# Project 3 Proposal

## By Sujin Guo, Yang Yang, Yan Xu and Erik Dubec

### Background

This project aims to explore how a variety of factors can be used to predict the salary of the NBA players from 1999-2019. Our goals for this project:

1. Salary prediction (discover which factors are the best predictors of an NBA player’s salary, using regression model to predict salaries).
2. Using cluster model to find out similarity or categorizing players (for example, utilizing physical characteristics data to predict performance of NBA players).
3. Determine which players have been overvalued and undervalued according to their given vs. predicted salary.

### Input datasets

We plan to get out input data from the following sources:

1. Physical characteristics of the 1996~2016 NBA players

<https://www.kaggle.com/justinas/nba-players-data/downloads/nba-players-data.zip/2> (CSV format)

1. Data for injuries

<https://www.kaggle.com/ghopkins/nba-injuries-2010-2018> (CSV format)

1. NBA Player Stats since 1950

<https://www.kaggle.com/drgilermo/nba-players-stats> (CSV format)

1. NBA Player Salary Data 2002/03 - 2017/18

<https://www.kaggle.com/hultm28/nba-player-salary-data-2002-2017> (CSV format)

The average NBA salary has drastically increased over the past 20 years. The solution is to normalize salary data by putting it as a percentage of the league’s salary cap, the total salary limit that a team can spend on its players in a given season. The salary cap has risen as player salaries have.

### Possible skills

* Python Pandas
* Python Matplotlib
* Power point ☺