Practical Data Science with Python

Assignment 1: Data Cleaning and Summarising

Submitted by
Md Abir Ishtiaque
S3677701

Data Preparation

After reading in all the values from the csv file, I wrote functions to detect and print out all bad values.

Typos:

There were typos in the ‘Team’ and ‘Position’ columns. Inside the functions I made two lists to hold all the valid values for team and position. I then looped over the respective columns and checked if the values were in those lists. If not I printed out the values to see the typos and manually fixed them with the replace function.

Whitespaces:

Whitespaces can be detrimental to our data analysis. I have used python’s strip() method to remove all leading and trailing whitespaces.

Casting to uppercase:

Since all positions and team values are in upper case, I have used string class’s inbuilt upper() method to convert them all to upper case.

Incorrect summation of points:

In some rows the summation of the players’ points were incorrect and exceeded the maximum amount i.e. 2000. The function I wrote which detected this, also correctly summed the columns and printed out the proper amount. Then I was able to fix it using the replace() method.

Negative values:

There were data entry errors where the collector mistakenly entered a negative sign in front of the value such as -19 for age.

Missing values:

The missing values in the 3P% column were due to a divide by 0 error. Since the missing values were not affecting my analysis and using 0 might not reflect the proper picture of the data I have decided to let the values be NaN instead.

Sanity checks:

After fixing the values, for sanity checking I once again called the printing bad values functions to see if I have fixed all the erroneous values.

Data Exploration

Task 2.1:

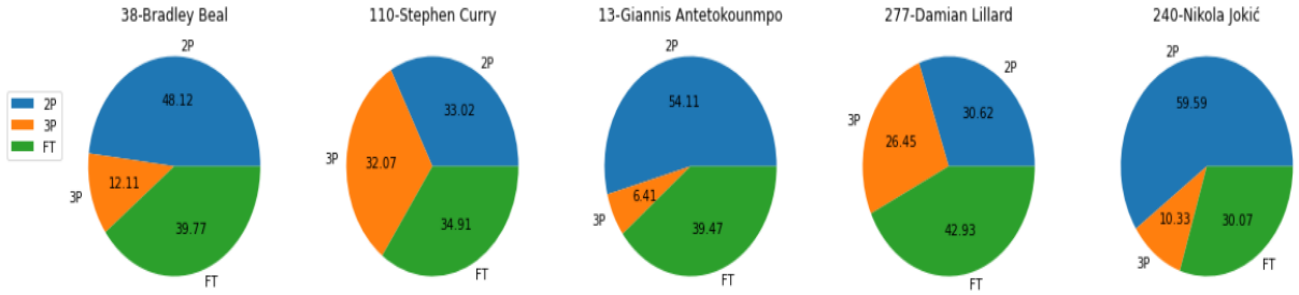

Figure 1 : Composition of points of top 5 players.

I created a dataframe capturing relevant information such as the columns that make up the points, excluding the rows that had 'TOT' as according to NBA rules TOT is not a team but a total summation of player points.[1]
After sorting the values from largest points, I have utilized pie charts for displaying the composition of player points. Referring to figure 1 above, we can clearly extract information such as Nikola scoring the largest amount of 2 pointers and Stephen Curry scoring the largest amount of 3 pointers.

Task 2.2

For figuring out the mistakes here, I have utilized boxplots to identify erroneous data.

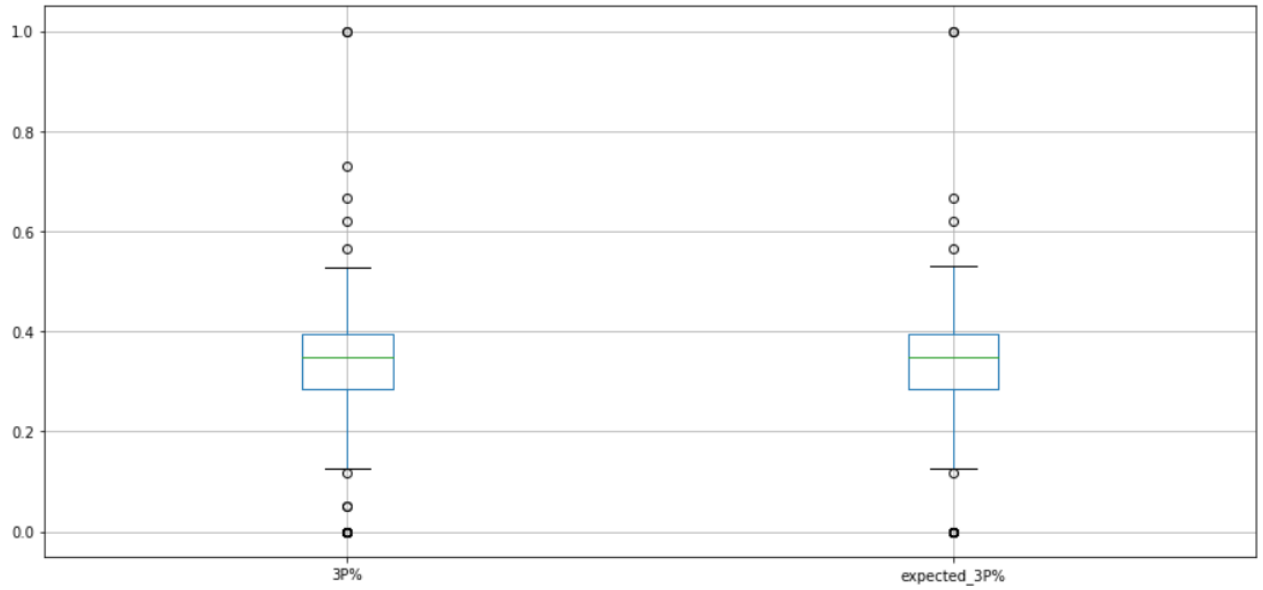

Figure 2 : Boxplots of 3P% with expected 3P%

From the above figure 2, we can see that there are outliers in 3P% that are not expected and hence can conclude there is an error in the 3P% column. The error cannot be in the 3P column as I have used that to calculate the total pts when checking for bad pts values. If an error was present

there it would have flagged the PTS value when I called the function. One of the errors found here was a transposition error where the value of 0.73 was corrected to be 0.37.
For sanity checks I have also written the print_bad_3pts_percentage() function that prints out the bad 3P data after comparing expected 3P% values using numpy's isclose() function which is used to compare floats with a tolerance value of 0.1.

Task 2.3

For this task I have used a number of graphs against the PTS values to see if there was a relationship. I first used a scatter matrix to get an overview of strong and weak relationships and then drew scatter plots of the prominent ones.

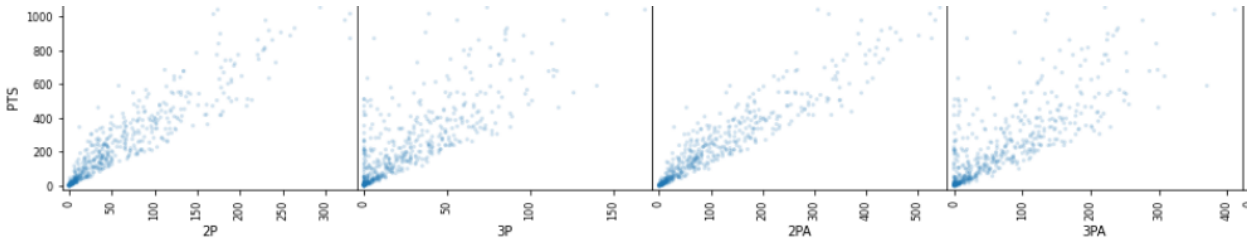

Figure 3: Relationship of PTS and 2P, 3P and their respective attempts

As expected we see an increase in total points as players make more 2P and 3P shots and attempts, and so there is a correlation.

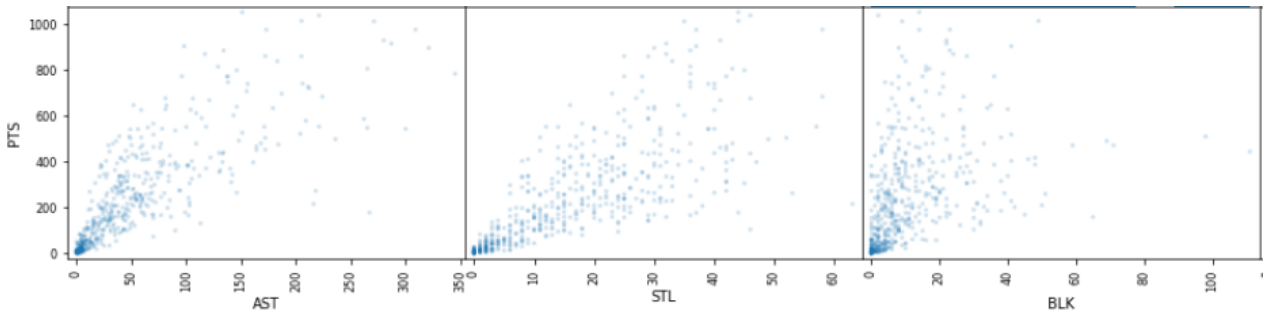

Figure 4 : Relationship of PTS and AST, STL, BLK.

For other player metrics such as Assists, Blocks and Steals when we visualize the data we can observe a weak correlation between AST and STL but not BLK.

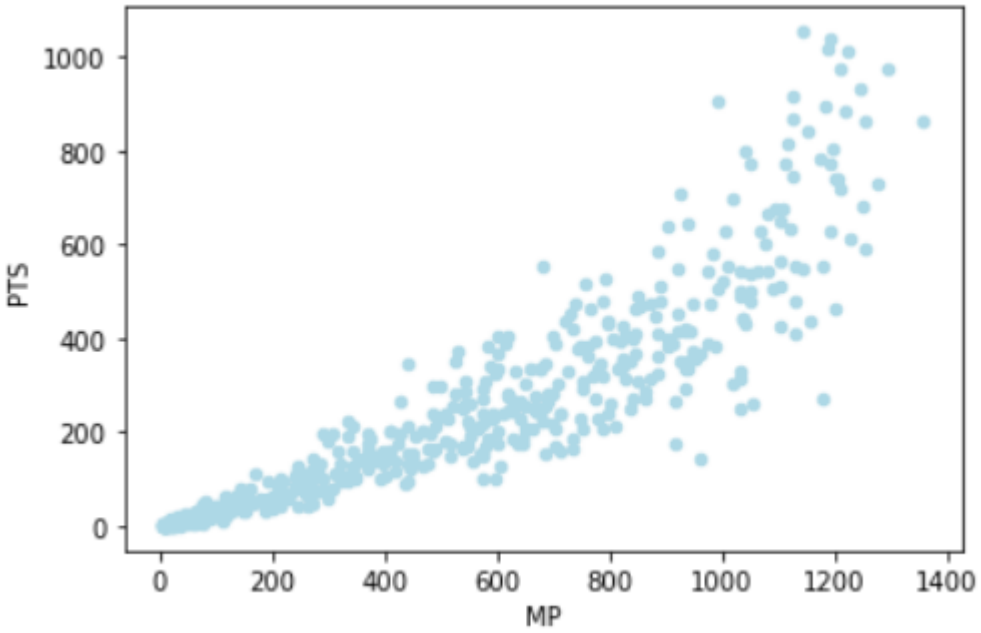

Figure 6 : Relationship between PTS and MP.

Referring to figure 6, we can see as a game runs longer and minutes played increase the points also increase in an exponential manner.

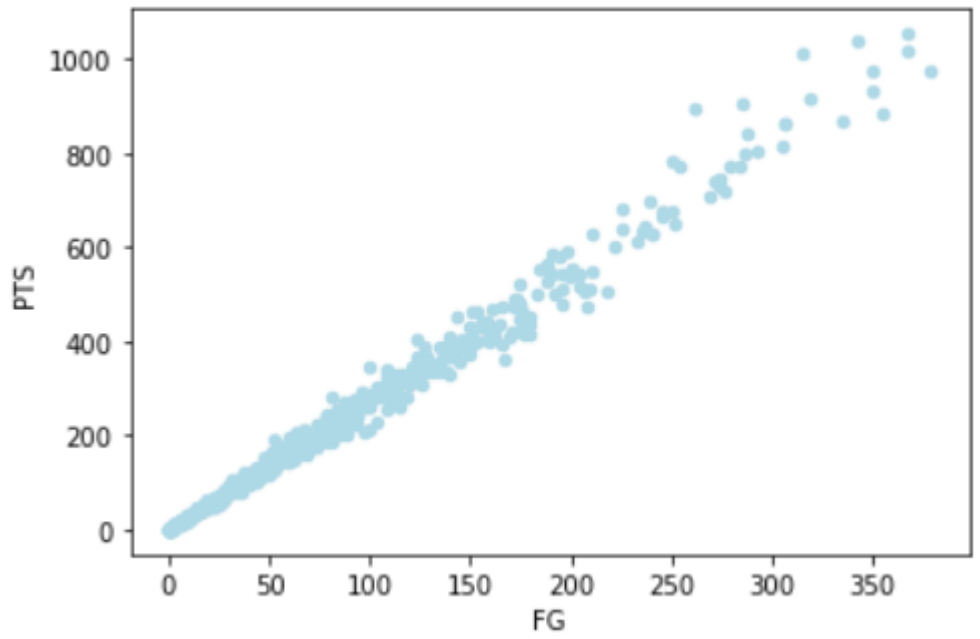

Figure 7 : Relationship between PTS and FG

As expected, field goals and points are proportionally related because the more goals a player scores the higher his overall points will be.

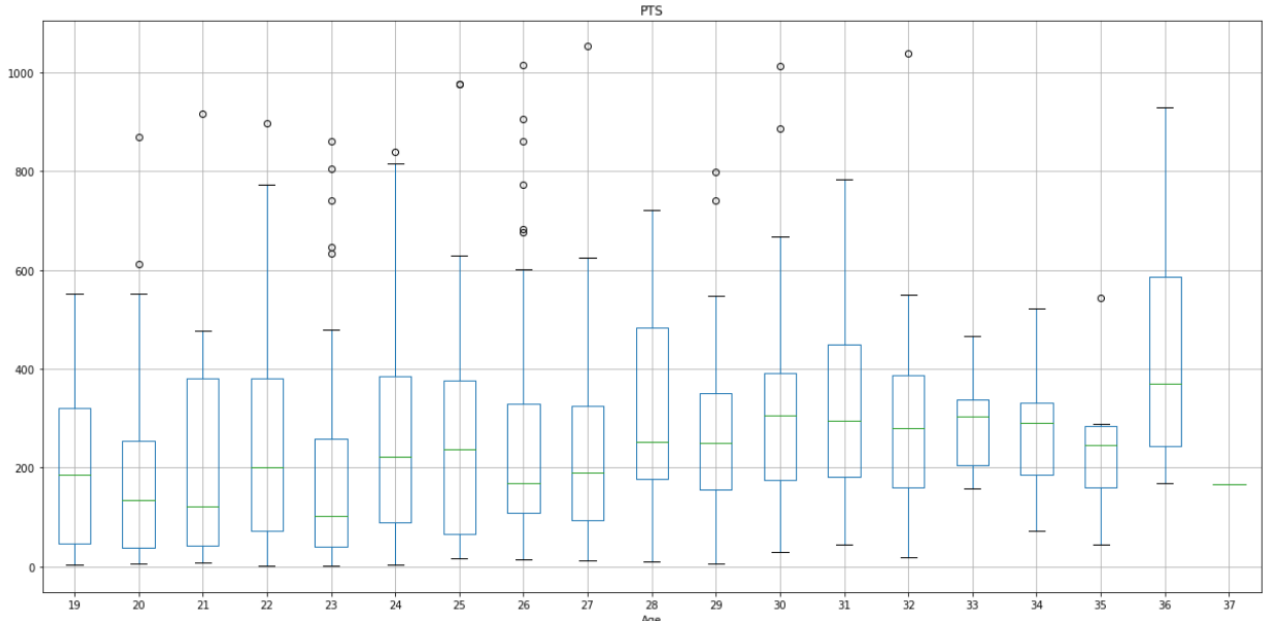

Figure 8 : Boxplot grouped by Age

When exploring the relationship between a player's age and points one could assume there is a correlation between age and points but referring to figure 8, there seems to be no plausible relationship even in the median values. I even drew a scatter diagram (figure 9) to further check for a strong or weak relationship.

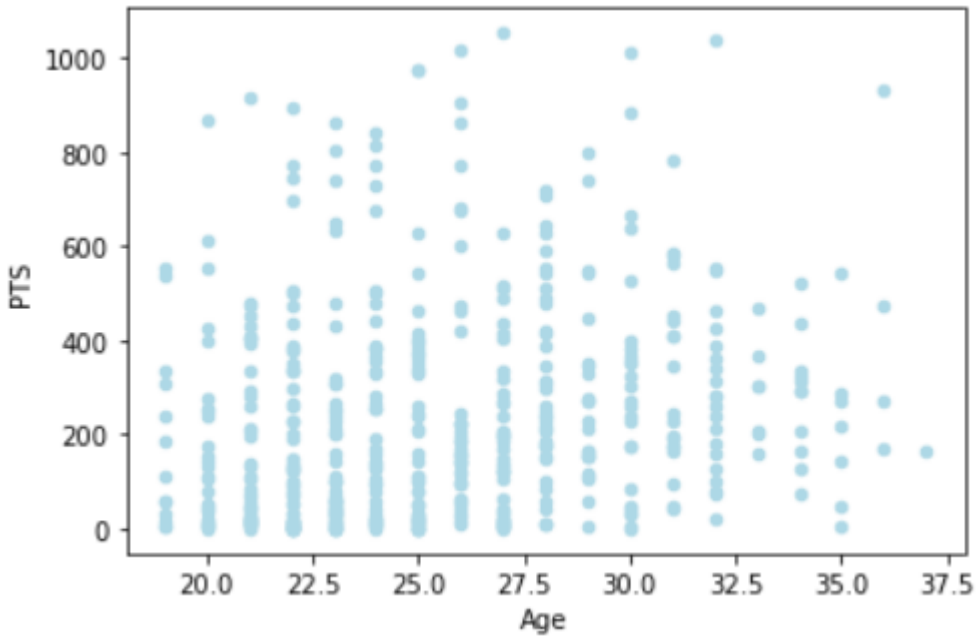

Figure 9 : Shows no relationship between player's Age and PTS

References

1. Wikipedia <https://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Baseball/Team_abbreviations>, viewed 16th April 2021.