RMIT University
Practical Data Science with Python
Assignment 1: Data Cleaning and Summarising

Task 1: Data Preparation

The aim of this task was to examine, cleanse and analyse the provided data set of NBA player stats. Using pandas functions, I loaded the data from the CSV file and cross checked it against the pre-existing data in the source CSV file. Following on from this, I proceeded to identify discrepancies and errors in the data such as typos, extra whitespaces and missing values and re-write the data into an updated CSV file. The values inside the each column was checked using values_counts() function, The top rows were displayed using the head() function, the datatype was checked using dtypes function. The dimensions of the data was checked using the shape() function which showed that there were 512 rows and 29 columns in the dataset.

Typos:

In the data set there were two sorts of typo errors, one where the spelling was incorrect, other where the case was wrong. The pandas function '.replace()' and '.str.upper()' were used to correct those errors which made sure the count of each value was correct.

Extra Whitespaces:

The extra whitespaces were removed with '.strip' function, this was much need as it was seen that a same string is repeated more than twice which was clearly because of some extra space and by removing them we had a proper count of each value.

Missing Values:

The missing value were found in the columns like FG%,3P%,2P% and FT%, I analysed them and saw that those values were only missing because FGA,3PA,2PA and FTA have 0 in their values and the formula to compute the percentage is like $\frac{FG}{FGA}$ and if there 0 in the denominator the value will be undefined so I just imputed them to 0 for the consistency of the data set as we cannot drop them as they are in huge amount and could lead to information loss.

Sanity Checks for impossible values:

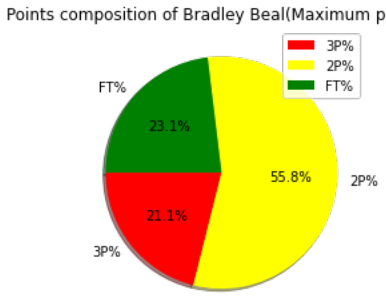
I sorted the values by using value_counts().sort_index() which allowed me to get an eye ball for some impossible values. The columns Age had some impossible values such as -19 and 280, for those I examined the whole row for them and did not find any other error, and also compared with other rows of same values, they were quite less but were enough to make me sure that we can keep these values. hence it was evident that they were just typos. The name column had some special characters, and I removed the characters which are not allowed in a name but there were still few names with some special characters, so I googled the player's name and saw it, It was same in the external sources and hence I kept those characters. The columns with percentage of scored/attempt had come values which were not rounded and hence I used the formula to compute them and rounded them to 3 places at the first place to get rid of errors and have consistency. The maximum minutes a player can play regardless of overtime is computed by G * maximum minutes per game, so I checked that the values of MP column are less than that. TRB column can't have more value than ORB + DRB so I checked the sum of that and saw if there was any bad value. There was a chance that fouls column could have a bad value as maximum foul a player can do is 6, so I calculated the average fouls per game by dividing fouls column by games played column and made a check that maximum value of that is less than or equal to 6. I also checked that the total rebounds are equal to offensive rebounds + defensive rebounds. I observed that PTS column had some impossible values as the maximum value can't be more than 2000 and that column had values more than that, hence I imputed them with the formula of total points which includes 3 pointers + 2 pointers + free throws.

Task 2: Data Exploration

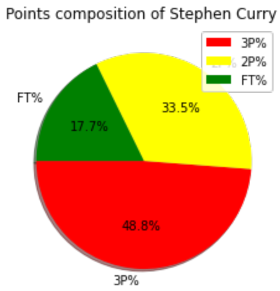
2.1

For this question I made a new data frame which only consists of top five players which included their points and the percentage of 3 pointers, 2 pointers and free throws from total points.

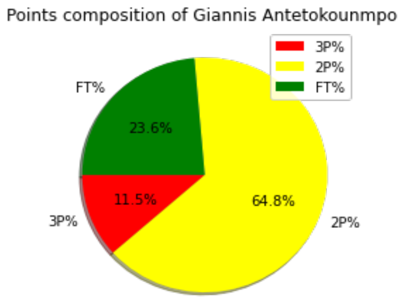
Then I plotted pie chart for each of the following players and analysed that



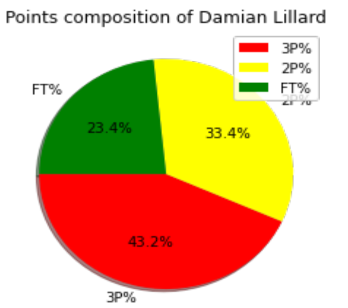
Bradley Beal had the maximum points which had the maximum percentage of 2 pointers and rest of the points constituted of 3 pointer and free throws which had the percentage of 21.1 and 23.1 respectively.



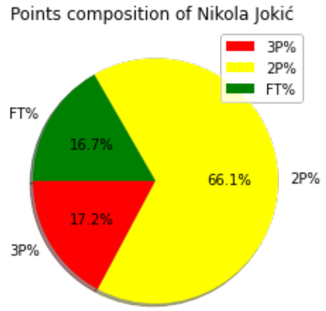
3 pointers contributed most to Stephen Curry's total points which was approximately 49% and free throws and 2 pointers contributed to rest of his total points.



3 pointers and free throws only covered around 35 % of Giannis Antetokounmpo's total points from which we analysed that majority number of points were scored through 2 pointers.



Points composition of Damian Lillard was quite similar to Stephen Curry's. The majority of points are scored by 2 pointers and 3 pointer and the least by free throws.

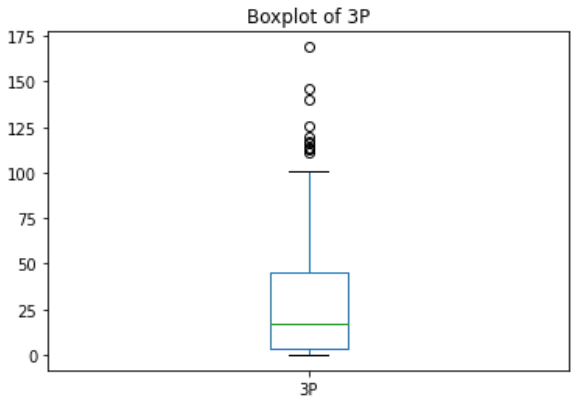


Similar to Giannis Antetokounmpo, Nikola Jokić had his majority of points from 2 pointers which is almost 66% and the rest is constituted by 3 pointers and free throws.

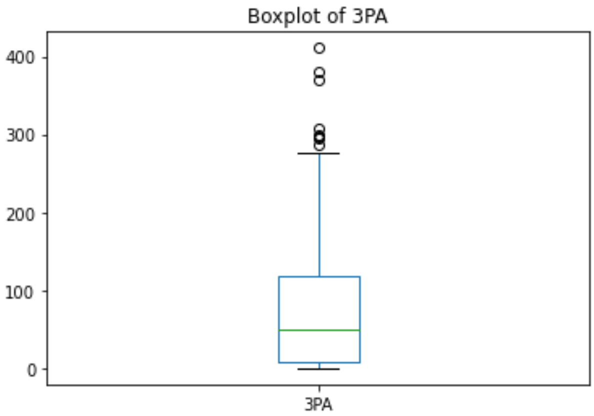
The analysis provides the conclusion that field goals accounted for the majority of points across all players.

2.2)

To check errors in 3P column we plot a box plot which gives us some outliers as see below.

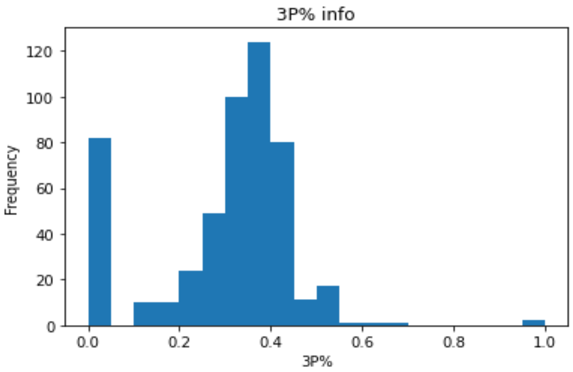


The plot shows us some values outside the mean range; hence I checked those values and explored the whole columns there were quite a few players and hence I checked whether the number of 3P are difference of total field goals and 2 pointers, all the values were true hence I don't believe there is any error.



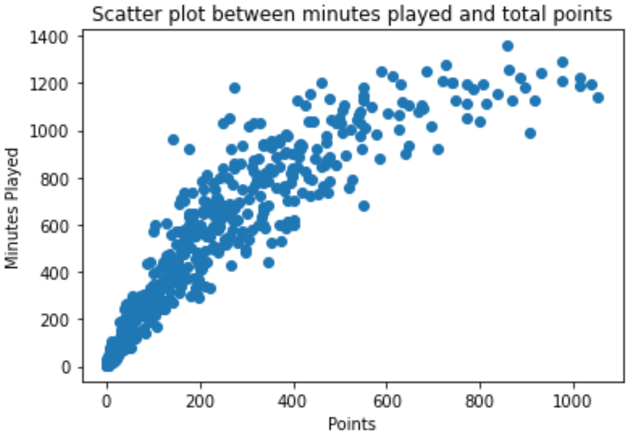
For the column 3PA I used the same approach by plotting aa box plot as seen below. It helps me to see the outliers and hence I explored those players by using the mask feature and found them valid and checked them in the similar way as done above by checking 3PA is the difference of total field goal attempts and 2 pointers attempted. All the values came true and hence it shows there is no error in that.

For the column 3P% I made a histogram and observed some outliers hence by masking them and checking whether 3P% can be obtained from number of total attempted and number of successful attempts, hence all the values were obtained just instead of the players having 0 attempts, but they can be ignored here. Hence, I feel I am able to get rid of errors while cleaning the data.

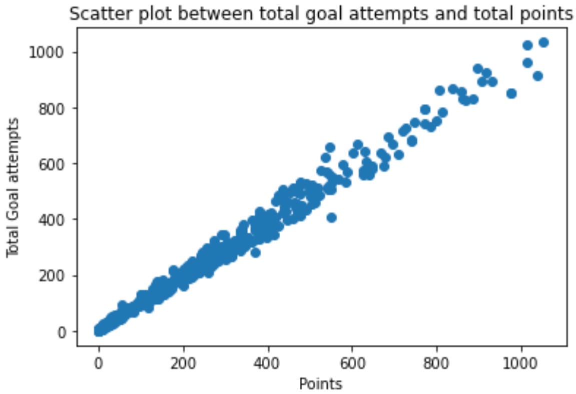


2.3)

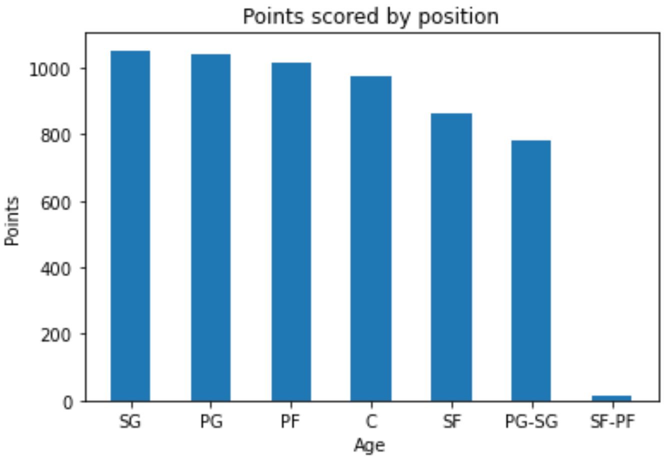
Firstly, I had a releationship between number of minutes played and total points and it is observed a linear scatter plot but having moderate density at the top because there are not many players having high number of minutes played. We can analyse here that more the player plays in a game there is more probability of having more points.



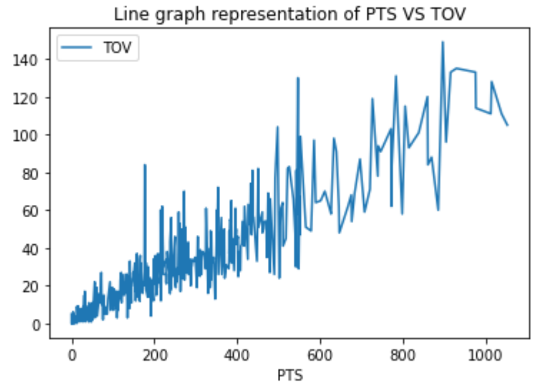
I also saw releationship between total number of goal attempts which included number of field goal attempts and number of free throws attempts, here I observed a strong linear scatter showing more the player attempts to score, the more he has his total points.

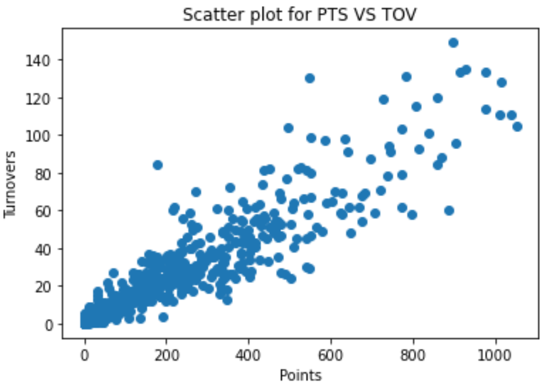


It is also observed by looking that the trend below that shooting guards (SG) and point guards(PG) have the maximum amount of points as their main objective is to score more and more points and on the other hand small forward(SF) and power forward (PF) have less points as compared to shooting guards and point guards as their main motive is to block the oppositions shots and defend the rim.



Moderately it is also analysed that Turnovers has also contributed towards the total points as more the player wins possession over the ball the more, he is likely to score more.





Reference:

Dr. Yongli Ren; 2021,'Practical Data Science: Data Curation', Lecture slides, COSC 2670, RMIT University, Melbourne.

Dr. Yongli Ren; 2021,'Practical Data Science: Data summarisation', Lecture slides, COSC 2670, RMIT University, Melbourne.

Pandas.pydata.org. 2021. *pandas documentation — pandas 1.2.4 documentation*. [online] Available at: https://pandas.pydata.org/docs/ [Accessed 17 April 2021].

Vegibit. 2021. *Matplotlib In Jupyter Notebook - Vegibit*. [online] Available at: https://vegibit.com/matplotlib-in-jupyter-notebook/ [Accessed 17 April 2021].

Basketball-Reference.com. 2021. *Basketball Statistics and History* | *Basketball-Reference.com*. [online] Available at: https://www.basketball-reference.com/ [Accessed 18 April 2021].