**NBA Players Stats and 2017-2018 Salaries**

Victoria Grose and Geyang Ye

ETL Project

Northwestern Data Science Bootcamp 2019

**EXTRACTION**

The data sources we extracted from were two separate data collections found on Kaggle, which comprised four total data sets: two tables containing information on player statistics (name, years, position, height, weight, birth date, birth city, birth state and college), a table containing information on aggregate individual statistics for 67 NBA seasons (basic box-score attributes such as points, assists, rebounds etc., to more advanced money-ball like features such as Value Over Replacement), and a table containing information on player salaries for the 2017-2018 season. All data was scraped from Basketball Reference and all datasets were available for download as csv files.

Links: <https://www.kaggle.com/koki25ando/salary>

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

**TRANSFORMATION**

The salary table contained duplicates when a player had played for multiple teams over the course of the season. The table was grouped by player name and the maximum player salary information was chosen.

The columns selected from the player statistics table were "G" (games),"FG" (field goals),"FGA" (field goal attempts),"AST" (assists), and "PTS"(points). These statistics were then grouped by player to obtain career totals. Then, assists and points were divided by total games to obtain averages per game and field goals were divided by field goal attempts to obtain career field goal success average. The calculated values were placed in new columns. The column “year” was deleted.

The columns selected from the player table were “name” and “position”. These columns were renamed “Player” and “Pos” respectively.

Link: <https://github.com/vgrose/ETL_project/etl_data.ipynb>

**LOAD**

Our transformed data was then loaded into a nonrelational database (mysql).

The final tables included in the database were a table containing the aggregate summary career statistics for each player active in the 2017-2018 season (nba\_career\_data), a table containing their highest 2017-2018 yearly salary (nba\_salary\_data) and a table containing each player’s positions (nba\_position\_data). These tables were joined by player name to create an aggregate table of all data.

Our final tables can be used to compare NBA player career performance and salary. The data could be further grouped by team or position to observe salary trends.