**Data Analysis of Movies and TV Shows using IMDB and Netflix Dataset.**

By

**STANLEY KINGSTON**   **19BEC1254**

**GAURAV PRASANNA**  **19BEC1315**

**VIGNESH S 19BEC1391**

A project report submitted to

**Dr SATHIYA NARAYANAN S**

**SCHOOL OF ELECTRONICS ENGINEERING**

in partial fulfilment of the requirements for the course of

**CSE3505 – Foundations of Data Analytics**

in

**B.TECH. - Electronics and Communication**

****

**Vandalur – Kelambakkam Road**

**Chennai – 600127**

**DECEMBER 2021**

**BONAFIDE CERTIFICATE**

Certified that this project report e­­­­ntitled “**Data Analysis of Movies and TV Shows using IMDB and Netflix Dataset.”** is a bonafide work of **Stanley Kingston-19BEC1254, Gaurav Prasanna-19BEC1315 and Vignesh S-19BEC1391** who carried out the project work under my supervision and guidance for **CSE3505 - Foundations of Data Analytics.**

**Dr SATHIYA NARAYANAN S**

Assistant Professor (Senior Grade 1)

School of Electronics Engineering (SENSE),

Vellore Institute of Technology (VIT Chennai)

Chennai – 600 127.

**ABSTRACT**

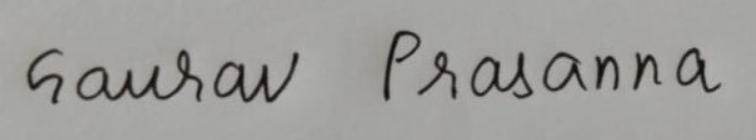
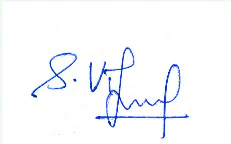
In this project, we have done exploratory data analysis on IMDB, Netflix and Over the Top platform datasets. A basic overview of how this works and its criteria, including its main assumptions and inferences are discussed. These tasks are done using various data analytics and visualizations tools such as RStudio, Excel and Tableau. Various pre-processing techniques and libraries have been used to implement this project in RStudio. This will enable the users and audience to have better understanding on which OTT platform to invest in for better content after looking at the visuals or the plots provided. The obtained results have shown promising potential for acting as the features which can be used as a recommendation system through collaborative filtering and Reinforcement Learning

**ACKNOWLEDGEMENT**

We sincerely thank and express our gratitude to our respected teacher **Dr Sathiya Narayanan S, SENSE Assistant Professor, VIT Chennai**, for his guidance and constant support throughout the course of this project work.

We would also like to extend our heartful gratitude to all the open-source programming communities for providing us with resources.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

 ****

**STANLEY KINGSTON GAURAV PRASANNA VIGNESH S**

**19BEC1254 19BEC1315 19BEC1391**

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO** |  | **TOPIC** | **PAGE NO.** |
| **1** |  | **INTRODUCTION** | **6** |
|  | **1.1** | **BACKGROUND AND MOTIVATION** | **6** |
|  | **1.2** | **PROBLEM STATEMENT AND OBJECTIVES** | **6** |
| **2** |  |  | **7-21** |
|  | **2.1** | **PROPOSED/IMPLEMENTED METHOD** | **7** |
|  | **2.2** | **ADVANTAGES** | **8** |
|  | **2.3** | **CHALLENGES FACED** | **9** |
|  | **2.4** | **CODE** | **9** |
| **3** |  | **ANALYSIS OF RESULTS** | **23** |
|  | **3.1** | **MAIN RESULTS** |  |
|  | **3.2** | **INFERENCE** |  |
| **4** |  | **CONCLUSION AND FUTURE WORK** | **23** |
|  | **4.1** | **CONCLUSION** | **23** |
|  | **4.2** | **FUTURE WORK** |  |
| **5** |  | **REFERENCES** | **24** |

**CHAPTER 1**

**INTRODUCTION**

**1.1 BACKGROUND AND MOTIVATION**

According to a recent survey most of the Indian population are not keen enough to spend lot of money on OTT subscriptions and this issue is primarily addressed in our project and also, we wanted to study on how the evolution of technology has influenced the Film Industry and this brings us to our major motivation behind choosing this project.

**1.2 PROBLEM STATEMENT AND OBJECTIVES**

1. To perform Exploratory Data Analysis on IMDB and Netflix dataset.
2. To understand and visualize various plots to see the evolution of Movies and TV shows through the years.
3. To know which OTT platform provides more no of quality content in terms of IMDB rating.
4. Having understanding about which country makes what kind of genres and also the most upvoted writers in their respective country.

**CHAPTER 2**

* 1. **PROPOSED/IMPLEMENTED METHOD**

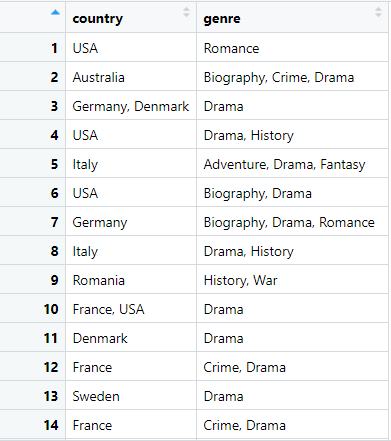
Using the dataset downloaded from Kaggle, we perform Exploratory Data Analysis for the IMDB dataset using RStudio for Data Analysis and data preprocessing. Using libraries such as ggplot, dplyr and few other plotting libraries were used for plotting and preprocessing in RStudio. Microsoft Excel and Tableau were used for further data visualization and make more inferences from the respective software.

**2.2 ADVANTAGES**

* Helps to know which OTT platform to spend on.
* Knowing more about World Cinema for the past 110 years.
* Helps identifying the factors affecting the Film Industry.

**2.3 CHALLENGES FACED**

One of the main challenges faced was during the pre-processing step. In the IMDB movies dataset we were not able to have the count of the individual countries and genres as there were multiple countries contributing to multiple genres and vice versa. So, in order to tackle this problem, we use the concepts of for loop and if statements.

(a) (b)

Figure 1: (a) The filtered original dataset (b) Pre-processed dataset

**2.4 CODE**

**Count of the no of countries and Genres:**

rm(list=ls())

setwd(‘C:\\Users\\VIGNESH S\\Desktop\\Fall 21-22\\J Components\\Data Analytics JC\\Datasets’)

library(plyr)

library(dplyr)

library(stringr)

a <- read.csv(“IMDb movies.csv”, na.strings=c(“”,”NA”))

names(a)

a <- select(a,”country”,”genre”)

a<- na.omit(a) #Removing all the rows having NA

#..........................Unique Countries...............................

ux<-unique(a$country)

x<-unlist(strsplit(ux, “,”))

ux<-unique(x)

for (j in 1:length(ux))

{

ux[j]<-gsub(“ “,””,ux[j]) # Removing all white spaces in the string.

}

unique\_countries<-unique(ux) # 194 individual countries.

#............................Unique Genres...............................

uy<-unique(a$genre)

y<-unlist(strsplit(uy, “,”))

uy<-unique(y)

for (I in 1:length(uy))

{

uy[i]<-gsub(“ “,””,uy[i]) # Removing all white spaces in the string.

}

unique\_genres<-unique(uy) # 25 individual genres.

**Most voted Writer:**

rm(list=ls())

setwd('C:\\Users\\VIGNESH S\\Desktop\\Fall 21-22\\J Components\\Data Analytics JC\\Datasets')

library(plyr)

library(dplyr)

library(stringr)

library(ggplot2)

a<-read.csv("IMDb movies.csv")

names(a)

a <- select(a,"country","title","writer","avg\_vote","votes")

df1 <- a %>% group\_by(country) %>% filter(avg\_vote>9.0) %>% arrange(country)

arr<- unique(df1[[1]])

#write.csv(df1,"Task2.xls")

df2<- count(df1)

write.csv(df2,"writers.csv")

#barplot

options(repr.plot.width=5, repr.plot.height=2)

ggplot(df2, aes(x = country, y = n, main="Most voted writers in each country")) +

geom\_bar(stat = "identity") +

coord\_flip() + scale\_y\_continuous(name="No of writers") +

scale\_x\_discrete(name="Countries") +

theme(axis.text.x = element\_text(face="bold", color="#008000",

size=12, angle=0),

axis.text.y = element\_text(face="bold", color="#008000",

size=12, angle=0))

#pie-chart

library(ggplot2)

df3<-select(df1,"country","writer","votes")

df3 <- df3 %>% mutate(writer = replace(writer, writer =="",NA))

df3<-na.omit(df3)

arr2=c()

#write.csv(df3,"Most voted writer.csv", row.names = FALSE)

df4a<-filter(df3,country=="USA")

df4b<-filter(df3,country=="India")

df4c<-filter(df3,country=="Bangladesh")

ba<-ggplot(data = df4a, aes(x = "", y = votes, fill = writer)) +geom\_bar(stat = "identity") +coord\_polar("y")

bb<-ggplot(data = df4b, aes(x = "", y = votes, fill = writer)) +geom\_bar(stat = "identity") +coord\_polar("y")

bc<-ggplot(data = df4c, aes(x = "", y = votes, fill = writer)) +geom\_bar(stat = "identity") +coord\_polar("y")

library(ggpubr)

ggarrange(ba, bb, bc + rremove("x.text"),

labels = c("USA", "India", "Bangladesh"),

ncol = 2, nrow = 2)

**Country wise total movie created and Rating improvement:**

rm(list=ls())

setwd('C:\\Users\\VIGNESH S\\Desktop\\Fall 21-22\\J Components\\Data Analytics JC\\Datasets')

library(plyr)

library(dplyr)

library(stringr)

library(quanteda)

a<-read.csv("IMDb movies.csv")

names(a)

a <- select(a,"country","year","avg\_vote")

a$year<-as.numeric(a$year)

s1<-table(a$country)

df1<-as.data.frame(s1)

df2<-filter(df1,Freq>2000)

df3<-filter(a,country==df2$Var1)

df3<-arrange(df3,country)

s2<-max(df3$year)

df4<-filter(df3,df3$year>1980)

df4<-select(df4,"country","year","avg\_vote")

df4$year<-as.numeric(df4$year)

df5<-filter(a,a$year>1980)

s4<-table(df5$year)

all\_country\_movies<-as.data.frame(s4)

all\_country\_movies$category<-"All Countries"

s3<-table(df4$year)

Total\_movies\_created<-as.data.frame(s3)

Total\_movies\_created$category<-"Top 6 countries"

#..........................Total\_movies\_created plotting..................

library("ggplot2")

ggplot(Total\_movies\_created, color="red",aes(x=Var1,y=Freq))+geom\_point()

plot(s3,type = "o", col = "blue", xlab = "Year", ylab = "Total movies",

main = "Total movies created in past 40 years by the top 6 countries ")

plot(s4,type = "o", col = "blue", xlab = "Year", ylab = "Total movies",

main = "Total movies created in past 40 years by all countries ")

final\_total\_movies<-rbind(all\_country\_movies,Total\_movies\_created)

ggplot(final\_total\_movies, aes(x=Var1, y=Freq, group=category, colour = category)) +geom\_line(size=0.75)+ggtitle("All countries vs Top 5 Countries - movie creation" )

#.................Country wise Total\_movies\_created plotting..............

df1a<-filter(df4,country=="France")

df1b<-filter(df4,country=="India")

df1c<-filter(df4,country=="Italy")

df1d<-filter(df4,country=="Japan")

df1e<-filter(df4,country=="UK")

df1f<-filter(df4,country=="USA")

df1a<- df1a %>% group\_by(country,year) %>% summarise(avg\_vote\_calc = mean(avg\_vote, na.rm=TRUE))

df1b<- df1b %>% group\_by(country,year) %>% summarise(avg\_vote\_calc = mean(avg\_vote, na.rm=TRUE))

df1c<- df1c %>% group\_by(country,year) %>% summarise(avg\_vote\_calc = mean(avg\_vote, na.rm=TRUE))

df1d<- df1d %>% group\_by(country,year) %>% summarise(avg\_vote\_calc = mean(avg\_vote, na.rm=TRUE))

df1e<- df1e %>% group\_by(country,year) %>% summarise(avg\_vote\_calc = mean(avg\_vote, na.rm=TRUE))

df1f<- df1f %>% group\_by(country,year) %>% summarise(avg\_vote\_calc = mean(avg\_vote, na.rm=TRUE))

rating\_improvemnet<-rbind(df1a,df1b,df1c,df1d,df1e,df1f)

ggplot(rating\_improvemnet, aes(year,avg\_vote\_calc)) + geom\_line(aes(colour = country))

ggplot(rating\_improvemnet, aes(x=year, y=avg\_vote\_calc, group=country, colour = country)) +

geom\_line(size=0.75)

par(mfrow=c(3,2))

plot(df1a$year, df1a$avg\_vote\_calc,pch = 20,col = "dark green",main = "France\_Rating\_analysis",xlab = "Year",ylab = "avg\_vote")

abline(lm(df1a$avg\_vote\_calc ~ df1a$year))

plot(df1b$year, df1b$avg\_vote\_calc,pch = 20,col = "blue",main = "India\_Rating\_analysis",xlab = "Year",ylab = "avg\_vote")

abline(lm(df1b$avg\_vote\_calc ~ df1b$year))

plot(df1c$year, df1c$avg\_vote\_calc,pch = 20,col = "red",main = "Italy\_Rating\_analysis",xlab = "Year",ylab = "avg\_vote")

abline(lm(df1c$avg\_vote\_calc ~ df1c$year))

plot(df1d$year, df1d$avg\_vote\_calc,pch = 20,col = "orange",main = "Japan\_Rating\_analysis",xlab = "Year",ylab = "avg\_vote")

abline(lm(df1d$avg\_vote\_calc ~ df1d$year))

plot(df1e$year, df1e$avg\_vote\_calc,pch = 20,col = "black",main = "UK\_Rating\_analysis",xlab = "Year",ylab = "avg\_vote")

abline(lm(df1e$avg\_vote\_calc ~ df1e$year))

plot(df1f$year, df1f$avg\_vote\_calc,pch = 20,col = "purple",main = "USA\_Rating\_analysis",xlab = "Year",ylab = "avg\_vote")

abline(lm(df1f$avg\_vote\_calc ~ df1f$year))

**Count of Movies and TV Shows:**

rm(list=ls())

setwd("G:\\5th Sem\\Data Analytics\\Project Files\\Imdb")

setwd("G:\\5th Sem\\Data Analytics\\Project Files\\netflix\_titles.csv")

library(dplyr)

library(plyr)

netflix\_data = read.csv("netflix\_titles.csv")

summary(netflix\_data)

df2 <- select(netflix\_data, c(type ,title, rating))

x = sum(is.na(df2$title))

x1 = sum(is.na(df2$rating))

df3 <- mutate\_all(df2, list(~na\_if(.,"")))

y = unique(df3$rating)

length(y)

rating\_count <- table(df3$type, df3$rating)

barplot(rating\_count, main="Plot between No Of Movies and Shows with different Rating",

xlab = "Movie and TV Ratings" , beside = TRUE, col = c("yellow", "red"),

legend = rownames(rating\_count))

**Countries producing the most number of movies:**

rm(list=ls())

setwd("G:\\5th Sem\\Data Analytics\\Project Files\\Imdb")

library(dplyr)

library(plyr)

library(stringr)

library(ggplot2)

library(plotrix)

df1 <- read.csv("IMDb movies.csv")

df2 <- select(df1, c(title, genre, year, country))

s1<-table(df2$country)

x1 <- as.data.frame(s1)

x2 <- filter(x1, Freq > 2000)

pie3D(x2$Freq, labels = x2$Var1, explode = 0.1, main = "Pie chart showing countries creating

maximum contents")

**Count of OTT platforms with good content:**

rm(list=ls())

setwd("G:\\5th Sem\\Data Analytics\\Project Files\\Imdb")

data = read.csv("OTT-Movies.csv")

library(dplyr)

library(plyr)

df1 <- select(data, c("Title", "IMDb", "Netflix", "Hulu", "Prime.Video", "Disney."))

df2 <- df1 %>% filter(IMDb >= "8.0/10")

write.csv(df2, "filteredOTT.csv", row.names = FALSE)

**USA vs World Gross income:**

rm(list = ls())

setwd("E:/collage/5th sem/foundation of data analytics/project/dataset")

getwd()

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_import\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

data=read.csv("task7")

data=subset(data,select=-c(X))

library(dplyr)

library(ggplot2)

#\_\_\_\_\_\_\_\_\_\_\_\_\_finding the mean world wide and us gross \_\_\_\_\_\_\_\_\_\_\_\_\_\_

X=data%>%group\_by(year)%>% summarise(mean\_worlwide\_gross\_income = mean(worlwide\_gross\_income),mean\_usa\_gross\_income = mean(usa\_gross\_income))

dim(X)

D=data.frame(year=NA,type=NA,value=NA)

for (i in 1:96){

for(j in 1:2){

f=data.frame(year=X[[1]][i],type=colnames(X[j+1]),value=X[[j+1]][i])

D=rbind(f,D)

}

}

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_piloting comment\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ggplot(D, aes(x=year, y=value, group=type)) +

geom\_line(aes( color=type))+

geom\_point(aes(color=type))+

theme(legend.position="top")

**Genre Analysis:**

rm(list = ls())

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_import\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

setwd("E:/collage/5th sem/foundation of data analytics/project/dataset")

getwd()

library(dplyr)

library(ggplot2)

X=read.csv("yearvsgener.csv")

# \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_filtering the comedy and drama \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

C=X%>%filter(gener!="Drama")%>%filter( gener!="Comedy" )

ggplot(X, aes(x=year, y=total, group=gener)) +

geom\_line(aes( color=gener))+

geom\_point(aes(color=gener))+

theme(legend.position="top")

ggplot(C, aes(x=year, y=total, group=gener)) +

geom\_line(aes( color=gener))+

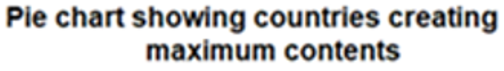
geom\_point(aes(color=gener))+

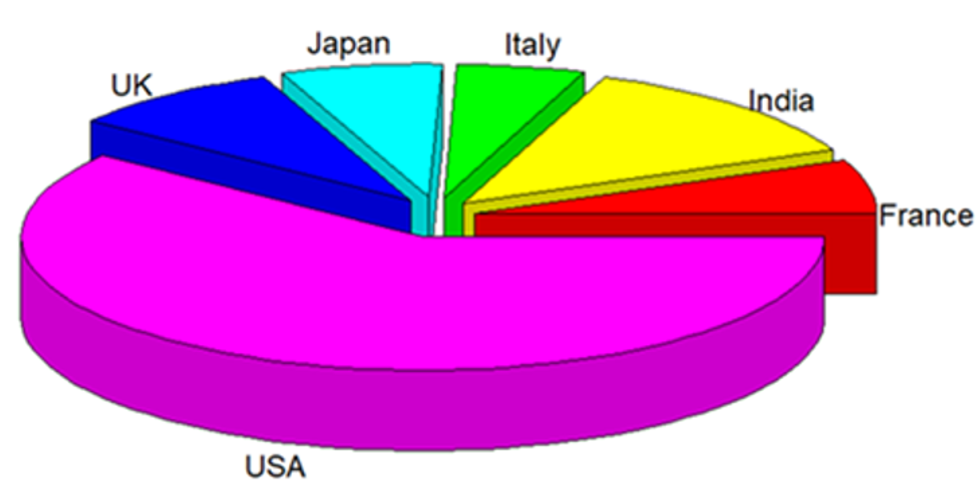
theme(legend.position="top")

**CHAPTER 3**

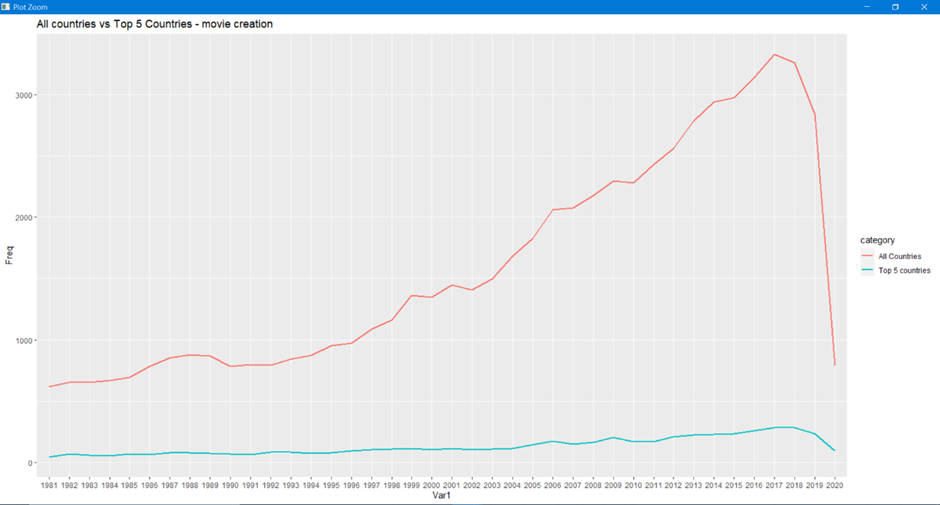
**ANALYSIS OF RESULTS**

**3.1 MAIN RESULTS**

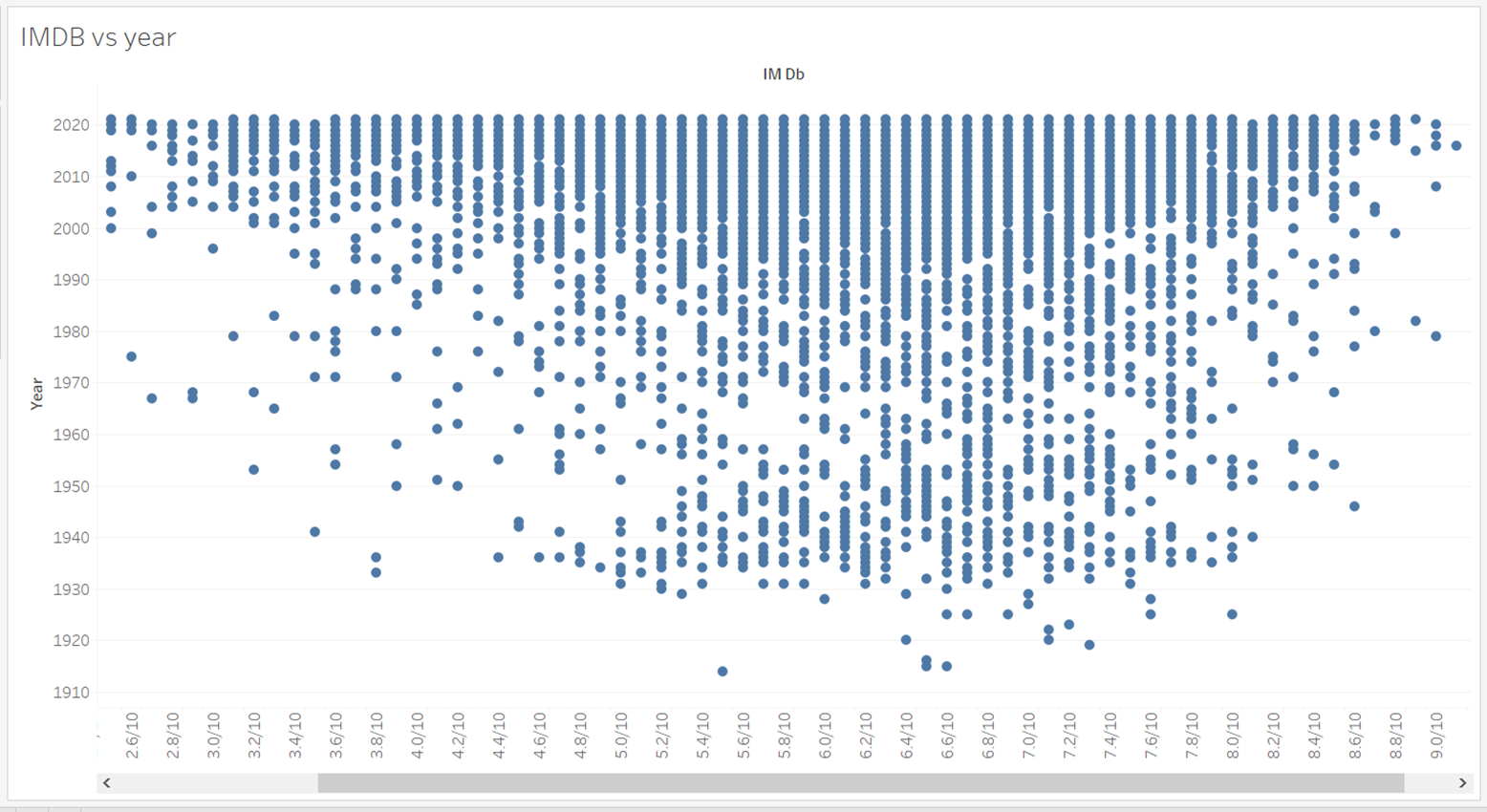




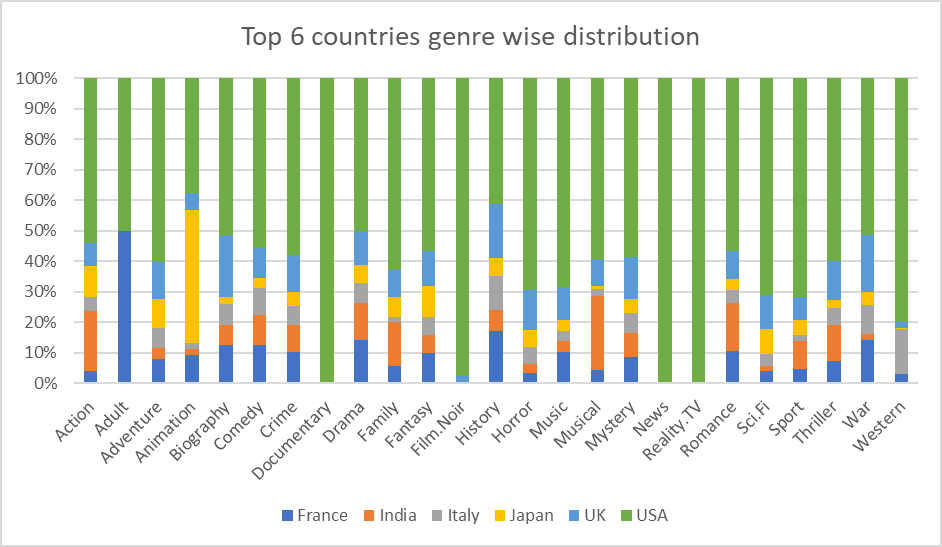
**Figure 2:** Countries producing the most content in recent times



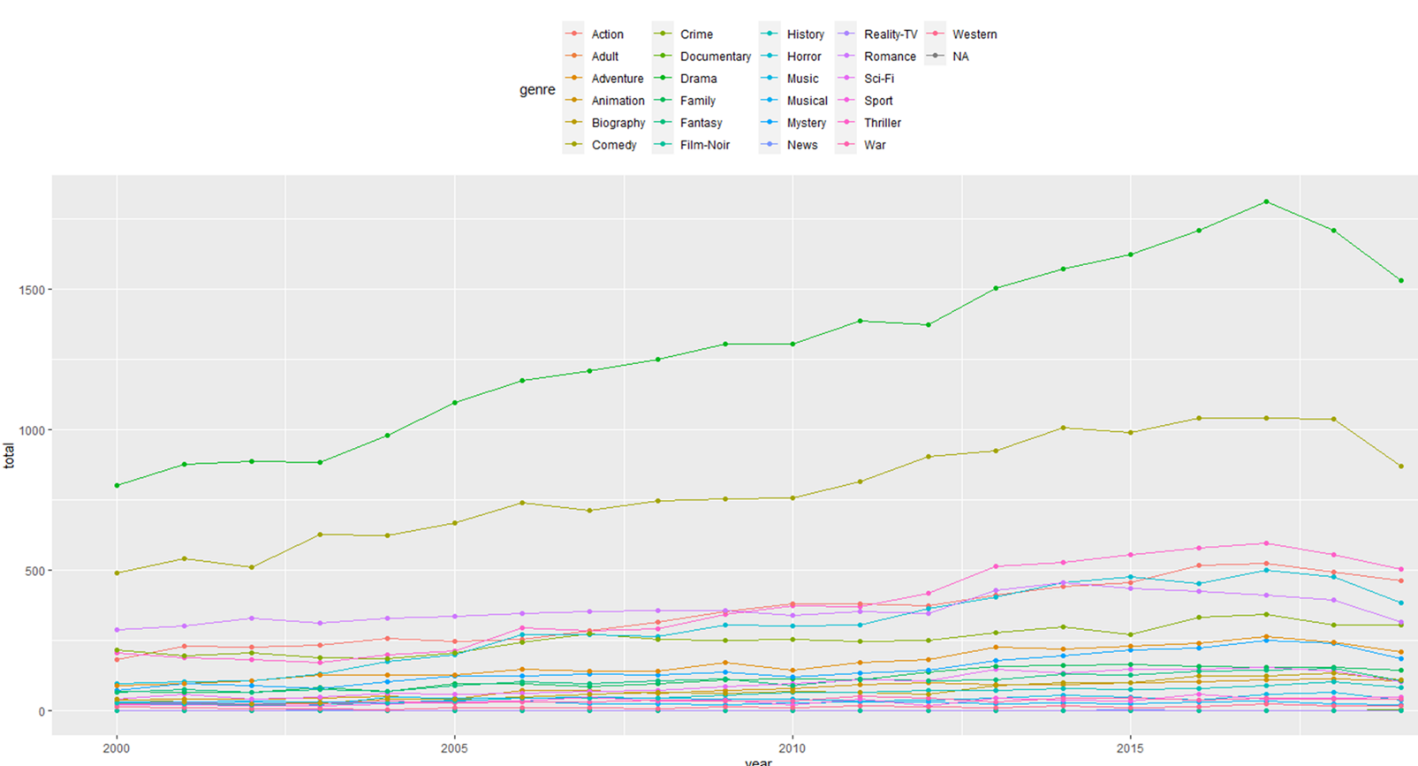
**Figure 3:** Comparing the total movies created by the whole world vs top 6 contributors for the past 40 years.



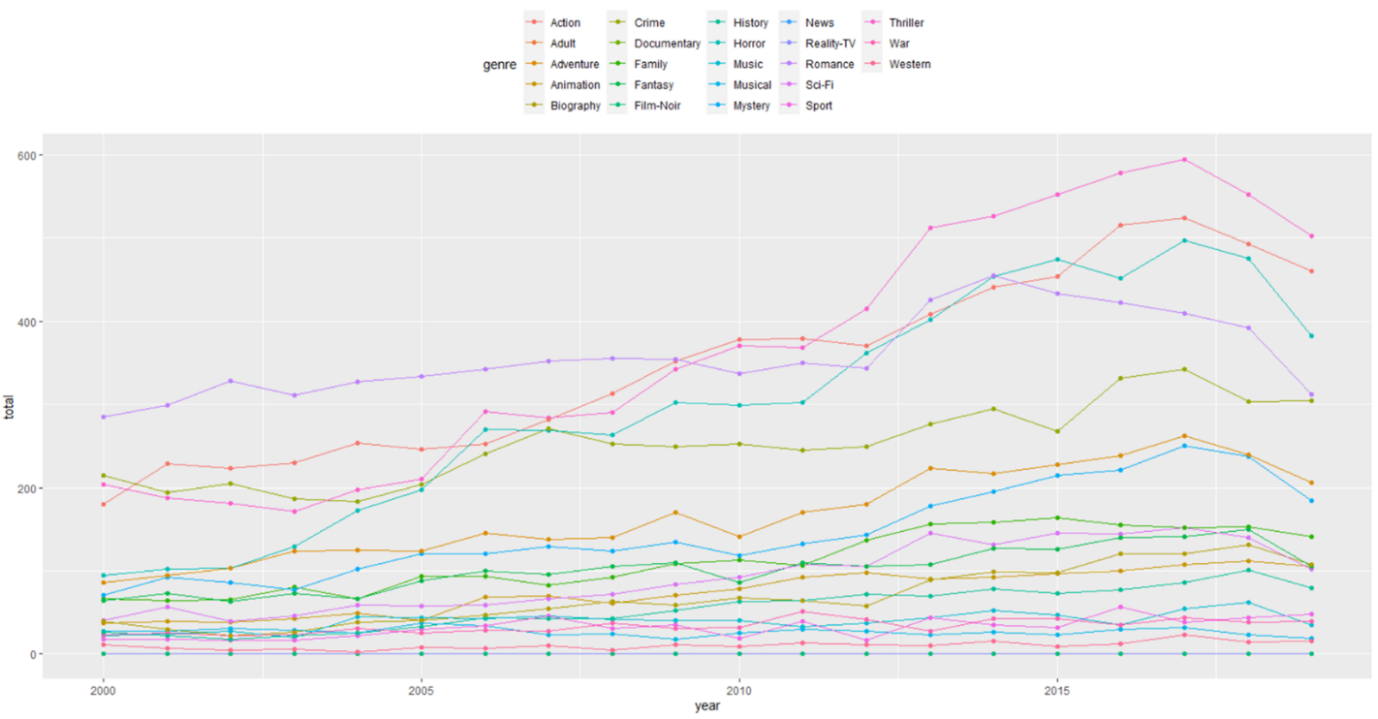
**Figure 4**: Evolution of Movie and Film industry throughout the years

****

**Figure 5:** Contribution by the top 6 countries towards each genre



(a) Plotting All 25 genres

****

(b) Plotting all 23 genres except comedy and drama

**Figure 6:** Comparing the total annual grossing of all genre

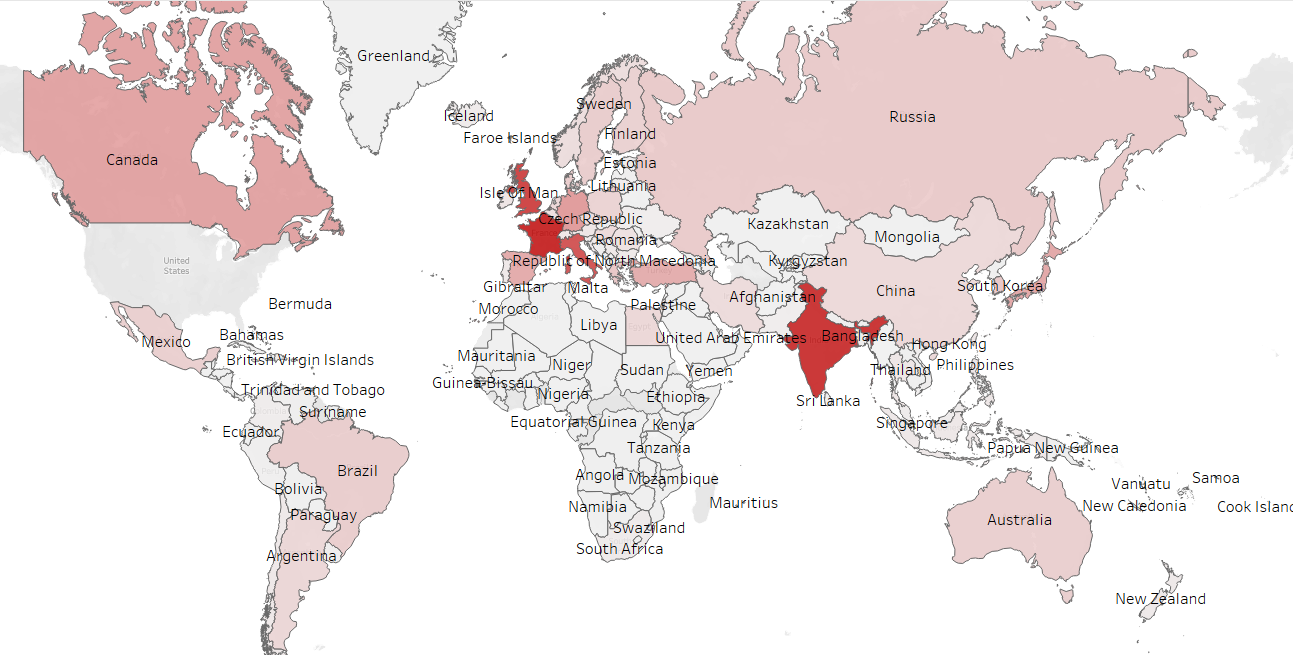
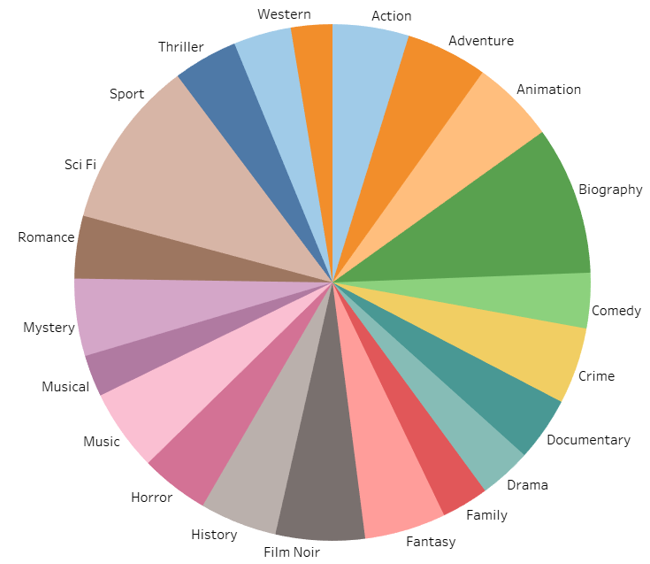
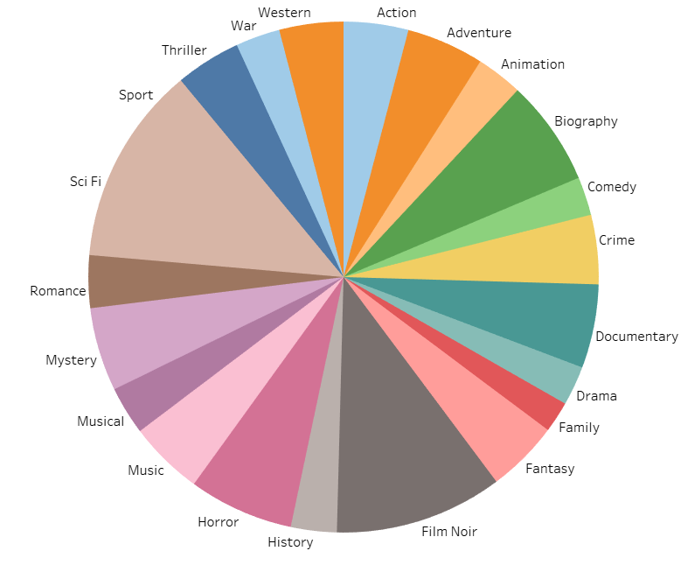


Figure 7: Highlighting the countries with most number of comedy content movies

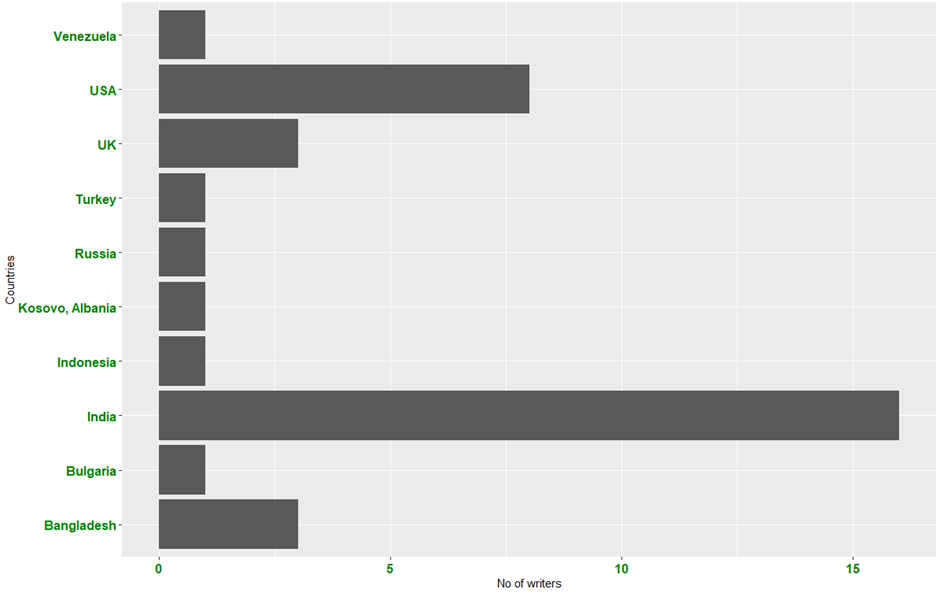


Figure 8: Japan highlighted for producing most number of movies in Animation genre

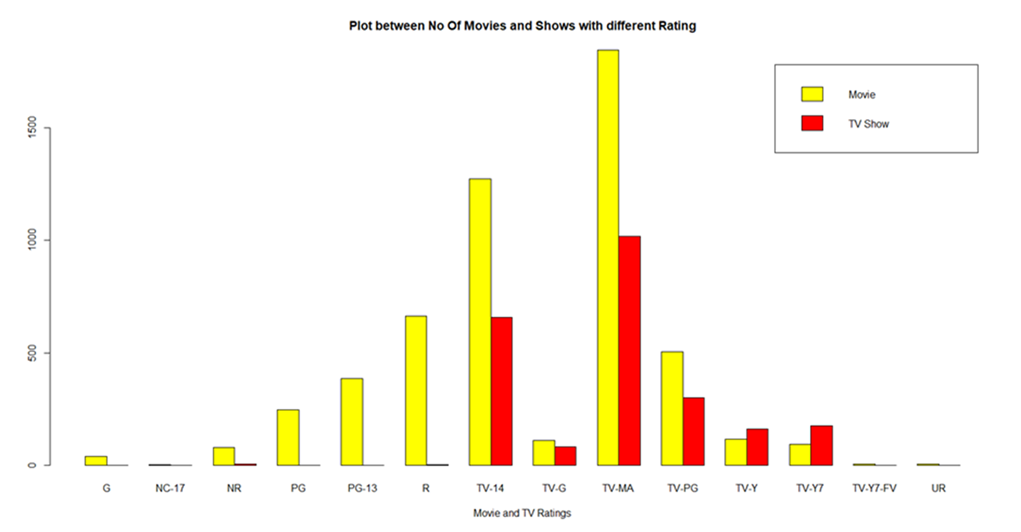


1. (b)

Figure 9: Plots comparing the most popular genres in (a) USA (b) World wide

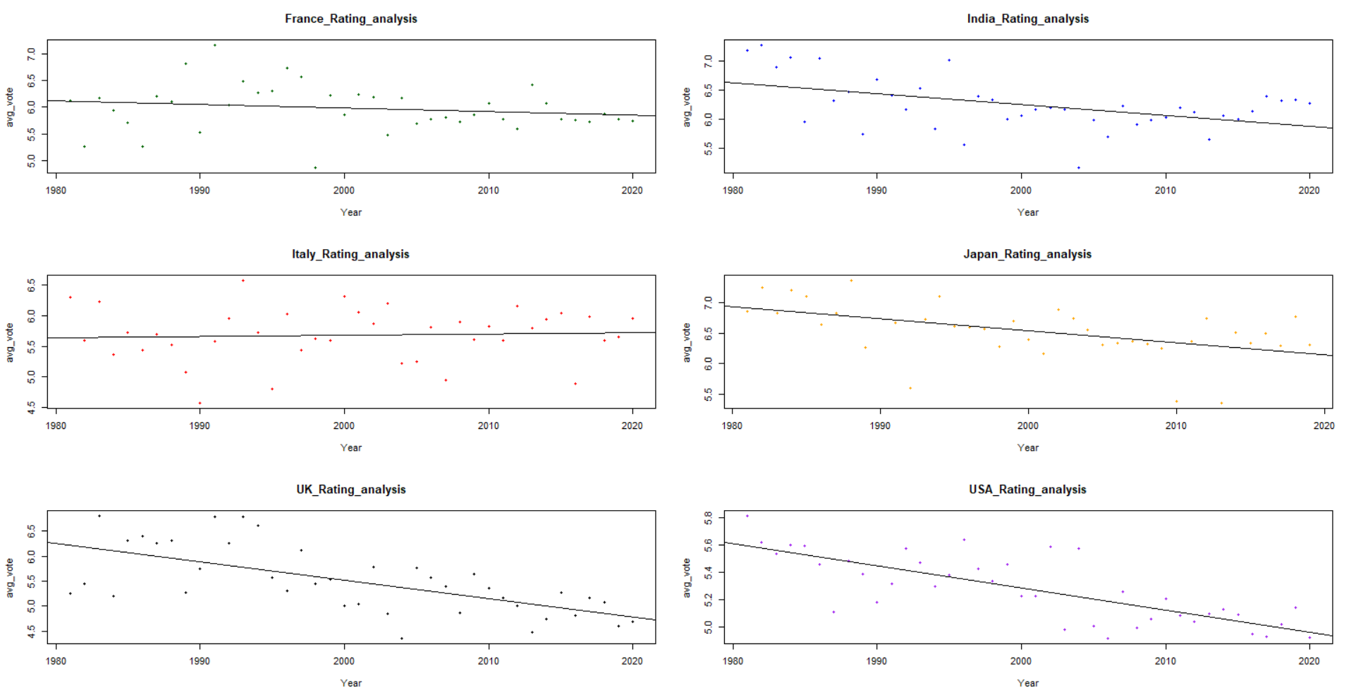
****

**Figure 10:** Count of most voted writer across the world



**Figure 11:** Count of Movies and TV Shows rating in Netflix

**Figure 12**: Count of OTT platforms with content with IMDB rating greater than 8.0



**Figure 13:** Rating Improvement for the top 6 contributing countries from 1980-2020

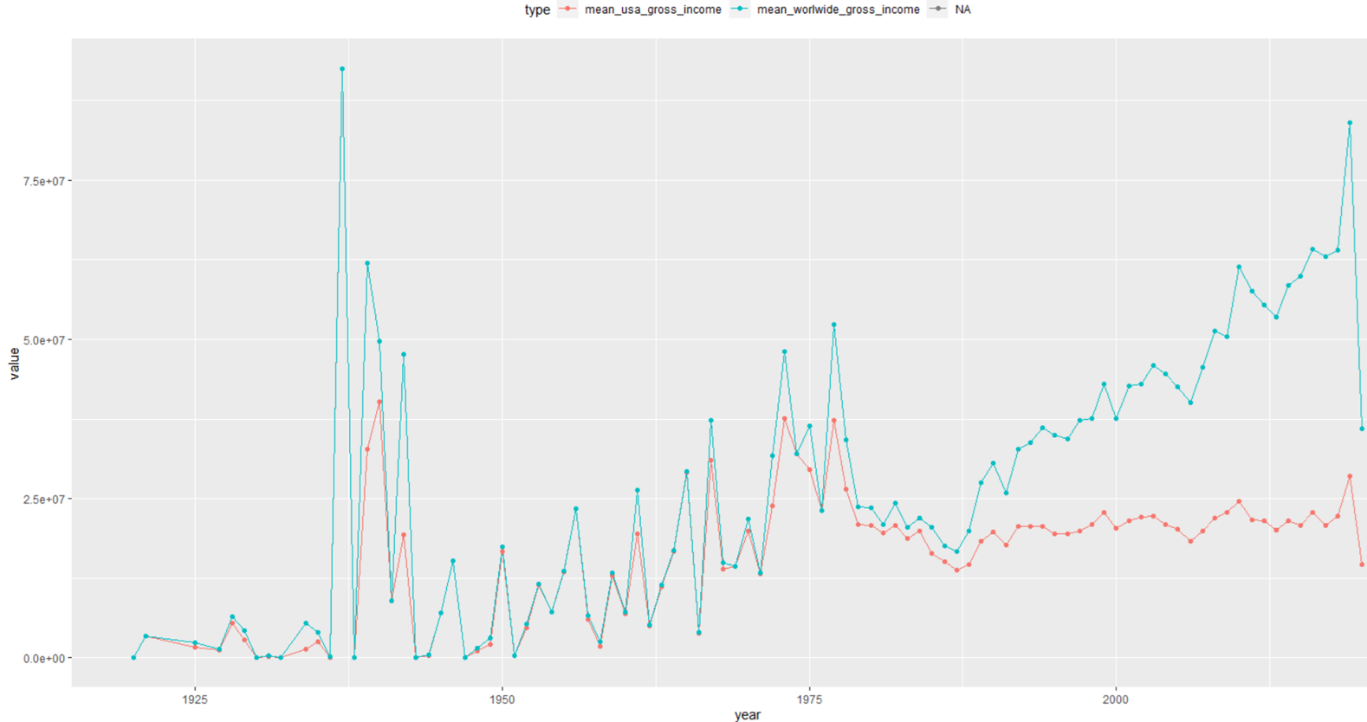


Figure 14: Comparing the annual world-wide grossing and USA grossing

**3.2 INFERENCES**

Our Exploratory analysis in the IMDB and Netflix datasets can be categorized into 3 groups: Genres, movie count and Grossing and Rating analysis. Therefore, from the following visualizations we can infer that, **Fig2**. among all the countries mentioned in the IMDB dataset, we are able to see these six countries that is USA, UK, Japan, Italy, India and France are the ones making maximum no of films since the earliest days, and USA accounts to almost more than 50% of the films made, making it the largest country in the world with respect to Production and creation of Movies and films. This provides us enough evidence to preform most of the visuals for the top 6 contributing countries for better understanding. **Fig3**. Comparing the total movies created by the whole world vs top 6 contributors for the past 40 years. We can also observe the drastic decline in the movie count from 2018 to 2020. The pandemic affected the Film industry Soo hard that the total movies created in 2020 is almost equal to that of movies created in 1980.

Whereas in case of **Fig5** to **Fig9.** Explain more about the Genre analysis. **Fig5** was plotted using Excel’s Visualization tool, we were able to visualize the highest contributor towards each genre in the dataset. Which in this case is USA for most of the genres except for Japan which contributes around 43.5% followed by USA in Animation category. Which is visualized in **Fig8** using Tableau. **Fig7** Compares the total annual grossing of all genre in which we could see that the entire box office is mostly contributed by comedy and drama content. But the horror content is having rapid expansion. But there is a great dip in the entire grossing in 2019. But in the case of **Fig9** The plots are representing popularity of the of all genre for both in USA and worldwide comparatively. Frow this plot is in USA documentary and horror Content are preferred by the audience. But the rest of the world prefer biography and animation more.

The final part of our EDA **Fig10** to **Fig14.** represents the Rating analysis. By filtering out the movies having IMDB rating >= 9.0, **Fig10.** helps us to clearly visualize the count of the most voted writers in each country. **Fig11** and **Fig 12** gives us a deep understanding of each OTT platform by providing the visuals of different movies and TV shows content based on ratings and IMDB scores. And by applying the abline() function on the IMDb movies dataset to plot the line of best fit to the cluster if points referring to the average votes for each year. In can be observed from **Fig13** that Italy’s performance towards rating is much better when compared to that of the UK’s plot where it drastically decreases from 6.3 to 4.6 over 40 years. Finally with the help of all out plots and results we can say that the World Cinema was dominated by the USA until 1972, but after that period around 194 countries started to significantly contribute in the field of the Film industry pushing the USA from the only contributor to one of the contributors to the World Cinema. This can be clearly visualized in **Fig14.**

**CHAPTER 4**

**CONCLUSION AND FUTURE WORK**

**4.1 CONCLUSION**

So, from the above plots we have performed Exploratory data analysis on IMDB and Netflix dataset, which had wide variety of columns giving us lot of insights about what content people prefer to watch in which region, and the most preferred genres and most voted writers in their region and also have visualized the various OTT platform for their performance with IMDB ratings and from the plots we have concluded that “Netflix” is the currently best OTT platform in terms of producing more number of good quality content as compared to Amazon prime, Hulu or Disney+. Apart from the inbuilt plotting and various plotting library that is used in RStudio we have also used Microsoft Excel and Tableau; the data pre-processing was done in R and then later the tools mentioned above were used for plotting and data visualization.

**4.2 RECOMMNDATIONS FOR FUTURE WORK**

1. In this project we have analysed and visualized the data from IMDB and Netflix dataset, so from this we can expand the base of our project to not just data visualization but to make movie or series recommender systems.
2. Collaborative filtering (CF) is a technique used by recommender systems. Collaborative filtering has two senses, a narrow one and a more general one. In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users.
3. In our current data we have wide variety of information ranging from most voted writers to the watch genres and also ratings for various genres, so keeping in mind with all this data, we can make an application where in a user could enter his/her preferences and the application could recommend a movie or tv show of their liking.
4. The application can also store the previously entered data and when the user opens the application to recommend new content based on the previously known data, this task can be achieved by applying Reinforcement Learning.

**REFERENCES**

List all sources you referred and make sure you cite it in appropriate place(s) in the text.