PES University, Bangalore

Established under Karnataka Act No. 16 of 2013

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

Designed by Harshith Mohan Kumar, Dept. of CSE - harshithmohankumar@pesu.pes.edu

Exploring data with R

This worksheet aims to develop your understanding of summary statistics and basic visualizations through a pragmatic approach.

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 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.
```

Problems

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
```

Load the Dataset

```
library(tidyverse)
# Remember: You need to install tidyverse package to load it!
df <- read_csv(path_to_csv)
```

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 ?

Problem 2

Create a histogram where the x-axis contains the Audience Country and y-axis contains the total follower count of all users in that country. Which country has the most amount of followers? (Hint) Use a dictionary to maintain the sum of followers across countries. What is the total for India and what rank does it fall compared to other countries?

Problem 3

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

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?