Python for Data Science: Assignment 1

Simon Karumbi
S3455453

Data Preparation

During the Data Preparation stage, I navigated through the dataset column by column, ensuring that the data expectations were met as per the data information brief. I started with categorical features, and then moved on to the numerical features.

Error 1: Player Positions

The first error was that there were formatting issues with the player positions, including white space and capitalised letters. I created a function to adjust this. The assumption is that the '.' and 'a' could be removed to successfully get the player positions.

Error 2: Team Names

Similar to positions, there were issues with some formatting with team names including white space and and '0' in Houston's team name.

Error 3: Age Range

The brief stated that ages under 18 are not possible, and ages over 60 were also filtered.

Error 4: Points

The brief stated points of over 2000 should not be possible, where two players had the incorrect points allocated. This was easy to resolve by performing a calculation to correct this based on FT, 2P and 3P.

Error 5: Points Percentage

The calculations for some of the points percentages were incorrect. This was rectified by performing the 2P/2PA etc. calculations specified by the brief.

Error 6: NAs

There were several NaN values in the data set, due to divide by 0 errors. This was rectified by changing all NaN values to 0 in the FT%, 2P% and 3P% columns

Data Exploration

2.1 Top 5 Players

The Top 5 players were examined, after mistakenly identifying James Harden as a potential frontrunner. After researching online, the team name 'TOT' refers to the aggregated statistics of a player from all teams that they played in, if it was more than one.

The top 5 players were relatively close in points scored across the season, however, the composition of these points differed greatly. Stephen Curry scored approximately the same number of Free Throws, 2 Pointers and 3 Pointers, whereas the other players either favoured 2 Pointers or 3 Pointers.

No player scored mostly 3 pointers, with Stephen Curry scoring the most of the top players. Giannis scored significantly more 2 pointers than other players.

2.2 Data Errors

No Data Errors were found in this section, although this may have been rectified while fixing the Point Attempted columns.

2.3 Further Exploration

Point/ Shooting Guards seem to be the highest scoring Players across the league, with Small/ Power Forwards not contributing many points to their teams. Most teams hover around the 4000 points scored per team mark, with Brooklyn Nets scoring the most points of the season.

In general, players tended to play more than half of the games per season, with younger players playing less games than their mature counterparts. There are a small smattering of players who played a few games, but scored many points.