**CS 240 Exploratory Data Analysis**

**Project**

*May 12,2017 (23:59 pm)*

**Introduction**

For the project, you will conduct your own data analysis and create a report to share that document your findings. You should start by take a look at your dataset and brainstorming what questions you could answer using it. You just need to complete following parts, you can use Python, Pandas, NumPy, Thinkplot and Thinkstats modules to answer the questions you are most interested in, and create a report sharing the answers.

The data you will analyze is;

**Baseball Data** - A data set containing complete batting and pitching statistics from 1871 to 2016, plus fielding statistics, standings, team stats, managerial records, post-season data, and more. This dataset contains many files, but you can choose to analyze only the one you are most interested in.

This project is open-ended in that we are not looking for one right answer. Just you must answer following questions in parts.

PART 1

Brainstorm some questions you could answer using the data set you chose, then start answering those questions. Here are some ideas to get you started:

* What is the relationship between different performance metrics? Do any have a strong negative or positive relationship?
* What are the characteristics of baseball players with the highest salaries?

Finally, you must state one question that you want to analyze. Then give your hypothesis for that question to test last part

PART 2

Show the columns that you are going to use then clean and organize your data to start analysis show your codes and explain it what it does.

PART 3

According to your question first show some relevant statistics then plot; 1 Histogram, 1 PMF and 1 CDF show your codes and explain it what it does.

PART 4

Use one of modelling distributions, model your data and explain it why you used that and what it does explain with showing also your codes

PART 5

Built one relationship according to your question and choose 2 variables in your data explain and show their correlation then visualize this correlation. Also, show your codes and explain what it does.

PART 6

Test your hypothesis step by step, show your codes, and explain what it does.

PART 7

Write a conclusion that describe your analysis and what you get end of the analysis.

**Report File**

After you finish your work upload following 3 files to your GitHub account and send me to your GitHub account’s link. No other alternative way allowed.

There must be 3 file that must be uploaded by you

1. Data File: The Data file that you use in your analysis
2. Codes File:That must show your codes that you use for each part and explain it why you used in comments. This file may be a Jupyter Notebook(Recommended) or Python File.
3. Report File: That must be at least 3 page with visuals that must require above conditions in each part and it must be PDF file

For any questions contact with assistant