**CS544 Final Project**

**Picking the Data Set**

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

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

**Preparing the data (15 points)**

* Import the data set into R.
* Document the steps for the import process and data preparation, processing, and cleaning procedures had to be done. Any R code used in the process should be included.

**Analyzing the data (50 points)**

* Do the analysis as in Module 3 for at least one categorical variable and at least one numerical variable. Show appropriate plots and properly label the plots. (10 points)
* Do the analysis as in Module 3 for at least one set of two or more variables. Show appropriate plots for your data. (10 points)
* Draw various random samples (using at least 3 different sample sizes) of the data and show the applicability of the Central Limit Theorem for at least one variable. (15 points)
* Show how various sampling methods (using at least 3 sampling methods) can be applied on your data. What are your conclusions if these samples are used instead of the whole dataset. (15 points).
* Implementation of additional feature(s) not mentioned above (20 points)

**Presenting the Project (15 points)**

* You will present your project on Wednesday 4/21 and Monday 4/26 in Zoom
* Each presentation should be no more than 7 minutes (this leaves us some time for Q&A).

**Submitting the Project**

Upload a zip file (CS544\_project\_yourName.zip) containing all the code (R file), the presentation document (PDF or PPT), and other results if any in a Word/PDF Document.

**Grading Rubric:**

* **Clear documentation of the data import and preprocessing steps (15 points)**
* **Correct data analysis methods and results; Appropriate visualization techniques and results. (70 points)**
* **The overall quality and organization of your slides and how well you present your work and answer questions. (15 points)**