PES University, Bangalore

Established under the Karnataka Act No. 16 of 2013

UE20CS312 - Data Analytics

Worksheet 1a - Part 1: Exploring Data with R

Harshith Mohan Kumar - harshithmohankumar@pesu.pes.edu

Anushka Hebbar - anushkahebbar@pesu.pes.edu

Exploring Data with R

Prerequisites

This worksheet aims to develop your understanding of summary statistics and basic visualizations through a pragmatic approach. To download the data required for this worksheet, visit this Github link.

Resources:

- Check out this beautifully comprehensive resource for everything you need to get started with R.
- This online book provides guided explananations about visualizations in R using the `ggplot2` library.

1. About the Data: Top 1000 Instagrammers

To make this worksheet a bit interesting for you all, we have picked a dataset from Kaggle which comprises of the details of the top 1000 influencers on Instagram. If you are on this list, send me an email ;P

This dataset has been taken from this Kaggle dataset by Syed Jafer.

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.
```

2. Assignment Submission Format

The following problems are to be completed using the R programming language and should be submitted as a R markdown file (`.rmd`). Since the dataset is public and many of you students will have the same numerical answers, the grades are allocated on the analysis of the problems and personalized answers within the conclusion section.

The markdown file should follow this format:

```
---
title: "UE20CS312 - Data Analytics - Worksheet 1a - Part 1 - Exploring data with R"
subtitle: "PES University"
author:
  - 'INSERT_NAME, Dept. of CSE - INSERT_SRN'
output: pdf_document
urlcolor: blue
editor_options:
  markdown:
    wrap: 72
---

## Solutions

### Problem 1
INSERT SOLUTION CODE IN MARKDOWN
INSERT SCREENSHOT OF R OUTPUT
INSERT ANALYSIS

### Problem 2
INSERT SOLUTION CODE IN MARKDOWN
INSERT SCREENSHOT OF R OUTPUT
INSERT ANALYSIS
(etc)

### Conclusion
INSERT SUMMARY
```

3. Loading the Dataset

Step 1: Ensure you are on the right working directory and the CSV exists in this directory.

```
# Get and print current working directory
print(getwd())

# Set current working directory
setwd("/UR/WORKING/DIRECTORY")

# Get and print current working directory
print(getwd())
```

Step 2: Read CSV File

```
# Load CSV
data <- read.csv("top_1000_instagrammers.csv")
```

4. Preliminary Guided Exercises

Make sure you have the R programming language installed on your system. It is also recommended to make sure RStudio, the popular IDE for R, is installed. RStudio provides a lot of useful functionality like R markdown, a script editor and GitHub integration. Use RStudio Projects as a great way of keeping each week's assignment work organized.

Data Import To import data from CSV files into a DataFrame:

```
data <- read.csv('top_1000_instagrammers.csv', header=TRUE)
```

The `header = TRUE` argument specifies that the first row of your data contains the variable names. If this is not the case you can specify `header = FALSE` (this is the default value so you can omit this argument entirely).

Compact Summary Use the `str()` function to return a compact and informative summary of the DataFrame.

```
str(data)
```

```
## 'data.frame':  1000 obs. of  8 variables:
##  $ X                   : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Name                : chr  "cristiano" "leomessi" "kendalljenner" "arianagrande" ...
##  $ Rank                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Category            : chr  "Sports with a ball" "Sports with a ballFamily" "ModelingFashion" "Musi
##  $ Followers           : num  4.63e+08 3.47e+08 2.48e+08 3.21e+08 1.47e+08 ...
##  $ Audience.Country    : chr  "India" "Argentina" "United States" "United States" ...
##  $ Authentic.Engagement: num  5500000 3600000 3000000 2400000 4300000 1700000 2400000 1200000 140000
##  $ Engagement.Avg.     : num  6600000 4800000 4900000 3400000 5800000 2500000 3200000 1900000 190000
```

Here we see that flowers is a 'data.frame' object which contains 1000 rows and 8 variables (columns). Each of the variables are listed along with their data class and the first 10 values.

Summary Statistics To access the data in any of the variables (columns) in our data frame we can use the $ notation. Indexing in R starts at 1, which means the first element is at index 1. Access the first 10 values of the `Name` column:

```
data$Name[1:10]
```

```
##  [1] "cristiano"     "leomessi"      "kendalljenner" "arianagrande"
##  [5] "zendaya"       "kimkardashian" "taylorswift"   "kyliejenner"
##  [9] "selenagomez"   "thv"
```

We can assign a column to another variable and calculate a mean of a numeric variable or get a summary of a variable using the summary() function.

```
names <- data$Name
summary(names)
```

```
##    Length     Class      Mode
##      1000 character character
```

```
auth_eng <- data$Authentic.Engagement
mean(auth_eng)
```

```
## [1] 566199.2
```

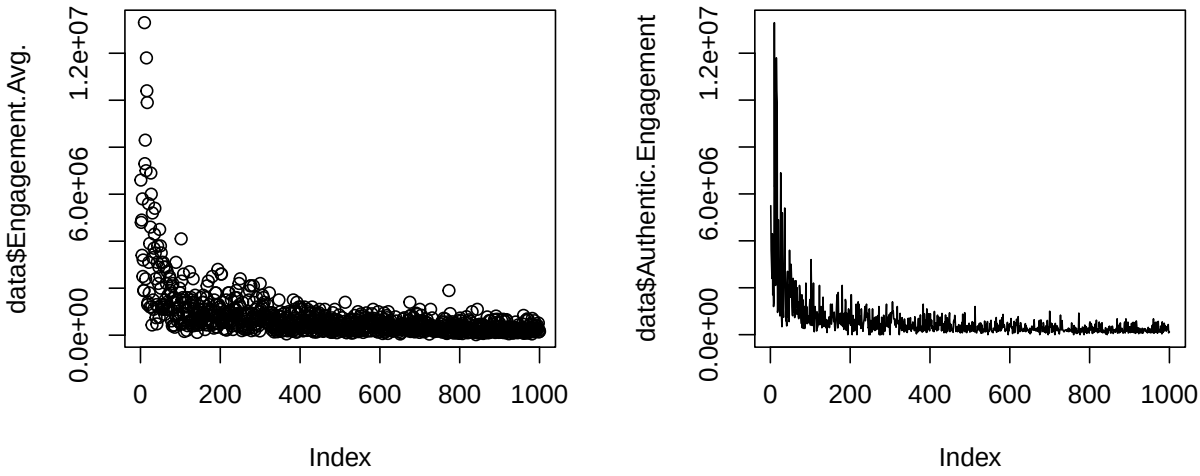
```
summary(auth_eng)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0   169000   316050   566199   604275 13300000
```

Notice how the behavior of the `summary` function changes with different types of variables. Let's now try to explore how we can visualize our data.

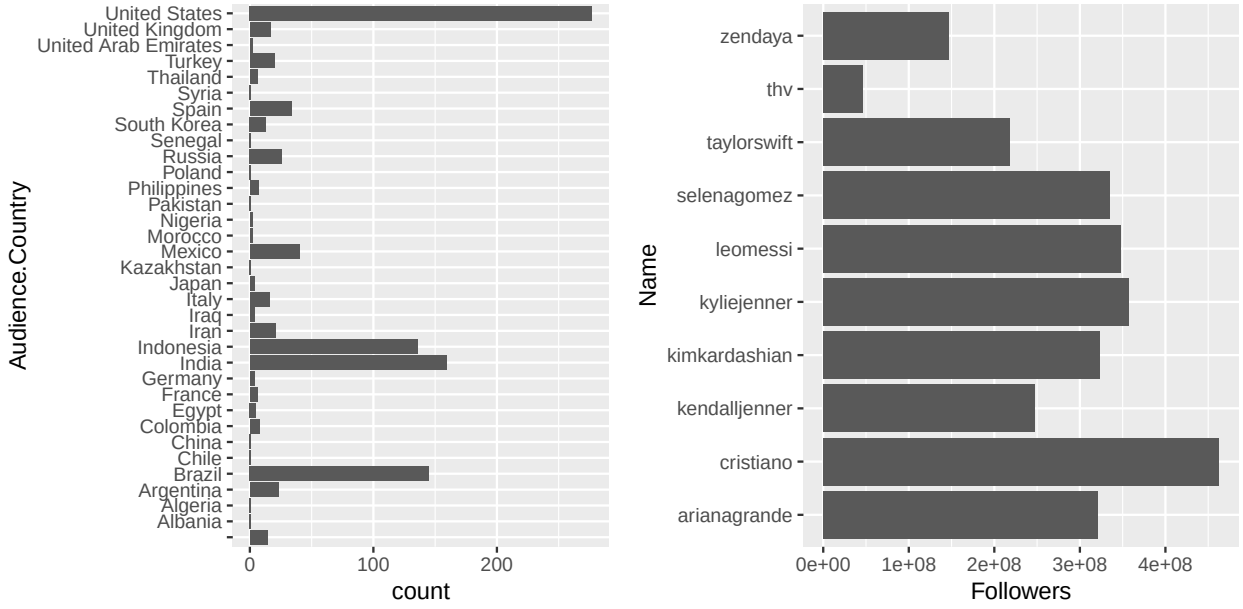
Scatterplots The most common high level function used to produce plots in R is the `plot` function.

```
plot(data$Engagement.Avg., type="p")
plot(data$Authentic.Engagement, type="l")
```



Using the ggplot2 Library Create a bar graph for the distribution of the categorical variable `Audience.Country`.

```
# Import the library for visualization
library(ggplot2)
# Create a bar graph
ggplot(data, aes(x=Audience.Country)) + geom_bar() + coord_flip()
# Number of followers of the top 10 most followed instagrammers
ggplot(data[1:10,], aes(x=Name, y=Followers)) + geom_col() + coord_flip()
```



5. Points

The problems for this part of the worksheet are for a total of 8 points, with a non-uniform weightage.

- *Problem 1* : 1 point
- *Problem 2* : 2 points
- *Problem 3* : 1 points
- *Problem 4* : 3 point
- *Conclusion* : 1 point

6. 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 ?

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.

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?

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?

Conclusion (1 point)

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?