# ETL Project Proposal

## By Yang Yang and ZiAng Zhang

### Background

The purpose of this project is to find the factors that correlate with 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. Flat file (CSV) dataset from Kaggle:
2. SQL dataset from Kaggle:
3. Structured data from web scrapping

### 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).

### 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.