**NBA Problem Statement**

In the current NBA landscape, there appears to be a growing number of international players (i.e. coming to the NBA from countries outside of US). Foreign players are not only increasing in population but also seemingly performing better as time progresses as evidenced by 4 out of the 5 most promising MVP candidates being international players. This project’s goal is two-fold:

1. Identify trends in total player distribution by ethnicity/nationality
2. Identify the top 10 NBA players supported by various statistics

Regarding 2nd goal, the 8 advanced statistics we will use are:

* Player Efficiency Rating (PER): Summarises a player’s overall statistical contributions to a single number
* Win Shares (WS): Estimates number of wins a player contributes to a team
* Box Plus-Minus (BPM): A box score-based metric that estimates a player’s overall impact on a team performance per 100 possessions
* True Shooting Percentage (TS%): A shooting efficiency metric that considers field goals, three-pointers, and free throws
* Value Over Replacement Player (VORP): Measures a player’s overall contribution compared to a replacement-level player (a hypothetical player who is readily available)
* Usage Rate (USG%): Measures the percentage of a team's possessions a player uses while on the court
* Defensive Rating (DRtg): Estimates the number of points a player allows per 100 possessions
* Offensive Rating (ORtg): Estimates the number of points produced per 100 possessions

To identify the top 10 players in the NBA over the last 5 seasons, I will:

1. Aggregate the data across 5 seasons, summing statistics that are cumulative and averaging those that are not
2. Create a formula that outputs a composite score that serves as a weighted average of all 8 statistics
3. Find players with top 10 composite scores

Currently,

Dataframes obtained from scraping data from basketball-reference and fetched from official NBA API are being saved as csv.

We are using SQL queries to create respective tables inside our already-made database and then using psycopg2 and SQLAlchemy to load csv data inside tables and then using pd.read\_sql() to retrieve query results which we then store again into df.

**Potential Issues:**

1. After data collection, need a separate section in the notebook called data cleaning for each df collected where you are looking for:
   1. Missing values
   2. Duplicate records
   3. Inconsistent formatting
   4. Outliers
2. Need a separate EDA section
3. For retrieval queries, since we collected nationality data, can obtain another dataframe for specifically that
4. Composite score weight balancing needs to be more fine-tuned. Currently, for simplicity, all metrics are equally weighted and this should not be the case

**POA (Plan of Action):**

1. Look back on the code and study each block. Really try to figure out and understand exactly what is going on.
2. After studying, look to solve above issues.