Course: DATA EXPLORATION AND PREPARATION

Course Code: CAP482

CA 2

Dated: - 24/Apr/2024

Submitted by

Name: Anureet Kaur
Roll No:9
Reg:12222064
Section: DE419, Group: 1

Submitted to

Ms. Ranjit Kaur Walia
UID: 28632
Assistant Professor
SCA, LPU

Lovely Faculty of Technology & Sciences

School of Computer Applications

Lovely Professional University

Punjab

Some steps of cleaning the dataset:

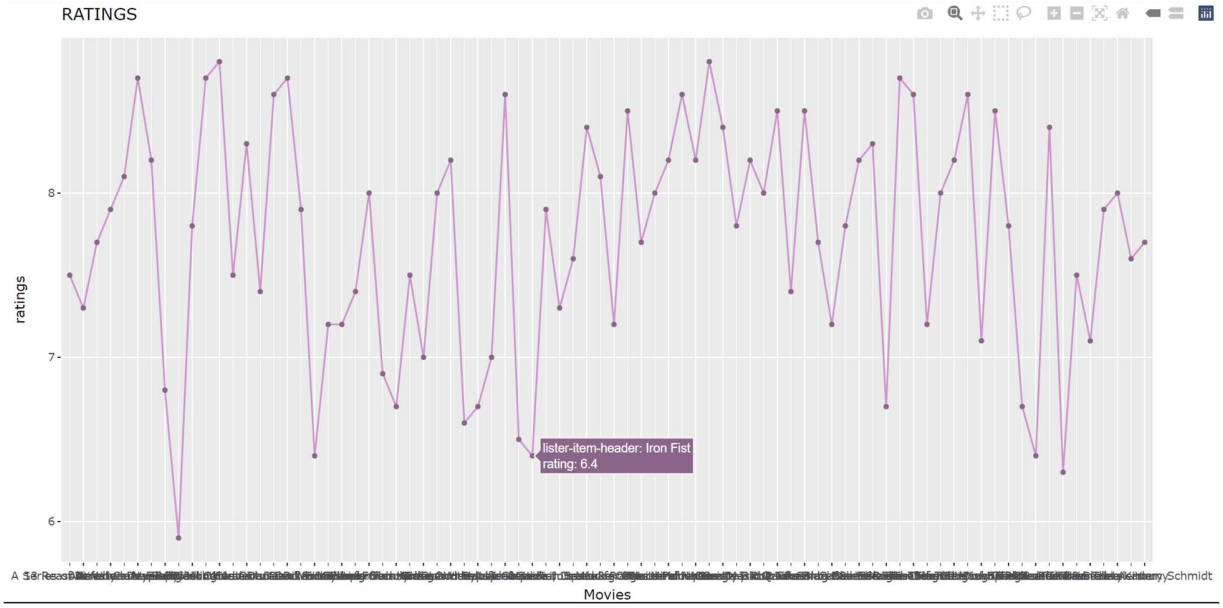
```r
#cleaning of data.....
View(imdb)
summary(imdb)
library(dplyr)
library(tidyverse)
imdb <-as.integer(imdb$runtime)

summary(imdb)
rm(imdb)

imdb <- imdb %>%
  mutate(certificate = replace(certificate, certificate=="PG", 0))
View(imdb)

summary(imdb)
str(imdb)
rm(imdb)
imdb <- imdb %>%
  separate(genre, into = c("genre1", "genre2" ,"genre3"), sep = ",")
view(imdb)

# Convert columns to appropriate data types
imdb <- type.convert(imdb, as.is = TRUE)
```

Q1) What are the total ratings of all movies?

Ans.) CODE:

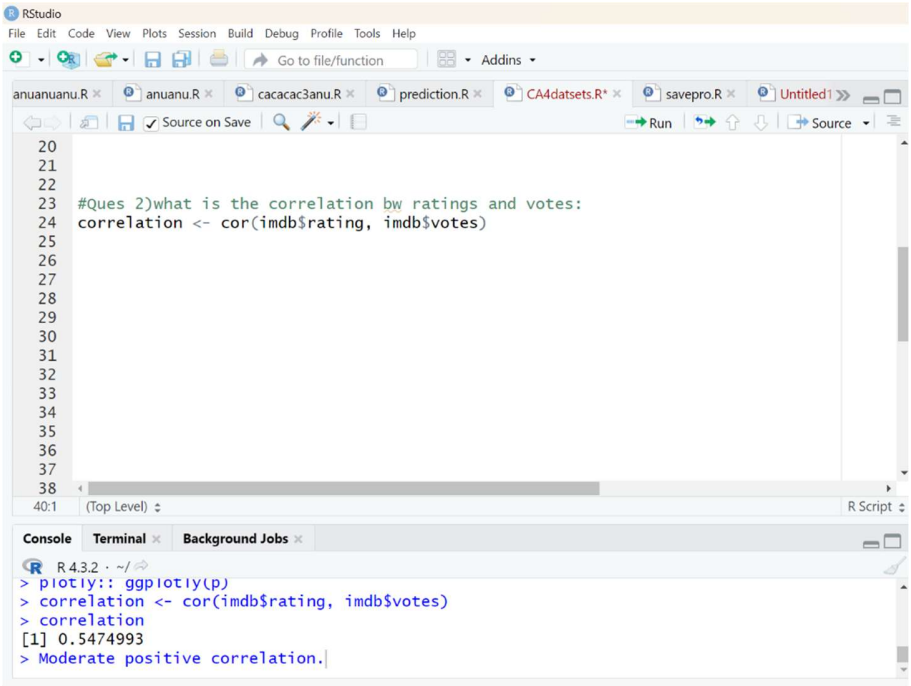
```r
library(tidyverse)
require(dplyr)
require(ggplot2)
require(ggpubr)
require(gridExtra)
read.csv(file.choose("imdb"),header = T)
View(imdb)
#QUES 1)Ratings of movies:
library(plotly)
anu <- ggplot(data = imdb, aes(x = `lister-item-header`, y = rating, group = 1)) +
  geom_line(color="plum3")+
  geom_point(size=1,color="plum4")+
  labs(title = "RATINGS",
       x="Movies",
       y="ratings")
plotly::ggplotly(anu)
```

OUTPUT:



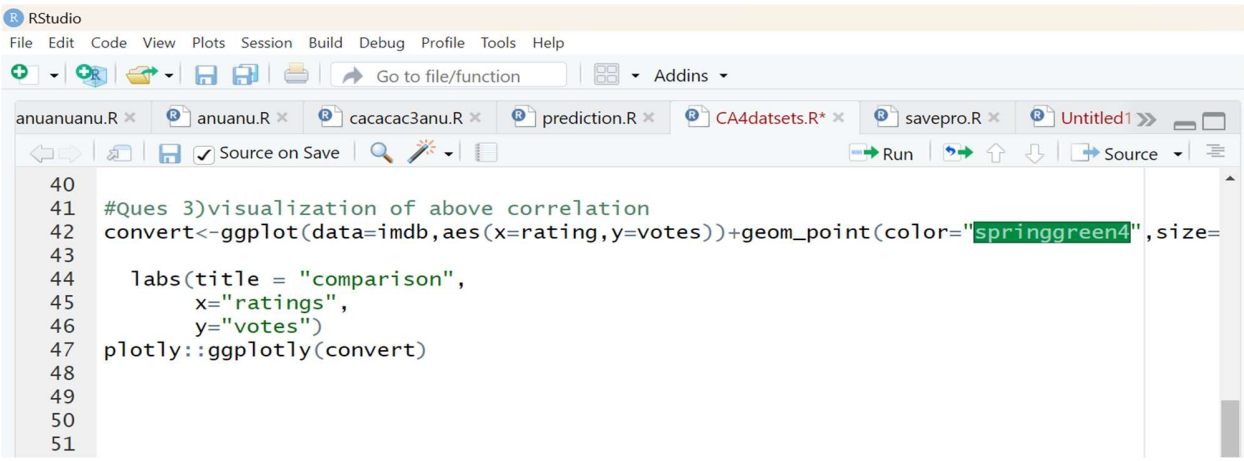
Q2) What is the correlation between ratings and votes.

Ans) CODE & OUTPUT

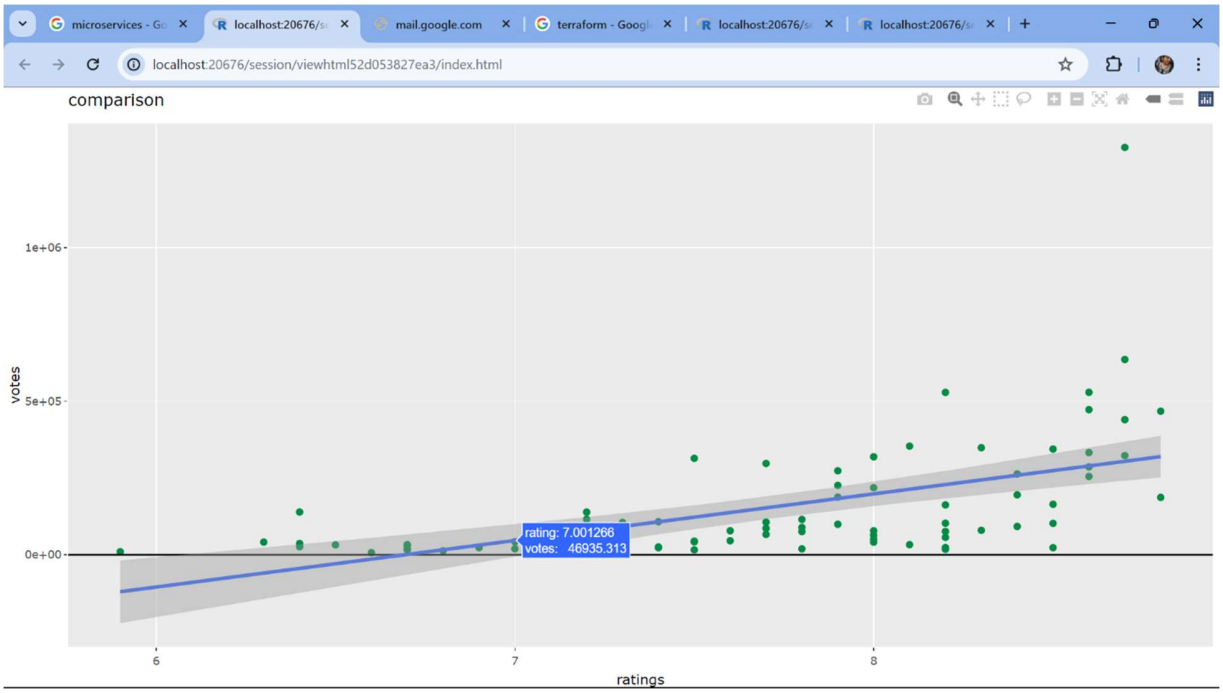


Ques 3) show the visualization of correlation.

Ans) Code:

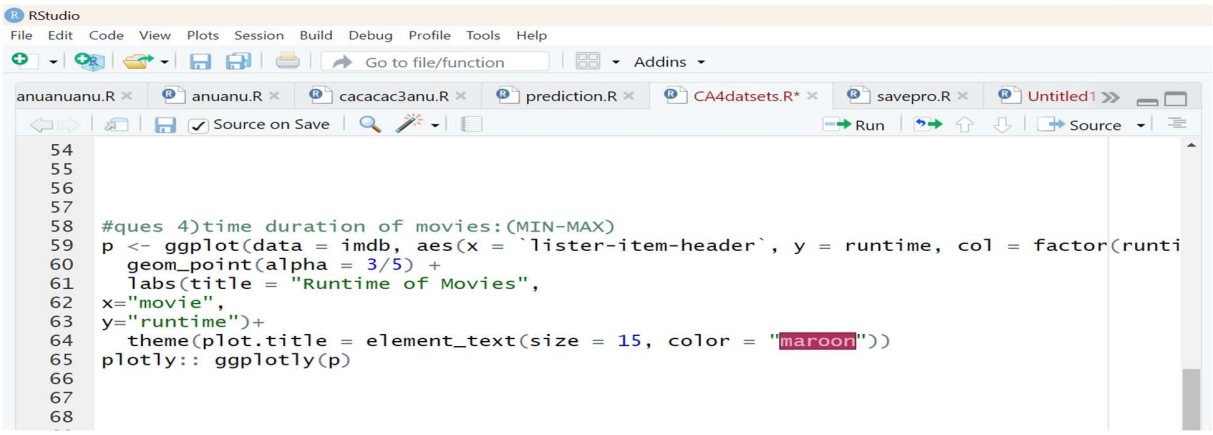


OUTPUT:



QUES 4) What is the runtime of movies from MIN-MAX.

ANS) code:



OUTPUT:

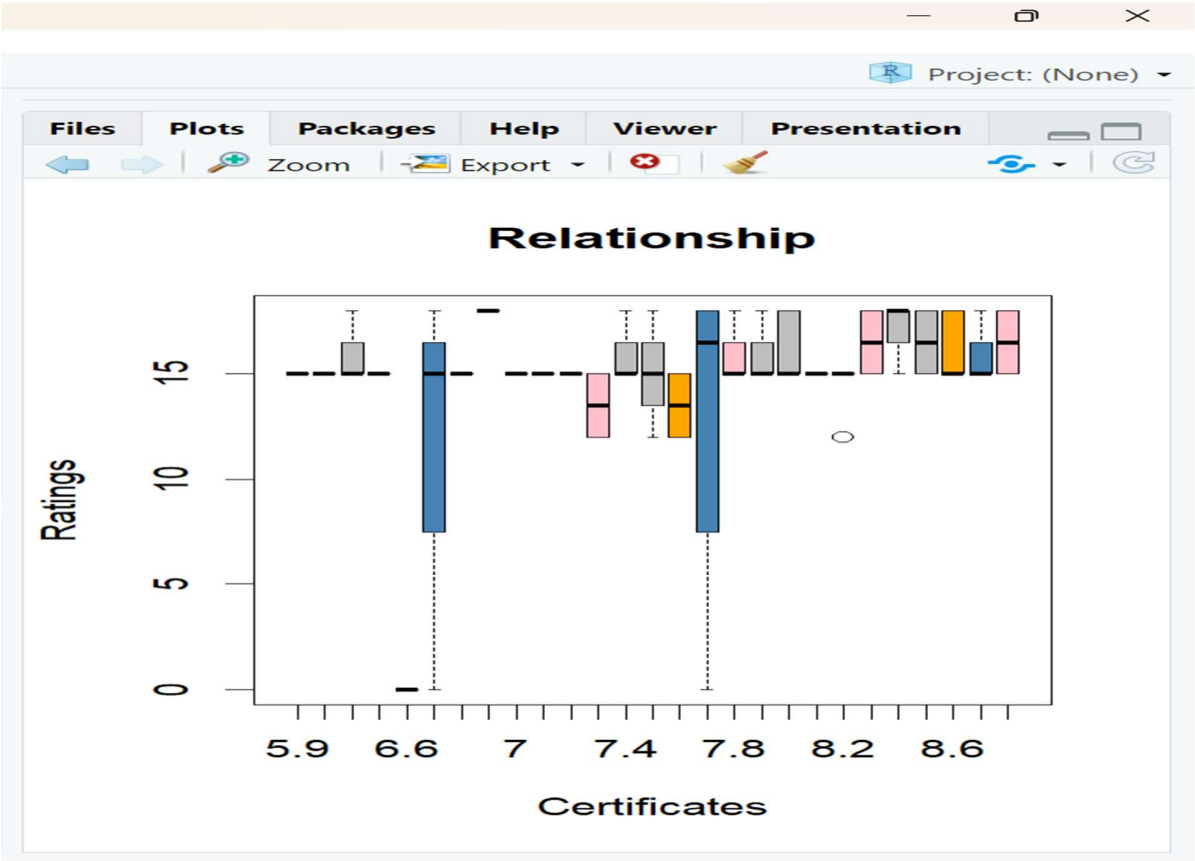
Ques 5) Create a boxplot of Relationship between ratings and certificates.

Ans) CODE:

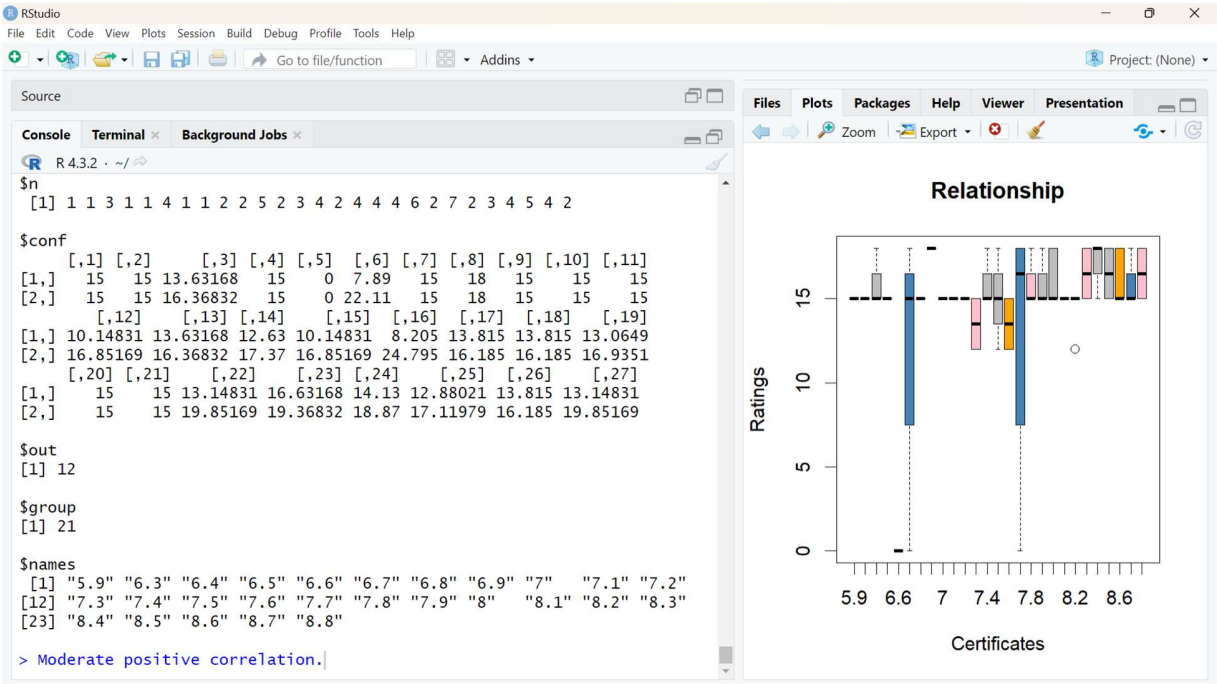
```
# Ques 6) Boxplot of the relation between certificate and rating:
boxplot(certificate ~ rating, data = imdb, main = "Relationship",
        xlab = "Certificates", ylab = "Ratings",
        col = c("steelblue", "pink", "grey", "grey", "orange"),
        border = "black")
```

86:1 (Top Level) R Scrip

OUTPUT:

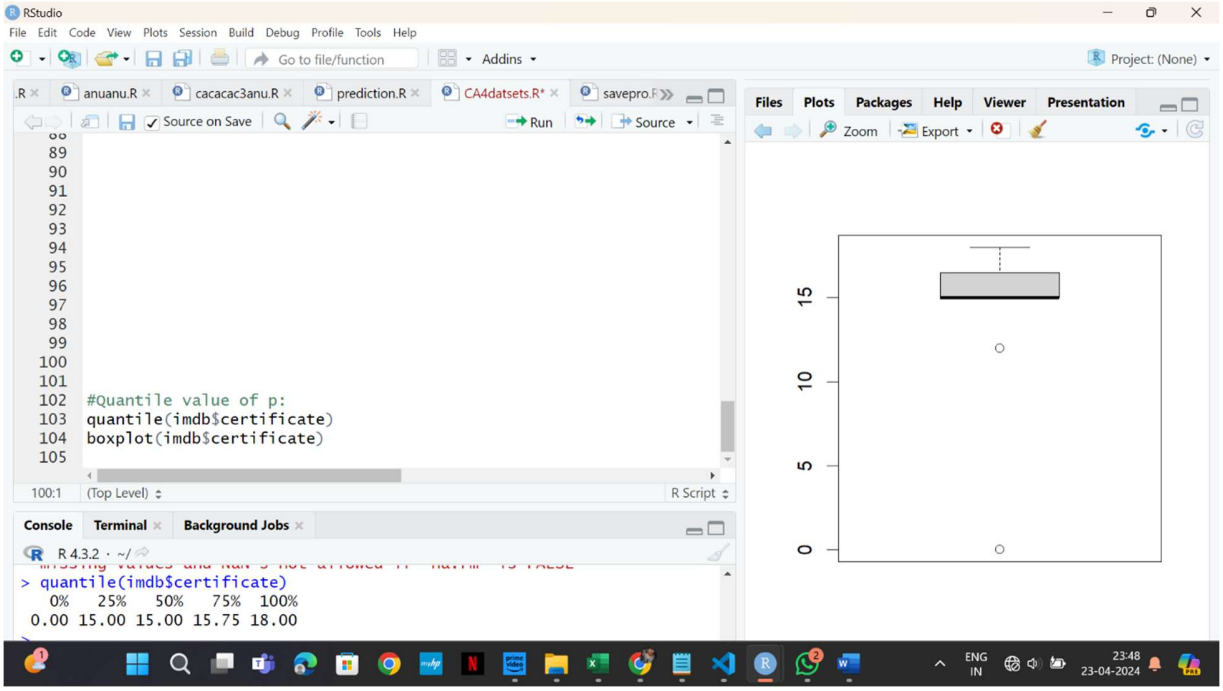


Another output:



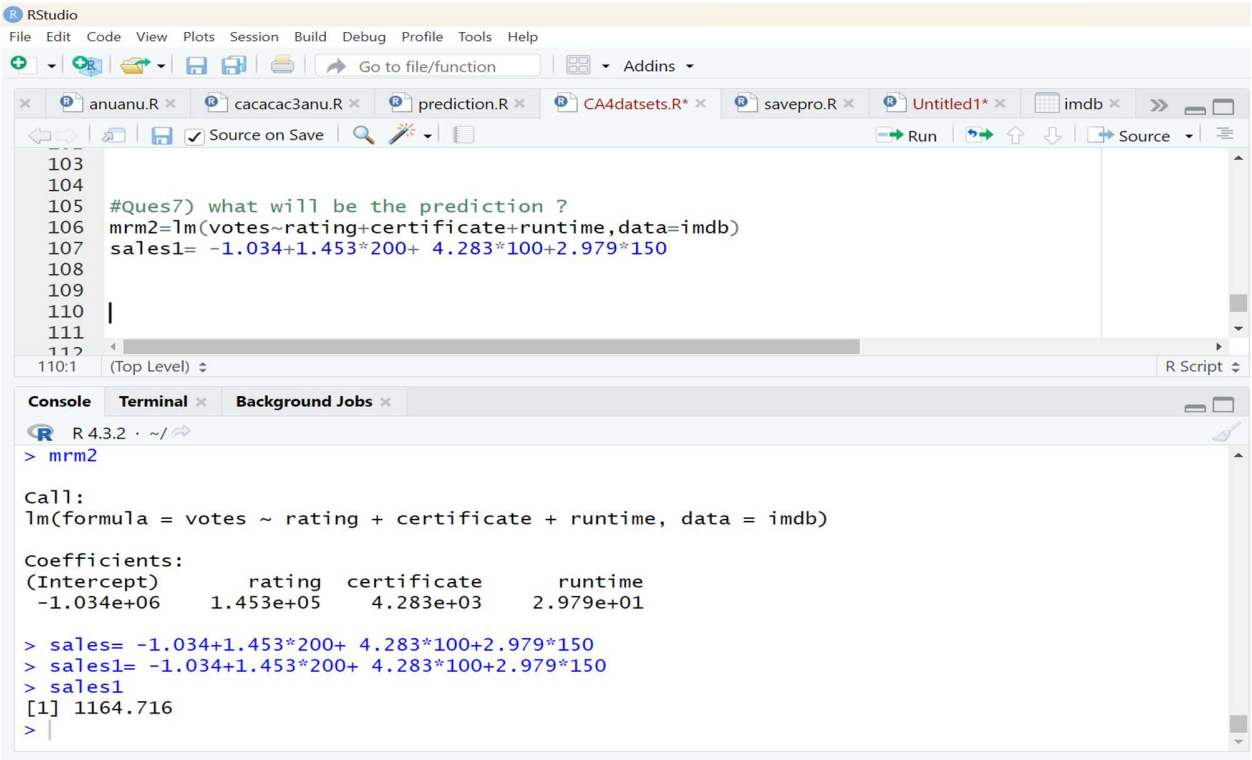
QUES 6) What will be the quantile value of column certificate?Visualize

Ans)



Ques 7) What will be the prediction?

Ans) code & output:



Ques 8) What is the maximum runtime of all the movies?

Ans) Code & Output:

```r
#Ques8)What is the maximum runtime of movies:
reet1<- imdb %>%
  group_by(runtime) %>%
  summarise(Movies = n())
max_genre1 <- reet1$runtime[which.max(reet1$Movies)]
```

```
# A tibble: 27 x 2
   runtime Movies
     <int>  <int>
 1      15      1
 2      22      1
 3      23      1
 4      25      3
 5      26      1
 6      27      1
 7      30     17
 8      34      1
 9      35      1
10      45      6
# i 17 more rows
```

Ques 9) What is the average and median value of all the votings?

Ans) Code & Output:

```r
#Ques 9) What is the average value and median of all the votings:
mean(imdb$votes)
median(imdb$votes)
```

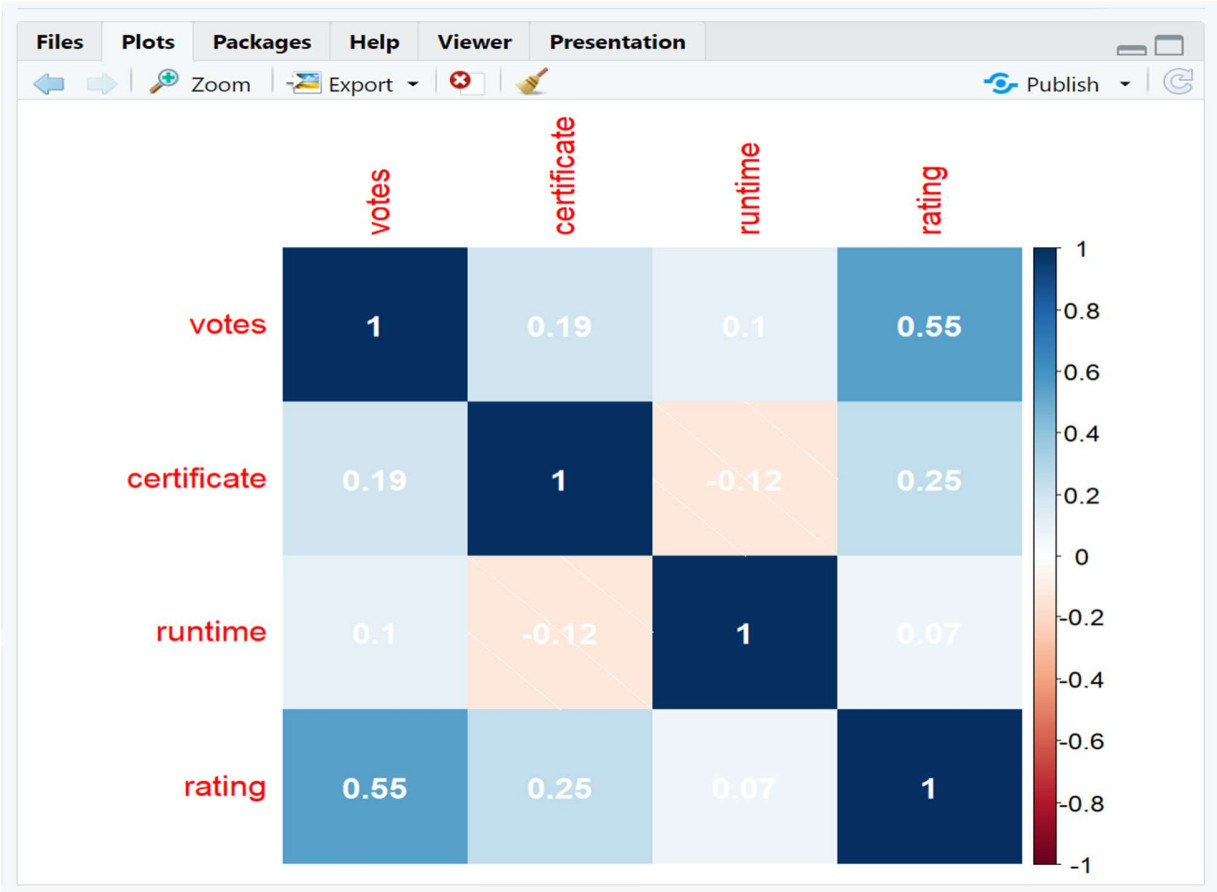
```
> mean(imdb$votes)
[1] 155994.6
> median(imdb$votes)
[1] 83033.5
>
```

Ques 10) Correlogram of all the numeric value columns:

CODE:

```
#Ques 10) Corrplot of all the numeric columns
library(corrplot)
 d2<- imdb[, c("votes","certificate","runtime","rating")]
correlation_matrix <- cor(d2)
corrplot(correlation_matrix, method = "shade", addCoef.col = "white", interactive = TRUE)
```

OUTPUT:

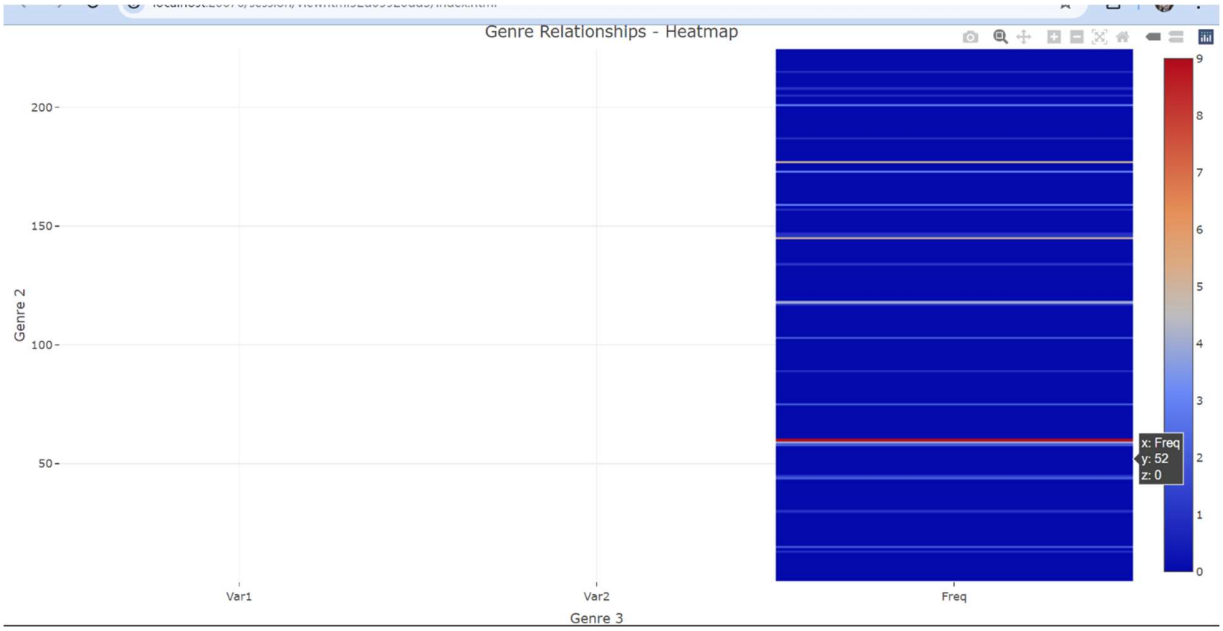


Ques 11) What is the relationship and distribution of genre 1 and genre 2.

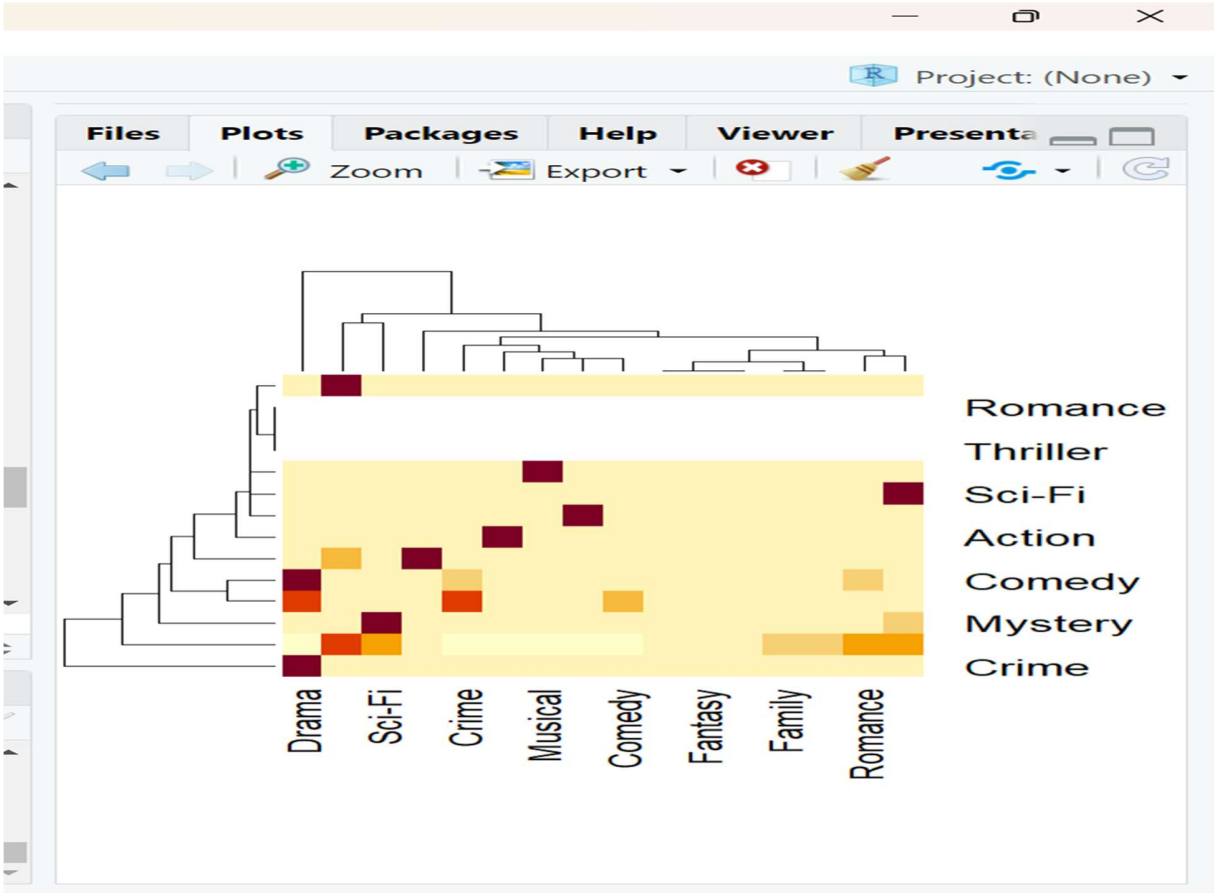
Ans)CODE:

```
genre2_distribution <- imdb %>%
  group_by(genre2) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
ggplot(genre2_distribution, aes(x = reorder(genre2, -count), y = count)) +
  geom_bar(stat = "identity") +
  labs(title = "Genre Distribution - Genre 2")
genre3_distribution <- imdb %>%
  group_by(genre3) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
ggplot(genre3_distribution, aes(x = reorder(genre3, -count), y = count)) +geom_bar(stat = "identity") +
  labs(title = "Genre Distribution - Genre 3")
genre_relationships <- table(imdb$genre2,imdb$genre3)
heatmap(genre_relationships)
genre_relationships_df <- as.data.frame(genre_relationships)
heatmap_plot <- plot_ly(z = ~as.matrix(genre_relationships_df),
                        x = colnames(genre_relationships_df),
                        y = rownames(genre_relationships_df),
                        type = "heatmap")
heatmap_plot <- heatmap_plot %>%
  layout(title = "Genre Relationships - Heatmap",
         xaxis = list(title = "Genre 3"),
         yaxis = list(title = "Genre 2"))
```

OUTPUT:HEATMAP

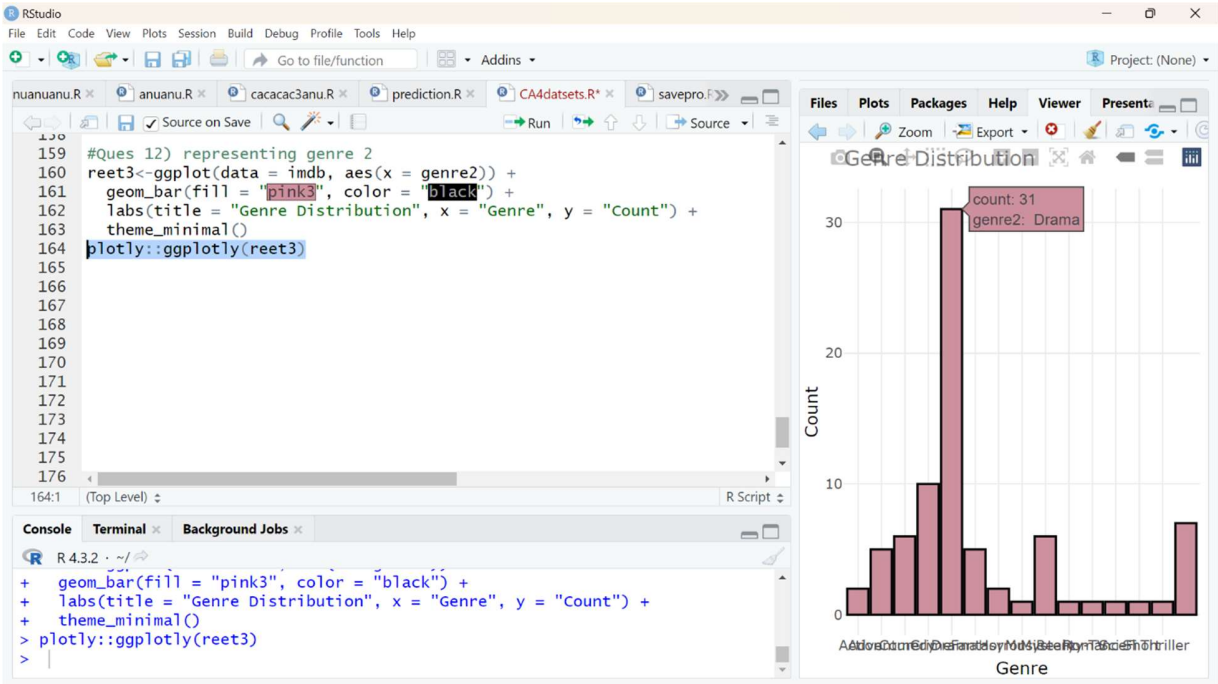


Without using plotly:



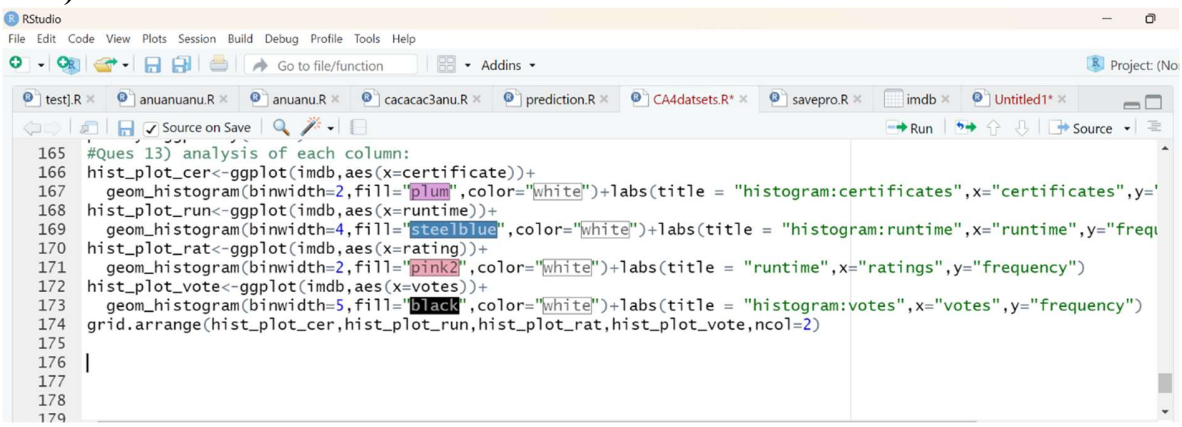
Ques 12) representing barplot of genre 2.

Ans- Code & output:



Ques-13) representing data analysis of all columns:

Ans) CODE:



OUTPUT:

Ques 14) What will be the prediction of voting based on certificate?

Ans) Code and output:

```r
#Ques 14)What will be the prediction of votings based on certificate:
result1<-lm(certificate~votes,data=imdb)
predict(result1)%>%round(0)
```

```
> result1

Call:
lm(formula = certificate ~ votes, data = imdb)

Coefficients:
(Intercept)        votes
  1.453e+01    3.274e-06

> predict(result1)%>%roun
Error in roun(.) : could not find function "roun"
> predict(result1)%>%round(0)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
19 16 16 17 16 16 16 16 15 15 16 15 16 16 15 16 15 15 15 15 15 15 15 15 15 15 15 15 15 16
31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
15 15 15 15 15 15 15 16 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
15 15 15 15 15 15 15 15 15 15 15 15 16 15 15 15 15 15 15 15
>
```

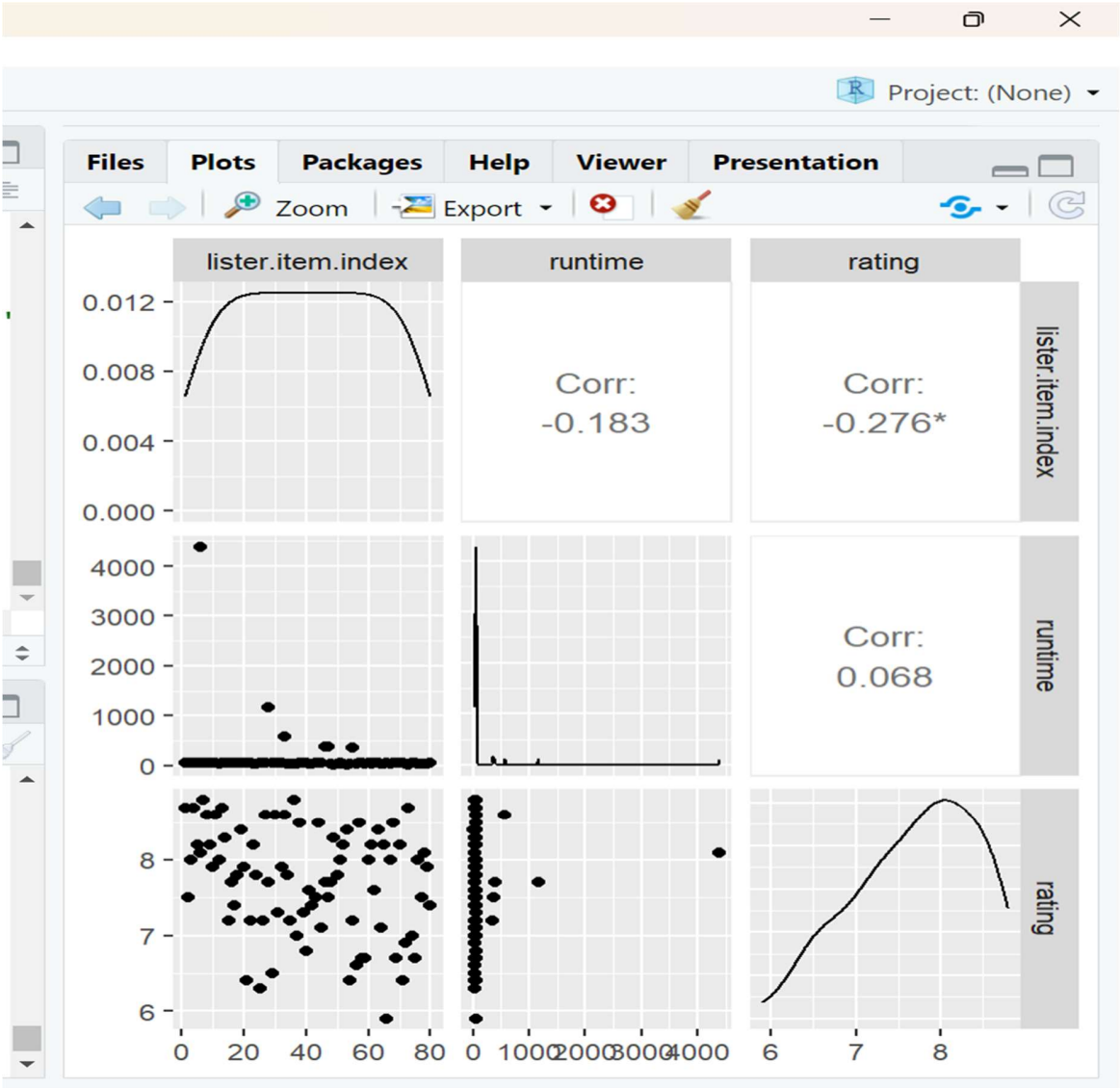
Ques 15) How the columns are related with each other.

Ans)

CODE:

```r
#Ques 15)SUMMARY:pairplot
install.packages("devtools")
library(devtools)
install.packages("GGally")
library(GGally)
columns_of_interest <- c("lister.item.index", "runtime", "rating")
subset_data1 <- imdb_subset[, columns_of_interest]
# Set a higher cardinality threshold
options(ggally_cardinality_threshold = 100)
ggpairs(subset_data1)
```

OUTPUT:


Project: (None)
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
lister.item.index
runtime
rating
Corr:
-0.183
Corr:
-0.276*
Corr:
0.068
0.012
0.008
0.004
0.000
4000
3000
2000
1000
0
8
7
6
0
20
40
60
80
0
1000
2000
3000
4000
6
7
8