# CS 240 Exploratory Data Analysis
# Project

*June 1,2018 (13:00)*

## 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 taking a look at your dataset and brainstorming what questions you could answer by using the data. You just need to complete following parts, which are decsribed detailly below. You can use Python, Pandas, NumPy, Matplotlib, Seaborn, Thinkplot and Thinkstats modules to answer the questions you are most interested in, and create a report with the answers that you come up with.

**The data you will analyze is;**
**Basketball Data** - 11 dataset containing complete statistics for NBA (National Basketball Association) in 2011-12 NBA season. These datasets contain many files such as players, teams, coaches, matches etc… and you can choose to analyze any datasets that you are mostly interested in.
**This project is open-ended, so we are not looking for one strict right answer but you have to write a report in the lights of what you have learnt from this course. The grades will be based on 2 part; first part is basically how well you understand questions and apply it to your analyze, then the most important thing is how you interpret your result and present it in report. The report must be written in section by section and answer following questions in REPORT PART which is 75 points of this project. Second part is about your code file (how clear your code is and how clear you express it with comments etc.). This part is 25 points worth.**

## A – REPORT PART (75 pts.)

## SECTION 1(10 pts.)

Only a brainstorming some questions you could answer using the datasets you chose, then start answering those questions. Here are some ideas in order to create those questions:

- What is the relationship between different performance metrics?
- Could it be any strong negative or positive relationship between variables?
- What are the characteristics of variables in datasets?

Finally, you must create at least 3 question that you can analyze. Then select **one** of among those three questions, create your hypothesis, and explain why you choose that question for hypothesis. That is necessary for the testing at last part.

## SECTION 2 (10 pts.)

According to your hypothesis, show the variables and datasets that you are going to use then clean and organize your data to start analysis, interpret your results and explain what it means in a clear way.

## SECTION 3 (10 pts)

According to your hypothesis first show at least 5 descriptive statistics then plot; 1 Histogram, 1 PMF and 1 CDF, interpret your results and explain what it means in a clear way.

## SECTION 4 (10 pts)

According to your hypothesis, use one of modelling distributions, model your data and interpret your results (how your model fits to your data etc.) and explain what it means in a clear way.

## SECTION 5 (10 pts)

Built one relationship according to your hypothesis and choose 2 variables in your data explain and show their correlation then visualize this correlation. Also, interpret your results and explain what it means in a clear way.

## SECTION 6 (10 pts)

Test your hypothesis step by step, show your steps and explain why you need that step and what needed for that step. Interpret your results (according to the p-value) at the end and explain what it means in a clear way.

.

## SECTION 7 (15 pts)

Write a conclusion that describes your analysis and what you get end of the analysis such as the result of hypothesis testing and some important findings out of your whole analysis.

## B – CODING PART (25 pts.)

Write your codes as an answer of sections in report part. Write your codes section by section (with the corresponding parts above) and explain with comments. The points will be given on how well you explain your codes in comment and consistant with related sections.

## What to SUBMIT

After you finish your work, upload the following 3 files to your GitHub account and send me your GitHub account’s link. **No other alternative way is allowed**.
Your repository should include the following files.

- Data Files: The Data files that you use in your analysis
- Code File:For part B. This file may be a Jupyter Notebook (Recommended) or Python File.
- Report File: That must be at least 5 page with visuals that must require above conditions in each section in part A and it must be in PDF file format.

## Interview

You have to make interview in order to get grade. There will be no more other interview date

For any questions contact with assistant