Data Preparation

My first step was to import the data using the pd.read_csv method. I used .dtypes to check the data type of each of the columns. For example, to make sure my ‘Age’ was not being imported as an object and was correctly being imported as in ‘Int’.

Errors in temp_NBA[‘Pos’]

When inspecting the data I used .unique() to check the possible values that were in the row of ‘Pos’. I noticed some case errors and white spaces. So I used .str.upper.str.strip() (see In[26] in Jupiter notebook) to remove white spaces and ensure we did not have case sensitivity errors. I then proceeded to run .unique() again. This showed a reduction in errors, but I still had some typo errors. It was obvious what the typos were (such as an ‘a’ or ‘.’ onto the end). This was simply a case of finding occurrences of the typo and replacing with the correct value. E.g: temp_NBA.loc[temp_NBA.Pos == 'PFA', 'Pos'] = "PF".

Errors in temp_NBA[‘Age’]

I used .unique() again to show all possible values for this column. This column possessed less errors, but I could see two impossible values. A negative number and a ‘280’ value. The absolute value for the negative number and assuming the ‘280’ had a ‘0’ typo on the end gave me reasonable ages for the ages of the players. It was assumed that both of these were cases of data mis-entry and were recorrected to their correct values.

Errors in temp_NBA[‘PTS’]

Here .unique() showed two impossible values ‘20000’ and ‘28800’. Given that it would be impossible for a player to get this total point score, we can assume that the data entry has made some kind of error. I was faced with either replacing the scores with a mean on the remaining points or trying to assume the error and replace. Given that both have trailing zeros, I assumed that the mistake was typos of trailing zeros. Removing these zeros gives us a reasonable score.

NaN errors

Looking in the data we can see that ‘3P%, 2P% and FG%’ have NaN values. I could see this was the result of players having 0 scores and 0 attempts inside the related columns comprising the formula that gives each of my % columns. However, we cannot use these NaN values, I decided to replace with 0. Given that a player who made no attempts still missed 100% of their shots for these respective categories and therefore made 0% of their shots.

TASK 2.1

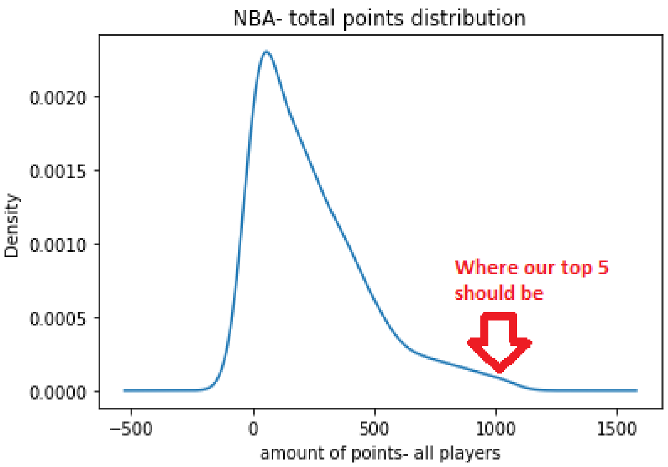


Figure 1 - above

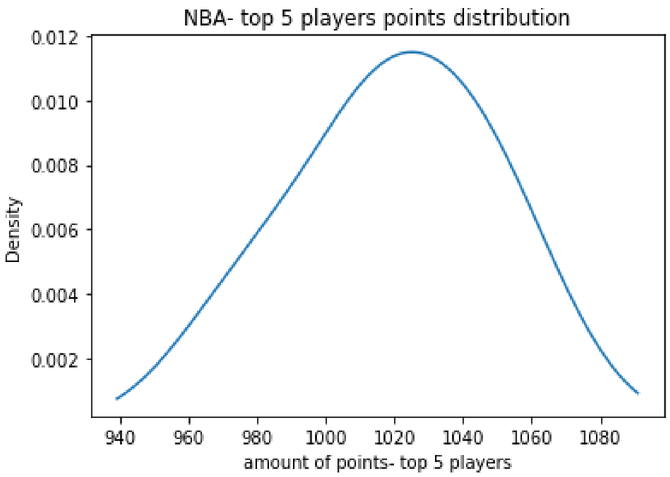


Figure 2 - above

First the top 5 players point distribution was checked (see figure 2). We can see that the median value sits roughly where our top players should be in figure 1.

When I analysed the composition of all players points based on 2P or 3P shots, I found that all players had various combinations of 2P and 3P shots. Some players seemed to show a preference towards 2P shots, some towards 3P shots and some a combination of both(*see figure 3).* This trend continued across low scoring players up towards high scoring players. We can see a typical looking scatter shotgun blast plot pattern.

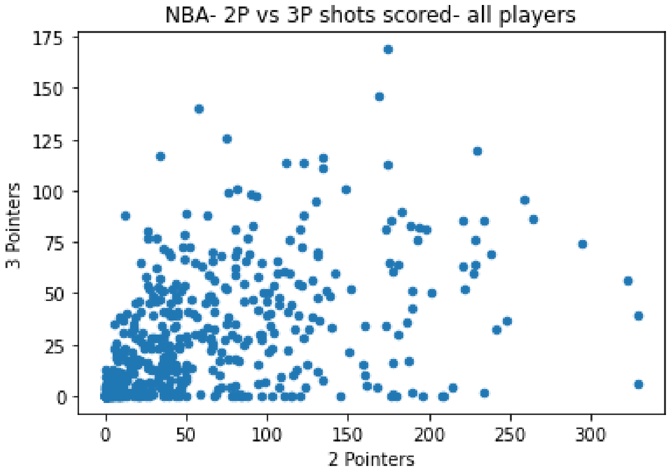


Figure. 3 - above

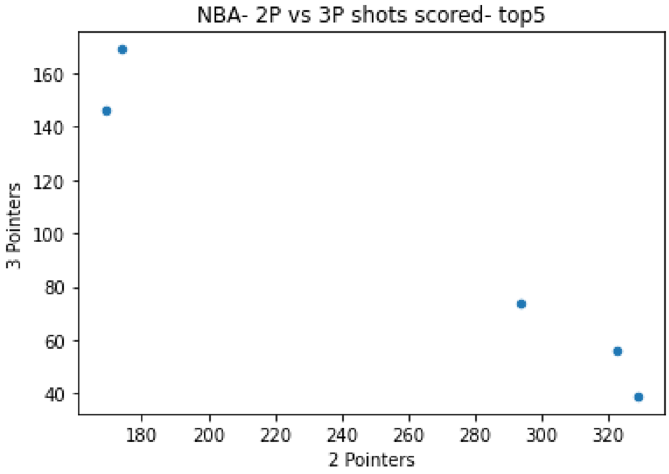


Figure. 4 - above

However, When the composition of the Top 5 scoring players was plotted (*see figure 4*) A clear bias was found towards 2P or 3P shots. This shows that the absolute top 5 best players specialise into a specific play style when it comes to a preference between 2P or 3P shots. We can deduce from these graphs, that players can be low, medium, or high scoring players based upon various point combinations. But to be one of the best 5 players, a player must specialise into prioritising 2P or 3P shots. While we must consider that the x axis is roughly twice the values of the Y axis in both figure 3 and 4, it must also be remembered that 3P shots are worth more points. So if a player scores roughly the same amount of 3P shots as he does 2P shots we regard that player as having a preference for 3P shots given their contribution to the total point value(*the total amount of points compromises a larger majority from 3P shots than 2P shots- even if they have equal 3P shots and 2P shots made*).

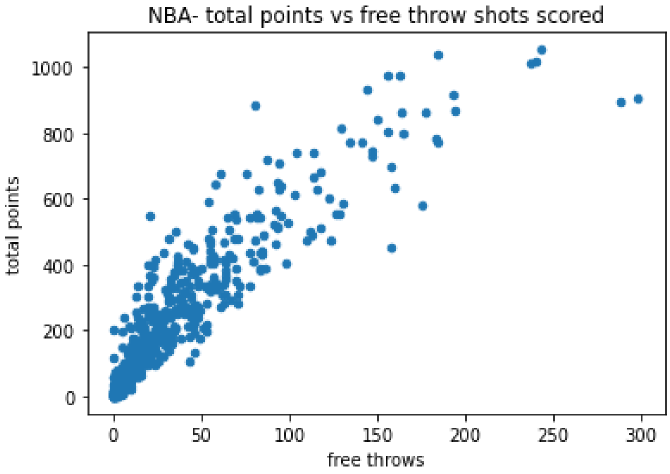


Figure 5- above

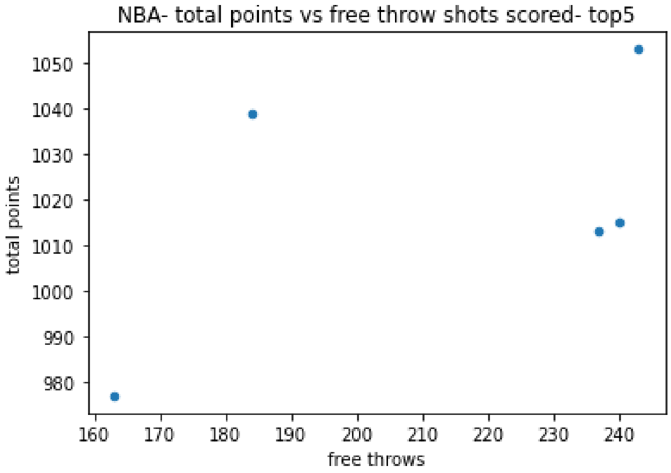


Figure 6 -above

When looking at the relationship between total points and free throws, there did not seem to be any specific pattern, other than the more total points a player had, the more free throws were had. There was a slight bias towards having less free throws in proportion for total points amongst all players, but the top 5 players seemed to range only roughly between 160-240 free throws for 980-1050 total points(*see figure 6*) . Or 16%-23% total score composition (*free throw percentage for makeup of total points*). There was no notable variation between the top 5 players and the percentage of total points made up by free throws.

TASK 2.2

To approach this task I first looked at the relevant columns relating to 3 point shots in the data. On index 5 for player 'Bam Adebayo' We can immediately see an entry mistake. '0.05' for two 3P shots and six 3PA. I first formed the assumption that the error was in the 3P% column(*with the intention of confirming*). I created a new column called 'my3P%' which was made by using python to automatically calculate a new 3P% that we can compare against the original 3P%.

Ideally, if there are no errors in 3P% we can plot 3P% on a scatter plot against my3P% and we would achieve a straight line like that seen in figure 8. However, when we do (*see figure 7*) we can see some points that do not follow the scatter plot pattern expected. These points are our errors. We can replace our 3P% column data with data from my3P%, but first we will confirm that the error has not occurred in another column. Given that 3P% = 3P/3PA, we can deduce that 3P = 3P% * 3PA. Thus we can calculate my3P using that formula, and plot it against 3P-see figure 8. Given that we have no plot points that do not follow the linear pattern, we can deduce that the error did indeed occur on the 3P% column and replace the values in 3P% with the values in our other column my3P%.

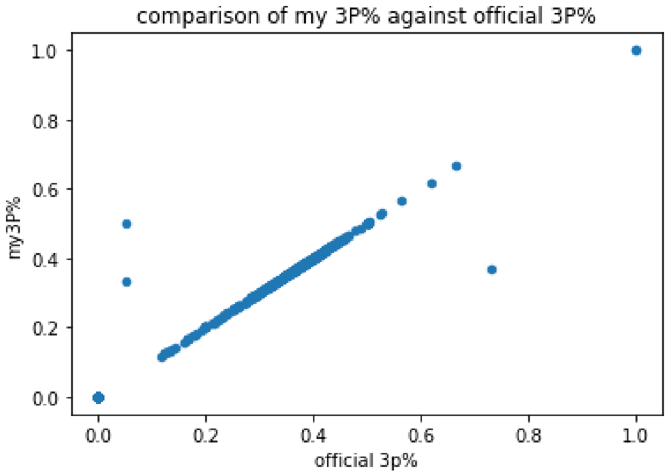


Figure 7- above. Plots that do not conform to the x=y graph pattern are errors. We can see 3 errors.

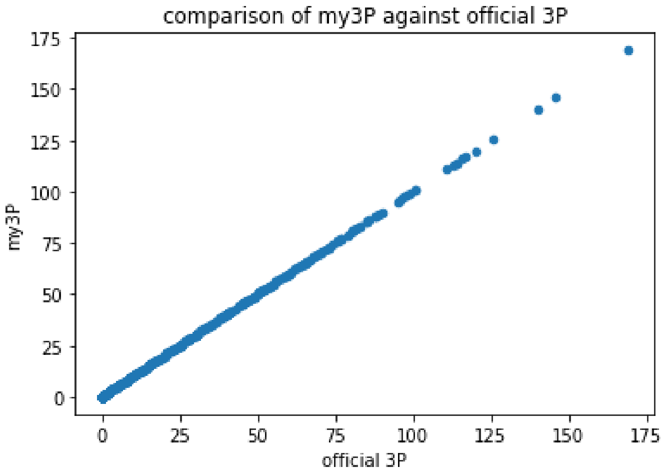


Figure 8- above- our comparison of 3P against my3P shows that there are no errors in 3P.

Lastly, we can also check that there are no errors inside the 3PA column by creating a my3PA using my3PA =my3P/my3P% and comparing against the official 3PA. We can see in figure 9 that there are no errors, so we can confirm that the errors were only in the 3P% column. 3P% was set to equal My3P% then written to the cleaned_NBA_players_stats.csv file.

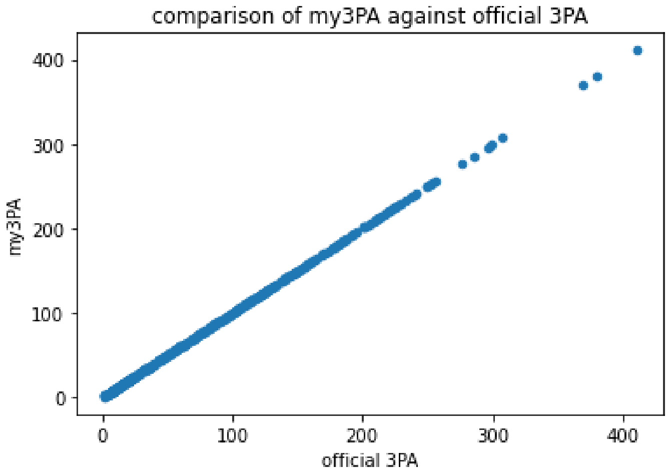


Figure 9 – above

Task 2.3

For this task, the goal set was to determine the relationship between what features caused more points to be scored by a player. The first two features chosen to be investigated were ones that seemed the most obvious, the number of games played, and number of minutes played. Given that more time spent on the court would allow more opportunities to score points. In figure 10 we can see somewhat of an increase in number of points per games played, and that the highest scoring players played over 30 games in the season, but we can see that most of the players scored below 400 points regardless of games played, which shows that more games doesn't necessarily mean more points given that the entire range of 0-35+ games is filled with players who scored less than 400 points. We can conclude that players who scored the greatest number of points played the most amount of games, but there is no linear relationship between amount of games played and amount of points scored.

The next time related feature was the amount of minutes spent playing. Looking at figure 11 we can see that there is a more linear relationship between amount of time spent on the court and the amount of points scored. It can be said that more minutes spent on the court leads to more points being scored. After 800 minutes we can see an upward curve suggesting that players who spend more than 800 minutes on the court begin to score more points per minute spent on the court than those who spend less than 800 minutes on the court. It can be concluded that for a player to perform better in terms of total points scored, they must spend at least 800 minutes on the court.

The next feature to be examined was the effect the players position had on the amount of points scored. Looking at figure 12 it can be deduced that the 'PG' position was the highest scoring position, with its top 25% percentile being the largest. The impact of the position did not seem to have a serious impact on the points scored by a player given that the median of each position was very similar. But it can be deduced that the top scoring players were more likely to be PG than any other position.

The feature investigated next was the number of personal fouls in proportion to total points(see figure 13). This was found to have a somewhat exponential relationship. This seemed to suggest that players who committed more fouls and in turn took more risks scored more points. However to make sure, personal fouls were then plotted against minutes played(see figure 17). This showed that personal fouls were in proportion to minutes played, and that personal fouls do not have an impact on the amount of points scored by a player, as personal fouls are a bi product of time spent on the court and do not bear an independent relationship that could indicate total points.

The next player feature investigated was rank. This showed no clear relationship between player rank and player points(see figure 14).

The next player feature investigated was age(see figure 15). This showed no clear relationship between player age and player points.

The next player feature investigated was Assists. This again seemed to suggest a relationship between player assists and player total score(see figure 16). But just as in personal fouls, when assists was also plotted against minutes played(see figure 18), a similar graph was found as when fouls was plotted against minutes played. Suggesting that there was no relationship between assists and points, given that assists was determined by minutes played.

In conclusion it was deduced that the biggest impact on a player's score was the amount of minutes that the player spent on the court. Players who spent more minutes on the court tended to score more points. Other potentially impacting features were found to also rely on the amount of minutes spent on the court. So, reducing minutes played would also reduce other impacting features. Any feature investigated would proportionally decrease as player minutes on court were reduced. It was however deduced that to make a top performing player, the ideal situation would be to have a player who spends at least 800 minutes on the court in the 'PG' position.

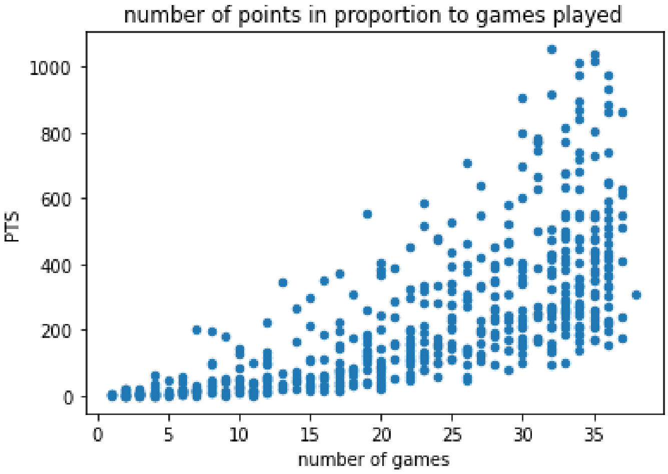


Figure 10 - above

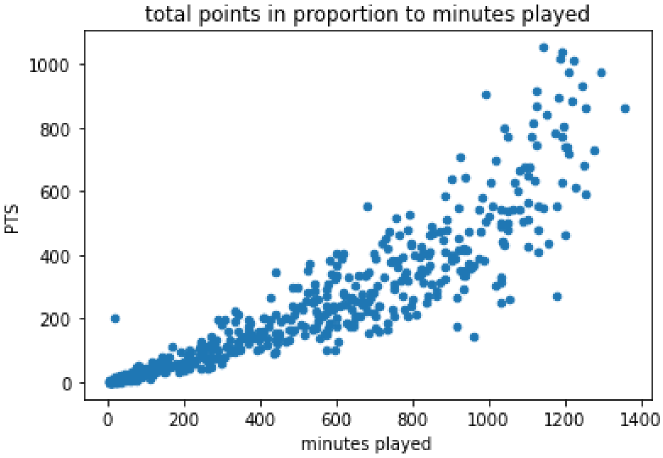


Figure 11 - above

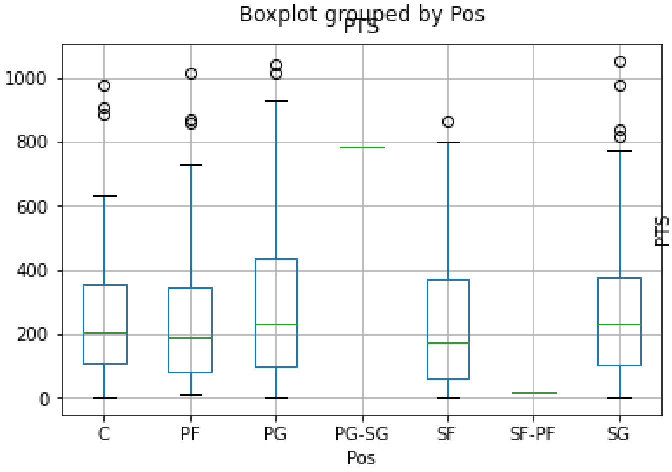


Figure 12- above

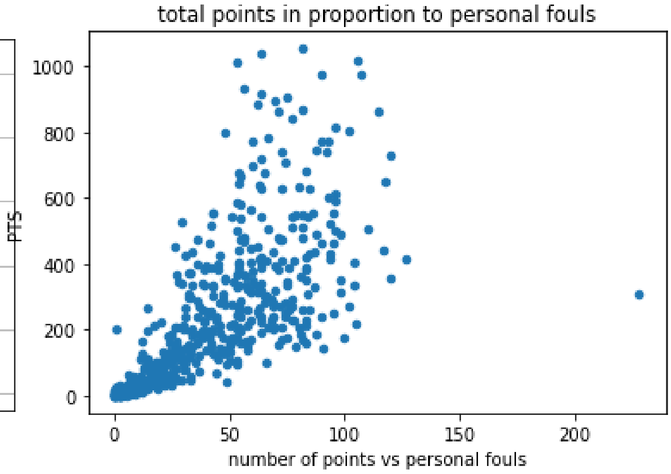


Figure 13- above

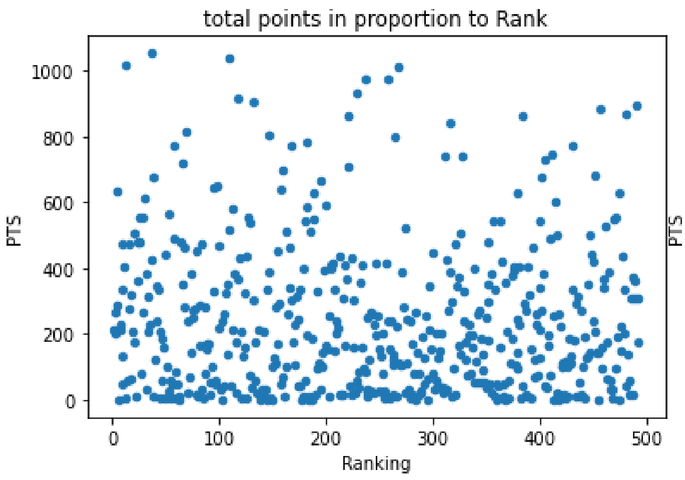


Figure 14- above

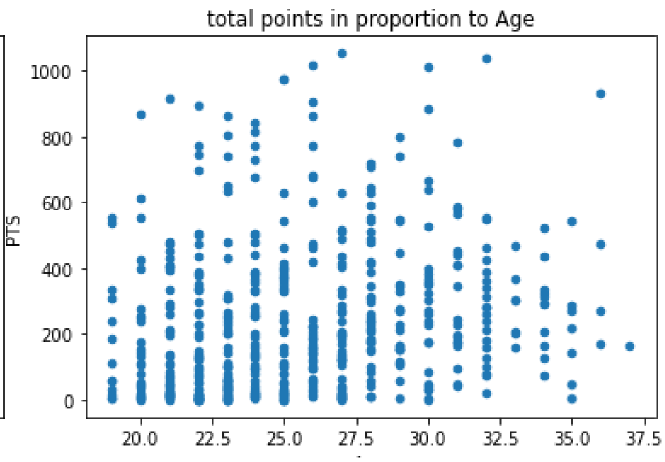
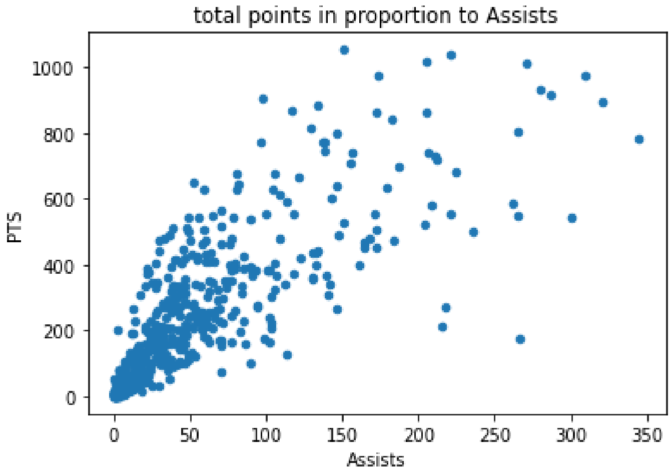


Figure 15 – above



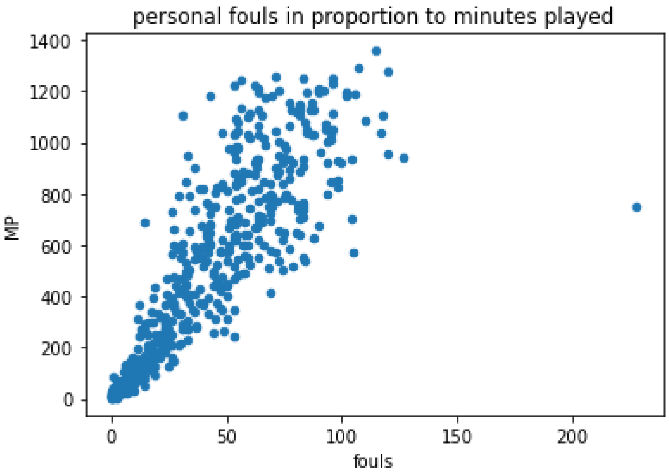


Figure 17 - above

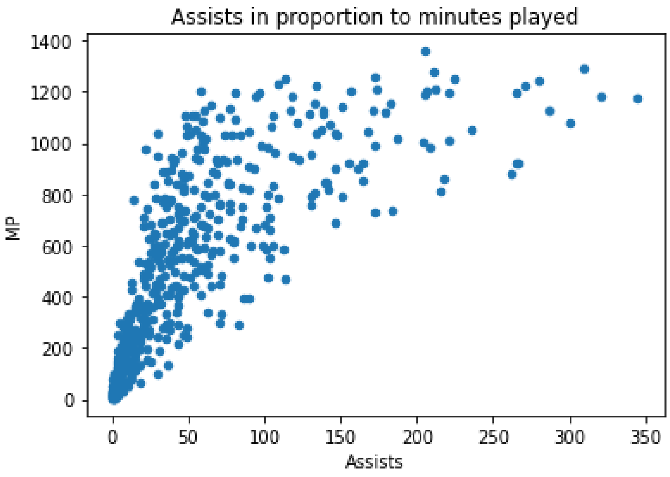


Figure 18 -above