Assignment Overview

In this week, we have introduced the concept of “hot hand” in sports and learned to test the existence of “hot hand” with both summary statistics and regression analysis using a 2016-2017 NBA shot log dataset. In this assignment, you will perform similar analyses using shot log information in the 2014-2015 NBA season. The shot log dataset that you will explore contains information on the outcome of each individual shot attempted by each player. It also has other relevant information on the shots and the game. Not all information in the 2016-2017 shot log dataset is present in this dataset. The 2014-2015 shot log dataset contains other useful information about individual shots.

To complete the assignment you will need to perform similar-steps in the lecture.  You will be asked to answer several multiple choice questions based on your data and result. After passing each quiz, you will be provided sample code that produces the correct answer for that portion of the assignment. You should use this sample code as the basis to move onto the next part of the assignment if your code differs greatly.

**Beware:** even though your code might get you to the correct answer at a given point, it is sometimes possible that the way you write it might interfere with completing a further step. So even if you get the answer right, you should look at the code we supply to check if you are going the same way. In practice, there are often many ways to get to answer in Python, and we do not insist that you follow our approach exactly – but simply warn you to be aware that differences could turn out to be problematic later.

# Assignment Instructions - Part 1

## **I. Data Preparation and Exploration**

1. Import the “Shotlog\_14\_15.csv” data file as “Shotlog\_1415” into Jupyter Notebook. Import “Player\_Stats\_1415.csv” data file as “Player\_Stats” into Jupyter Notebook. Descriptions of the datasets and selected variables
   1. a) In the dataset “Shotlog\_14\_15,” each observation represents an attempt of a shot. In the dataset “Player\_Stats\_14\_15,” each observation represents a player.
   2. b) The “average\_hit” variable in both dataframes indicate the average success rate of a player making a shot over the season. It is defined and calculated the same way in both dataframes.
   3. c) The variable “home\_away” indicates whether the team that the player belongs to played at home or away.
   4. d) The variable “result” indicates whether the team that the player belongs to won or lost the game. The variable “final\_margin” represents the difference in final score between the team the player belongs to and their opponent’s.
   5. e) The variable “shot\_number” is the order of the shot the given player attempted at the given game.
   6. f) “game\_clock” is the countdown clock for each quarter. The game clock starts at 12 minutes. “shot\_clock” refers to the display of a countdown clock of the time within which the team possessing the ball must attempt a field goal. The shot clock starts at 24 seconds.
2. Convert the “date” variable to a date type variable and calculate summary statistics for the “shot\_clock” variable.
3. Create a lagged variable “lag\_shot\_hit” to indicate the result of the previous shot by the same player at the same game. Hint: In this dataset, the variable “match” may not be able to uniquely identify each game; you can use “game\_id” instead. You can sort the data by shot number for each player to create the lagged variable.
4. Create a variable “error” to indicate the prediction error for each shot and a variable “lagerror” for the prediction error for the previous shot. The “error” variable is defined as the difference between the outcome of the current shot and the average success rate (“average\_hit”) and the “lagerror” variable is defined as the difference between the outcome of the previous shot and the average success rate.
5. Calculate summary statistics for the “error” and “lagerror” variables.

# Assignment Instructions - Part 2

## **II. Conditional Probability and Autocorrelation**

1. Create a dummy variable “conse\_shot” that indicates a player made consecutive shots.
2. Create a dataframe “Player\_Prob” for the probability of making the previous shot and the joint probability for making both the previous and current shots. Name the probability of making the previous shot “average\_lag\_hit” and the probability of making both shots “conse\_shot\_hit.”
3. In the “Player\_Prob” dataframe, calculate the conditional probability “conditional\_prob” for a player to make a shot given that he made the previous shot.
4. Merge the “Player\_Prob” dataframe into the “Player\_Stats” dataframe.
5. Calculate summary statistics for the probability for a player to make a shot (“average\_hit”) and the conditional probability for a player to make a shot given that he made the previous one (“conditional\_prob”) and the probability of players making consecutive shots (“conse\_shot\_hit”).
6. Perform a t-test for the statistical significance on the difference between conditional probability and unconditional probability of making a shot.
7. Calculate the first order autocorrelation coefficient on making a shot (correlation coefficient between making the current shot and the previous shot) for the entire shotlog dataset.
8. Calculate the first order autocorrelation coefficient on making a shot for each player. Display the top ten players with the highest first order autocorrelation coefficient.

# Assignment Instructions - Part 3

## **III. Regression Analysis**

In this section, you will run several regressions to investigate the “hot hand.” In all the regressions, the dependent variable is “error” and the independent variable of interest is “lagerror.”

**Reg1**: Run a linear least squares regression using the entire shotlog dataframe. Include the following control variables:

* Shot distance
* Number of dribbles
* Touch time
* Type of shot (“points” variable)
* Quarter of the game (as a categorical variable)
* Home or away game
* Shoot\_player
* Closest defender
* Closest defender distance

**Reg2**: Run a weighted least squares regression using the entire shotlog dataframe. Include the same set of control variables as in Reg1. The regression should be weighted by the number of shot per game (weight=1/shot\_per\_game).

**Reg3\_player**: Run linear least squares regressions on individual players. Include the following control variables:

* Shot distance
* Number of dribbles
* Touch time
* Type of shot (“points” variable)
* Quarter of the game (as a categorical variable)
* Home or away game
* Closest defender distance

**Reg4\_wls\_player**: Run weighted least squares regressions on individual players. Include the same set of control variables as in Reg3. The regression should be weighted by the number of shot per game (weight=1/shot\_per\_game).