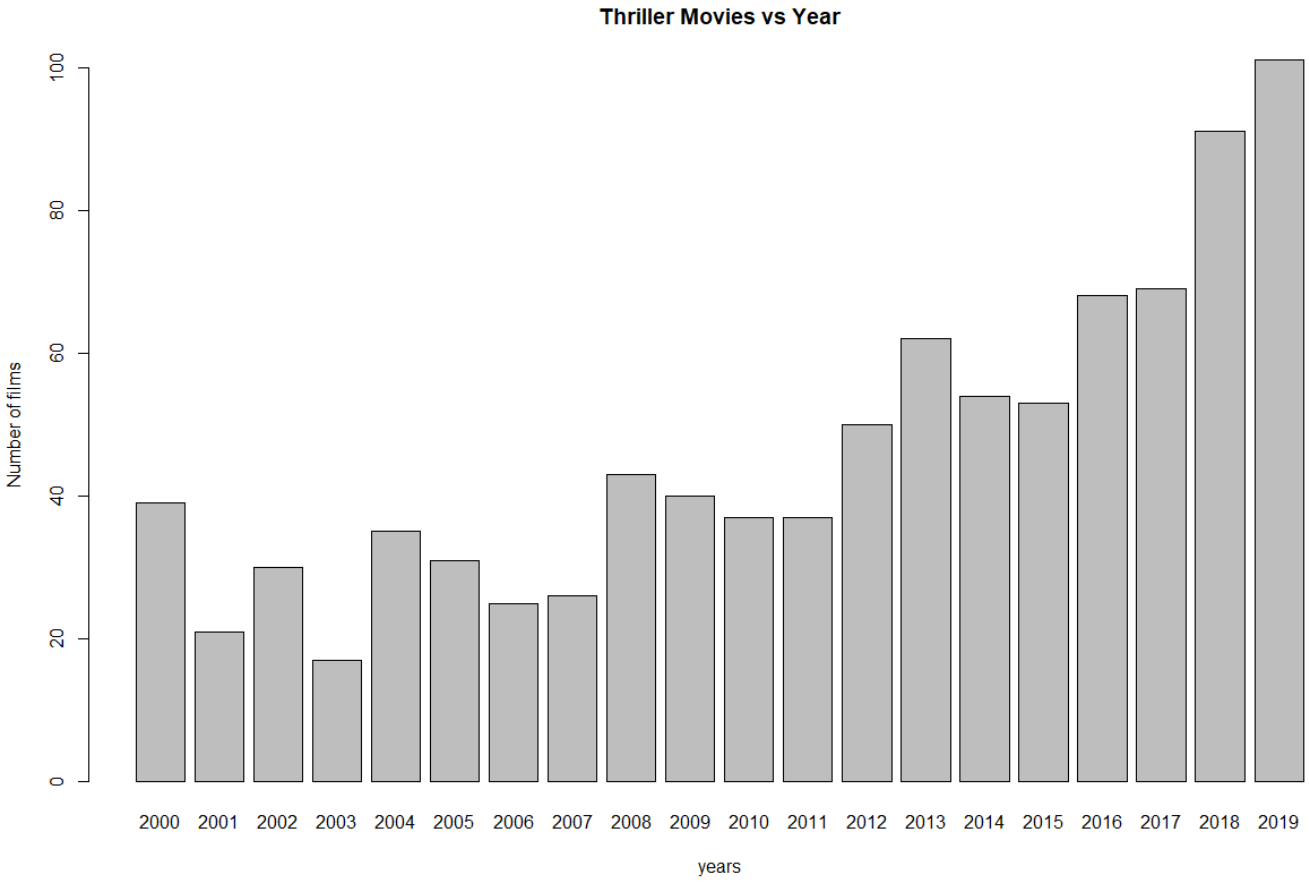


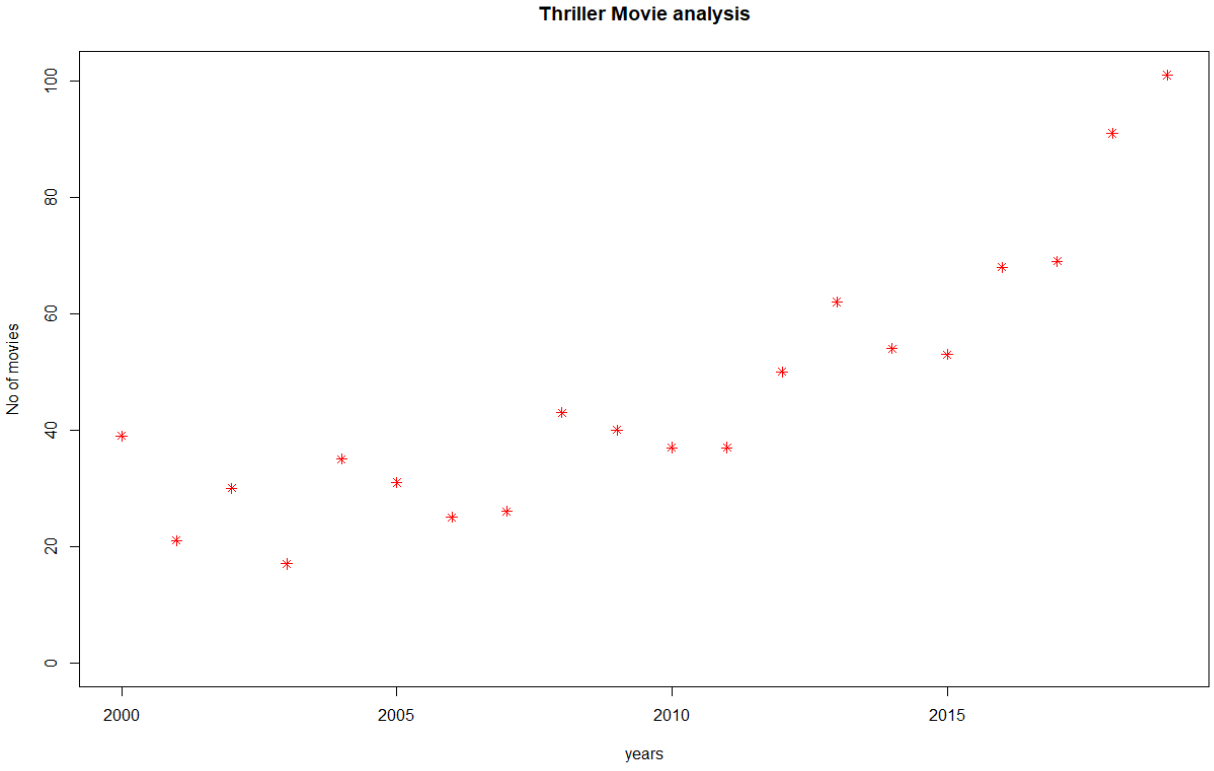
Case 4: Horror



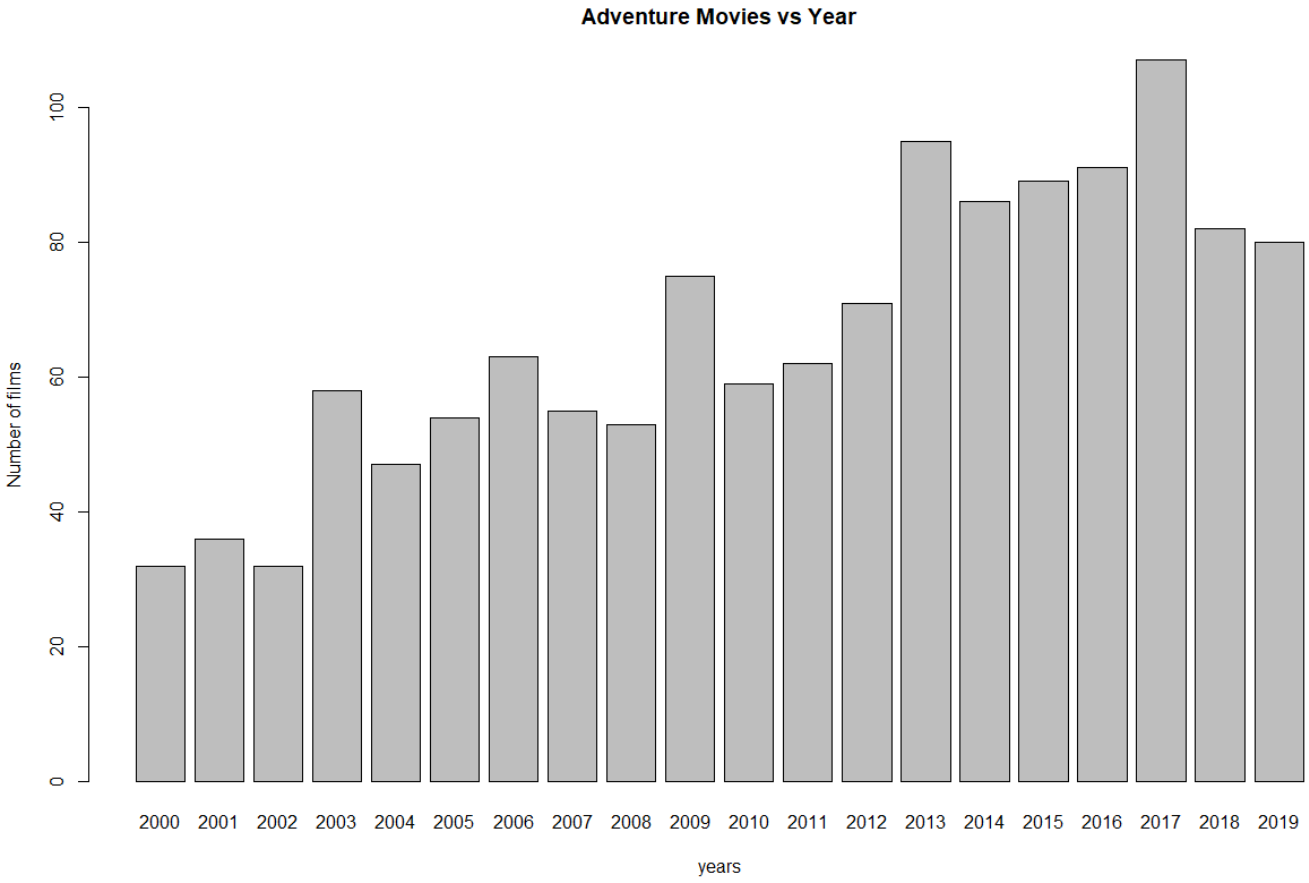


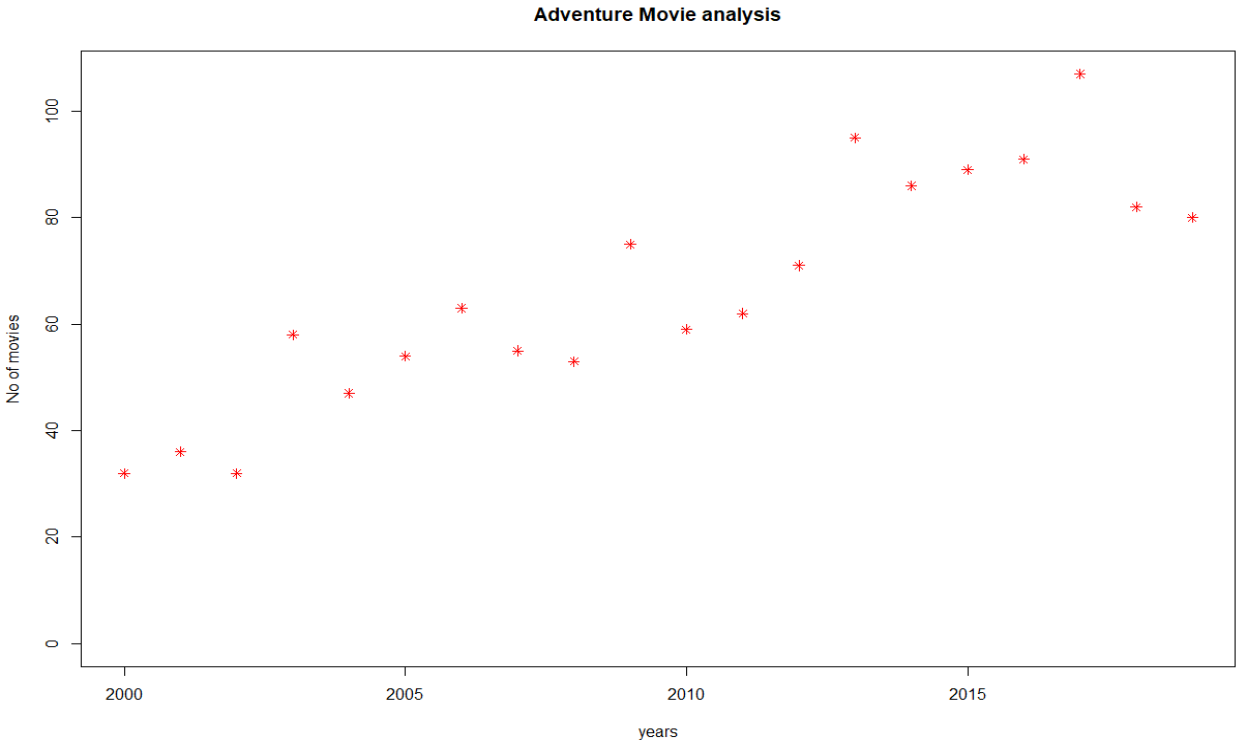
Case 5: Thriller



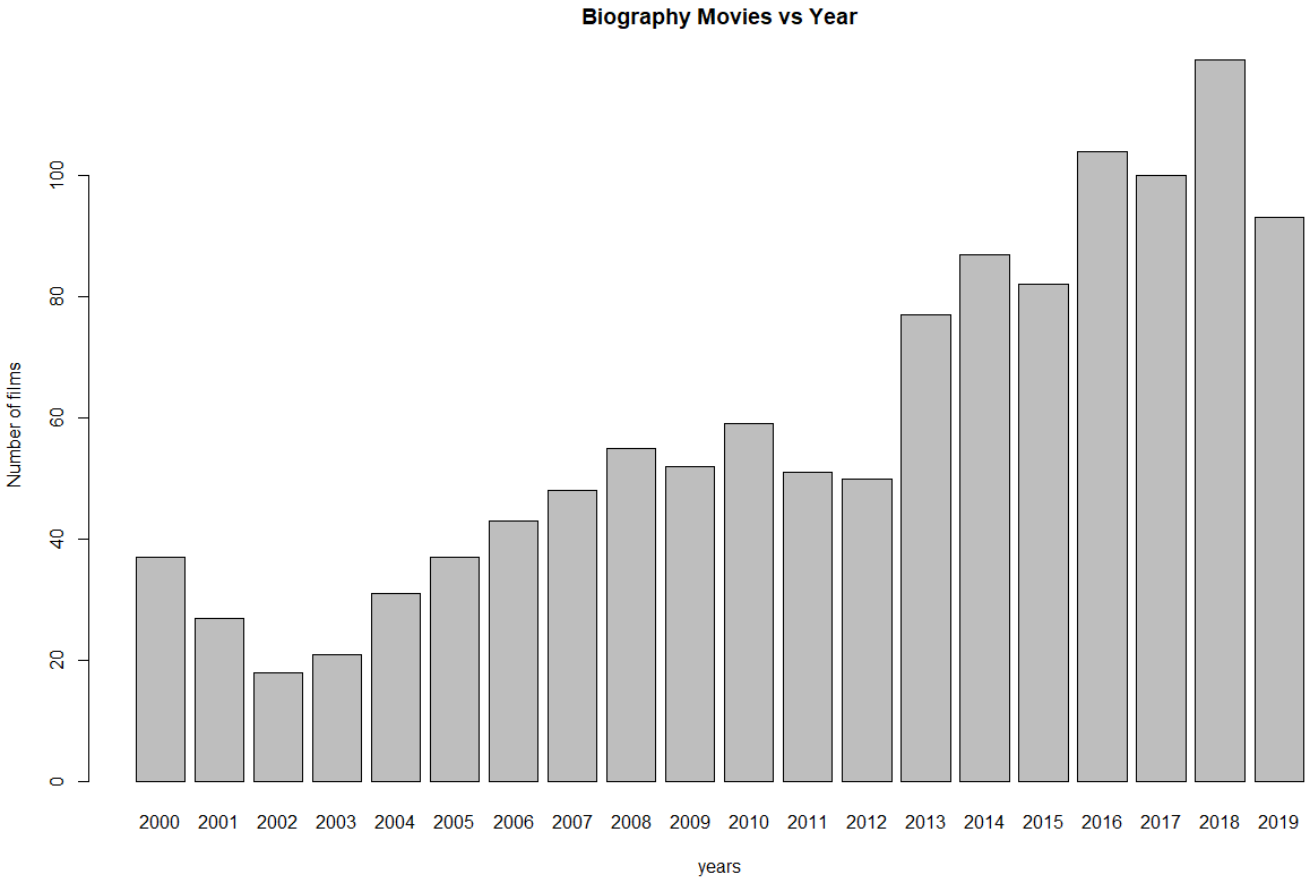


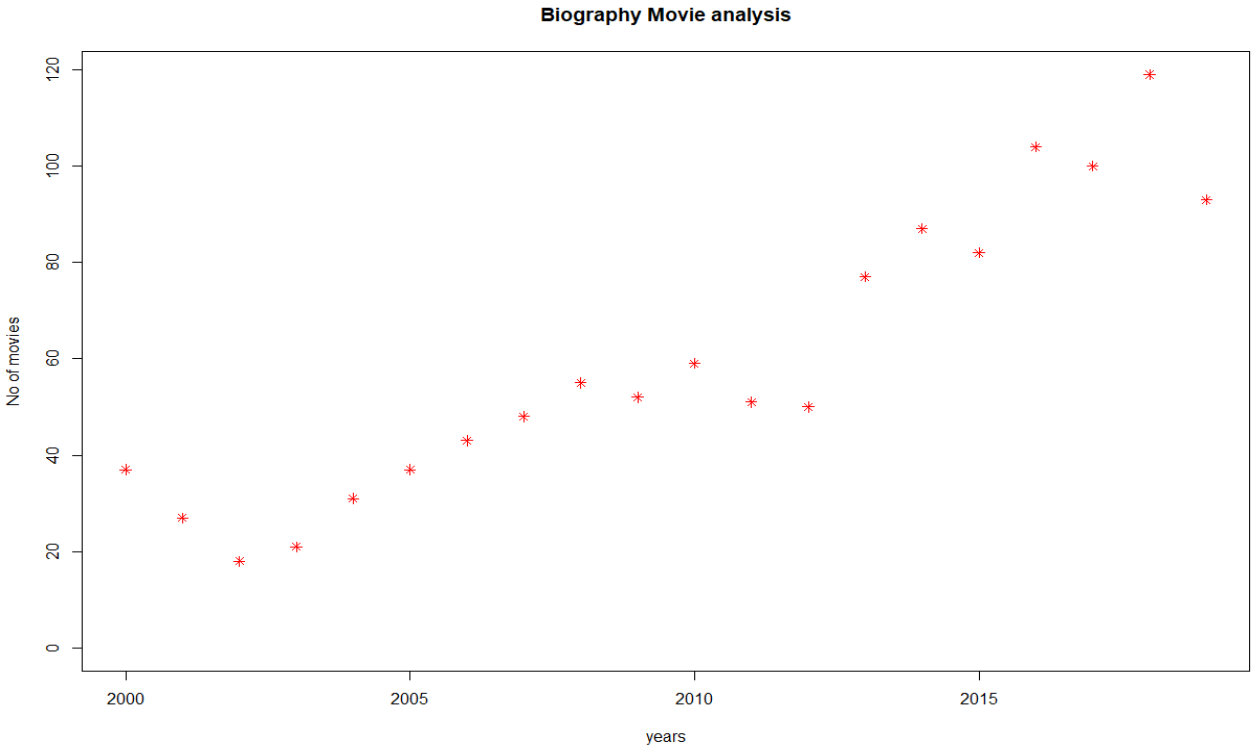
Case 6: Adventure



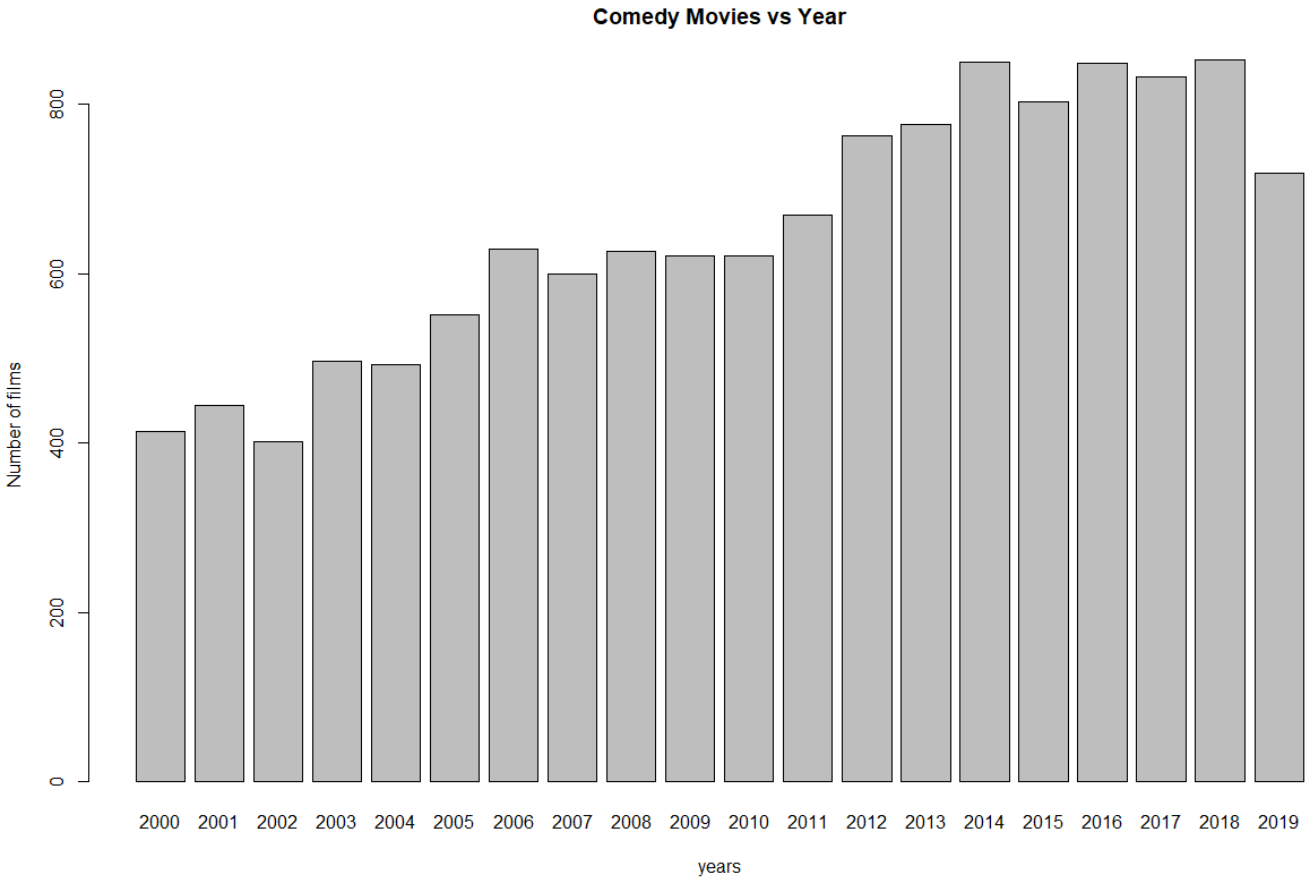


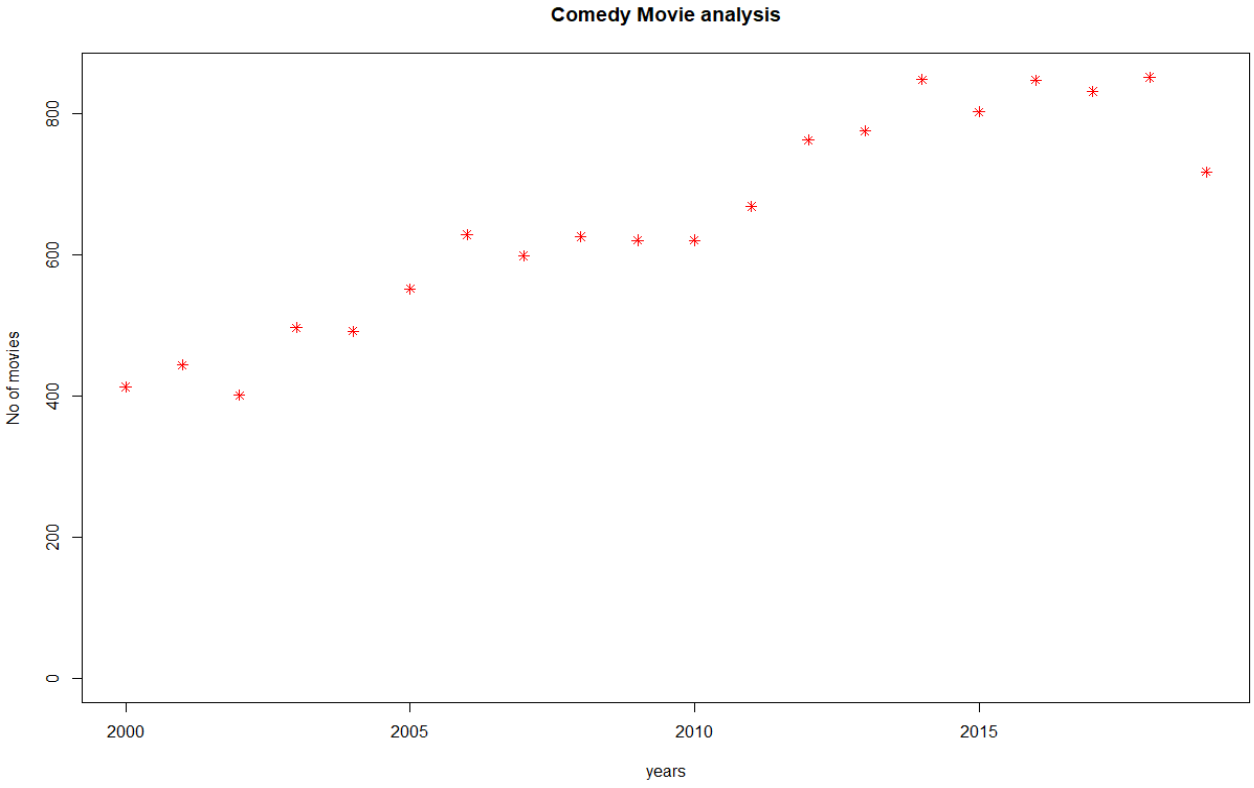
Case 7: Biography



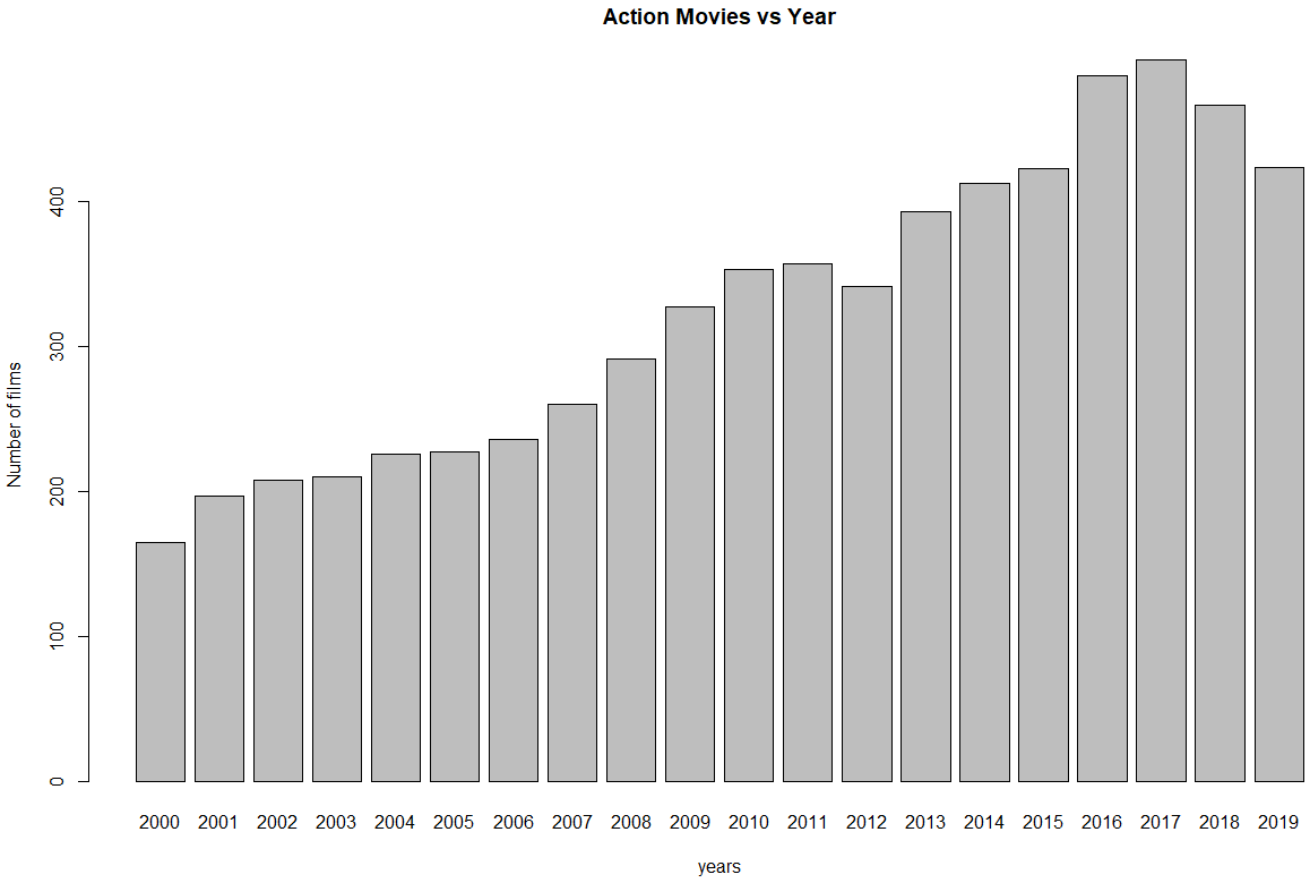


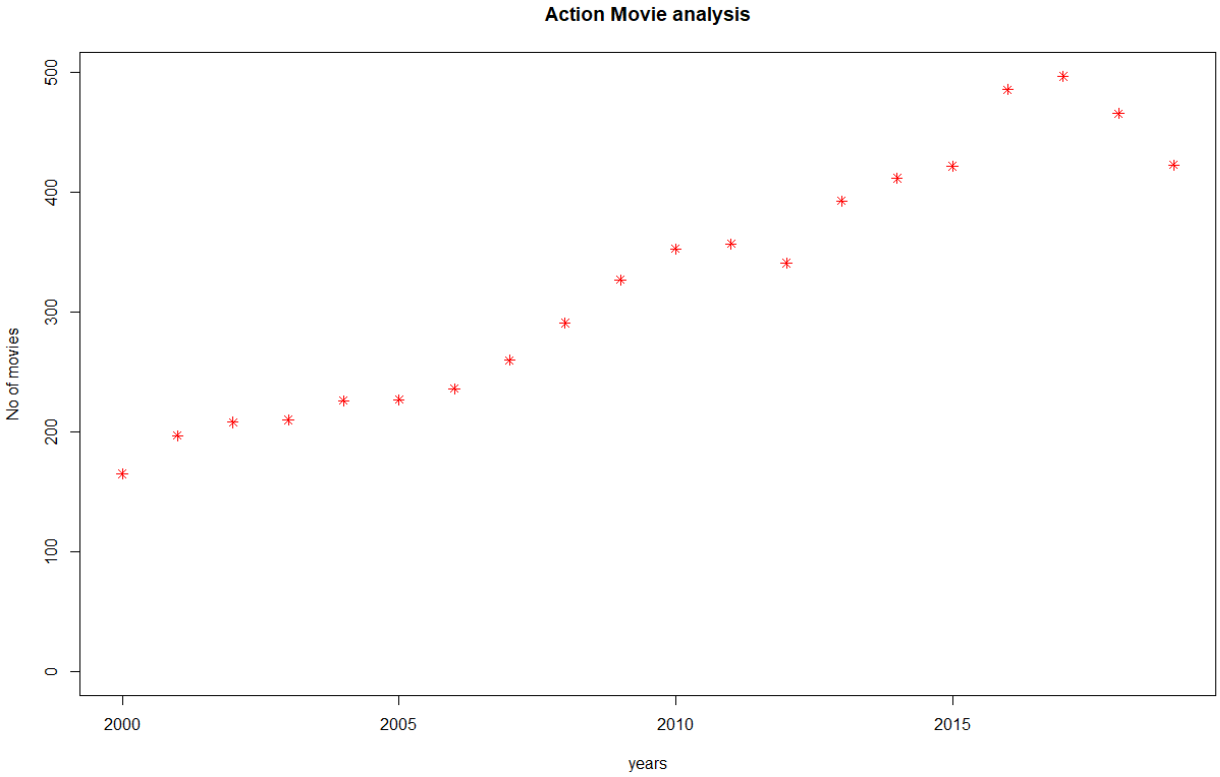
Case 8: Comedy



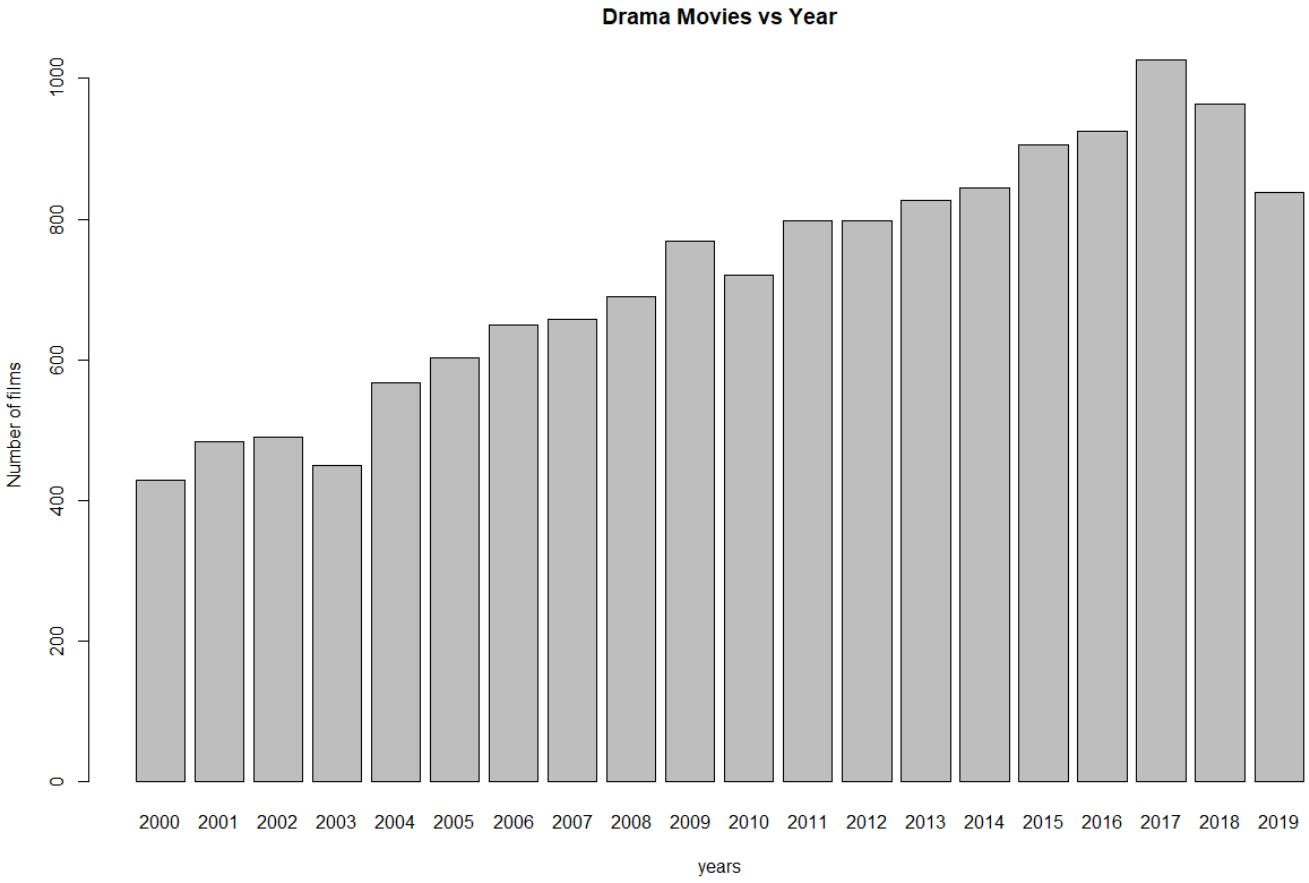


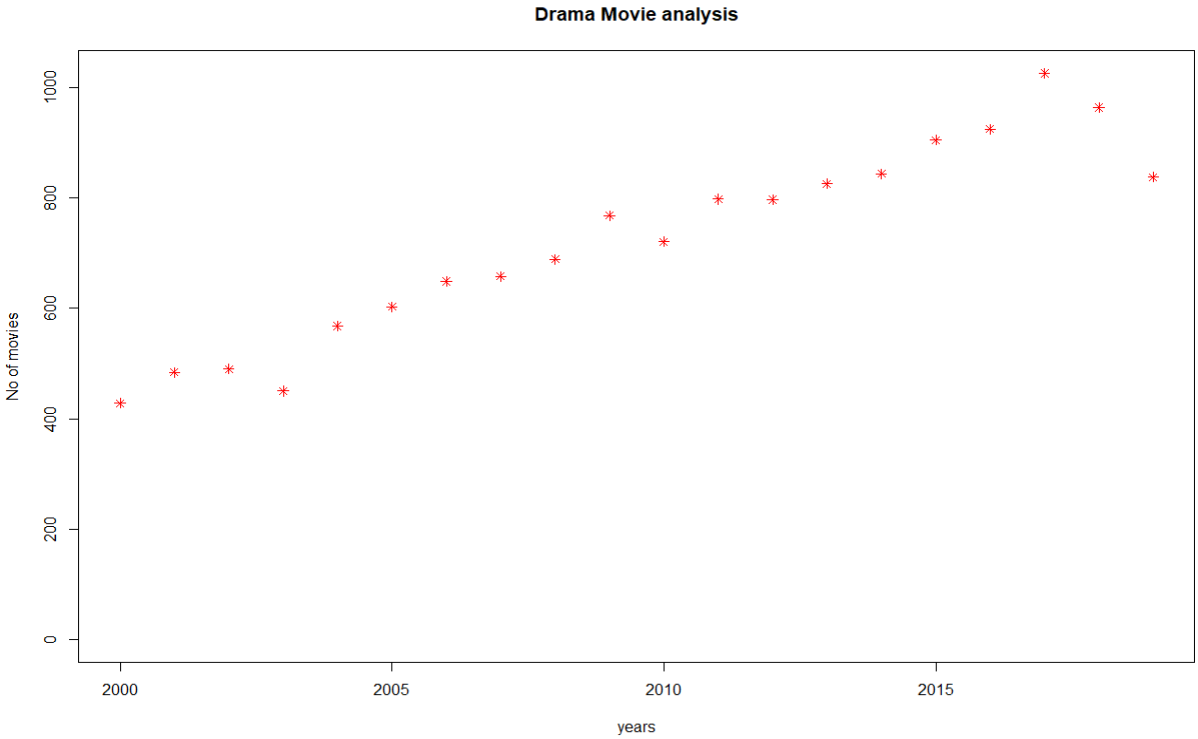
Case 9: Action





Case 10: Drama





Here we are considering different cases where users can give input genres based on which the data will be divided on which we can work for data analysis and get the result.

Graphs like Barplot and Scatter are used for getting the results.

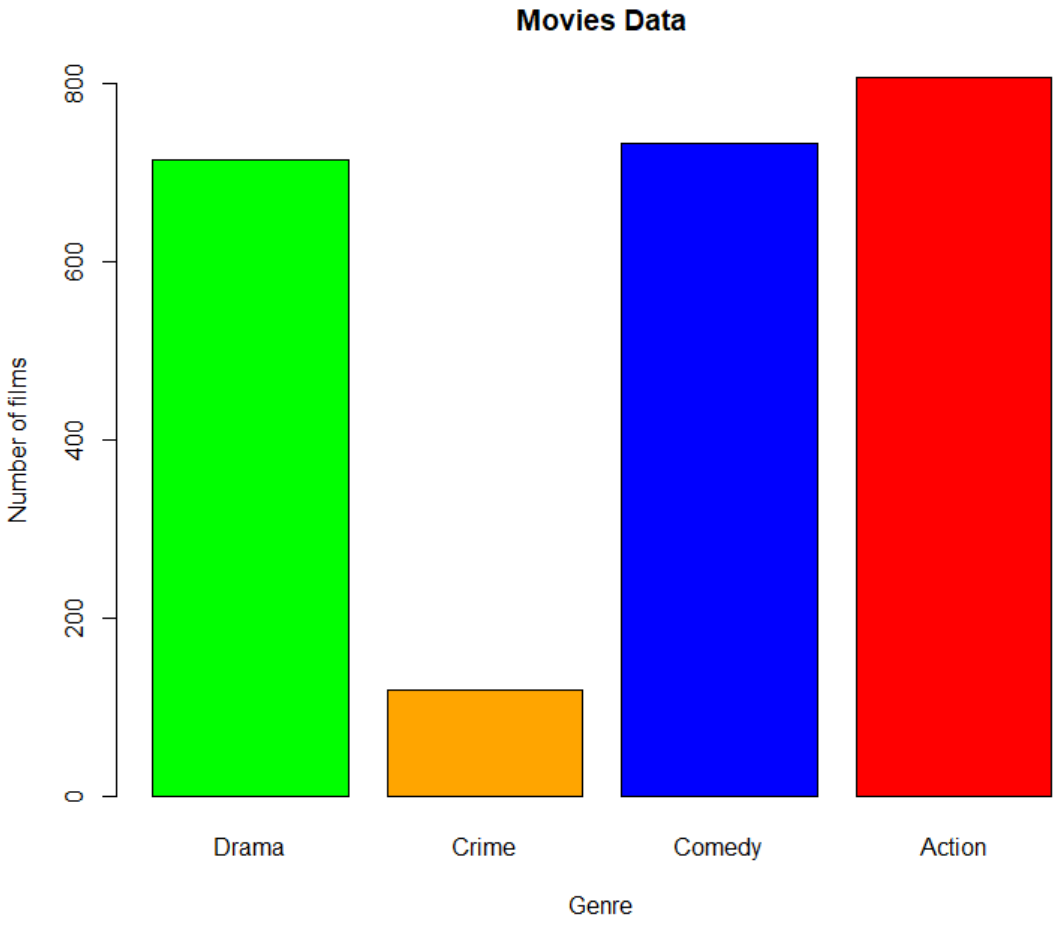
Results:

Case 1:

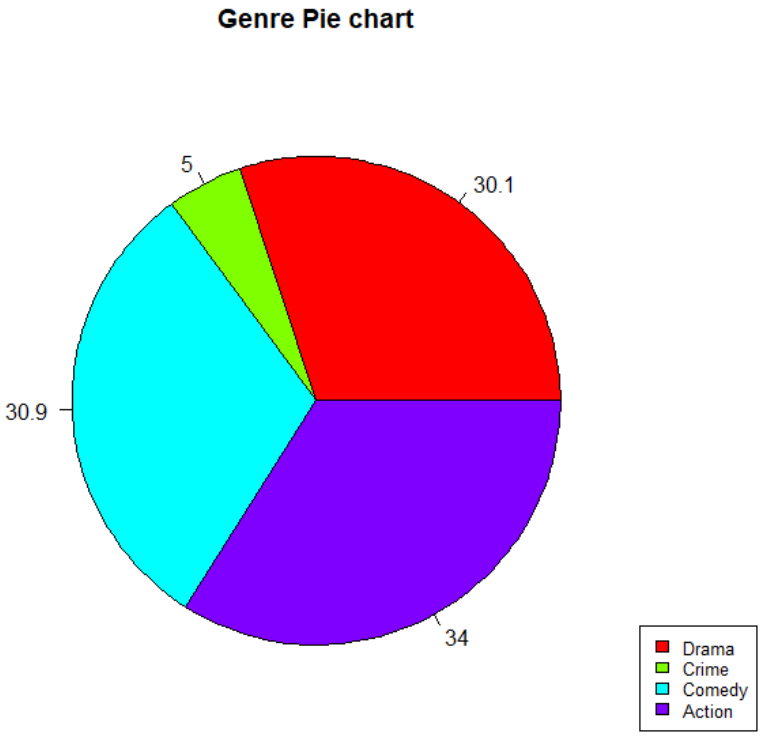
If the user selects-
year = 2010+
country = India
Rating = 4+

```
> temp_year = readline(prompt="Enter the year for where you want to consider movies  -> ")
Enter the year for where you want to consider movies  -> 2010
> temp_country =  readline(prompt="Enter the country name -> ")
Enter the country name -> India
> temp_rating = readline(prompt ="Enter the rating between 1-9 decimal is allowed -> " )
Enter the rating between 1-9 decimal is allowed -> 4
```

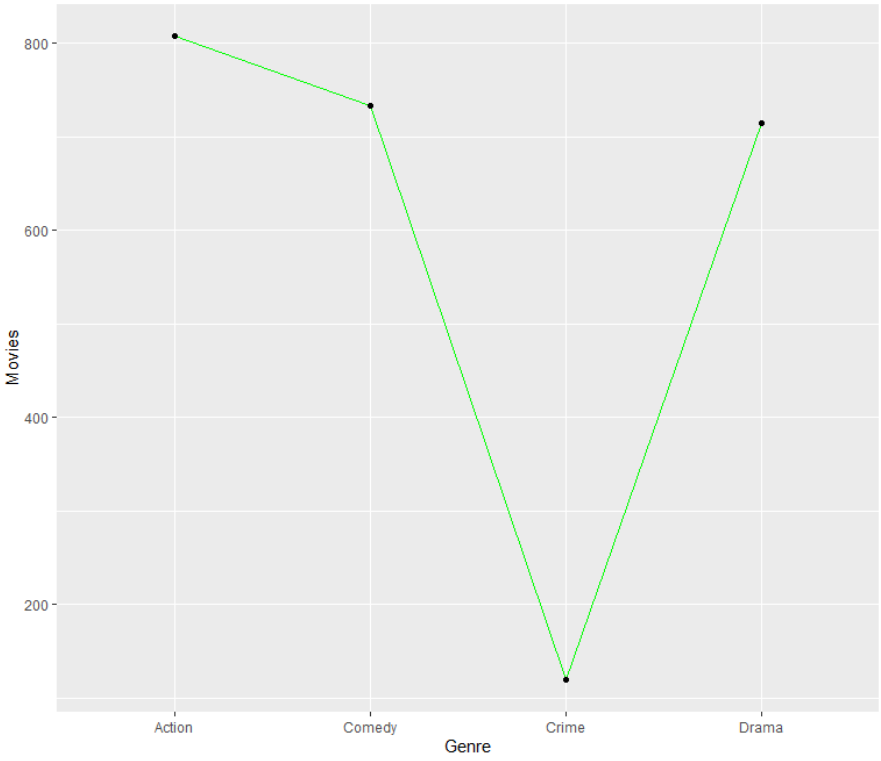
Barplot



Pie Chart



Line chart



Case 2:

If the user selects-

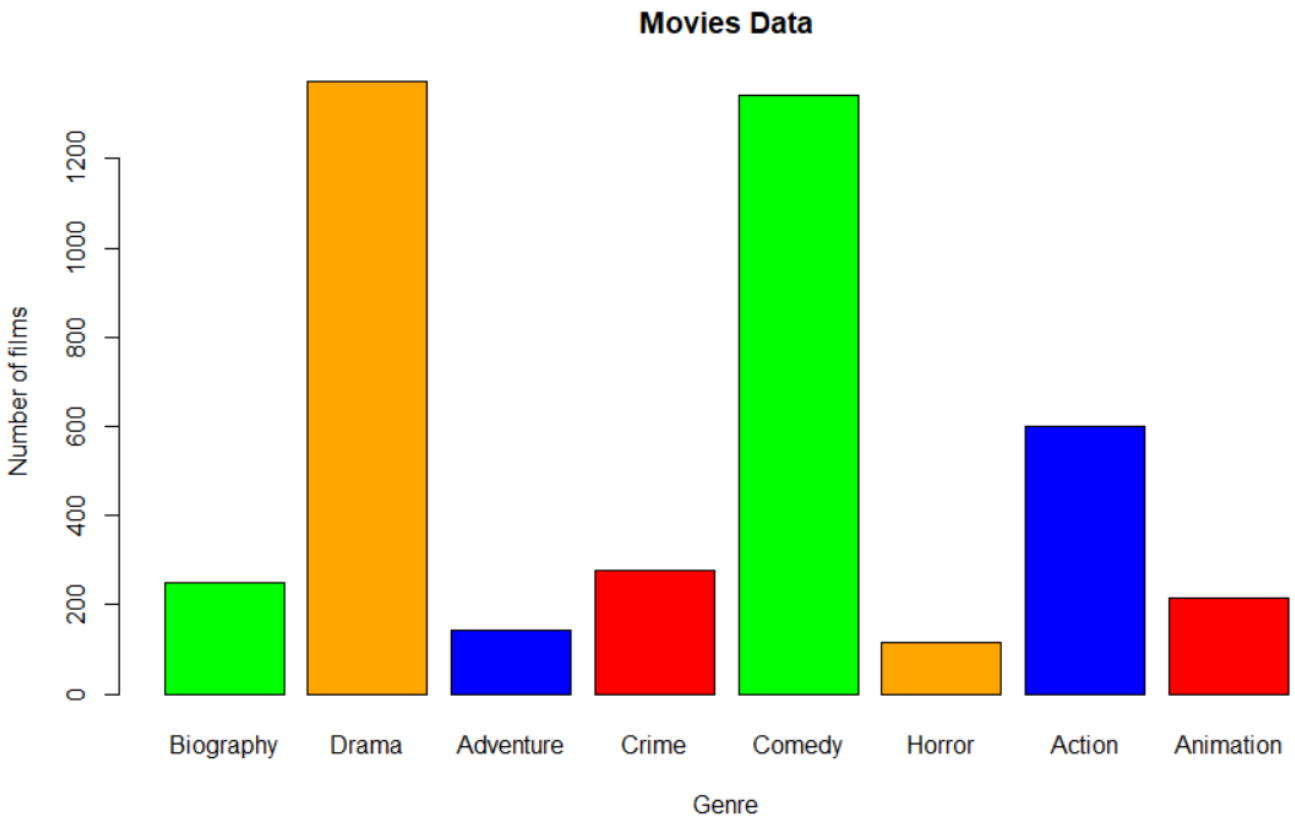
year = 2000+

country = USA

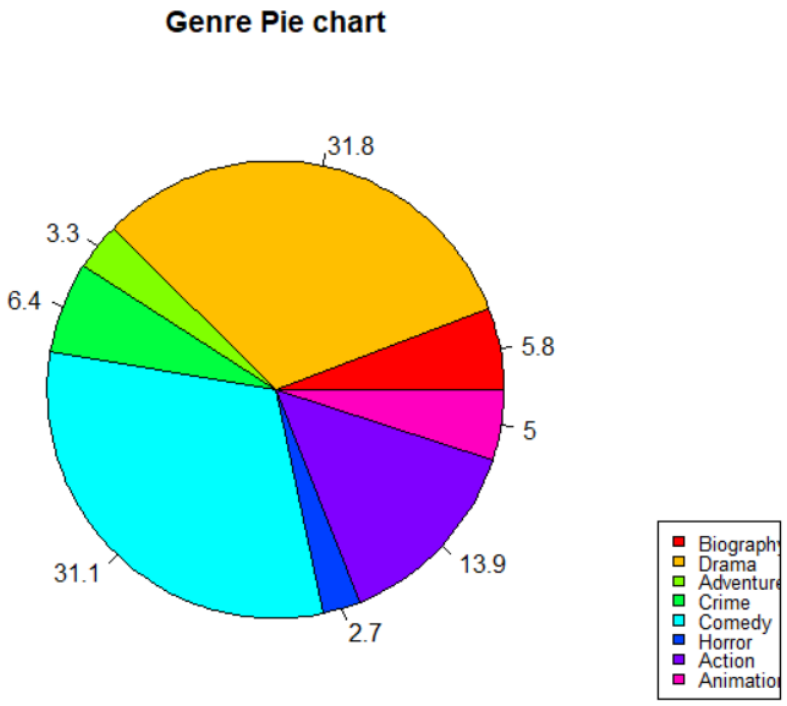
Rating = 6+

```
> temp_year = readline(prompt="Enter the year for where you want to consider movies  -> ")
Enter the year for where you want to consider movies  -> 2000
> temp_country =  readline(prompt="Enter the country name -> ")
Enter the country name -> USA
> temp_rating = readline(prompt ="Enter the rating between 1-9 decimal is allowed -> " )
Enter the rating between 1-9 decimal is allowed -> 6
> temp=subset(data,country==temp_country & year>temp_year & avg_vote>=temp_rating)
> temp
```

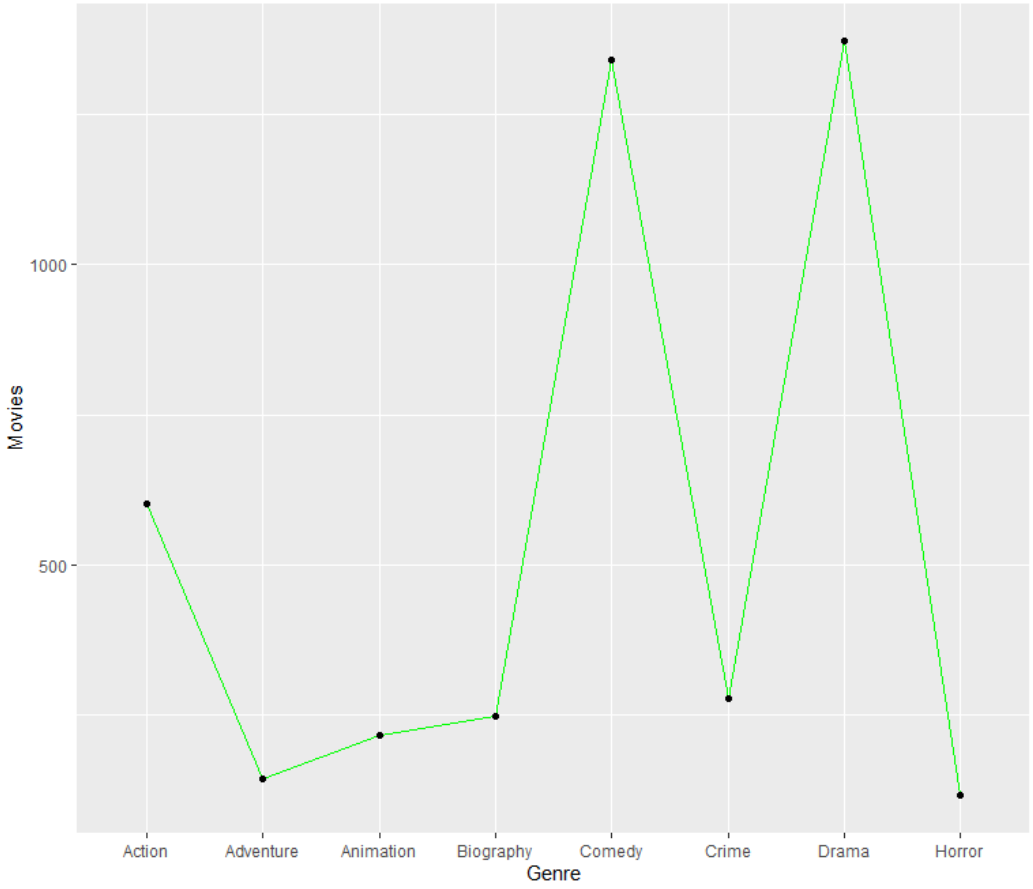
Barplot



Pie Chart



Line chart



Case 3:

If the user selects-

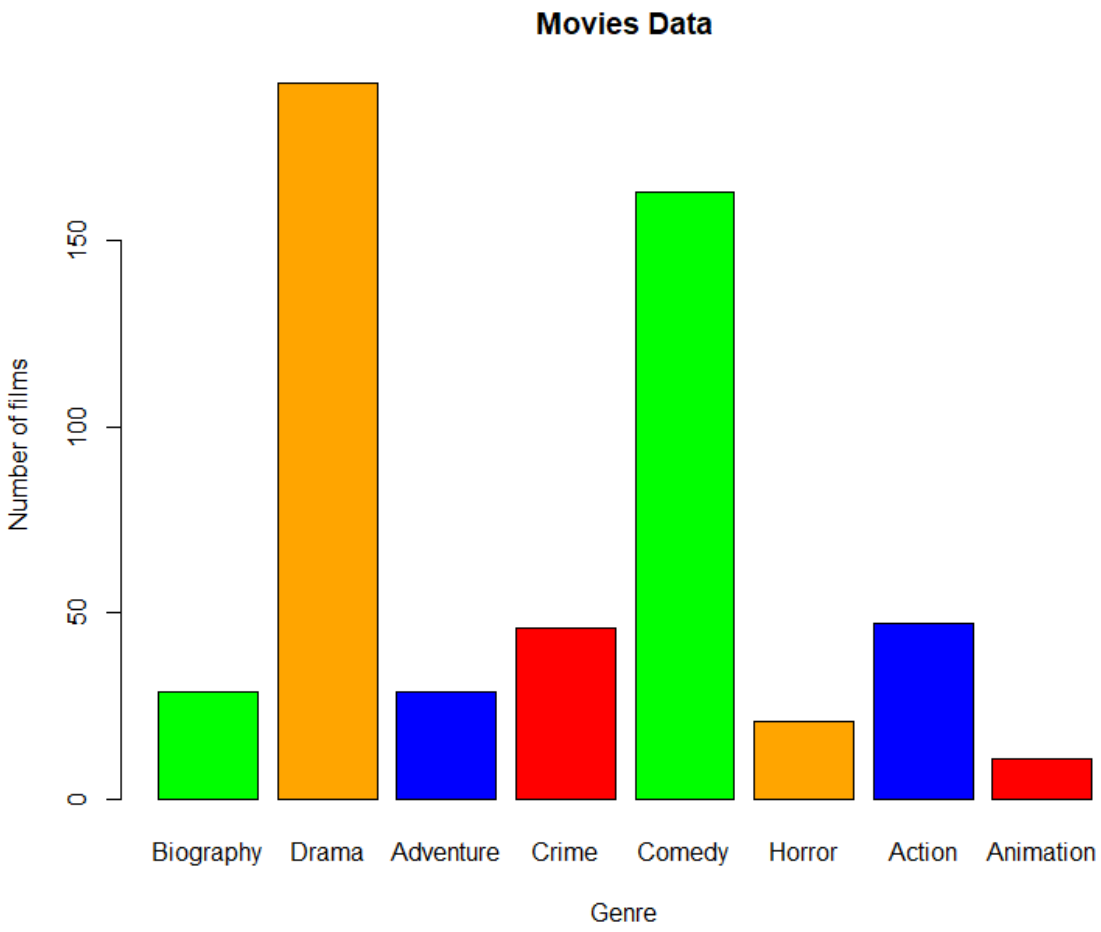
year = 1990+

country = Australia

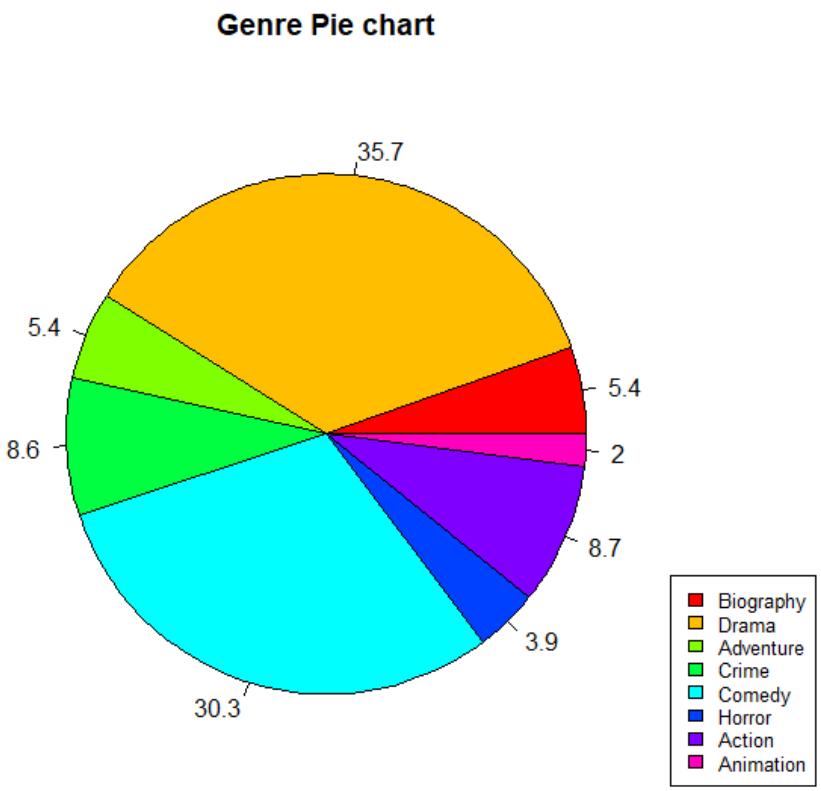
Rating = 5+

```
> temp_year = readline(prompt="Enter the year for where you want to consider movies  ->
 ")
Enter the year for where you want to consider movies  -> 1990
> temp_country =  readline(prompt="Enter the country name -> ")
Enter the country name -> Australia
> temp_rating = readline(prompt ="Enter the rating between 1-9 decimal is allowed -> " )
Enter the rating between 1-9 decimal is allowed -> 5
> temp=subset(data,country==temp_country & year>temp_year & avg_vote>=temp_rating)
> temp
```

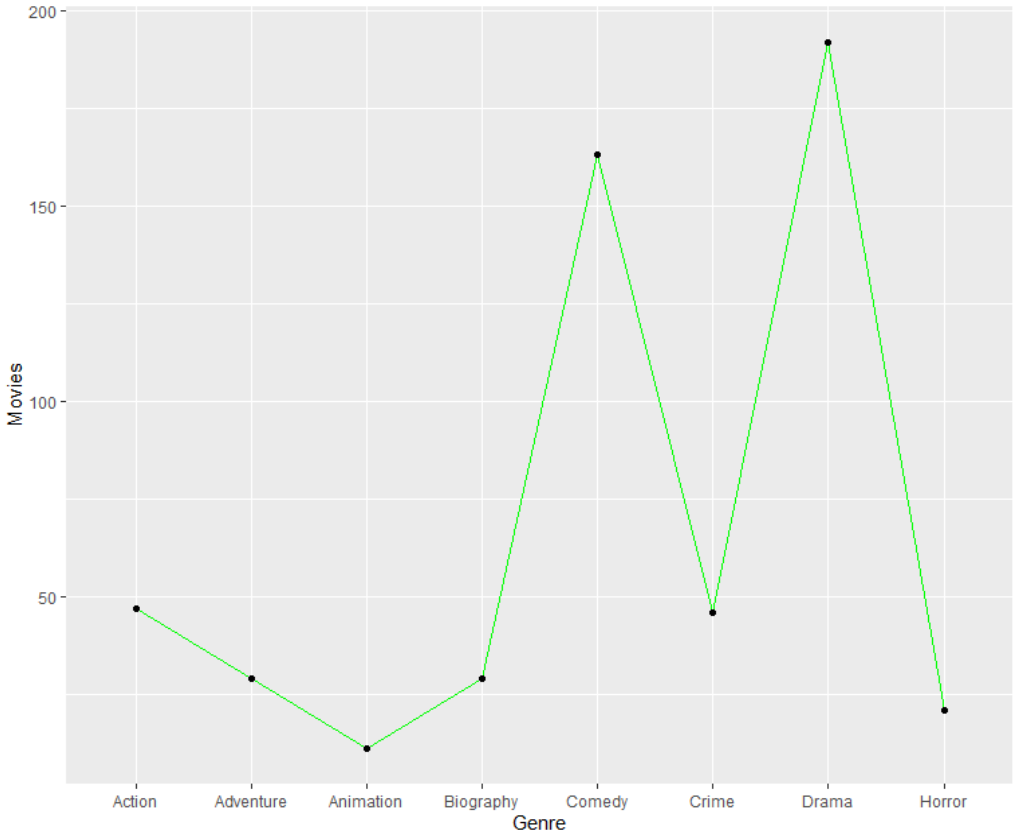
Barplot



Pie Chart



Line chart

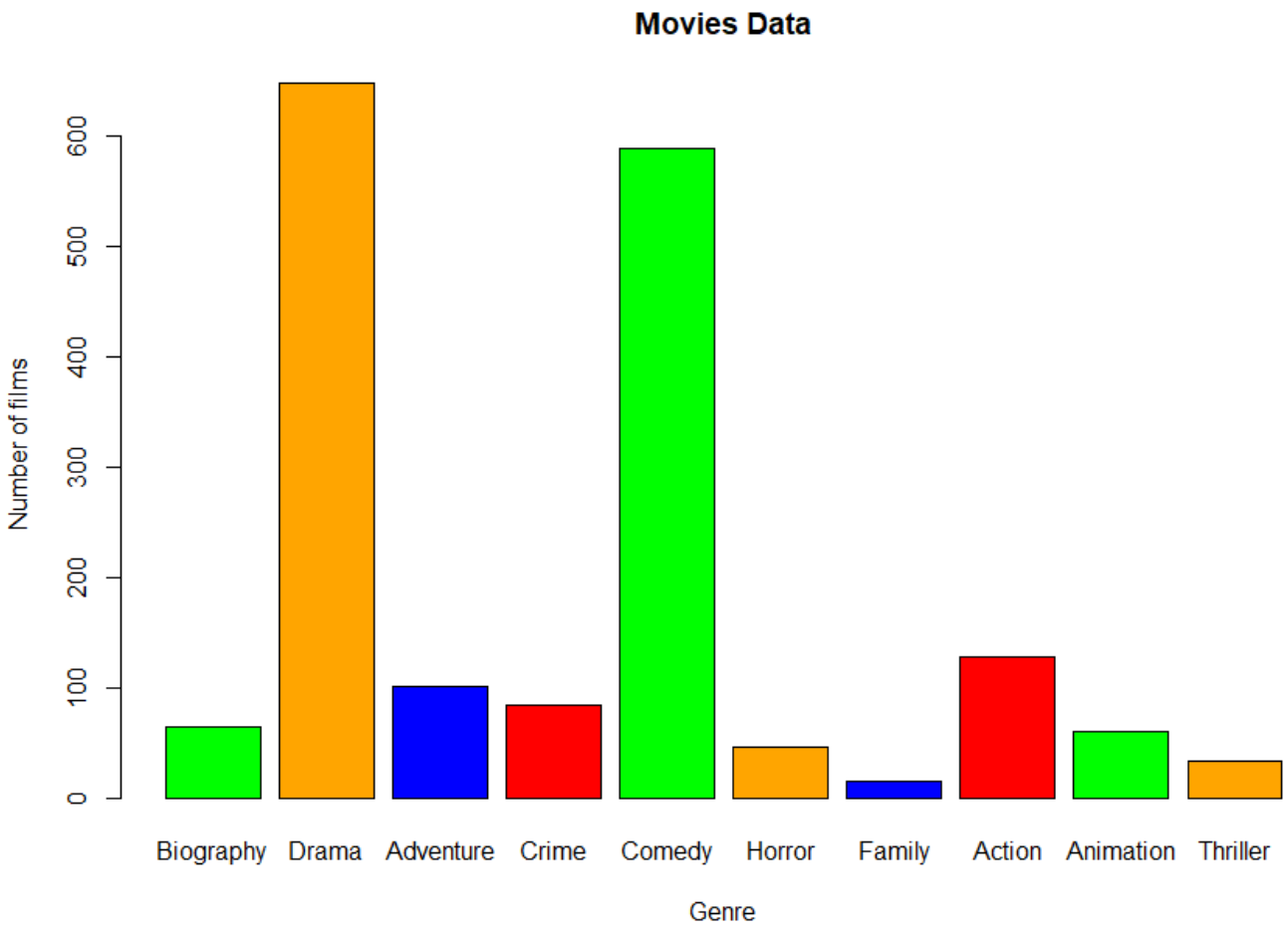


Case 4:

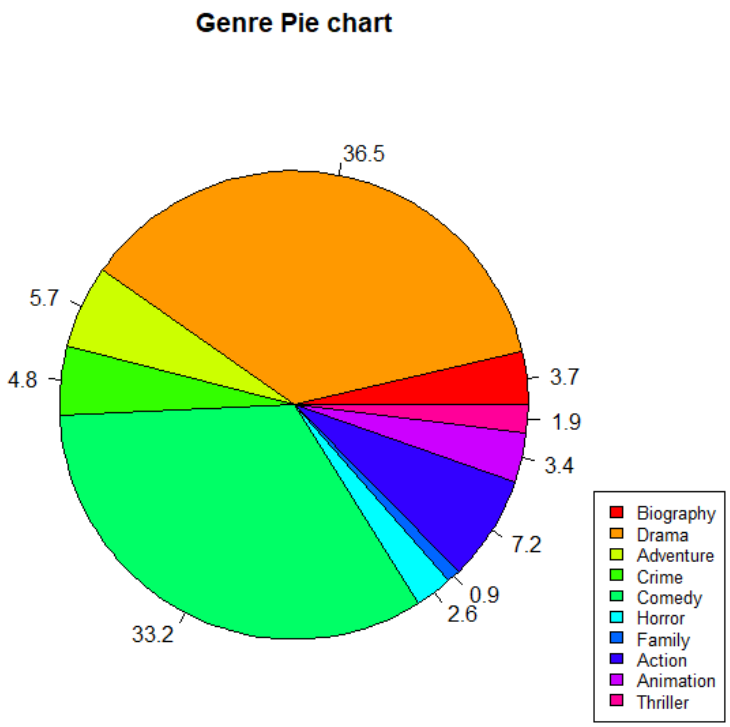
If the user selects-
year = 1980+
country = Germany
Rating = 3+

```
> temp_year = readline(prompt="Enter the year for where you want to consider movies  -> ")
Enter the year for where you want to consider movies  -> 1980
> temp_country =  readline(prompt="Enter the country name -> ")
Enter the country name -> Germany
> temp_rating = readline(prompt ="Enter the rating between 1-9 decimal is allowed -> " )
Enter the rating between 1-9 decimal is allowed -> 3
> temp=subset(data,country==temp_country & year>temp_year & avg_vote>=temp_rating)
> temp
```

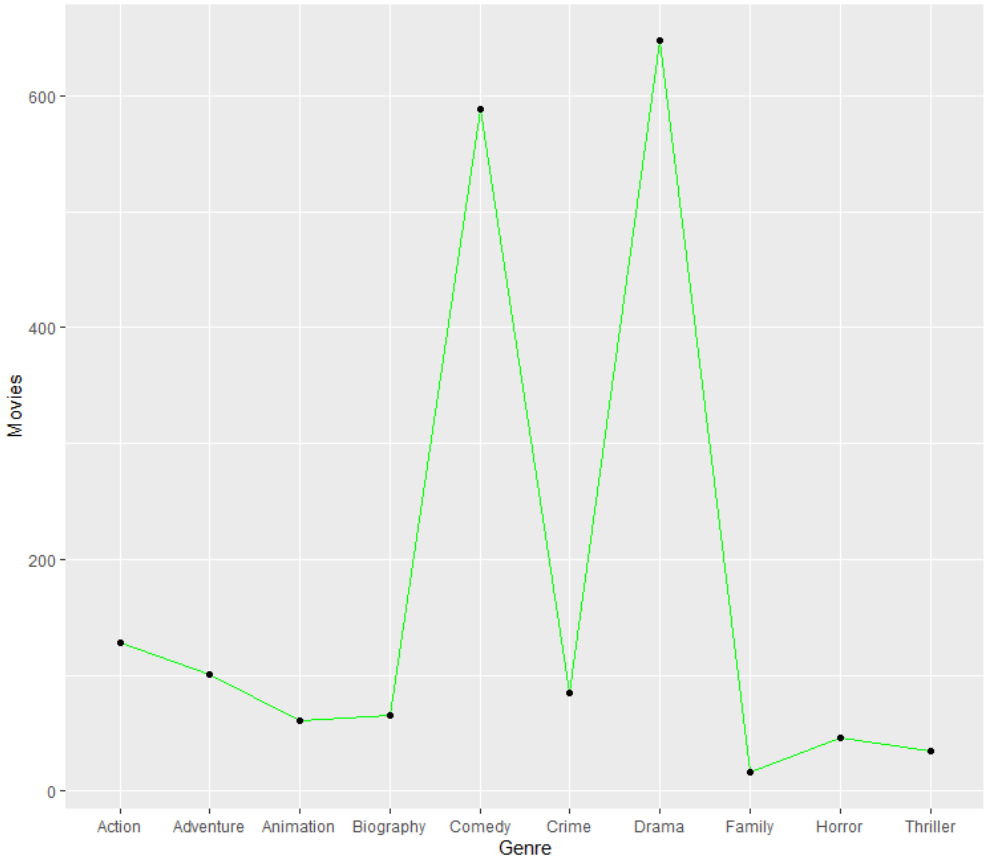
Barplot



Pie Chart



Line chart

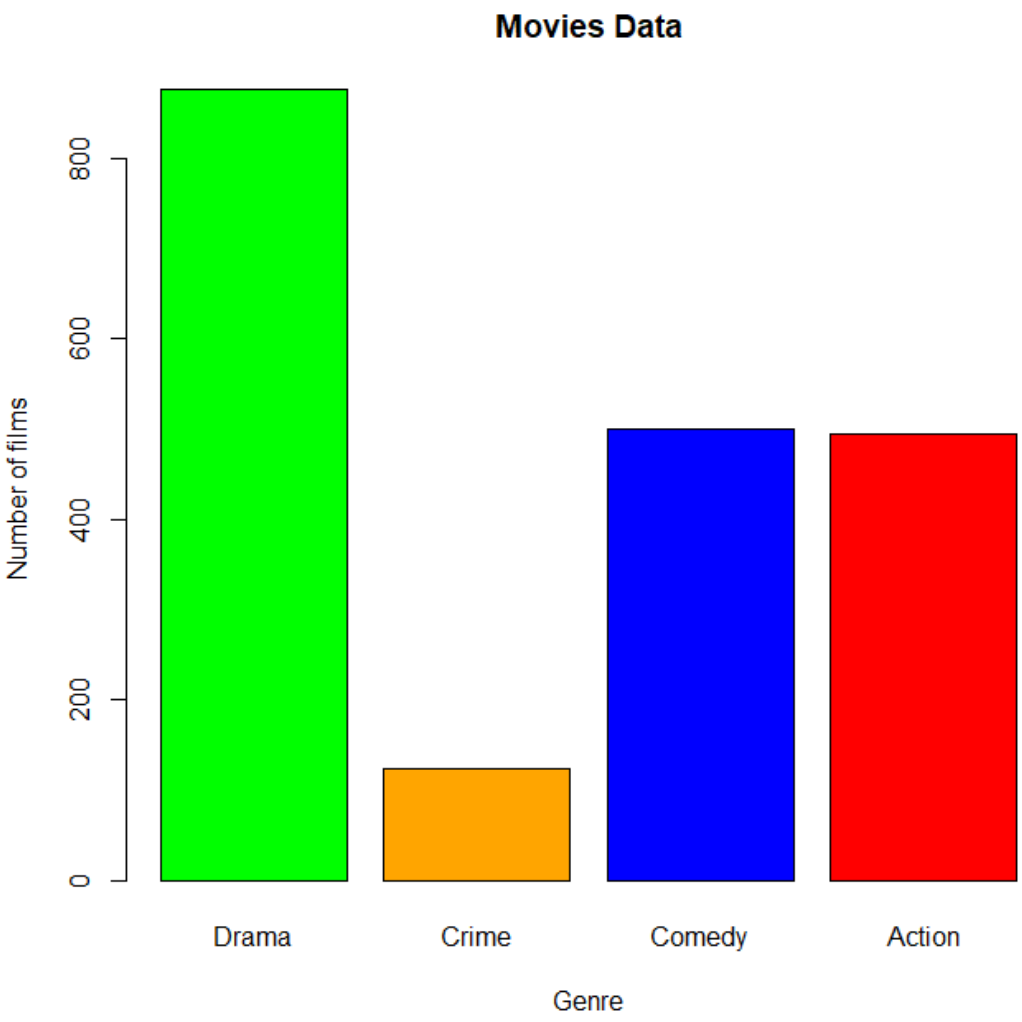


Case 5:

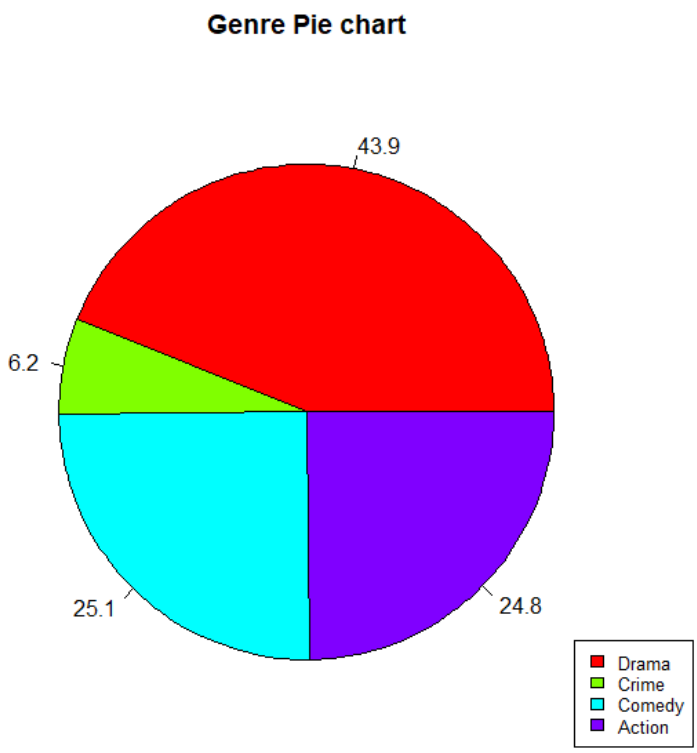
If the user selects-
year = 1950+
country = India
Rating = 7+

```
> temp_year = readline(prompt="Enter the year for where you want to consider movies  -> ")
Enter the year for where you want to consider movies  -> 1950
> temp_country =  readline(prompt="Enter the country name -> ")
Enter the country name -> India
> temp_rating = readline(prompt ="Enter the rating between 1-9 decimal is allowed -> " )
Enter the rating between 1-9 decimal is allowed -> 7
> temp=subset(data,country==temp_country & year>temp_year & avg_vote>=temp_rating)
> temp
```

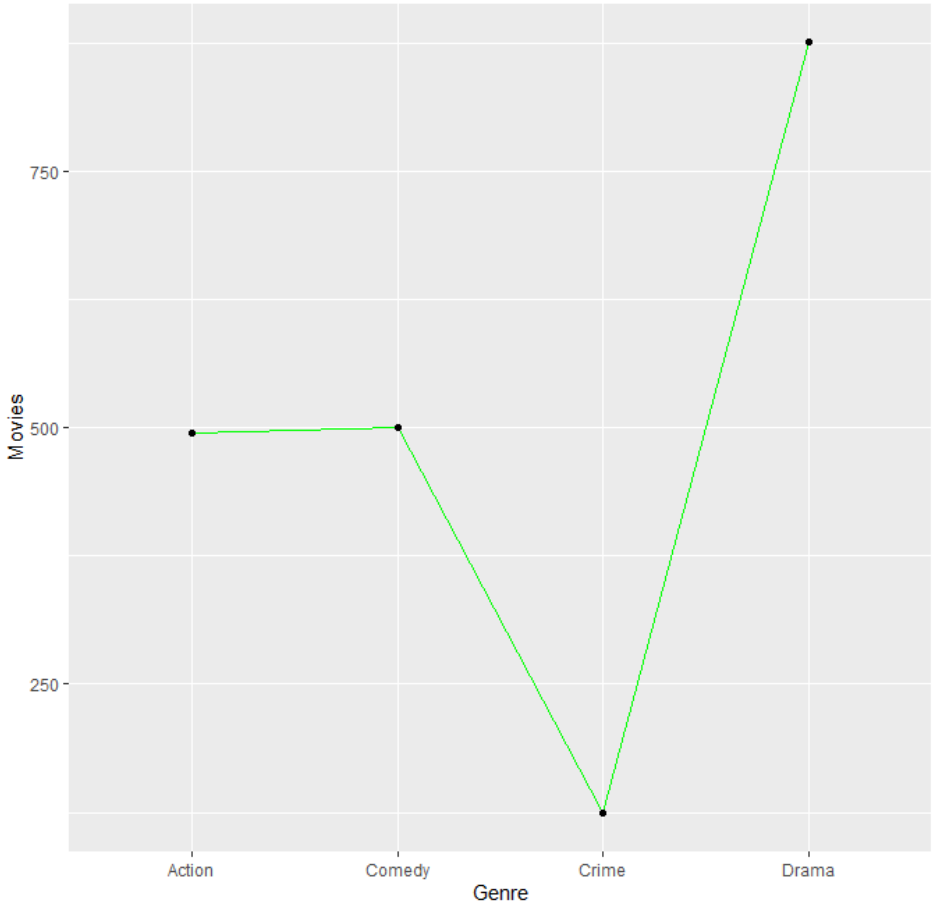
Barplot



Pie Chart


Genre Pie chart
43.9
6.2
25.1
24.8
Drama
Crime
Comedy
Action

Line chart


750
500
250
Movies
Action
Comedy
Crime
Drama
Genre

Here we are considering different cases where users can give different inputs like a year, country, and Rating based on which the data will be divided on which we can work for data analysis and get the result.

Graphs like Barplot, pie chart, and line chart are used for getting the results.

Conclusion:

We shall conclude this project with a specific case. Assume that if a user wants to make a movie he selects the country, year, and rating to analyze.

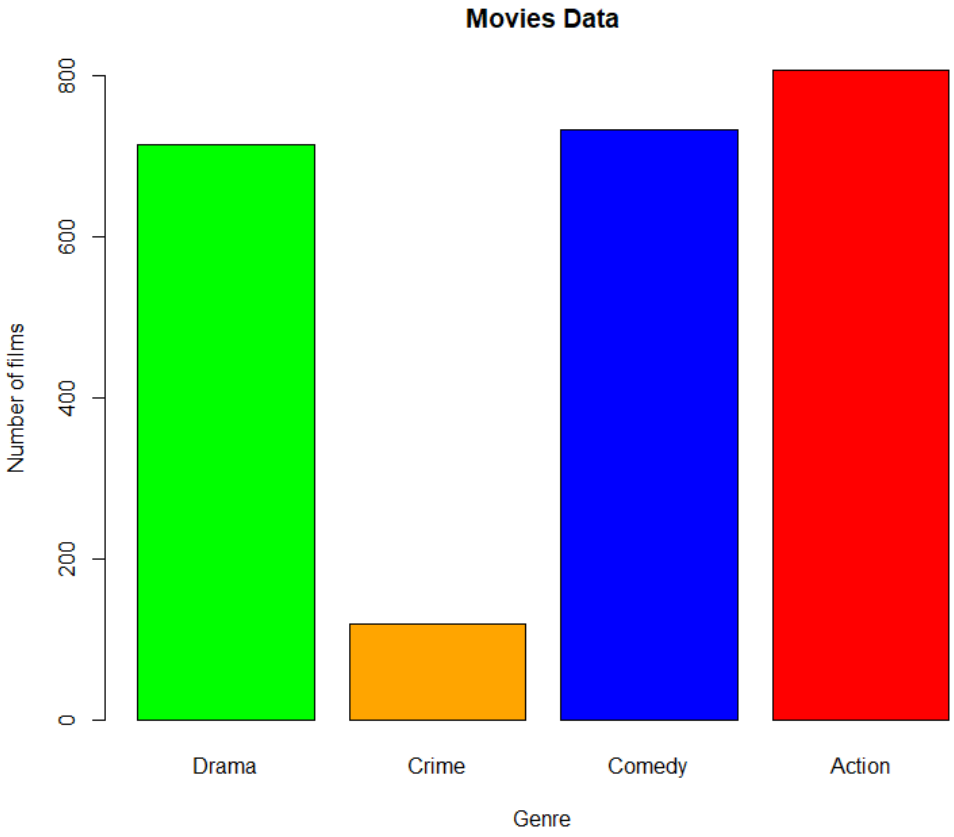
Let,

year = 2010+

country = India

Rating = 4+

Output:



Based on the graphs there is a high probability of making profits if the user directs an action film.

The probability order goes as Action→ Comedy→Drama→Crime.

References:

1.https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset this is used for downloading movie dataset

2.https://www.r-bloggers.com/2021/04/how-to-clean-the-datasets-in-r/ this is used for getting a basic understanding of data cleaning

3.https://analyticsindiamag.com/data-preprocessing-with-r-hands-on-tutorial/ this is used for getting a basic understanding of data preprocessing

4.https://www.javatpoint.com/r-data-visualization this is used for getting understanding about graphs