Title: Basketball Python Package for Importing, Manipulating, and Visualizing Data

Author: David Cuellar

Summary:

The purpose of this project is to develop a Python package that simplifies the process of importing, cleaning, and visualizing basketball data. The goal is to bridge data science techniques with sports analytics by creating a tool that allows users to visualize key performance metrics. This package will enable data scientists, analysts, statisticians, and sports enthusiasts to efficiently analyze player and team performance using publicly available data sources such as Basketball Reference, the official NBA website, and StatMuse. This project also reflects my personal interest in combining the technical skills and knowledge gained during my graduate studies with my passion for sports, particularly basketball.

Proposed design:

This package will be structured around two main modules. The first module, Module 1, will focus on importing and cleaning data from these public data sources. This module will contain a function that will load the data (ideally a CSV file) into a dataframe using pandas. A second function will handle the cleaning of the data, ensuring that there is no missing data and that all the numeric data types are consistent. A third and final function will be used to filter out any non-numeric data for better analysis. All of these functions will be encapsulated within a class, allowing users to manage and manipulate their data as an object for an easier workflow.

The second module, Module 2, will be used to provide different visualizations using matplotlib. This module will contain a function showing a player's stat progression over multiple seasons using a line plot. The second function will compare multiple players' averages for a specific season using a side-by-side bar chart. The third function will compare how teams rank against each other in efficiency and scoring using a scatter plot. Each function will handle further filtering, styling, and labeling.

The external libraries needed for this Python package include pandas and matplotlib. The pandas library is required for importing, structuring, and cleaning the basketball datasets. Matplotlib will help with the creation and customization of these visualizations.

A few potential challenges that can occur during the creation of this Python package include handling incomplete or large datasets, designing flexible functions that can handle player and team-leve data, ensuring visualizations are clean and scalable, and maintaining modularity for future extensions.