Data Wrangling Notes:

- Integration:
    - Downloaded two separate datasets from Kaggle: one containing 22-23 season stats across 11 .csv files and the other containing 23-24 season stats in a single .csv file.
    - Used Excel to add an extra column in each file indicating the season (22-23 or 23-24) to differentiate data
    - Used Python code to combine all .csv files into a single dataset 'nba.csv' housing a total of 30 columns and 6201 rows (excluding header row).
        - Every row is a player, and every column shows a different basketball metric such as FG (field goals made), FGA (field goals attempted), FG% (field goal percentage calculated by FG/FGA), player positions, and teams, etc.
- Cleaning:
    - Merged the original data ('nba.csv') with a subset containing only 'TOT' entries to flag players who played for multiple teams in a season (helps ensure that cumulative stats for such players were accurately represented).
        - Kept only the 'TOT' entries for flagged players within the same season, and retained all other entries for different seasons, preserving accurate season totals and avoiding duplication.
    - Filtered the dataset to include only rows where the 'Pos' column value is one of the top 5 positions (SG, SF, PF, C, PG) to focus on the most relevant positions and exclude positions with low value counts.
    - Encountered null values in percentage columns ('FG%', '3P%', '2P%', 'eFG%', and 'FT%'); decided to keep them null and impute their values during KNN and Random Forest modeling, rather than set them to 0 or the median out of concern of misrepresentation and bias.

EDA Notes:

- Summary statistics:
    - Though there is a presence of both young rookies and veteran players, the dataset predominately consists of young to mid-career players, reflecting the broader NBA landscape where athletic prime typically ranges from mid-20s to early 30s.

    - The shooting efficiency (FG%) is around 45.71%, aligning with typical NBA shooting averages. While the 3-point accuracy (3P%) is around 32.27% which is reflective of the varying proficiency across the league.
    - Minutes played (MP) has an average of18.85 minutes per game, meaning players average about half a game in playing time, indicating a mix of starters and rotational players. However, its range of 1 to 41 minutes implies that some players are on the court almost the entire game, while others play very limited minutes, emphasizing differences in player roles and stamina.
    - Points (PTS) shows that players score an average of 8.7 points per game, with significant variability, especially when there is a wide range in points scored in a single game (0-35).; this underscores the diversity in scoring roles, from high-scoring stars to defensive specialists and bench players.
    - The distribution of player positions shows that Shooting Guards (SG) are the most common, followed by Small Forwards (SF), Power Forwards (PF), Centers (C), and Point Guards (PG). This indicates a balanced representation of positions, with a slight emphasis on perimeter players (SG, SF, PG) who are crucial for scoring and defense.
- Distributions:
    - Used a histogram to show the distribution of points per game (PTS).
    - Used boxplots to show summary stats of FG% by Pos.
    - Used a scatterplot to show MP vs. PTS by Pos.
    - Used a scatterplot to show TRB vs. PTS by Pos.
    - Used bar chats to show the top positions for key metrics like assists, turnovers, and personal fouls.
- Correlation analysis:
    - Created a correlation matric for every numerical variable.
    - Created a correlation matrix for a subset of relevant variables.
    - Applied hierarchical clustering to the correlation matrix (groups similar variables together and reorders the matrix based on these groupings) to see how different metrics are related.
    - Filtered the correlation matrix to show only strong correlations (absolute value > 0.5)