UE20CS312 - Data Analytics - Worksheet 1a - Part 1 - Exploring data with R

PES University

'SUNDEEP A, Dept. of CSE - PES1UG20CS445'

2022-08-17

Solutions

Problem 1

```
#The libraries used:
library(readxl)
library(stringr)
library(modeest)
#reading the dataset:
dframe <- read_excel("C:\\Users\\HP\\Downloads\\top_1000_instagrammers.csv")
df<-dframe
#user defined function to convert M in million in numbers
str_numeric<-function(X){
  if(str_sub(X,-1)=="M"){
    X<-(as.numeric(str_sub(X,1,str_length(X)-1)))*1000000
  }
  else if(str_sub(X,-1)=="K"){
    X<-(as.numeric(str_sub(X,1,str_length(X)-1)))*1000
  }
  else{
    X<-as.numeric(str_sub(X,1,str_length(X)-1))
  }
}
#converting the the data in Followers, Authentic Engagement, Engagement Avg from Character to Numeric
df$Followers = unlist(lapply(df$Followers, str_numeric))
df$`Authentic Engagement` = unlist(lapply(df$`Authentic Engagement`, str_numeric))
df$`Engagement Avg.` = unlist(lapply(df$`Engagement Avg.`, str_numeric))
#Printing the Statistical values of the Numerical Columns
summary(df[c("Followers","Authentic Engagement","Engagement Avg.")])
```

```
##    Followers         Authentic Engagement Engagement Avg.
##  Min.   :  1600000   Min.   :   20600     Min.   :   21100
##  1st Qu.:  8600000   1st Qu.:  176025     1st Qu.:  252425
##  Median : 14100000   Median :  325050     Median :  463300
##  Mean   : 26044600   Mean   :  577754     Mean   :  772861
##  3rd Qu.: 25425000   3rd Qu.:  607300     3rd Qu.:  850125
##  Max.   :528400000   Max.   :13300000     Max.   :13300000
##                      NA's   :20
```

```
sprintf("The mode of Followers data is : %d",mfv(df$Followers))
```

```
## [1] "The mode of Followers data is : 7300000"
```

```
sprintf("The mode of Authentic Engagement is %d:",mfv(df$`Authentic Engagement`,na_rm = TRUE))
```

```
## [1] "The mode of Authentic Engagement is 1200000:"
```

```
sprintf("The mode of Engagement Avg is : %d",mfv(df$`Engagement Avg.`))
```

```
## [1] "The mode of Engagement Avg is : 1100000"
```

```
sprintf("The standard Deviation of Followers data : %f",sd(df$Followers))
```

```
## [1] "The standard Deviation of Followers data : 44098837.153357"
```

```
sprintf("The standard Deviation of Authentic Engagement data %f:",sd(df$`Authentic Engagement`,na.rm = T
```

```
## [1] "The standard Deviation of Authentic Engagement data 1002399.232644:"
```

```
sprintf("The standard Deviation of Engagement Avg data : %f",sd(df$`Engagement Avg.`,na.rm = TRUE))
```

```
## [1] "The standard Deviation of Engagement Avg data : 1117135.227353"
```

Analysis:

The analysis gives us an idea about the data.The analysis gives us information about the Min,1st Quartile,3rd Quartile,Median,Mean,Max,Mode,Standard Deviation.The things we can note from this analysis is the the average followers are 26044600, the average authentic Engagement is 577754 and the average Engagement is 772861. There is no point in comparing My Instagram stats with these influencers, My stats are a nobody when compared to their stats

Problem 2

```
library(ggplot2)
temp=df[!(df$`Audience Country`==''),]
Total_followers<-aggregate(temp$Followers,by=list(temp$`Audience Country`),FUN=sum)
print("The total followers from India are : ")
```

```
## [1] "The total followers from India are : "
```

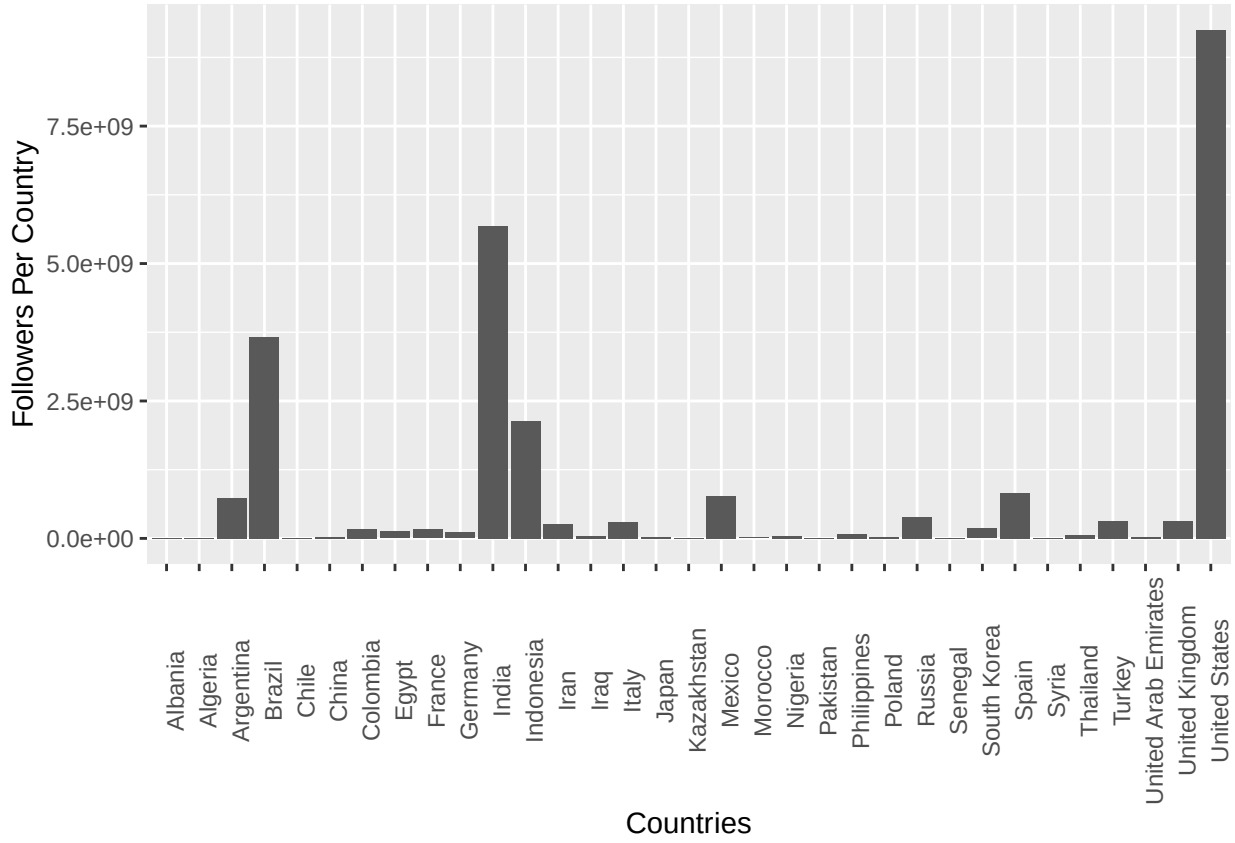
```
print(Total_followers[Total_followers$Group.1=='India',])
```

```
##    Group.1          x
## 11   India 5684300000
```

```
graph<-ggplot(Total_followers,aes(x=Group.1,y=x))+geom_histogram(stat = 'identity')+labs(y="Followers P
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
graph + theme(axis.text.x = element_text(angle = 90))
```

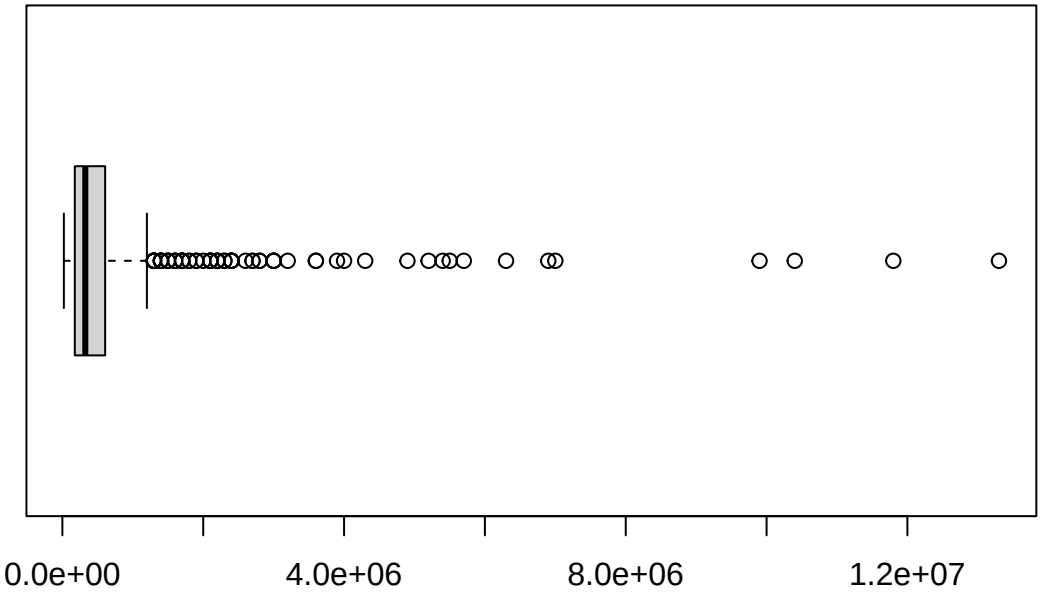


Analysis:

From the Histogram details we can see that United States has the most number of followers and India holds the second spot with 5684300000 followers

Problem 3

```
boxplot(df$`Authentic Engagement`, horizontal = TRUE)
```



Analysis:

From the box plot and whisker plot displayed above, we can see that the graph is Right-skewed.we can also observe that there are a lot of outliers.

In my opinion, Box plot and whisker plot are not giving a good inferences[conclusions] for the influencers data set.

Conclusion:

My instagram username is "sundeep___28" , this is a public account with 242 followers with an estimated engagement of about 40%.My account stats are not even close to 0.1% of the stats obtained for the top 1000 instgram influencers list.Basically I am a foodie, so If I were to tasked to become an influencer, then I would basically share stories,posts about the local foods of India cooked in family run hotel/restaurants . This way the people will get to experience the local foods of India and the hotels will be saved from closing their business due to heavy competition from fast-food chains.