18021570012

Date

Examination Roll. No: 18021570012
Name of Student : Arnav Kumar Jain
Course : B.Sc(H) Computer Science
Semester: VI
Subject: Introduction to Data Science
Unique paper code: 32347608

From the Bollywood dataset given as .csv file, write R script for the following:
i) Find total number of missing values and replace NA with median value of the same column.
ii) Find the average opening collection according to the column "Genre". Create a suitable plot to visualize the result. x-label: "Average"; y-label: "Genre" and title: "Exam roll number".
iii) Create a subset of the dataset for all the "Super Hit" movies.
iv) Write a R function to return the name of movies whose "Total Collection" is more than the average Total Collection.

i)

```
df <- read.csv ('Bollywood.csv')
```

```
print(paste ("Total number of NA values :",table(is.na(df))[2]))
```

```
[1] "Total Number of NA values: 2"
```

```
# As NA values are in only one column

df [is.na(df$Openingcollection),4]= median (df$Openingcollection,
na.rm = TRUE)
```

Date 18021570012

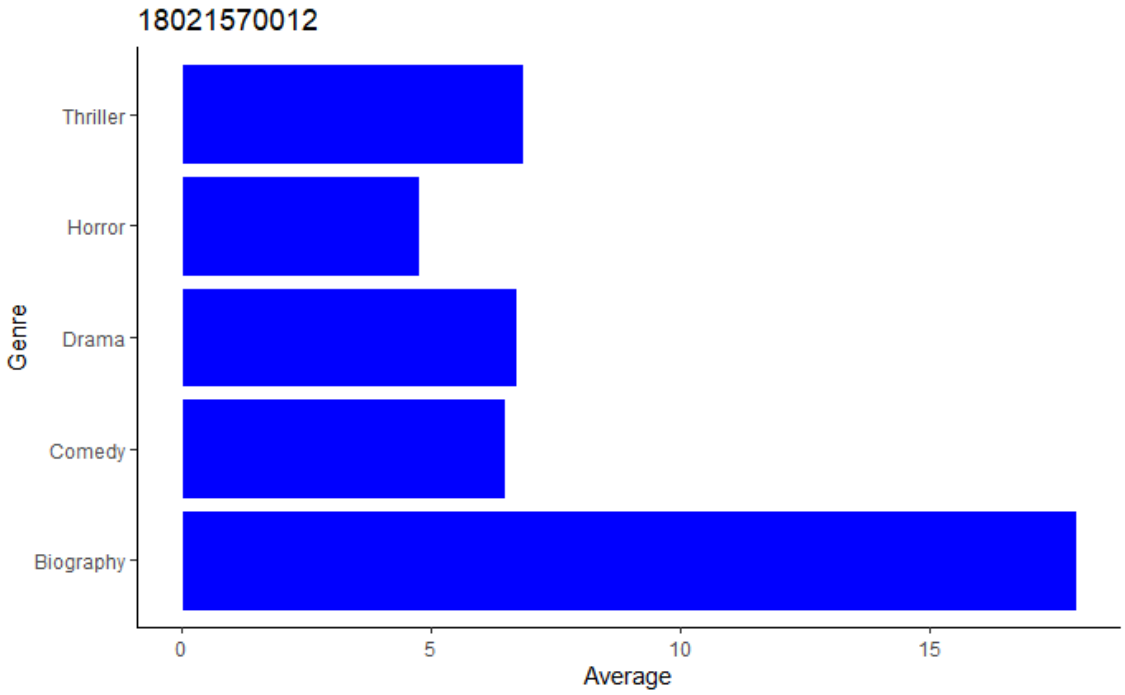
ii) library (dplyr)
library (ggplot2)

```
gdf = group_by (df, Genre)
avg_collection = summarise (.data=gdf, aoc = mean
(Openingcollection))
ggplot (avg_collection, aes (x="Average"
```

```
ggplot (avg_collection , aes (x=aoc, y = Genre )) +
geom_bar (stat='identity', color= "white", fill = "blue") +
theme_classic () + ggtitle ('18021570012') + labs (
x = "Average , y= "Genre")
```

iii) sdf = subset (df, Verdict == 'Super Hit')

iv)
```
greater_average <- function (df)
{
    return (df [df [,5] > mean (df [,5]), ])
}
print ("Movies with total collection greater than average
 total collection are:")
print (greater_average (df))
```




```
> print(sdf)
         Movie    Lead     Rdate Openingcollection TotalCollection   Verdict     Genre
1       Dangal   Aamir 23-Dec-16             29.78          386.68 Super Hit Biography
8  M. S. Dhoni  Sushant 30-Sep-16             21.30          132.85 Super Hit Biography
11         Pink Amitabh 16-Sep-16              4.32           61.83 Super Hit     Drama
16       Rustom  Akshay 12-Aug-16             14.11          127.13 Super Hit Biography
>
```

```
> print("Movie with total collection greater than average total collection are:")
[1] "Movie with total collection greater than average total collection are:"
> print(greater_average(df))
[1] "Dangal"       "Shivaay"      "M. S. Dhoni " "Rustom"       "Housefull 3"  "Baagl
>
```

Write a R script to simulate a sample of 20 random numbers in the range- 10 to 500 and store it in a vector. Create a scatter plot for the vector, in blue color and draw vertical lines in red and green color corresponding to mean and median, respectively.

1802157O012 Date

```
R = sample (10:500, 20)
axis_ticks = axis (1, R, labels = R)
plot (R, 1:20, col = 'blue', xaxt='n')
abline (v = mean(R) , col = 'red')
abline (v=median (R), col = 'green')
for (i in axis_ticks) axis (side = 1, at = i, labels
= i, las = 2, cex.axis = 0.5)
# to avoid skipping tick labels as much as possible
```

```
> print(R)
 [1] 273 141 163 443 364 436 394 455 149 220 462 215 424 376  24 415 218 145  73 136
```