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

PES University

Name: Adarsh S Nayak, Dept. of CSE - PES1UG20CS620

SRN : PES1UG20CS620 Roll No : 54

1.Data fetched is about the Top 1000 Instagrammers.

Data Dictionary

```
Name: Name of the account
Rank: Overall rank in the world.
Category: Stream of the account (Music, Games, etc..)
Followers: Number of followers
Audience Country: country of the majority of audience.
Authentic Engagement: Engagement with the users.
Engagement Avg.: Average engagement of the users.
```

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6     ✔ purrr   0.3.4
## ✔ tibble  3.1.8     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.1
## ✔ readr   2.1.2     ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
#get the current directory.
print(getwd())
```

```
## [1] "C:/Users/Hp/Desktop/Data-Analytics---Elective/Assignment 2"
```

```
#load the given dataset
df <- read_csv("top_1000_instagrammers.csv")
```

```
## Rows: 1000 Columns: 7
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (6): Name, Category, Followers, Audience Country, Authentic Engagement, ...
## dbl (1): Rank
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
print(df)
```

```
## # A tibble: 1,000 × 7
##    Name              Rank Category                Follo…¹ Audie…² Authe…³ Engag…⁴
##    <chr>            <dbl> <chr>                   <chr>   <chr>   <chr>   <chr>
##  1 cristiano            1 Sports with a ball      462.9M  India   5.9M    6.0M
##  2 leomessi             2 Sports with a ballFamily  347.2M  Argent… 3.6M    4.8M
##  3 kendalljenner        3 ModelingFashion         247.6M  United… 3M      4.9M
##  4 arianagrande         4 Music                   321.4M  United… 2.4M    3.4M
##  5 zendaya              5 Cinema & Actors/actresse… 147M  United… 4.3M    5.8M
##  6 kimkardashian        6 FashionBeauty           323.6M  United… 1.7M    2.5M
##  7 taylorswift          7 Music                   210.2M  Brazil  2.4M    3.2M
##  8 kyliejenner          8 FashionModelingBeauty   357M    United… 1.2M    1.9M
##  9 selenagomez          9 MusicLifestyle          334.9M  United… 1.4M    1.9M
## 10 thr                 10 <NA>                    46.3M   United… 13.3M   13.3M
## # … with 990 more rows, and abbreviated variable names ¹Followers,
## #   ² `Audience Country`, ³`Authentic Engagement`, ⁴`Engagement Avg.`
## # ℹ Use `print(n = ...)` to see more rows
```

Problems

##Problem 1 (1 point)

Get the summary statistics (mean, median, mode, min, max, 1st quartile, 3rd quartile and standard deviation) for the dataset. Calculate these only for the numerical columns [Audience Country, Authentic Engagement and Engagement Average]. What can you determine from the summary statistics? How does your Instagram stats hold up with the top 1000 :P ?

```
summary(df)
```

```
##      Name               Rank         Category          Followers
##  Length:1000        Min.   :   1.0   Length:1000        Length:1000
##  Class :character   1st Qu.: 250.8   Class :character   Class :character
##  Mode  :character   Median : 500.5   Mode  :character   Mode  :character
##                     Mean   : 500.5
##                     3rd Qu.: 750.2
##                     Max.   :1000.0
##  Audience Country   Authentic Engagement Engagement Avg.
##  Length:1000        Length:1000          Length:1000
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##
##
##
```

-->This is the summary for whole dataset.

Conclusion : ANALYSIS 1.Authentic engagement is less than Audience engagement 2. The distribution is right skewed.

```
df=df[!(is.na(df$Name) | df$Name==""),]
df=df[!(is.na(df$Rank) | df$Rank==""),]
df=df[!(is.na(df$Followers) | df$Followers==""),]
df=df[!(is.na(df$Category) | df$Category==""),]
df=df[!(is.na(df$`Audience Country`) | df$`Audience Country`==""),]
df=df[!(is.na(df$`Authentic Engagement`) | df$`Authentic Engagement`==""),]
df=df[!(is.na(df$`Engagement Avg.`) | df$`Engagement Avg.`==""),]
```

```
summary(df$`Audience Country`)
```

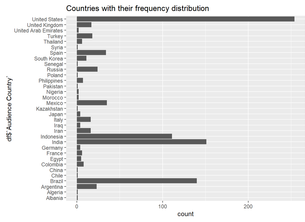
```
##    Length     Class      Mode
##       909 character character
```

##Problem 2 (2 points) What are the top 3 audience countries that follow most of the top 1000 instagrammers? Hint: Go back to bar graph created earlier. Use R to calculate the percentage of the top 1000 instagrammers that have the top 1 audience country.

Making the bar graph

```
library(ggplot2)
ggplot(data=df,aes(x=df$`Audience Country`))+geom_bar()+coord_flip()+ggtitle("Countries with their frequency distribution")
```

```
## Warning: Use of `df$`Audience Country`` is discouraged. Use `Audience Country`
## instead.
```




```
frequency_table_of_the_country=table(df$`Audience Country`)
#Create a frequency table with the countries and their occurences.
after_sorting=frequency_table_of_the_country[order(frequency_table_of_the_country,decreasing=TRUE)]
#Sorting the distribution in decreasing order
cat("The top 3 audience countries that follow most of the top 1000 instagrammers are ")
```

```
## The top 3 audience countries that follow most of the top 1000 instagrammers are
```

```
print(after_sorting[1:3])
```

```
##
## United States         India        Brazil
##           254           151           140
```

```
#printing the top 3 countries
cat("Percentage of the top 1000 instagrammers that have the top 1 audience country is ",after_sorting[1]/sum(after_sorting))
```

```
## Percentage of the top 1000 instagrammers that have the top 1 audience country is  0.2794279
```

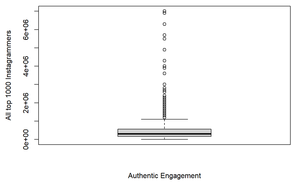
##Problem 3 (1 point) Create a horizontal box plot using the column Authentic.Engagement. What inferences can you make from this box and whisker plot?

```
a<- df[["Authentic Engagement"]]
a <- ifelse(grepl('m', ignore.case = TRUE, a), as.numeric(gsub("[$m]", "", a)) * 10^6,as.numeric(gsub("[$k]", "", a)) * 10^3)
```

```
## Warning in ifelse(grepl("m", ignore.case = TRUE, a), as.numeric(gsub("[$m]", :
## NAs introduced by coercion

## Warning in ifelse(grepl("m", ignore.case = TRUE, a), as.numeric(gsub("[$m]", :
## NAs introduced by coercion
```

```
boxplot(a , xlab = "Authentic Engagement", border = "black",
ylab = "All top 1000 Instagrammers")
```



From the box plot we can observer

that there are some points which exist as outliers.

1. Median ~ 0.3 million
2. 1st quartile ~ 0.2 million
3. 3rd quartile ~ 0.6 million
4. IQR ~ 0.4 million.
5. outliers present beyond 1.25 million.
6. Most of the top 1000 instagrammers have an engagement between thousands to 1.25 million. Distribution is positively skewed as median is closer to lower quartile

##Problem 4 (3 points)

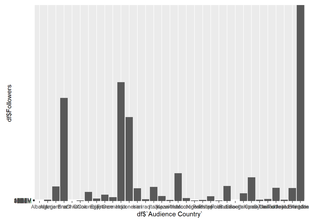
Create a histogram where the x-axis contains the Audience Country and y-axis contains the total follower count for accounts with that Audience Country. Which country is associated with the most amount of followers? Hint: Recall the concept of groupby() in Pandas. Try using the aggregate() function in R to achieve the same goal. What is the total for India and what rank does it fall compared to other countries?

```
library(ggplot2)
ggplot(data=df, aes(x =df$`Audience Country`, y = df$Followers)) +geom_bar(stat="identity",FUN=max)
```

```
## Warning: Ignoring unknown parameters: FUN
```

```
## Warning: Use of `df$`Audience Country`` is discouraged. Use `Audience Country`
## instead.
```

```
## Warning: Use of `df$Followers` is discouraged. Use `Followers` instead.
```



##analysis from the histogram :

United States of America has the highest number of the follower count, Brazil and India follow upon. India stands in 2nd position with 5684300000 Followers.

##Conclusion : In a few short sentences, describe your Instagram profile (category, followers, estimated engagement). Compare your profile to the analysis done of the top 1000 profiles. If you were tasked to becoming an influencer, what would be the best way for you to increase your followers and user engagement?

I'm Adarsh S Nayak having an instagram account with **964 followers**,it's a private account I created this account nearly four years back.

My account falls under the category of **Photography** , as I am a photography enthusiast, I love taking pictures developing them and posting it on instagram. Most of my followers are my friends from school and college.Coming to the user engagement, as the account has a less number of followers it's very less compared to the top 1000 instagram influencers.

User engagement would increase when we post things regularly so that the followers could follow along with the posts from the account, having better content with good photographs would make my followers take interest in stuff that I post. It would further help me increase the number of my followers as well.

As we can see from the analysis through box plots and histograms, account with lesser followers also have more user engagement. For example Jennie Ruby Jane(68200000 followers with authentic engagement=7e+06) has more user engagement than Joe with 69e+05 followers. So the quality of the content has a major role to play in user engagement.