**R PROJECT**

**ANALYSIS OF SERIES DATASET**

**PROVIDED BY IMDB**

Komal Jolaniya Kalal

20/49035

Hansraj Sinwar

20/49009

The dataset consist 2000 rows representing information related to series aired on television

**Columns:**

1. Poster Link
2. Series Name
3. Runtime of series
4. Certificate
5. Average runtime of episodes
6. Genre
7. IMDB rating
8. Cast
9. Votes

**Information we are trying to extract :**

* Plot the votes gained by show as bar graph

* Plot the ratings gained by shows

* Number of series corresponding to each certificate

* Series which has rating above 9

* Series which has rating between 8 and 9

* Find out series which went underrated and were not appreciated much as compared to their ratings

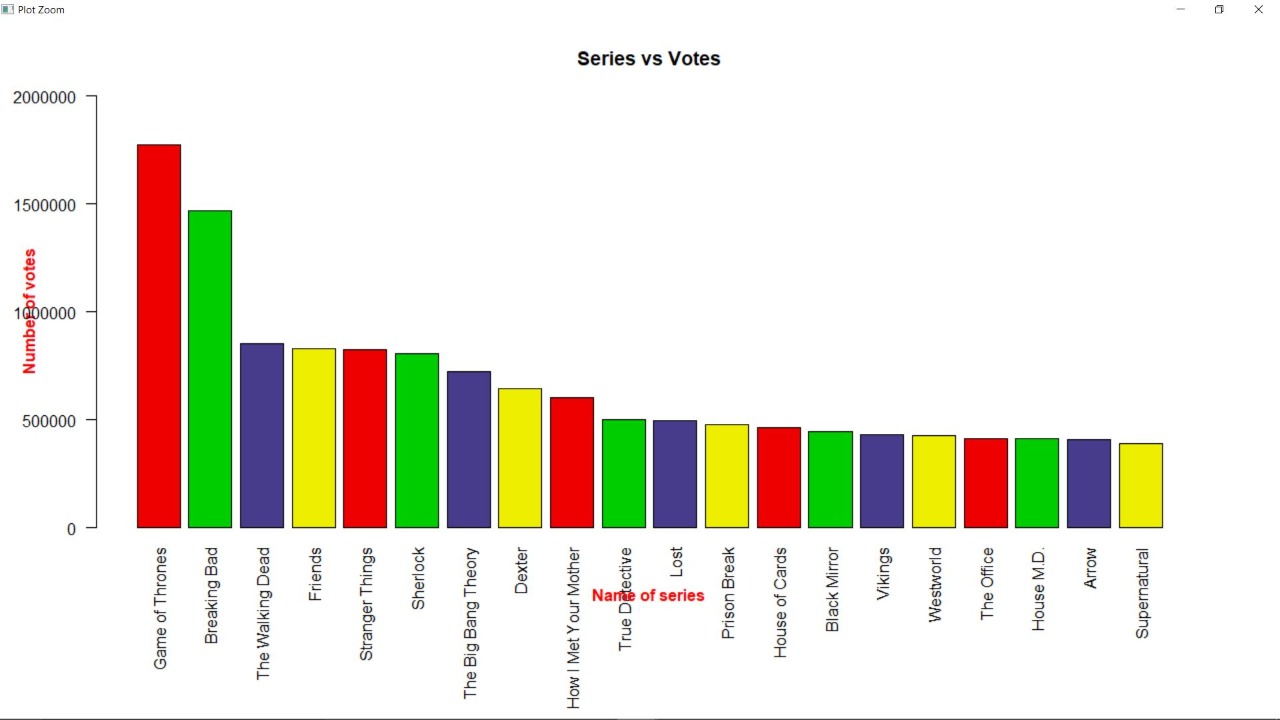
**R SCRIPT**

df<-read.csv("D:/MyData/4th Sem/R programming/archive/series\_data.csv")

par(mar=c(10,5,4,4))

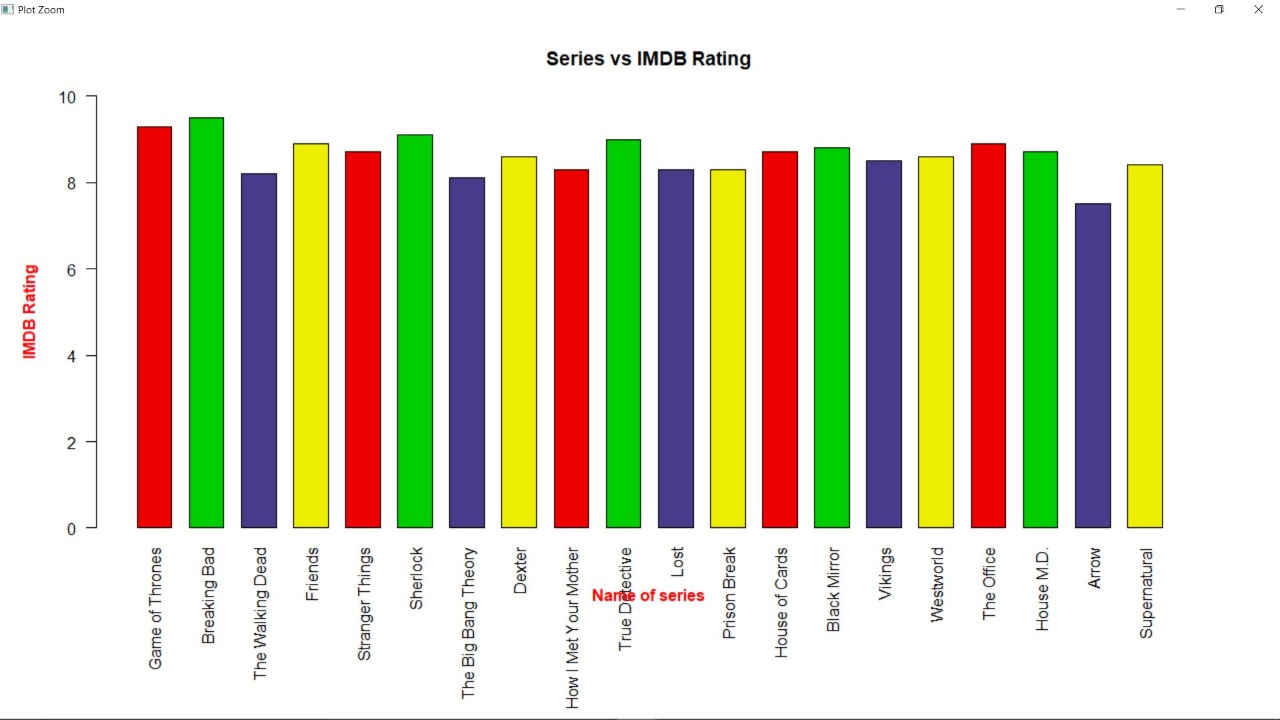
**#votes vs series**

barplot(height = head(df$No\_of\_Votes,20),ylim = c(0,2000000),las=2,names.arg=head(df$Series\_Title,20),col=c("red2", "green3", "slateblue4", "yellow2"),xlab = "Name of series",ylab = "Number of votes",main="Series vs Votes",font.lab=2,col.lab="red")



**#2)rating vs series**

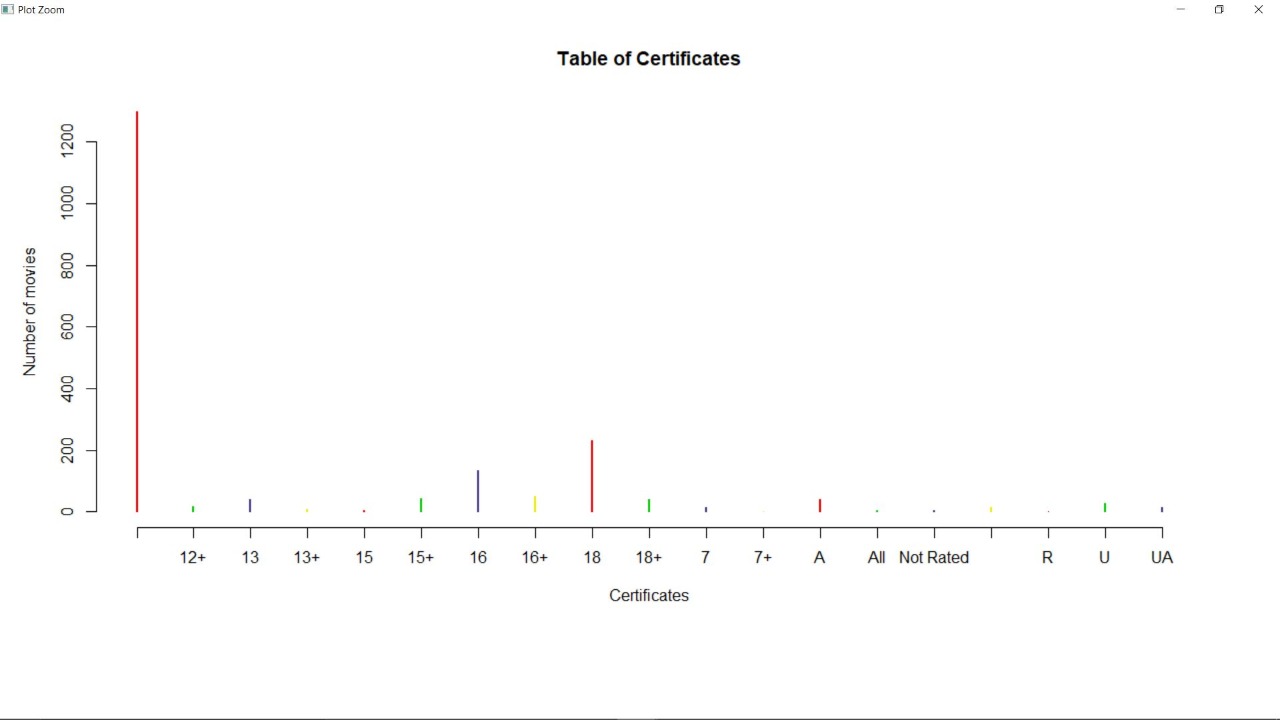
barplot(ylim = c(0,10),height = head(df$IMDB\_Rating,20),las=2,names.arg=head(df$Series\_Title,20),col=c("red2", "green3", "slateblue4", "yellow2"),xlab = "Name of series",space=0.5,ylab = "IMDB Rating",main="Series vs IMDB Rating",font.lab=2,col.lab="red")



**#3)certificates corresponding to number of series**

plot(table(df$Certificate),col=c("red2", "green3", "slateblue4", "yellow2"),

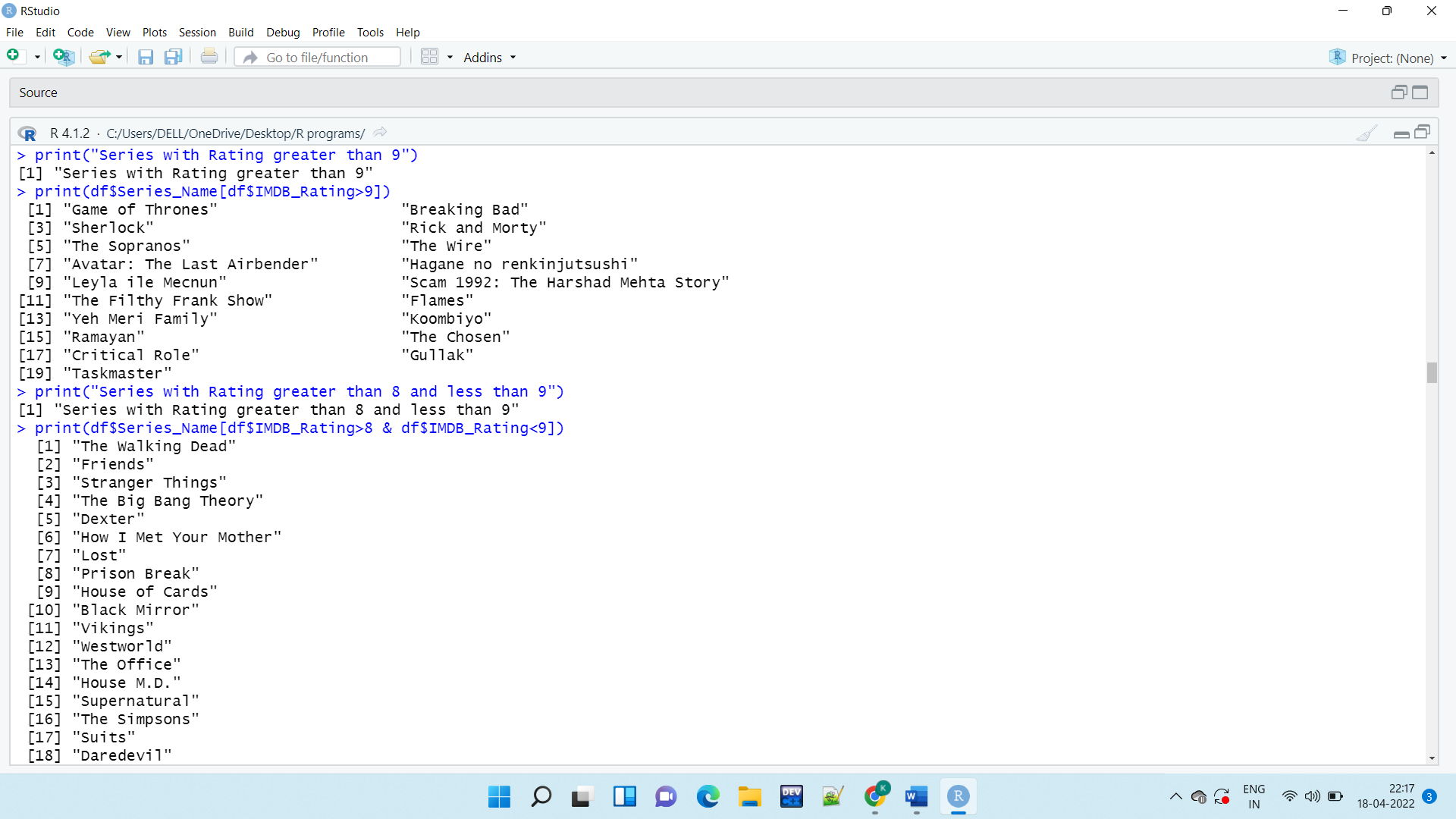
          xlab="Certificates",ylab="Number of movies",main="Table of Certificates")



**#4)series with rating above 9 between 9 and 8 and less than 8**

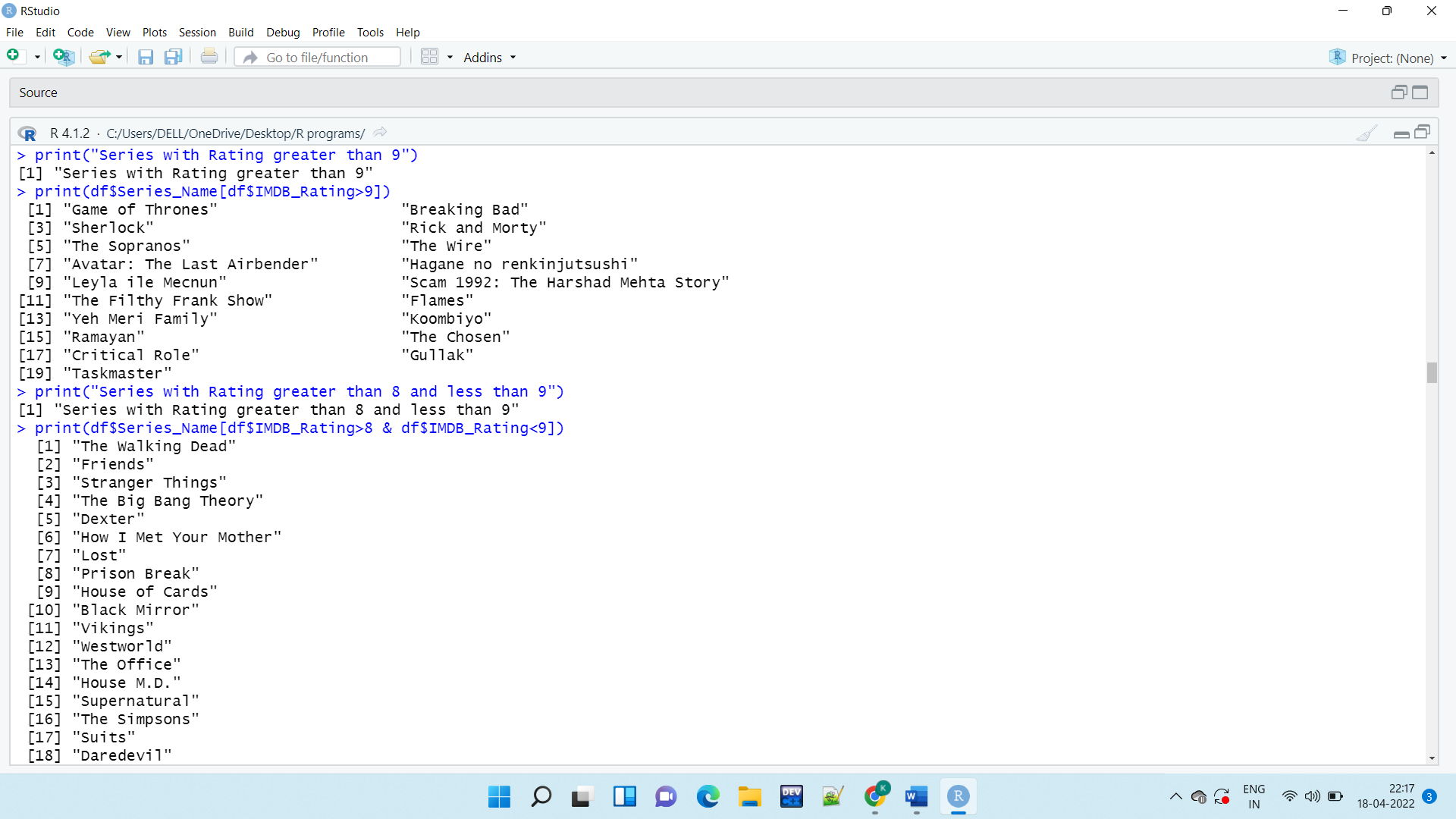
 print("Series with Rating greater than 9")

 print(df$Series\_Name[df$IMDB\_Rating>9])



 print("Series with Rating greater than 8 and less than 9")

 print(df$Series\_Name[df$IMDB\_Rating>8 & df$IMDB\_Rating<9])

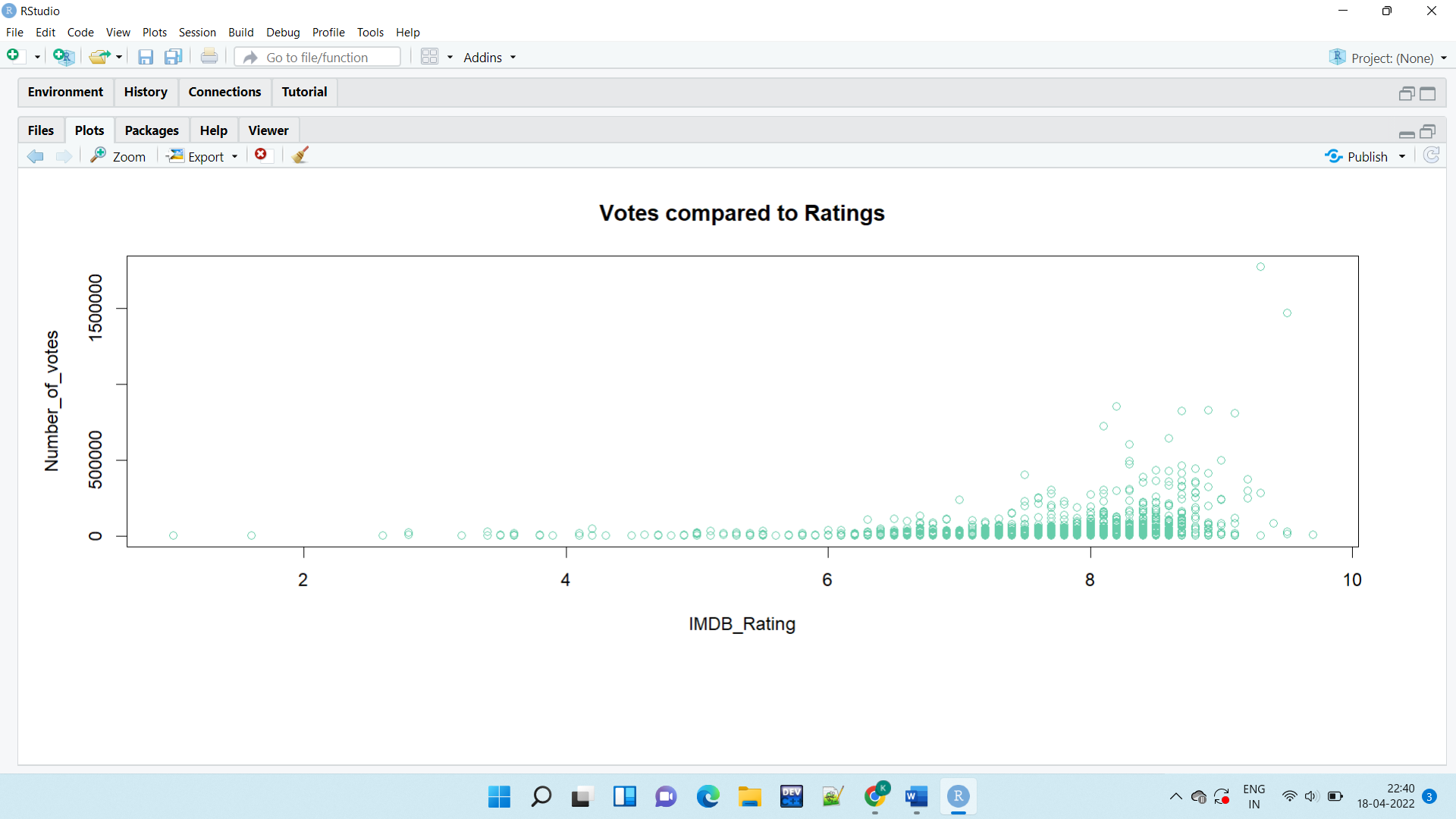


**#5)how accurate is imdb rating corresponding to ppl's vote**

 plot(df$IMDB\_Rating,df$No\_of\_Votes,,xlab="IMDB\_Rating",ylab="Number\_of\_votes",main="Votes compared to Ratings", col="#66cdaa")

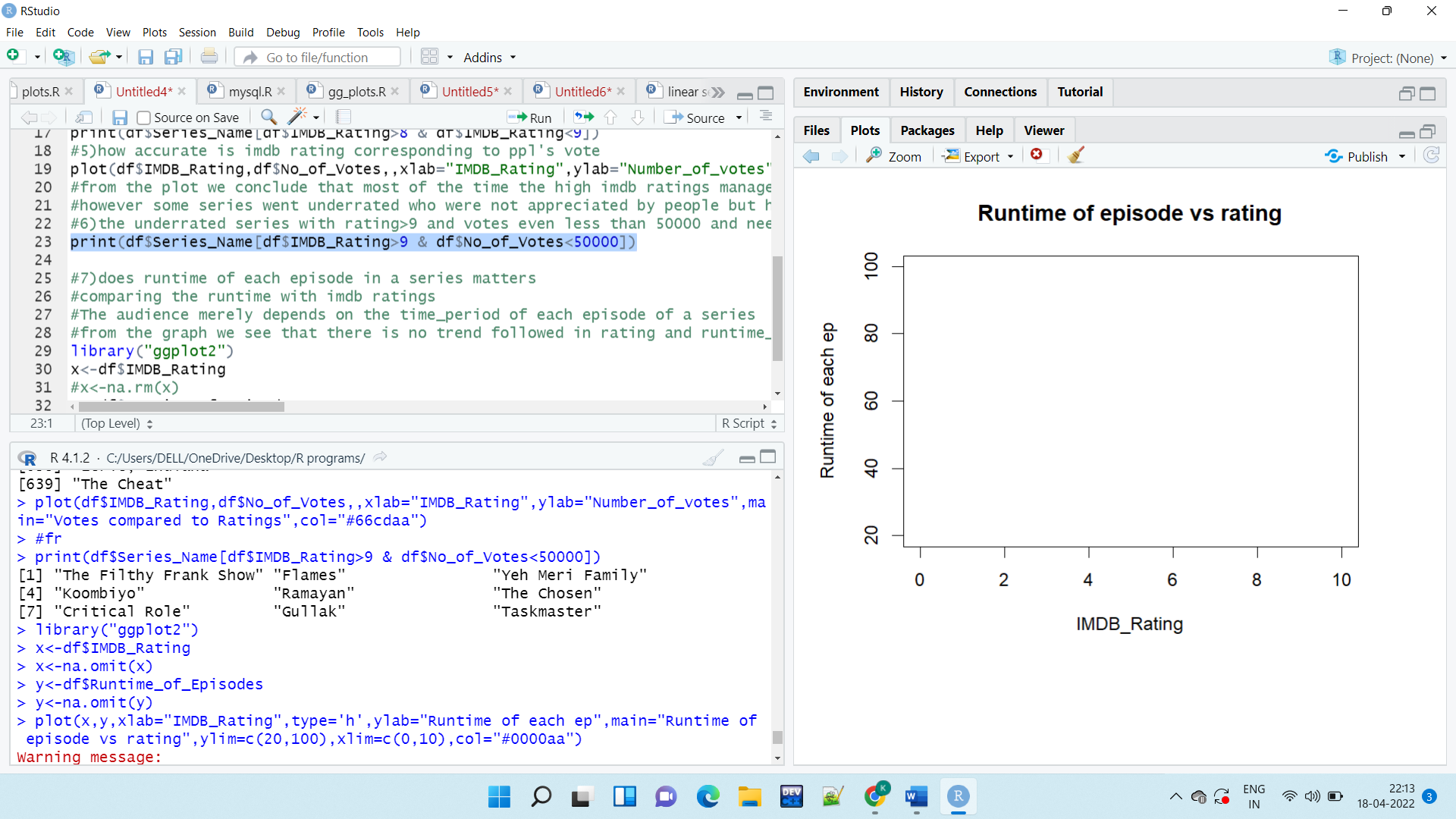
#from the plot we conclude that most of the time the high imdb ratings managed to get highest votes

#however some series went underrated who were not appreciated by people but had high ratings



**#6)the underrated series with rating>9 and votes even less than 50000 and need recognition**

print(df$Series\_Name[df$IMDB\_Rating>9 & df$No\_of\_Votes<50000])



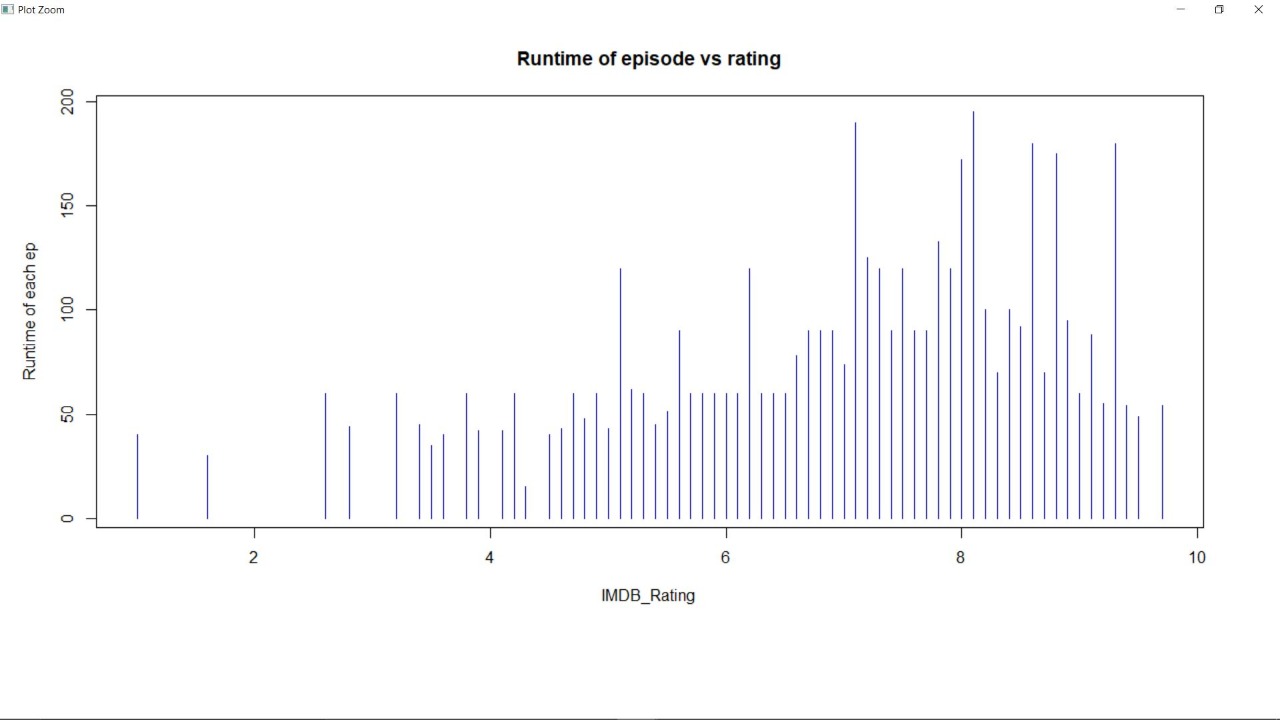
**#7)does runtime of each episode in a series matters**

**#comparing the runtime with imdb ratings**

**#The audience merely depends on the time\_period of each episode of a series**

**#from the graph we see that there is no trend followed in rating and runtime\_of\_each\_ep**

plot(df$IMDB\_Rating,df$Runtime\_of\_Episodes,xlab="IMDB\_Rating",type='h',ylab="Runtime of each ep",main="Runtime of episode vs rating",col="blue")



**CONCLUSIONS :**

* From the sample dataset, **Breaking Bad** has got the maximum IMDB rating of 9.5, followed by **GoT** and **Sherlock**.
* The show with maximum votes has got rating around 9.5
* **Game of Thrones** is the longest running series among the sample dataset.
* **Game of Thrones** has secured the highest number of votes around 1.75million, followed by **Breaking Bad** (around 1.5 million) and **The Walking Dead** (~0.85 million).
* Certificate type **A** has a maximum number of series around 300.
* The runtime vs rating graph is almost scattered from rating 5 to 9, which means the runtime of the episode does not play a significant role in the ratings achieved.