# ETL Project Proposal

## By Yang Yang and ZiAng Zhang

### Background

The purpose of this project is to find the factors that could determine the income of basketball players, including team, player position, field goal percentage, length of career, and (even) height, etc. Strong correlation may indicate possible causation. Please see below for a detailed description for choice of input data, processing steps, and format of our final output database.

### Input datasets

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

1. NBA Player Stats since 1950

<https://www.kaggle.com/drgilermo/nba-players-stats>

*Flat file (CSV) dataset from Kaggle*

1. NBA Play Salaries 1999-2019

<http://www.espn.com/nba/salaries>

*Structured dataset scrapped from web*

1. NBA Salary Cap by History

<https://www.basketball-reference.com/contracts/salary-cap-history.html>

*SQL dataset (originally CSV, exported to SQL externally using Python Alchemy)*

This way we have input data from three sources of three different types.

### Data processing

We plan to conduct the following data processing steps on our input data: preprocessing (necessary cleanup for conversion), data read-in (convert uniformly into SQL or Pandas), data merging (combining data into one table, or a few well-connected tables), data exportation (export data into target database).

### Analysis

After we finish processing the data, we will carry on with the data analysis. We intend to discover the factors with the strongest correlation to our dependent variable – player salary, then apply statistical tests to verify the correlation. The end goal is to acquire the list of important factors, with the significance and manner of correlation of each identified.

### Final database

Our final database will be in SQL format, with a Flask API for responsiveness. We plan to use a relational data format due to the highly structured nature of our data as required by our analysis goals.