MATH 652 Final Project Proposal

Chris McMinimy, Juna Luzi, Nathan Bick

Project Scope

We will investigate basketball data available at Basketball-Reference.com. This includes in-game data for men's and women's leagues, as well as professional and college leagues. It appears that some data goes back several decades, with the current set of columns available for approximately the past forty years.

Data Model

The data is available across many pages on the Basketball Reference website. We will need to gather the data from the website either through API or scraping, as needed. Then, we will need to merge these various data sets into a convenient form for our analyses.

We also observed other possible relevant data like audience attendance, player salary, etc on the same website, as well as on ESPN. Depending on our final choice of research questions, we will also attempt to join these datasets. In the end, we will unify the schemas and create a unified dataset.

The datasets are tabular, with each row representing the game statistics for a single player in a single game (this is a unique combination). It is highly multivariable (see list of columns). There are hundreds of players across thousands of games, giving us confidence that there is sufficient data to complete our analyses.

- What are some of the columns
    - **Position** - includes one of the following: center, power forward (PF), small forward (SF), point guard (PG), and shooting guard (SG)
    - **Height** in cm
    - **Weight** in pounds
    - Season Game
    - Player's **age** on February 1 of the season
    - Own team and opponent **team names**
    - **Games Started** - binary variable with 0 indicating a specific player did not start in the game and 1 indicating that they did
    - **Minutes Played** - integer, 0 to 48 minutes in NBA. Other max values for other leagues
    - **2-PT, 3-PT Field Goals, Attempts, Percentage** - counts and rates of (un)successful shots
    - **Free Throws, Attempts, Percentage** - counts and rates of (un)successful free throw shots

- **Offensive, Defensive, Total Rebounds** - integer, count of rebounds
- **Assists** - integer, count of times a player passes to teammate who shortly thereafter scores
- **Steals** - integer, count of times the player forces a turnover by opponent
- **Blocks** - integer, count of times the player swats the ball, preventing a score
- **Turnovers** - integer, times the player loses possession of ball to other team
- **Personal Fouls** - integer, 0-6, count of illegal
- **Points** - positive integer, number of points scored by player in game
- **Player salary** - annual compensation for the player in dollars

Potential Questions

Our first task will be to determine the appropriateness of normality assumptions for the data. We can use methods like the Q-Q plots, running the Shapiro-Wilks tests, and visualizing the univariate and bivariate distributions (hoping to see bell curves and ellipsoids, respectively). We will use the outlier detection methods from the course, if needed.

We are considering some player-level and team-level analyses, including the following:

- Are there underlying but unobserved variables (skill level, fitness, etc) which determine the statistics?
- Can we predict player salary, winning, season awards? Can we predict team success?
- Can we show statistical significance between the in-game stats of groups of high achievers compared to other peers?
- Can we show statistical significance between in-game stats of different teams at the pro-level? Compared to college level? Can we say that a change of coaches make a significant difference in player performance?

We may consider the population of Georgetown players specifically

Potential Methods

In order to answer these questions, we will likely make use of these topics from our coursework:

- Multivariate multiple regression, including variable selection (e.g. using AIC)
- Multivariate anova (MANOVA) - groupings could be teams, pro vs college
- PCA and/or factor analysis
- Paired comparisons, i.e.T test and confidence region/simultaneous intervals

Depending on the success of these methods, we may make use of non-course topics such as K-means or KNN clustering.