INFO332 Project Proposal
Names: Kevin Shi, Richardson Chhin, Sparsh Jain, Keith Cho, Brandon Hung

Objective: The primary goal of this project is to analyze the performance statistics of NBA rookies from 1980 to 2020 and identify patterns or predictors associated with becoming a top NBA player.

Importance: Understanding the statistical makeup of NBA rookies can provide insights for analysts and decision-makers in basketball. This project uses sports analytics to evaluate how early-career performance translates to long-term success. Identifying trends and patterns in rookie performance can help teams, players, and analysts understand the early signals of which rookies will become the rising stars.

Methodology:

Data Collection:

- **Main Kaggle dataset: https://www.kaggle.com/datasets/thedevastator/nba-rookies-performance-statistics-and-minutes-p**
  - **This dataset contains data of NBA Rookies Performance Statistics and Minutes from 1980 to 2016. The Kaggle link has detailed descriptions of each of the columns.**
  - **We will be joining the datasets below with this main dataset.**
- Kaggle dataset: https://www.kaggle.com/datasets/thedevastator/nba-rookies-performance-statistics-and-minutes-p?select=NBA+Rookies+by+Year_Hall+of+Fame+Class.csv
  - This dataset contains data of NBA Rookies inducted into the Hall of Fame.
  - We will join the Hall of Fame column with the main dataset.
- Kaggle dataset: https://www.kaggle.com/datasets/ignaciovinuales/nba-rookies-stay-longer-than-2-years
  - This dataset contains data of NBA Rookies 1979-2020.
  - We will add the rookies from years 2017-2020 to the main dataset so we have more recent rookie data to work with.
  - We will add the columns 'Team', 'Conf', 'Age', 'Target' from this dataset to the main dataset by joining on the rookie's name.
- Kaggle dataset: https://www.kaggle.com/datasets/ethankeyes/nba-all-star-players-and-stats-1980-2022?select=final_data.csv
  - This dataset contains data of NBA All Star Players and Stats 1980-2022.
  - We will join this all star data to the main dataset on the names of the rookies to label which rookies were all stars and which years they were all stars.

Summary of Joins

- Add 'Hall of Fame Class' column from hall of fame data to the main dataset by joining on the name of the rookie.
- Add the columns 'Team', 'Conf', 'Age', 'Target' from “NBA Rookies 1979-2020” to the main dataset by joining on the name of the rookie.
- Add all star data from “NBA All Star Players and Stats 1980-2022” to the main dataset by joining on the name of the rookie.

Data Cleaning

- Check for missing values and impute or remove as necessary
- Address duplicate values
- Convert columns to appropriate data types
- Feature engineering
- Feature scaling

Statistical Analysis

- Use descriptive statistics and plots to explore trends.
- EXTRA:
    - t-tests: Compare means between all-stars and non-all-stars across key metrics (PTS, MIN, FG%, etc.) to assess statistical significance.

Predictive Modeling

Answer <u>one</u> of the three research questions below:

1. NBA all-star Prediction
    a. What performance statistics best differentiate all-stars from non-all-stars?
    b. Can all-star selection be accurately predicted based on rookie-year statistics?
    c. Use logistic regression, decision trees, or random forest to predict all-star selection
2. Cluster Rookies by Performance
    a. Can rookies be clustered into performance tiers or player positions based on their rookie stats?
    b. Do certain clusters contain more successful players than others?
    c. Use k-means clustering to group rookies into performance tiers or playing positions
3. Efficiency & Points Prediction
    a. Which rookie stats most strongly predict efficiency rating and total points?
    b. Can we create an accurate model to forecast rookie performance metrics like EFF or PTS?
    c. Use linear regression or random forest to predict efficiency rating or points

Evaluation

- Use accuracy, precision, recall, F1-score for classification
- Use RMSE, MAE, MAPE, r^2 for regression
- Utilize visualizations/plots (ggplot2, plotly)