CS544 Final Project

Picking the Data Set

Look into the following sites as an example and select a data set that interests you.

1. https://www.kaggle.com/datasets
2. https://github.com/fivethirtyeight/data
3. http://www.kdnuggets.com/datasets/index.html
4. Any other source of your choice

Preparing the data

- Import the data set into R.
- Document the steps for the import process and any preprocessing had to be done prior to or after the import. Any R code used in the process should be included.

Analyzing the data

- Do the analysis as in Module3 for at least one categorical variable and at least one numerical variable. Show appropriate plots for your data.
- Do the analysis as in Module3 for at least one set of two or more variables. Show appropriate plots for your data.
- Pick one variable with numerical data and examine the distribution of the data.
- Draw various random samples of the data and show the applicability of the Central Limit Theorem for this variable.
- Show how various sampling methods can be used on your data. What are your conclusions if these samples are used instead of the whole dataset.
- **Implementation of any feature(s) not mentioned in the above specification.**

Presenting the Project

- **You will do your project presentation with the Facilitator using Zoom.**
- **Each presentation is for at most 10 minutes. Signup sheet will be provided later.**

Grading Rubric:

- **Preparing the Data and documenting the data preparation (15 points)**
- **Analyzing the Data and documenting the same (50 points)**
- **Implementation of any feature(s) not mentioned in the specification (10 points)**
- **Presenting the project in the Live Classroom with Facilitator (25 points)**

Submitting the Project

Upload a zip file (CS544Final_lastName.zip) containing all the code as RMarkdown (Rmd file), the presentation document (PDF or PPT, if any), and all the results in a RMarkdown HTML.