# IMDB SERIES RANKING

A FINAL PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE
STAT 291 – STATISTICAL COMPUTING I
DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GROUP_19
2429397 Ecehan Vergiliel
2501955 Emine Doğa Aşkın
2428902 İlayda Berra Belim
2428936 Can Bulut
2429090 Mehmet Egemen Gündür

January 2022

## ABSTRACT

This report gives information about peoples' preferences of TV series worldwide in the 21st century. The report helps us analyze which genres have been watched and which countries of origin voted more. Intending to study this topic, we used "Series Ranking" data. The data contains the title of the series, country of origin, release year, rating, number of votes, director, genres, runtime, dashboard link, IMDb link, and plot summary. We cleared the data for better analysis and a more accessible study.

The research questions give the purpose of the project: analyzing the watching choices of people, understanding the relationship between rates and votes, and country preferences about types of TV series. After reconfiguring our data, we observed the popular genres, relations between rates and votes, and producing countries. The project concluded with three main findings: there is no relationship between rates and votes, the most popular genre is drama, and the most productive country is the USA.


## 1. Introduction

To put the research superficially, some statistical interpretations and inferences were tried to be made on the "Series Rankings" obtained by filtering IMDb data. The data contains information about the genres and ratings of the series voted by many countries. Furthermore, the data includes information about the ratings of the series, how long the runtime is, and the summary of the subject. Here are some variables from the "Series Rankings" data to help you understand it better: title, country, year, rating, votes, director, genres, runtime, dashboard link, link, plot summary.

Firstly, to understand our data, we constructed a histogram and a line graph for the distribution of ratings of tv series according to years to determine whether the rating differs in different years or not and how. After that, we calculated the correlations between ratings and votes variables and constructed a simple linear regression model to analyze the relationship between ratings and votes. Finally, to understand the differences and similarities between countries' preferences to produce TV series, we created two different bar plots of countries and their most made genres. One of them contains all nations, and the other does not have the US analyze other countries' preferences more precisely.

To sum up, we do statistical analysis to achieve these three goals, including lots of data visualization, correlation coefficient calculations, and linear regression.

### 1.1. Data description

The URL of the data is http://metabase.intellimenta.com/public/dashboard/eae564a4-d9a3-46b1-9cd4-1f95ab5b1b18?item_type=Series&years_%253E_%253F=2000&rating____=8&__votes____=50000.

The data is interactive data created based on the IMDb ranking list. We filtered the dataset via type, time, rating, and vote number. We sifted by the class: series, the years: between 2000-2021, rating: between 8-10, vote number: >50000. The runtime has two different kinds of data. IMDb determines the running time by whether the series has been completed or not. If the series has been completed, the running time is the total time of all episodes; if not, it is the time of any episode.

Data contains 214 observations of 11 variables. The variables are title, country, year, rating, votes, directors, genres, runtime, dashboard link, link, and plot summary. The unmodified version of the dataset contains one continuous numeric variable and ten characteristic variables. The hyphens are changed to NAs. Years and votes were written with commas (e.g., '289,276' or '2,002'); we omit the commas and convert them to integer values. After we clear and organize the data, it contains one discrete variable (votes), two continuous variables (runtime and rating), three categorical variables (year, country, and genres), one logical variable (director), and one characteristic variable (title). The data has 214 NA values in director and only one NA value in runtime.

### 1.2. Research questions

The research questions are focused on what we want to analyze clearly.

**i.** What is the distribution of ratings of TV series according to years?

**ii.** Is there any relation between ratings and votes? If yes, how can we describe this relation?

**iii.** Which type of TV series do countries produce the most?

### 1.3. Aim of the study

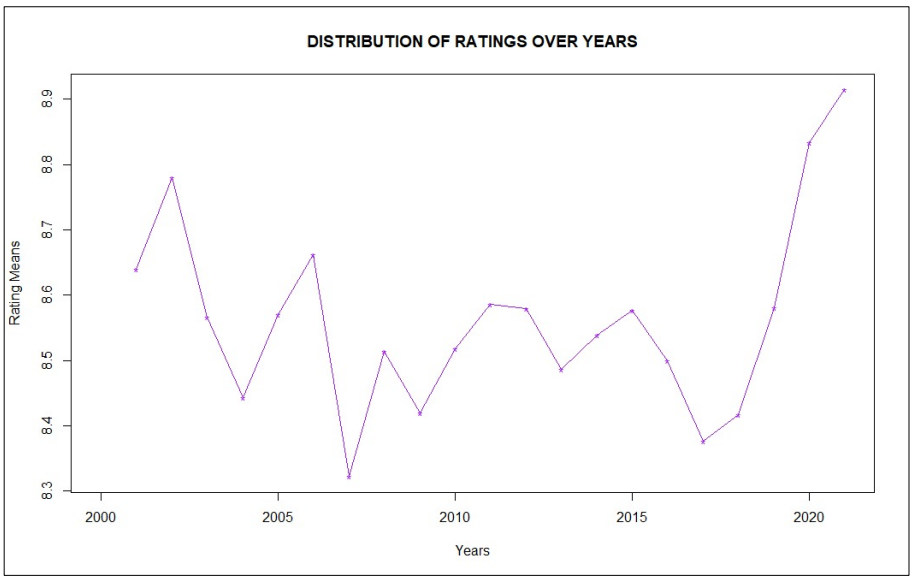
This project aimed to obtain our data findings and analyze them using applications of statistics. We asked ourselves three questions to get these findings that helped us understand different aspects of IMDb data to have a better understanding of countries' preferences of TV series, how ratings change, and relations between variables. To answer these questions, we created several plots such as the distribution of ratings over the years, histograms of rating means, the relationship between ratings and votes, and genres produced by countries. After we created these plots and interpreted them, we found answers to our questions and reached our goal of understanding the different characteristics of IMDb data.

## 2. Methodology/Analysis

Data is taken from http://metabase.intellimenta.com/public/dashboard/eae564a4-d9a3-46b1-9cd4-1f95ab5b1b18?item_type=Series&years_%253E_%253F=2000&rating____=8&__votes____=50000 .

Director, Votes and Years columns are reconfigured. While reconfiguring our data, we used gsub() function to omit commas. After determining our research questions, we planned which data structures we should use and coded our algorithms. In the first question, we used tables and nested loops to get the distribution of ratings over the years. Then the acquired data was visualized with a line graph and a histogram to be better understood. We calculated the correlation between Ratings and Votes in the second question to see their relationship. We used a scatter plot, and in our plot, we included the correlation coefficient and the regression line we got from the fitted linear model using simple linear regression. The plot better helped us see the outliers' effect on the regression line. In the last question, we used necessary functions from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/strsplit to get the particular genres as char values. We constructed bar plots to visualize unique genres for each country and the most produced genre. Since country names were too long for bar plots, we manually abbreviated the country names and genres. In addition, we used different colors in all our plots to visualize them better. As the US was in the lead, we constructed another bar plot to discern other countries' productions without the US. Different colors are used to read the plots easily, and we took the colors from two different web site sources.

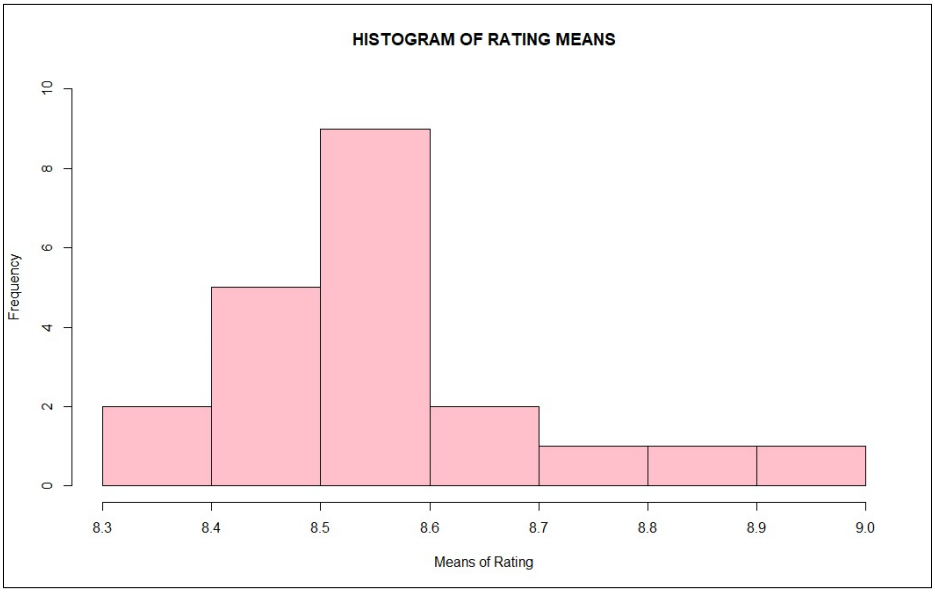
## 3. Results and Findings

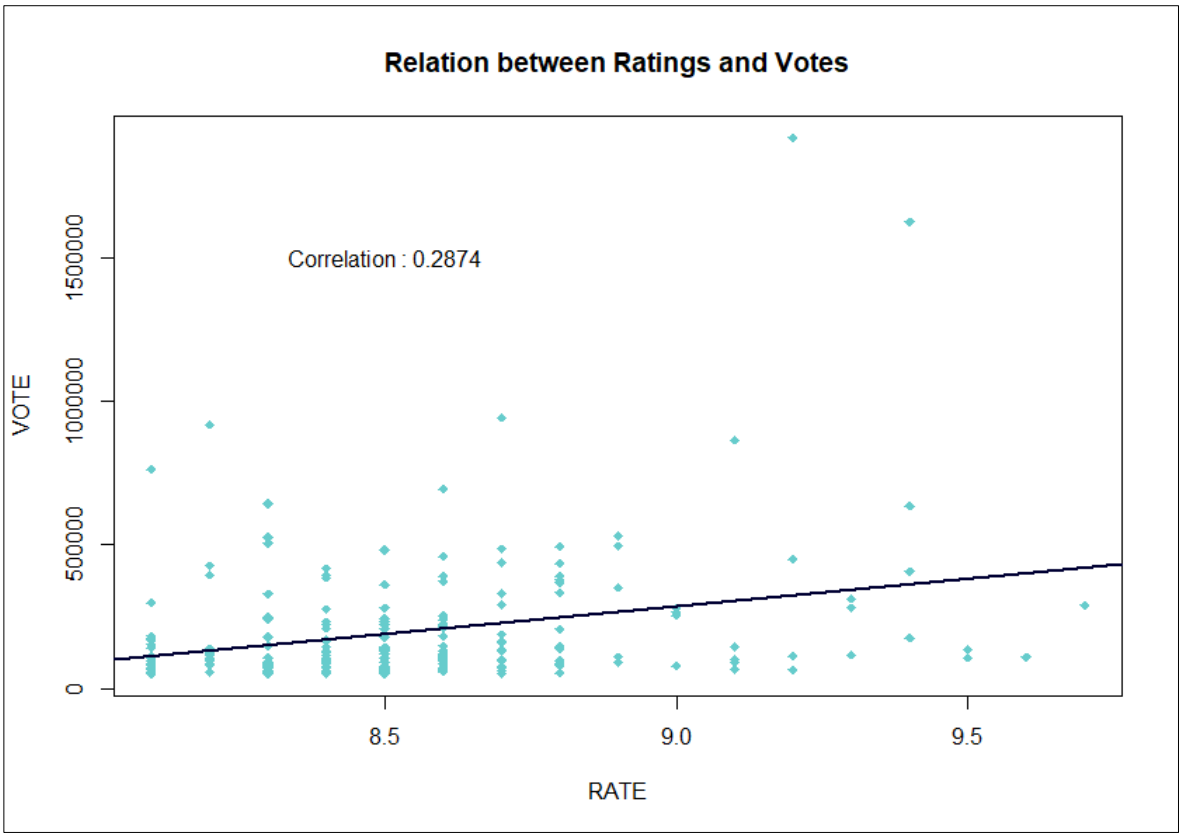



| years | means |
|---|---|
| 2001 | 8.64 |
| 2002 | 8.78 |
| 2003 | 8.566667 |
| 2004 | 8.442857 |
| 2005 | 8.57 |
| 2006 | 8.6625 |
| 2007 | 8.322222 |
| 2008 | 8.514286 |
| 2009 | 8.42 |
| 2010 | 8.518182 |
| 2011 | 8.585714 |
| 2012 | 8.58 |
| 2013 | 8.486667 |
| 2014 | 8.538462 |
| 2015 | 8.576923 |
| 2016 | 8.5 |
| 2017 | 8.376471 |
| 2018 | 8.416667 |
| 2019 | 8.58125 |
| 2020 | 8.833333 |
| 2021 | 8.914286 |

The first graph we used is a line graph. It is used to see the year-based means of the rating changes easily. According to the graph, there is no constant increase or decrease. Therefore, it is said that there is a continuous fluctuation. Moreover, the graph took an enormous value in 2021 and the smallest value in 2008.

We also plotted a histogram to understand the data better. The histogram shows the rating means and frequencies of them. As seen in the histogram, there is a pile on the left side, so we can say that this histogram is skewed right. That means the frequencies on the left side are bigger than the rest.






The scatter graph is used to show whether there is a relationship or not between 'ratings' and 'votes.' To determine the type of relationship between them, we first found the correlation coefficient, which is approximately 0.28. According to this, it is said that there is a moderate positive relationship between 'ratings' and 'votes.' In other words, there is a weak uphill linear relationship. Moreover, The graph with the regression line is proof of the expectations.

The summary of our linear model:

| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
|---|---|---|---|---|
| -291526 | -100277 | -51717 | 37345 | 1593524 |

| Coefficients | |
|---|---|
| (Intercept) | Rating |
| -1449011 | 192735 |

**Estimated Linear Model :**

***Votes = -1449011 + 192735 x Rating***

Summary of imdb_lm

Call:

lm(formula = Votes ~ Rating, data = imdb_data)

Residuals:

| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
|---|---|---|---|---|
| -291526 | -100277 | -51717 | 37345 | 1593524 |

Coefficients:

| | **Estimate** | **Std. Error** | **t value** | **Pr ( >\| t \|)** |
|---|---|---|---|---|
| (Intercept) | -1449011 | 376897 | -3.845 | 0.00016 *** |
| Sales | 192735 | 44113 | 4.369 | 1.95e-05 *** |

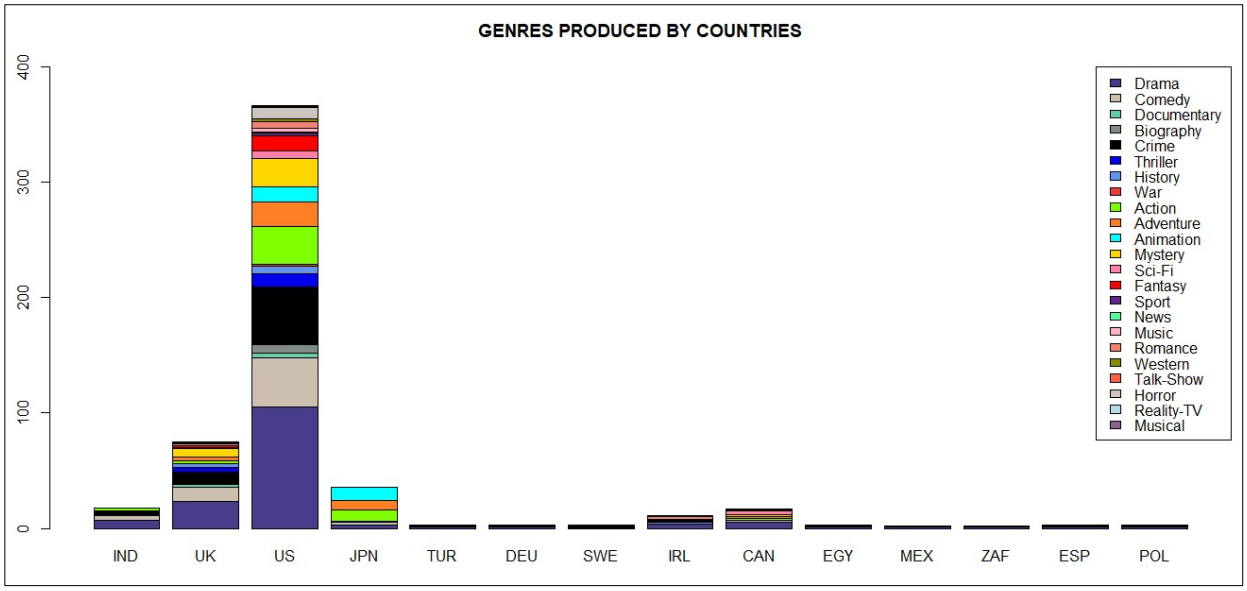
---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

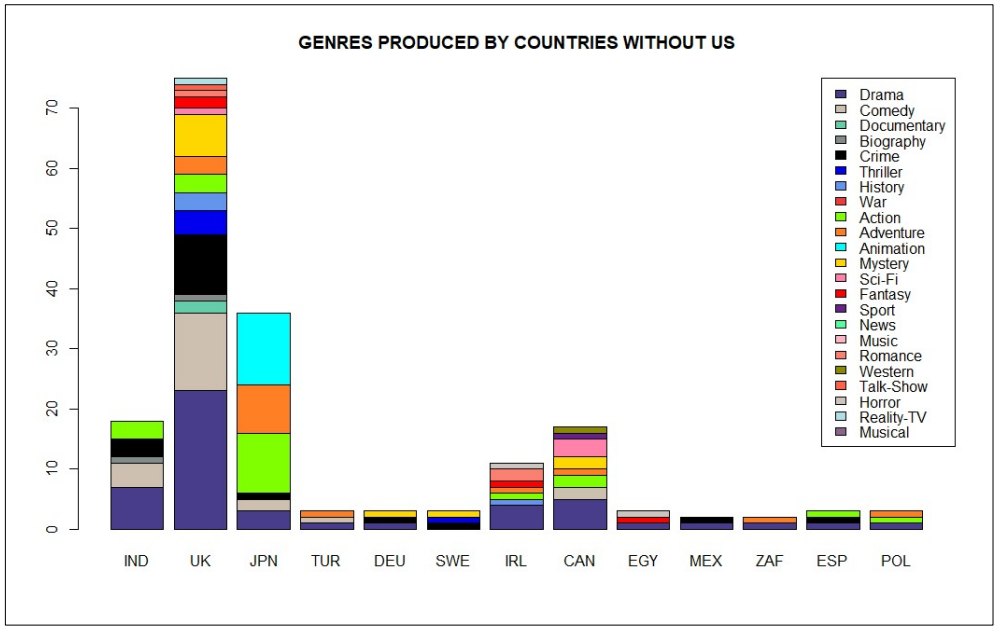
Residual standard error: 214700 on 212 degrees of freedom

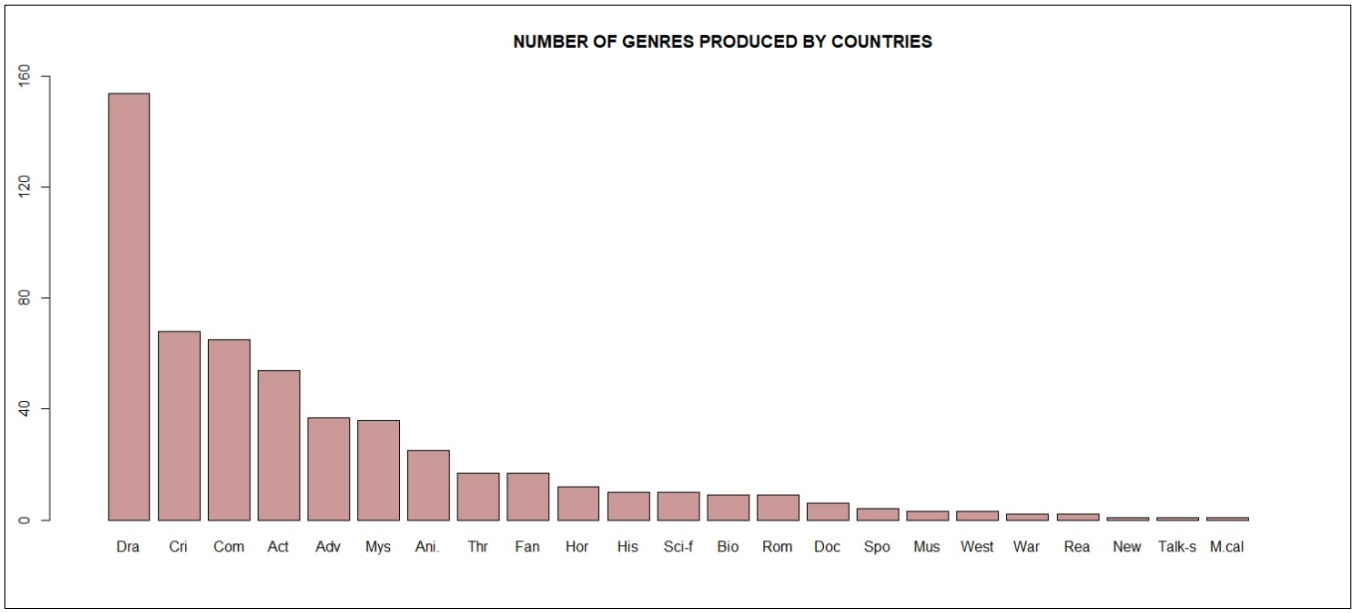
Multiple R-squared: 0.0826, Adjusted R-squared: 0.07828

F-statistics: 19.09 on 1 and 212 DF, p-value: 1.952e-05




According to the bar plots, it can be noticed with ease that The United States is ahead of producing TV series in every genre with a big difference from other countries. Behind the US, the second most productive country is the United Kingdom. When we observe these biggest producers, it can be seen that the first five preferred genres are the same. They produced drama, comedy, documentary, crime, biography, and thriller. Since the US is such an extensive data, it is difficult to see the changes in other data. To see and interpret all the data efficiently, we plotted another version of the bar graph in which we extracted the US data. After drawing the new diagram, it is observed that most countries chose to produce the genre of drama the most. Additionally, one of the things that stood out the most was Japan's most produced genre, 'Animation.' While japan creates a significant amount of animations, the animation is not preferred to be produced by any other country in our data.






| Genres | Number of productions | Sci-Fi | 10 |
|---|---|---|---|
| Drama | 154 | Biography | 9 |
| Crime | 68 | Romance | 9 |
| Comedy | 65 | Documentary | 6 |
| Action | 54 | Sport | 4 |
| Adventure | 37 | Music | 3 |
| Mystery | 36 | Western | 3 |
| Animation | 25 | War | 2 |
| Thriller | 17 | Reality-TV | 2 |
| Fantasy | 17 | News | 1 |
| Horror | 12 | Talk-Show | 1 |
| History | 10 | Musical | 1 |

Finally, we used a bar plot to show the number of series produced from each genre. As can be seen, the most created genre is drama, followed by crime and comedy. Moreover, this graph also confirms other information we interpreted in the previous two charts.

## 4. Discussion/Conclusion

With this project, we wanted to examine IMDb lists better and find out some answers for our research questions. We filtered our data as the 00s and 10s TV series mainly. First of all, we wanted to see the distribution of ratings over the years. We classified the ratings by

year and got the mean of each year. We visualize our data with two different graphs. One of them was a line graph which shows us the means of ratings for each year, and the second one was a histogram which gives us the information about the frequencies of ratings throughout the years. Mainly, we found out that there is no sustained behavior in the means of ratings, and the highest frequency of rating was 8.5-8.6 throughout the years. This shows us that we can not say that TV series are getting better ratings as time goes by. Secondly, we wanted to know if there is a relationship between ratings and vote numbers. We calculated the correlation, and we had a scatter plot drawn. It turns out that the correlation coefficient was minimal, and the scatter plot supports this result. Also, we build a linear model as "Votes = -1449011 + 192735 x Rating" with an error of 44113. So, we can interpret that the estimated value can vary 44113 votes from the actual observation. This linear model supports our result: there is no relationship between rates and votes. Furthermore, one can say that ratings do not increase while vote numbers are increasing. Thirdly, we tried to determine which type of TV series countries produce the most. We created a table with two factors: countries and genres. After creating our table and finding out which genres countries preferred, we drew our bar plot with all countries in the list. The USA is the most productive country with the UK and Japan following it. Furthermore, our last question tells us that the most preferable genre is drama, followed by crime and comedy.

There are some main comments which can be done to out IMDb data. Ratings of TV series are changing over the years. There is irrelevance between ratings and vote numbers. The most productive country in the TV series sector is the USA and the most preferable genre is drama.

## REFERENCES


1. *Running times*. IMDb Help Center. Retrieved from https://help.imdb.com/article/contribution/titles/runningtimes/G8UMSUYEDCN3M2K8?ref_=helpms_helpart_inline#

2. *Movies and Series Rankings (IMDB Data)*. Metabase. Retrieved from http://metabase.intellimenta.com/public/dashboard/eae564a4-d9a3-46b1-9cd4-1f95ab5b1b18?item_type=Series&years_%253E_%253F=2000&rating____=8&__votes____=50000

3. Wei, Y. (2006). *Colors in R* [PDF]. Department of Biostatistics Columbia University. from http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

4. *Strsplit Function - RDocumentation*. Rdocumentation.org. Retrieved from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/strsplit

5. *Colors in R - Easy Guides - Wiki - STHDA*. Sthda.com. Retrieved 30 January 2022, from http://www.sthda.com/english/wiki/colors-in-r

6. *Grep Function - RDocumentation*. Rdocumentation.org. Retrieved from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/grep

# APPENDIX

```
imdb_data <- read.csv("Series Rankings (IMDB Data).csv",
                      na.strings = c(NA,"-"))

imdb_data <- imdb_data[1:8]

# Some values cannot be changed to numeric as these char values have commas in them.
# We first omit the commas using gsub and then change the values to numeric.

imdb_data$Year <- gsub(",","",imdb_data$Year)
imdb_data$Year <- as.numeric(imdb_data$Year)

imdb_data$Votes <- gsub(",", "", imdb_data$Votes)
imdb_data$Votes <- as.numeric(imdb_data$Votes)

str(imdb_data)
```

```
'data.frame':    214 obs. of  8 variables:
 $ Title   : chr  "Aspirants" "Dhindora" "Planet Earth II" "Scam 1992: The Harshad Mehta
Story" ...
 $ Country : chr  "India" "India" "United Kingdom" "India" ...
 $ Year    : num  2021 2021 2016 2020 2006 ...
 $ Rating  : num  9.7 9.6 9.5 9.5 9.4 9.4 9.4 9.4 9.3 9.3 ...
 $ Votes   : num  289276 109717 107179 134387 175790 ...
 $ Director: logi  NA NA NA NA NA NA ...
 $ Genres  : chr  "Drama" "Comedy" "Documentary" "Biography,Crime,Drama" ...
 $ Runtime : int  45 25 298 54 538 49 330 594 59 23 ...
```

```
summary(imdb_data)
```

```
   Title              Country              Year          Rating          Votes
Length:214         Length:214         Min.   :2001   Min.   :8.100   Min.   :  50044
Class :character   Class :character   1st Qu.:2008   1st Qu.:8.300   1st Qu.:  79728
Mode  :character   Mode  :character   Median :2013   Median :8.500   Median : 115360
                                      Mean   :2012   Mean   :8.537   Mean   : 196440
                                      3rd Qu.:2017   3rd Qu.:8.700   3rd Qu.: 232310
                                      Max.   :2021   Max.   :9.700   Max.   :1917673
```

```
 Director          Genres             Runtime
 Mode:logical   Length:214         Min.   : 11.00
 NA's:214       Class :character   1st Qu.: 30.00
                Mode  :character   Median : 45.00
                                   Mean   : 81.11
                                   3rd Qu.: 60.00
                                   Max.   :594.00
                                   NA's   :1
```

# RESEARCH QUESTION 1: What is the distribution of ratings of tv series according to years?

```
rate_year <- table(imdb_data$Year, imdb_data$Rating)
unique_rates <- colnames(rate_year)
unique_years <- rownames(rate_year)
means <- c()

for (year in unique_years){
  sum_of_rating <- 0
  for (rate in unique_rates){
    current <- as.numeric(rate)*rate_year[year,][rate][[1]]
    sum_of_rating <- sum_of_rating + current
  }
  means <- c(means, sum_of_rating/sum(rate_year[year,]))
}

plot(unique_years,means,
     type = "o",
     pch = "*",
     xlim = c(2000,2021),
     xlab = "Years",
     ylab = "Rating Means",
     main = "DISTRIBUTION OF RATINGS OVER YEARS",
     col="purple")


hist(means,
     ylim = c(0,10),
     main = paste("HISTOGRAM OF RATING MEANS"),
     xlab = "Means of Rating",
     col = "pink")
```

# RESEARCH QUESTION 2: Is there any relation between ratings and votes? If yes, how can we describe this relation?

```
corimdb <- round(cor(imdb_data$Rating,imdb_data$Votes),4)

imdb_lm <- lm(Votes ~ Rating, data = imdb_data)

summary(imdb_lm)

plot(imdb_data$Rating,imdb_data$Votes,
     pch = 18,
     col = '#66CCCC',
```

```
    xlab = "RATE",
    ylab = "VOTE",
    main = "Relation between Ratings and Votes")

abline(imdb_lm,col= "#000033",lwd=2) #regression line

text(x = 8.5, y = 1500000, paste('Correlation :', corimdb))
```

# RESEARCH QUESTION 3: Which type of TV series do countries produce the most?

```
countries <- imdb_data$Country
genres <- strsplit(imdb_data$Genres, ",")

max_len = 0
for (item in 1:length(genres)){
  if (max_len < length(genres[[item]])){
    max_len <- length(genres[[item]])
  }
}

countries_genres <- data.frame()
countries_genres[1, 1:(max_len+1)] <- rep(NA, max_len+1)


for (item in 1:length(genres)){
  countries_genres[item, 1:(length(genres[[item]])+1)] <- c(countries[item], genres[[item]])
}

unique_genres <- array()
for (level in 1:length(genres)){
  for (genre in 1:length(genres[[level]])){
    if (genres[[level]][genre] %in% unique_genres == FALSE){
      unique_genres <- c(unique_genres, genres[[level]][genre])
    }
  }
}

unique_genres_clean <- unique_genres[!is.na(unique_genres)]
unique_countries <- unique(countries)

count_countries_genres <- matrix(0, length(unique_countries),
                                 length(unique_genres_clean),
                                 dimnames = list(unique_countries, unique_genres_clean))

for (row in 1:length(countries_genres[, 1])){
  for (col in 1:max_len){
```

```
    if (is.na(countries_genres[row, col+1]) == FALSE){
      count_countries_genres[countries_genres[row, 1], countries_genres[row, col+1]] <-
        count_countries_genres[countries_genres[row, 1], countries_genres[row, col+1]] +1
    }
  }
}
count_countries_genres
```

| | Drama | Comedy | Documentary | Biography | Crime | Thriller | History | War | Action | Adventure | Animation | Mystery | Sci-Fi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IND | 7 | 4 | 0 | 1 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| UK | 23 | 13 | 2 | 1 | 10 | 4 | 3 | 0 | 3 | 3 | 0 | 7 | 1 |
| US | 105 | 43 | 4 | 7 | 50 | 12 | 6 | 2 | 33 | 21 | 13 | 25 | 6 |
| JPN | 3 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 10 | 8 | 12 | 0 | 0 |
| TUR | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| DEU | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| SWE | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| IRL | 4 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| CAN | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 2 | 3 |
| EGY | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MEX | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ZAF | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ESP | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| POL | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |

| Fantasy | Sport | News | Music | Romance | Western | Talk-Show | Horror | Reality-TV | Musical |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 13 | 3 | 1 | 3 | 6 | 2 | 0 | 10 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

```
production <- sort(colSums(count_countries_genres), decreasing = T)

barplot(production,
        main = "NUMBER OF GENRES PRODUCED BY COUNTRIES",
        col = '#CC9999',
        xlim = c(0,30),
        ylim = c(0,160),
        yaxt='n')
axis(side = 2, at= seq(0, 160, by=40))

rownames(count_countries_genres) <-
c("IND","UK","US","JPN","TUR","DEU","SWE","IRL","CAN","EGY","MEX","ZAF","ES
P","POL")



barplot(t(count_countries_genres),
        main = "GENRES PRODUCED BY COUNTRIES",
        col = c("darkslateblue","antiquewhite3","aquamarine3","azure4","black",
                "blue2","cornflowerblue","brown2","chartreuse1","chocolate1","cyan",
                "gold","palevioletred1","red1","darkorchid4","seagreen1","pink1",
                "salmon","yellow4","tomato","seashell3","powderblue","plum4"),
        ylim = c(0,400))
legend("topright",
```

```
      unique_genres_clean,
      fill = c("darkslateblue","antiquewhite3","aquamarine3","azure4","black",
                              "blue2","cornflowerblue","brown2","chartreuse1","chocolate1","cyan",
                              "gold","palevioletred1","red1","darkorchid4","seagreen1","pink1",
                              "salmon","yellow4","tomato","seashell3","powderblue","plum4"))


barplot(t(count_countries_genres[rowSums(count_countries_genres) < 300,]),
       main = "GENRES PRODUCED BY COUNTRIES WITHOUT US",
       col = c("darkslateblue","antiquewhite3","aquamarine3","azure4","black",
             "blue2","cornflowerblue","brown2","chartreuse1","chocolate1","cyan",
             "gold","palevioletred1","red1","darkorchid4","seagreen1","pink1",
             "salmon","yellow4","tomato","seashell3","powderblue","plum4"),)

legend("topright",
      unique_genres_clean,
      fill = c("darkslateblue","antiquewhite3","aquamarine3","azure4","black",
             "blue2","cornflowerblue","brown2","chartreuse1","chocolate1","cyan",
             "gold","palevioletred1","red1","darkorchid4","seagreen1","pink1",
             "salmon","yellow4","tomato","seashell3","powderblue","plum4"))
```