Student Name: Matthew Moloney Student ID: s3717566

Data Preparation

For task one, I took a brief look at the initial 50 records to identify any obvious errors before beginning to scan each column using value_counts() to identify values outside the specified sets in the data description. For numerical columns, I sorted the columns and checked the head and tail of the data to see if there were any numerical anomalies, such as values exceeding the limit or negative numbers. Then I checked the value_counts() on columns that should be unique and make sure that no data was duplicated.

Error 1: White space

The first error I look to fix is any extra whitespace in each value. This error is easy to spot using value_counts() as these values will be displayed separately from the properly formatted value. Below is an example of the value_counts() of the column 'Pos', ' PG' is displayed separately from 'PG'.

This error can be fixed by using strip() which will remove whitespace before and after the string. However, it is important to examine the valid data and ensure that there are no values where leading and trailing whitespace is apart of the valid string. In the case of 'Pos', there was no valid whitespace values, and this held true for all columns.

Error 2: Letter capitalisation

The next most obvious and simple error were mistakes in the capitalisation of the string. value_counts() would return data considered separate due to mismatching capitalization. This was handled with upper() which converted the full string to upper case to match the allowed values for columns such as 'Pos' and 'Tm'. There were no cases in which strings needed to be converted into lower case.

Error 3: Duplicate players

This error required a lot of thought on whether a player on a new team should just be considered as a separate player, but ultimately, I decided that a player should only have a single record with all their combined stats for a few different reasons, such as task two is all about the players and that in a spreadsheet called NBA_players_stats all that players data should be grouped into a single record for ease of analysis.

However, combining the player records came with some careful considerations on how to merge certain columns that weren't numeric. These columns include 'Pos' and 'Team'. For Pos, for each player, I checked all the positions they had under their name, and found that their position either didn't change, or their was a position that had combined their positions into the values such as 'SF-PF' of which were the positions they had previously played. I resolved this column by choosing the longest string which would always be the combined values. For 'Team', it was a bit trickier but as data is added sequentially into a table, I chose the team associated with the latest record, aka the largest index, to use as the current team for that player.

Error 4: Numerical Typos

Numerical typos also had to be addressed. These typos usually lead the values to become extremely large, as so much as a single digit typo would lead to extreme max or min data values. Fortunately, these numerical extremes are easy to identify especially with strict rules on the range of the data. for the column 'PTS', it was stated that the value must be < 2000 and therefore a simple sort by ascending table easily identified incorrect records, such as the number '28800'. However, depending on the column, it can be difficult to identify which value such an error should be corrected to. Fortunately, in this case, the column 'PTS' can be acquired from calculations on other columns and therefore can be justly rectified, however this is a special case, and most columns will not be equal to calculations on other columns. For the 'Age' feature, using sort_values(), an extreme maximum and minimum were found. Fortunately, this errors had obvious assumptions to what the proper data was, and was fixed using abs() to get the absolute value of the negative number and a simple replace to removing an extra 0 from the maximum. Fortunately, no other numerical errors were identified, apart from that will be fixed in task 2.2.

Student Name: Matthew Moloney Student ID: s3717566

Error 4: Individual errors

There were multiple errors that only appeared in very low amounts, mostly a single record, and were not related to any of the previous errors and therefore needed specific attention to be rectified. These errors were fixed by targeting the value with a .replace() and changing it to the fixed value.

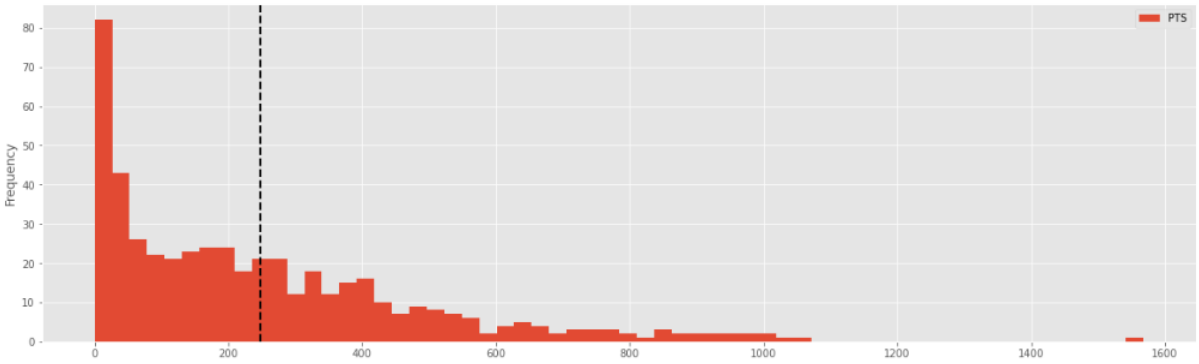
One example of this is the value 'SGA' and 'PFA' in the column 'Pos'. These values were not apart of the values that the specification specified that were apart of the column, and therefore needed to be fixed. These were remedied by removing the trailing 'A', as no position value contains an A and by removing the letter, they now had values that were valid. A similar error is the value 'SF.' in the same column, which was valid once the trailing period was removed.

Another error is the value 'SFPF', which was missing a middle dash (-) to become a valid value.

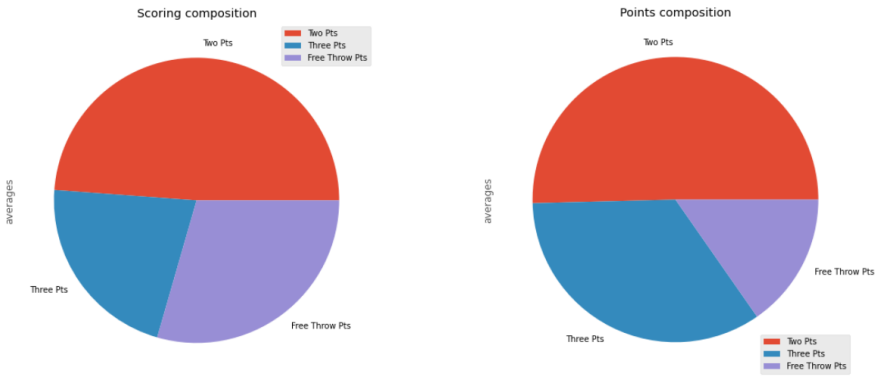
Data Exploration

Task 2.1: Explore the players' total points: Please analyse the composition of the total points of the top five players with the most points.

To get the fastest understanding of every player's total points, a histogram is a good way to show the point distribution as well as quickly see a rough minimum and maximum value. We can also create a line where the average value is, which will help us see how evenly distributed the data is on the upper half and the lower half.

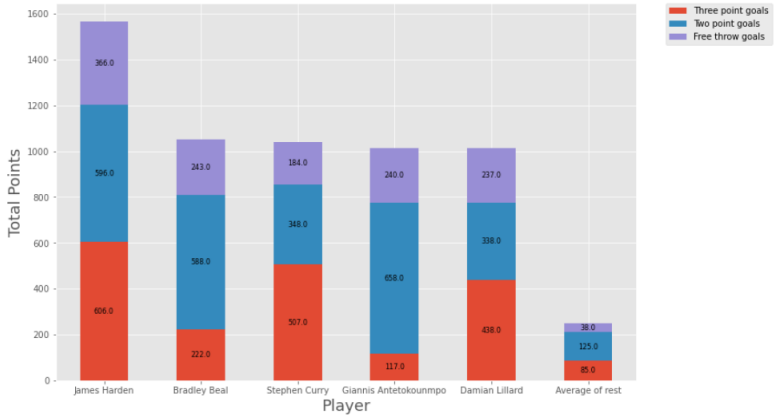


From the histogram, we can see that we have a positively skewed distribution of points, where the majority of players are sitting around 0 points and fairly consistently trickles down as the points get higher. Another important detail from this graph that we can take away is the huge margin between the top 0.1% of players against the majority, and this will be further analysed as we investigate the data of the top 5 players in terms of points. Before we do that, let us look at the composition of what makes up the total points.



To explore the total points of each player, it's important to know how the total points are calculated per player. The total points are a sum of 3 times the 3-point scores (3P) with 2 times the 2-point scores (2P) plus any free throw points (FT). The two above pie charts show the composition of the average total points. The scoring composition pie graph is showing the number of times that points has been gained that way. It does not account for the fact that a 3-

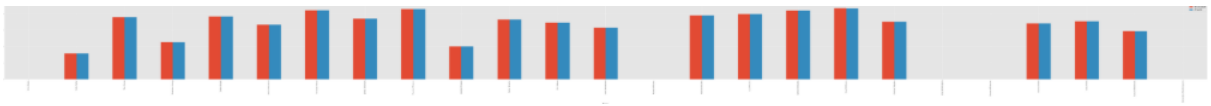
point shot is worth 3 points, etc, which is what the points composition pie chart on the right displays. We can see that the most predominant shot taken is the two-point shot, at roughly just over double the number of 3 point shots, and that points gained from free throw shots are slightly more common than 3 point shots. However, when considering the number of points gained for each shot, three point shots account for roughly double those gained from free throws, and interestingly two point shots maintains its percentage, accounting for marginally over half of all points gained. Let us see how this average composition compares to the top 5 players composition of their total points.



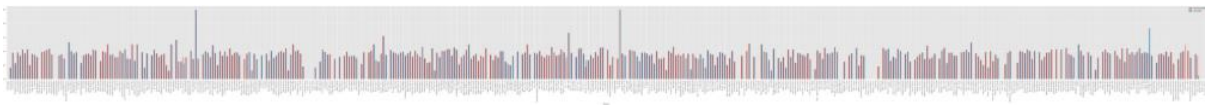
The first thing to notice with this bar graph is how tightly the top 2-5 players are, only emphasising the discrepancy that is the top player in terms of points scored. In terms of the composition of how the top 5 players points are made from, there seem to be two patterns used out of the 5 players. We can see that the players James Harden, Stephen Curry and Damian Lillard have between 1/3rd and ½ of their total points made up of 3-point shots, demonstrating their play style to go for significantly more 3 point shots than the other two players in the top 5, Bradley Beal and Giannis Antetokounmpo. The free throw points seem comparatively consistent amongst the top 5, at making up consistently roughly 1/5th the score.

Task 2.2: Assuming that the data collector makes an entry error when collecting data, it can be ensured that the error occurred in the 3P, 3PA and 3P% columns, but it is not sure which player's information the error lies on. Please try to explore the error by visualization to identify how many errors there are and try to fix it.

To start this task, we need to find the equation that 3P% is equal to, and thus do math on other columns to be able to get a correct result and verify the column. As stated in the spec, the equation to 3P% is 3P/3PA, or 3-point goals over 3-point attempts. Once we are able to calculate our own 3P%, we can graph it out on a bar graph to see the difference in values between our calculated one and the stored one. First we will graph out the first few players and see if there are any obvious errors.

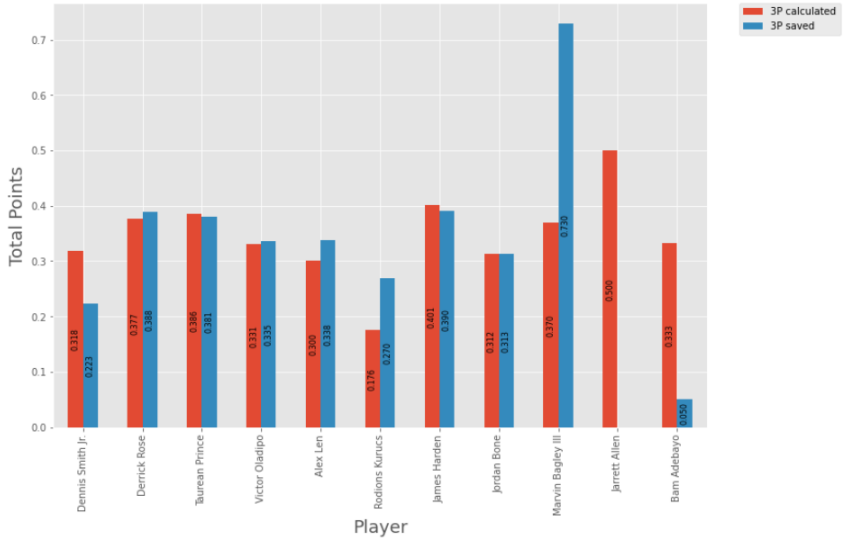


As seen above, the red column is our calculated 3P% and the blue is the stored data, therefore we can scan for a mismatch in column height. Scaling the height of the chart will also assist in noticing any differences, and if enlarged will make obvious any errors in the data. However, as with any graphical display of 500 players, as seen below it is arduous and borderline incomprehensible without being severely enlarged, so I looked for a filter to narrow down the columns such that it will only display errors.



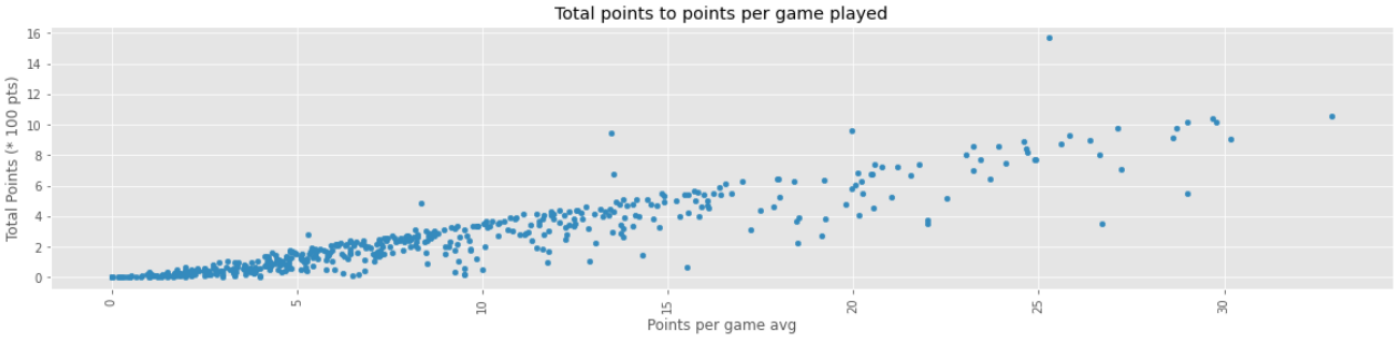
By adding a filter in the data to only display the players where the calculated 3P% is not equal to the stored 3P%, we are left with a much smaller chart displaying only incorrect player data of which we can use to target and fix.

Student Name: Matthew Moloney Student ID: s3717566

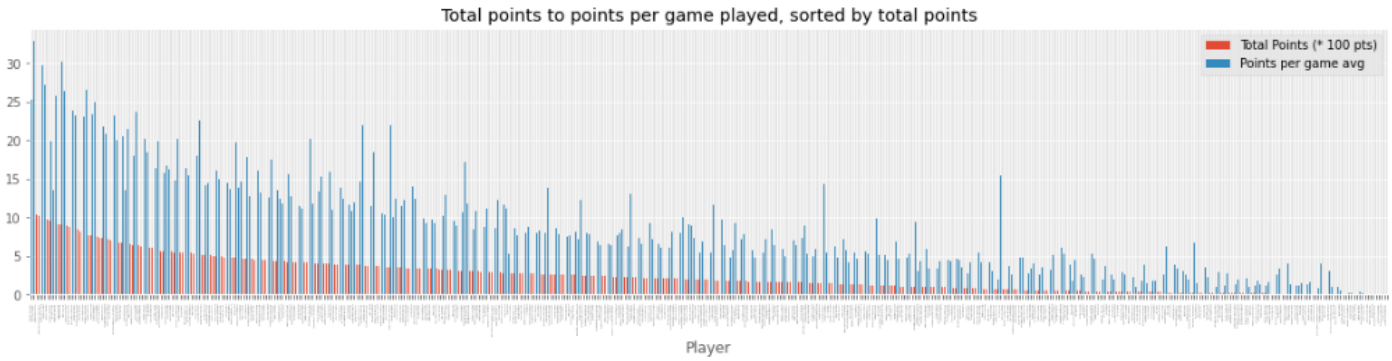


Task 2.3: Please analyse the relationship between the player's total points and the rest features (columns). Please use at least three other columns.

When asked to analyse the player's total points and their relationship to other features, I felt that the most important part of correlation that I believe we would find is the relationship between the total points and the number of games played. The first chart shows the relationship between the total points for that player (per 100 points) and the number of points scored per game on average.

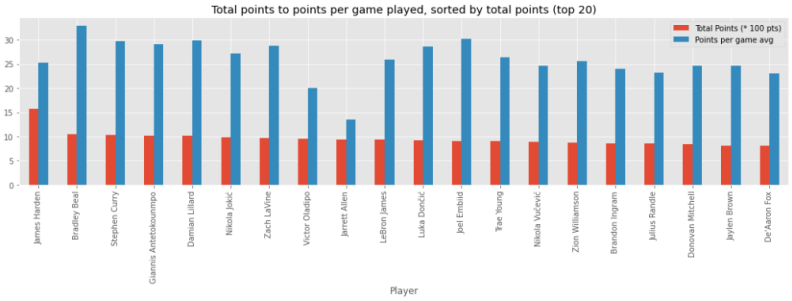


As seen from the scatter plot above, there is a clear rate of linear increase of points per game played as a player scores a higher total points. We are able to determine from this that most high scoring players are due to their ability to score high per game, and not just from shear number of games played. However, we can spot one outlier, having the highest total points by a significant margin yet not at a particularly high points per game average. To further analyse this information, we can explore other charts.

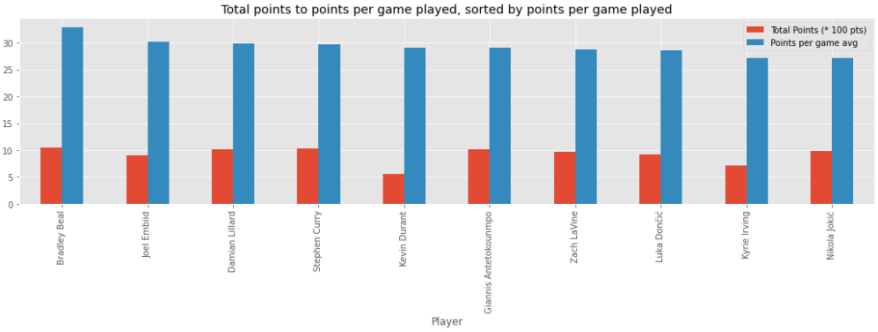


Although the chart is too small to see individual players, it gives a good understanding of the trends and the relationship between them. We can further see that there is a general relationship between the total points scored and total points scored per game. By zooming into the data and further analysing it we can determine the efficiency of players, where their average points per game is high and if there is correlation an efficient player and a top scoring player.

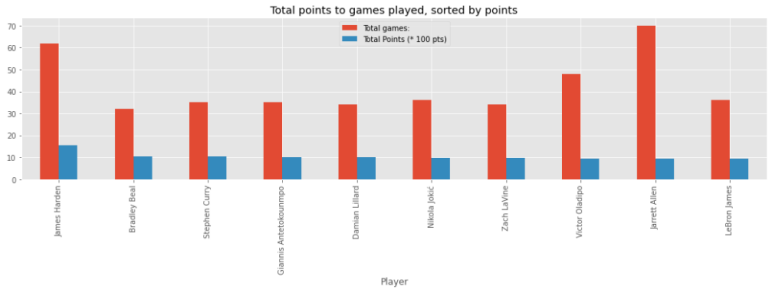
Student Name: Matthew Moloney Student ID: s3717566



Above are the top 20 players in terms of total points on the same graph. Let us look at the correlation in names when we now sort by points per game played.



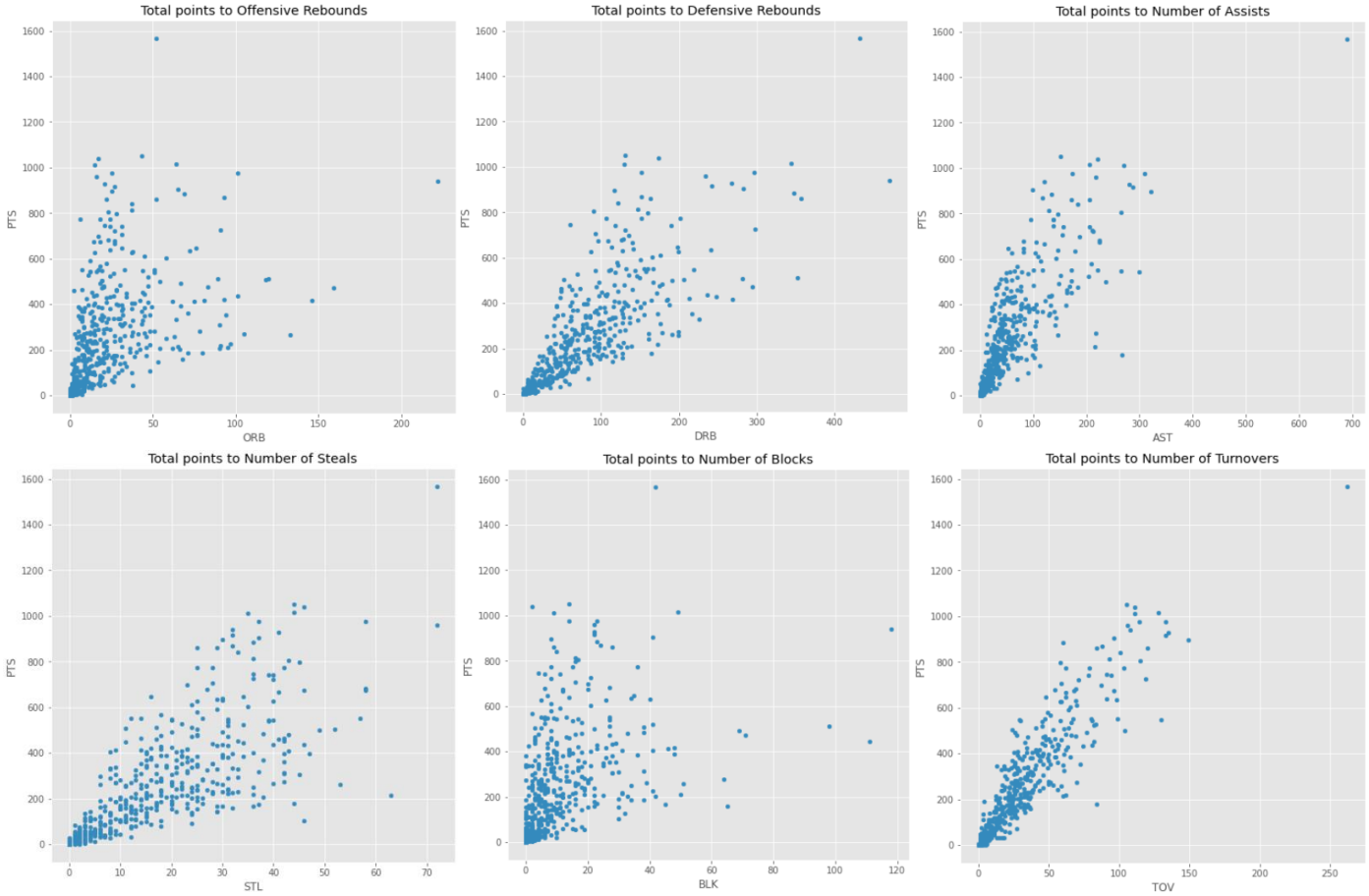
Looking at the above chart, we notice a few overlapping names between both charts. The number one player in the graph sorted by points per game played, which I will refer to as the Efficiency graph, Bradley Beal is ranked second in total points, giving us the information that while Beal is a high scoring player, he is also one of the most efficient, as opposed to James Harden, the top scoring player not appearing in the top 10 most efficient players. As Harden has such a huge lead in terms of points, it is interesting to see him not be a part of the top 10 most efficient. We can investigate this by simply creating another bar graph with the number of games played.



From the graph, it is discernible that although Harden lacks in efficiency, he is able to make up in sheer number of games played to obtain points.

Outside of analysing these top players efficiency, we can look to see if there are any other features that may correlate with the total points feature. The easiest way to show correlation is through scatter plots.

Student Name: Matthew Moloney Student ID: s3717566



From the above graphs, we can see that there are certain levels of correlation between points and the chosen features. The most prominent is the total points to number of turnovers, where there is significant correlation, as all the top players with high total points are all the players with a high number of turnovers. As opposed to this pattern, looking at the total points to number of blocks shows us that the highest scoring players do not necessarily have the highest blocks, in fact the majority of the top 5 in highest points are on the lower end of blocking, showing that blocking is not something a required at all to score high. Defensive rebounds are an interesting chart as it shows that the highest 10 scoring players are spread out almost the entire range of defensive rebounds, going to show that a player with a high amount defensive rebounds has the same chance to be a top 10 player than a low number of defensive rebounds.

References

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html

https://matplotlib.org/stable/tutorials/introductory/pyplot.html

https://pandas.pydata.org/docs/getting_started/intro_tutorials/04_plotting.html

https://www.w3schools.com/python/numpy/numpy_creating_arrays.asp