PES University, Bangalore Established under the Karnataka Act No. 16 of 2013

UE20CS312 - Data Analytics

Worksheet 1a - Part 1: Exploring Data with R

Ishita Bharadwaj, Dept. of CSE - PES1UG20CS648

Collaborated with Hita Juneja - PES1UG20CS645

Exploring Data with R

```
# Load CSV
library(tinytex)
library(ggplot2)
data <- read.csv("top_1000_instagrammers.csv", header=TRUE)
```

Solutions

Problem 1

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 ?

Null and 0 values from Audience Country, Engagement Avg. and Authentic Engagement have been removed as they cannot be replaced by any mean value. They are MCAR (Missing Completely At Random), the missing data records are independent of records with complete data.

There's a drastic difference between the median(305,900) and mean (diff= 203,959). In fact, the mean(509,859) tends to be closer in value to the 3rd Quartile, aka the 75th percentile. From the histogram, distribution is positively skewed as a lot of the influencers have a lesser engagement than the mean. Engagement Avg. summary stats are very similar to that of Authentic Engagement.

The highest number of influencers are from the United States (mode=US). Expectedly, my Instagram stats are nowhere near the top 1000. After all, all my time goes into "authentically engaging" in classes :)

```
#Removes 0 and NULL values for columns
#907 records. => 94 removed.
data=data[data$Authentic.Engagement != 0, ]
data=data[data$Category != '', ]
data=data[data$Audience.Country != '', ]
print(nrow(data))
```

```
## [1] 907
```

Summary Statistics for Authentic Engagement

```
x <- data$Authentic.Engagement
sum_auth <- summary(x)
```

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

```
std_auth  <- sd(x) * sqrt((length(x)-1)/length(x))
mode_auth <- getmode(x)
print(sum_auth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20600  169000  305900  509859  559050 7000000
```

```
print(std_auth)
```

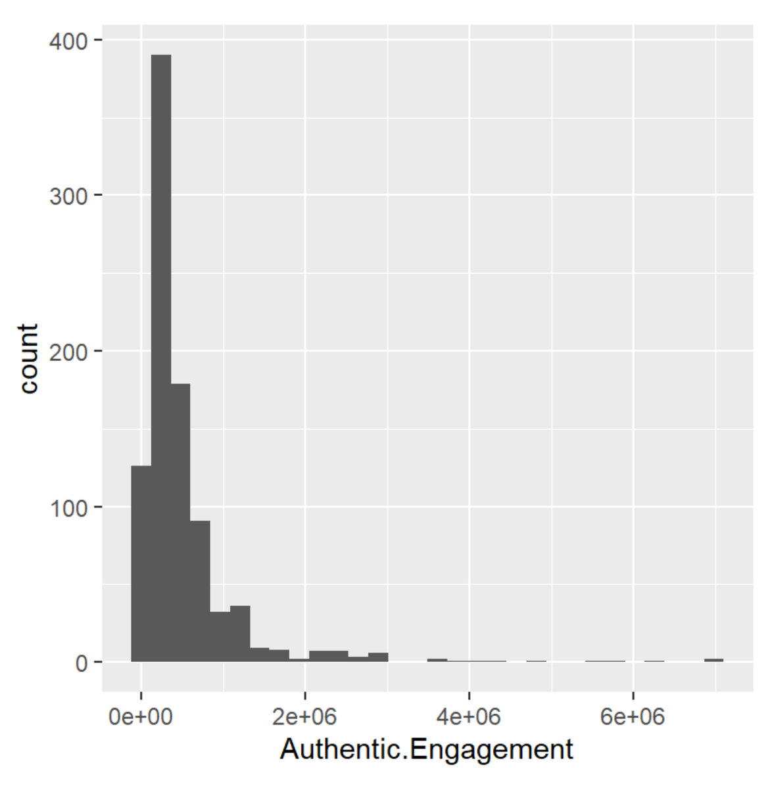
```
## [1] 698999.7
```

```
print(mode_auth)
```

```
## [1] 1200000
```

```
ggplot(data, aes(x=Authentic.Engagement)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Summary Statistics for Audience Country

```
# factoring categorical data to numeric data.
y<- factor(data$Audience.Country)
y_new=unclass(y)
print(y_new)
```

```
##   [1] 11  3 33 33 33 33  4 33 33 12 12 11 33 33 33 12 33  4 11 11 12 26 33 33 11
##  [26] 12 33 33 18 11  4 33 33 33 11 33 11 33 11 33 33 12 33 24 11 26 12 11  4  9
##  [51] 11 11 11 11 18 33  4  4 11 33 33 11 18 18 15  8 12 12 30 33  3 33 12 27  4
##  [76]  4 33 12 11 33 11 11 18 33  6 11 18 12 33  9 33  3 12 11 11 12 11  4  4 26
## [101] 33 33  4 11 24 11 11  4 33  4 11  4 11 33 33 11 26 33 27 11 33 11 33 27 11
## [126]  4 33 33 33 15 11 18 33 33 11 27 11  4 27 11 13 11 33 33 11 18 33 33  4  4
## [151] 11  4  4 27 13 11 12 12  7  7 33  3  4  4 18 11 30 30 12 33 33  4 33 33 33
## [176] 27 12 12  4 12 27 18 11  3  4 12 12  4 27 30 27 11 33 12  4 33 11  4 11 33
## [201] 11 11 11 33 12 12 12 33 33 12  4 18 12 11  3  4 11  4  3 22 12  3 11  4 12
## [226] 11 33 24 33 11 18 18 27  4  4 11 12  3 18 12 11 26 11 18  4 11 11 11 12 11
## [251] 33 12 11 33  4 12 33 11  4 33 33 11 33  4  3 12 12 11 13  4  3 23 12 12  4
## [276] 33 33 12 12 12 33 12 24 29 12 33 18 12 22 33 11 12 33 12 33 12 27 11 12  4
## [301] 33 32 32 33 15 33 12 12 18 33 32 33 26 33 33 11 33 11 27 33 33 33 12 18 33
## [326] 13  3 33 12 33 24 11  8 33 11 33  4  4 33 33  4  4  3 11 33 33 33  4 33 33
## [351] 33 26 12 33 32 12 12 33 33 33 11  4 33 11 12 33 11 11  4 11  3  4 12 33 24
## [376] 12 11 30 11 12 12 33  4 30 30 12 12 27 12 11 12 27 24 11 33 33  3 33  4 26
## [401] 33  4  4 11 33  3  4 12 15 24 24 11 11 33  3 18 27  4 33 33 14 33  4 18 12
## [426]  4 33 12  4 27 11 24 33  4 33 11  7 33  4 33 33 33 33  4 11 29  4 11 13 15
## [451] 32  4 12 33  4 10 12  4 13 33 11 11 33 27  4 26 24  4 12 30 33  4  7 33  4
## [476]  7  3 33 11 33  4  4  4 20 33 22 33 20 33  4  7  4 33 33 12  4 12 33  4 33
## [501]  4  4 33 33  4  4  4 32 11 32 13 33 11 33  9 11 27 33 32 33 33 33 33 33 33
## [526] 11 11 19  4 12 11 12 11  3 33  4 33 33 33 32 27 18 33 33 33  4 33  4 18 12
## [551] 33 32 33 11 12 33 10 33 12 18 11 12 33 12 11  4  4 28 33 33 11 11 12 11 15
## [576] 27 14 33  4 11 33 33  4 11 24 11 33  4 11 11 11  4 33 32 10 24  4 15 11 33
## [601]  4 31 16  4 11 33 29 22 33  4 24  2 24  4 11 12 27  4 33 11 12  4 33 33 33
## [626] 11 18  4 32 12 32 18 33 27 33 33 33 13 18 33  4 11 33 13 30 33 24 32 15 13
## [651]  3 15  3 13 11 33 11 30 11 33 33 26 15  4 33 33 15 33 13 11 11  4 11 24 18
## [676] 33  4 18 12 12  4 27 33 11 12  8  4 33 18 33  4 33 33  3  7  9 27 33 33  4
## [701]  3 25 15 27 12 12 33 33  4 32 33  4  4 11 12 33 33  4 26  4 11 33  4 19 33
## [726] 33 11 15 30 13 11 33 12 22  4 33 27 33  7 33 18 29  4 12 16 12 24 12 12 18
## [751] 12 27 24  4 33  4  4 33  5 33 32 27 33 33 12  4 11 11 24 12  4 11 11 24 17
## [776] 12  4  4 24 27 18 33 11 13 33 11  4 33 33 11 32  4 27 12 11  4 33 18  4 18
## [801]  4 33  8 29 12 12 33 27 18 30 16  4  4 33 33  4 33 11 33 29 30 33  4 24 33
## [826] 33 27 15 24  4  4  9 15 33 13 30 11 11 30 12  4  4 33 33 33 11 33 11 33 30
## [851] 12 11 30 15  1 33 12 27 33  4 33 33 12  4 12 33 33 33  4 12 33 11 11 11 30
## [876] 11 11  8  4 11 33 33 33  4 11 33 11 14 13 10 33 33 33 11 33 33 16 12 22 33
## [901] 12 21 22 11 12  9 14
## attr(,"levels")
##  [1] "Albania"              "Algeria"              "Argentina"
##  [4] "Brazil"               "Chile"                "China"
##  [7] "Colombia"             "Egypt"                "France"
## [10] "Germany"              "India"                "Indonesia"
## [13] "Iran"                 "Iraq"                 "Italy"
## [16] "Japan"                "Kazakhstan"           "Mexico"
## [19] "Morocco"              "Nigeria"              "Pakistan"
## [22] "Philippines"          "Poland"               "Russia"
## [25] "Senegal"              "South Korea"          "Spain"
## [28] "Syria"                "Thailand"             "Turkey"
## [31] "United Arab Emirates" "United Kingdom"       "United States"
```

```
sum_country<-summary(y_new)
std_country  <- sd(y_new) * sqrt((length(y_new)-1)/length(y_new))
mode_country <- getmode(y_new)
```

```
print(sum_country)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   11.00   13.00   18.51   33.00   33.00
```

```
print(std_country)
```

```
## [1] 11.30116
```

```
print(mode_country)
```

```
## [1] 33
```

Summary Statistics for Engagement.Avg.

```
z<-data$Engagement.Avg.
sum_engagement<-summary(z)
std_engagement  <- sd(z) * sqrt((length(z)-1)/length(z))
mode_engagement <- getmode(z)
```

```
print(sum_engagement)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21100  239850  422900  690407  780400 8300000
```

```
print(std_engagement)
```

```
## [1] 873746.3
```

```
print(mode_engagement)
```

```
## [1] 1100000
```

Problem 2

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.

Top 3 audience countries that follow top 100 Instagrammers are United States, United Kingdom, and United Arab Emirates

```
a<-data$Audience.Country
df1<-data.frame(a)
data_new1 <- unique(df1[order(df1$a, decreasing = TRUE), ])  # Order data descending
data_new1
```

```
##  [1] "United States"        "United Kingdom"       "United Arab Emirates"
##  [4] "Turkey"               "Thailand"             "Syria"
##  [7] "Spain"                "South Korea"          "Senegal"
## [10] "Russia"               "Poland"               "Philippines"
## [13] "Pakistan"             "Nigeria"              "Morocco"
## [16] "Mexico"               "Kazakhstan"           "Japan"
## [19] "Italy"                "Iraq"                 "Iran"
## [22] "Indonesia"            "India"                "Germany"
## [25] "France"               "Egypt"                "Colombia"
## [28] "China"                "Chile"                "Brazil"
## [31] "Argentina"            "Algeria"              "Albania"
```

```
top3 = data_new1[1:3]
print(top3)
```

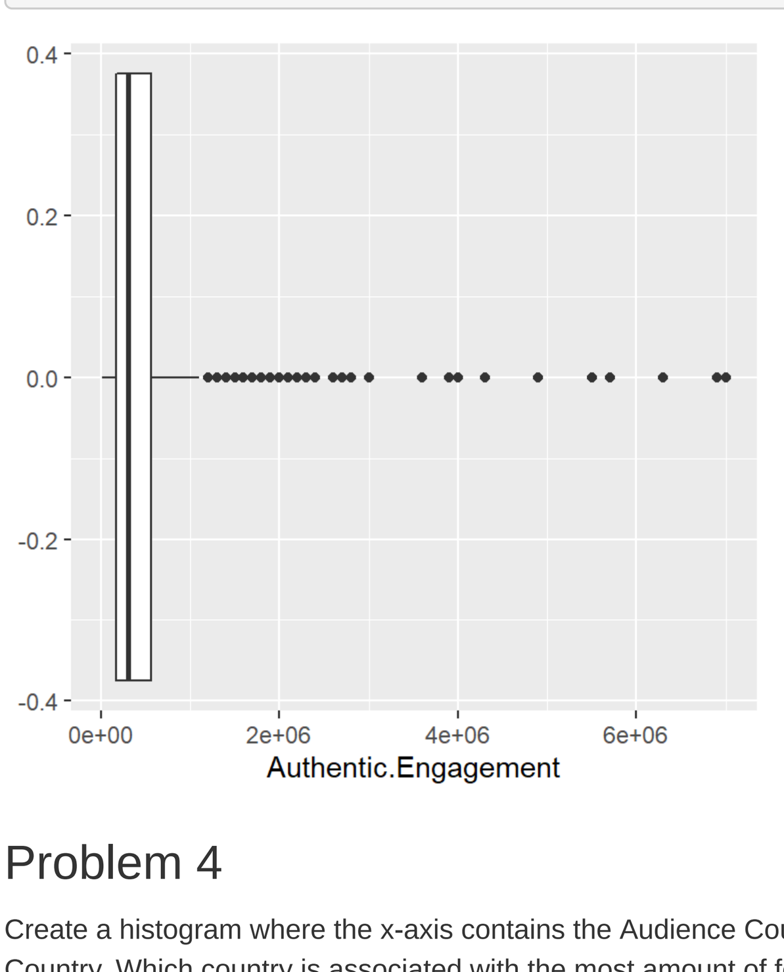
```
## [1] "United States"        "United Kingdom"       "United Arab Emirates"
```

Problem 3

Create a horizontal box plot using the column `Authentic.Engagement`. What inferences can you make from this box and whisker plot?

Most of data points are clustered left(lower values) of the boxplot.
Few records have very high (upto 7,000,000) values, which is why mean is comparatively greater than median. Distribution of this column is positively skewed.

```
library(ggplot2)
ggplot(data, aes(y=Authentic.Engagement)) +
  geom_boxplot() +
  coord_flip()
```



Problem 4

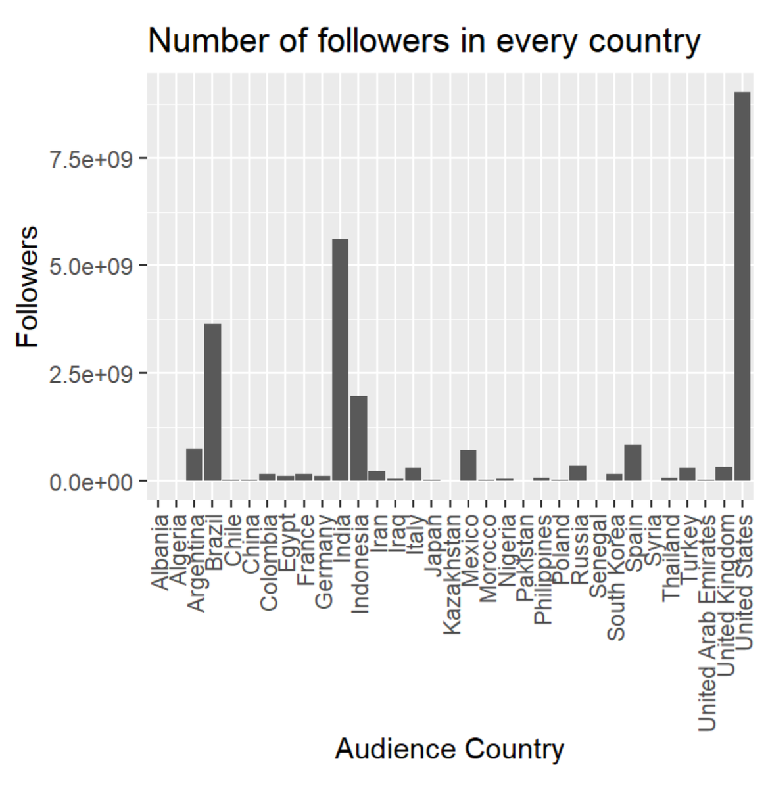
Create a histogram where the x-axis contains the Audience Country and y-axis contains the total follower count for accounts with that Audience Country. Which country is associated with the most amount of followers? *Hint:* Recall the concept of `groupby()` in Pandas. Try using the `aggregate()` function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

Total follower count for India is 5,619,600,000, and has 2nd largest number(rank) of followers.

```
agr <- aggregate(data$Followers, list(data$Audience.Country), FUN=sum)
print(agr)
```

```
##                 Group.1          x
## 1               Albania   10300000
## 2               Algeria    9100000
## 3             Argentina  741800000
## 4                Brazil 3632300000
## 5                 Chile   14100000
## 6                 China   28900000
## 7              Colombia  170200000
## 8                 Egypt  125500000
## 9                France  162400000
## 10              Germany  108100000
## 11                India 5619600000
## 12            Indonesia 1961500000
## 13                 Iran  220600000
## 14                 Iraq   45500000
## 15                Italy  296800000
## 16                Japan   24300000
## 17           Kazakhstan    6900000
## 18               Mexico  730000000
## 19              Morocco   17300000
## 20              Nigeria   39500000
## 21             Pakistan    6600000
## 22          Philippines   79500000
## 23               Poland   26800000
## 24               Russia  352300000
## 25              Senegal    1600000
## 26          South Korea  170200000
## 27                Spain  823300000
## 28                Syria    9000000
## 29             Thailand   60300000
## 30               Turkey  294000000
## 31 United Arab Emirates   14400000
## 32       United Kingdom  319100000
## 33        United States 9017700000
```

```
ggplot(agr, aes(x=Group.1, y=x)) + geom_col() +theme(axis.text.x = element_text(angle = 90,vjust = 0.5, hjust=1))
+ggtitle("Number of followers in every country") +xlab("Audience Country")+ylab("Followers")
```



Conclusion

My Instagram profile has been around for 2 years now. I have 700 followers with an approximate interaction with 200 followers. My account falls under the Lifestyle category. You'll find pictures of family, friends, trips and other fun things I've done on my account profile page. Unlike fan accounts, my account is a personal/private one with known followers. Hence, there can hardly be any comparison with top 1000 celebrities/instagrammers.

Best way to increase followers and user engagement: 1. Be part of the Instagram communities that follow top instagrammers/influencers. Most of them are either musicians, sportspersons or actors. It's highly probable I will interact/communicate with accounts that are like-minded or have similar content to the influencers I look up to/am a fan of.

2. Putting out content that's targeted to be comprehensible/relatable to the mass of the audience countries. They are well represented(US, India, Brazil).