Assignment 1

Hongjie Xu s3880497@student.rmit.edu.au

April 18, 2021

1 Data Preparation

To address the Task 1, I firstly read the data into Pandas dataframe and check each attributes separately. Then, I used value_counts(), isna(), unique(), etc. methods to find the potential errors and correct them one by one.

I will explain below how I solved with each type of errors found in the dataset.

1.1 Error 1 - Typos

Typos are found in [Pos], [Tm]. Identifying and Cleaning steps are shown as follows:

1. Using value_counts() method to show all possible value and check with the allowed value list, then clear them one by one.
2. Using str.replace(".","", regrex=False) to remove the redundant "." in [Pos]. (". " is not a regular expression, setting regrex as False)
3. Using replace() method to deal with the data entry error. In detail, The "H0U" should be "HOU" in [Tm]. The "PFa" should be "PF" and "SGa" should be "SG" in [Pos].

1.2 Error 2 - Redundant Whitespaces

Redundant Whitesespaces are found in [Pos], [Tm]. Cleaning steps are shown as follows:

1. Using value_counts() method to show the all possible values.
2. Using str.strip() method to remove the leading and trailing redundant whitespaces in [Pos], [Tm].

1.3 Error 3 - Capital Letter Mismatches

Capital Letter Mismatches are found in [Pos], [Tm]. Identifying and Cleaning steps are shown as follows:

1. Due to the number of possible value is not many, using value_counts() to check. All letters in [Pos] and [Tm] must be capital letter.
2. Using str.upper() to change all letters in [Pos], [Tm] to be Capital letter.

1.4 Error 4 - Missing Value

Missing Values are found in [3P%, 2P%, FT%]. Identifying and Cleaning steps are shown as follows:

1. Using isna() method to identify whether there are missing values NaN existing in columns.
2. Using fillna() method to fill missing values. To show the difference between players who really have 0 [3P%, 2P%, FT%, FG%] and players who have 0 [3PA, 2PA, FTA, FGA]. I decided to re-calculate these three columns based on the formula provided, then use fillna(-1) method to replace missing value NaN to solve this problem.

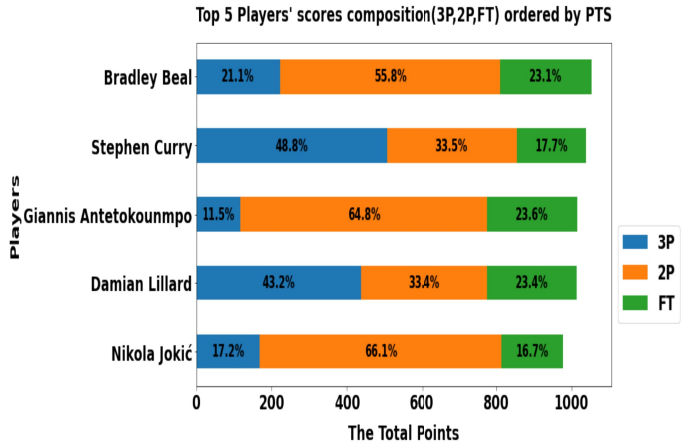
1.5 Error 5 - Impossible Values & Sanity Checks

Impossible values are found in [Age], [PF]. Identifying and Cleaning steps are shown as follows:

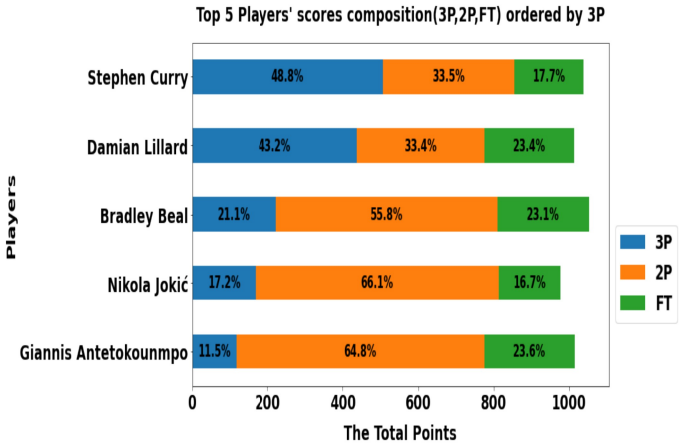
1. Using unique() method to list all unique values
2. 280, -19 are found in [Age], Obviously they are impossible value. The 0 and negative symbol are redundant. Using replace(280,28) and replace(-19,19) to correct them.
3. I also found a player who has 6 personal fouls in each game. Obviously it is not possible. After checked the data source cite, this player's personal faults per game from season 2017 to season 2021 is about 2.1, then the total estimated personal faults in this season is $38 \times 2.1 \approx 80$. Then using replace(228,80) to correct it.
4. The impossible values also exist in the PTS column, where the value is greater than 2000 because of incorrect calculation. I decided to re-calculate the PTS column so that all value should be correct. The formula is $3P \times 3 + 2P \times 2 + FT \times 1 = PTS$.

2 Task2. Data Exploration

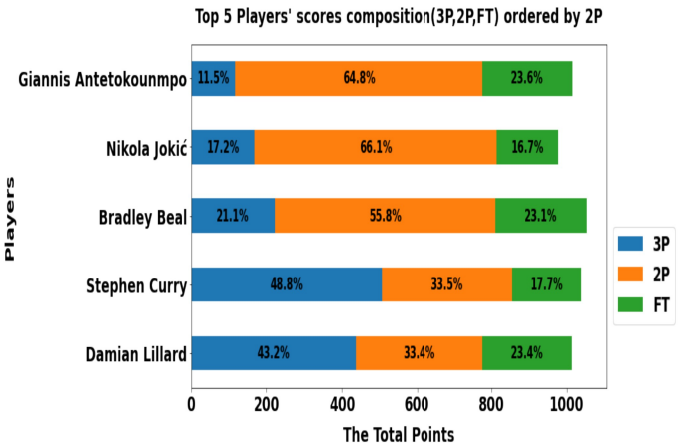
2.1 Task 2.1



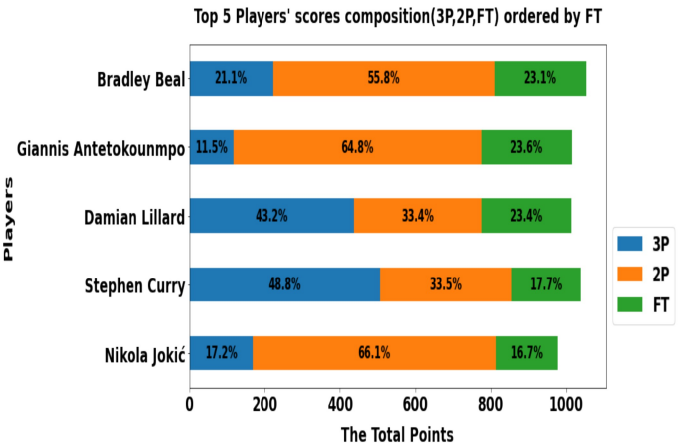
(a) Ordered by PTS



(b) Ordered by 3P



(c) Ordered by 2P



(d) Ordered by FT

Figure 1: The relationship between PTS composition and Players' Position

Note that 'TOT' in [Tm] means the player plays for more than one team, and the 'TOT' row shows the player's total performance. Meanwhile, those players' performance data in the specific single team also exists in dataset. This analysis is based on the Player but not team. Therefore, I removed those players' data in the specific single

team but kept those 'TOT' data to show these players' total performance.
Figure 1 shows the Top 5 Players' scores composition, ordered by different points. The percentage shows the player's percentages of 3P, 2P and FT in their total points.

1. (a) shows the total points composition of top 5 players ordered by PTS. Bradley Beal achieved the highest place, followed by Stephe Curry who is at 2nd Rank.

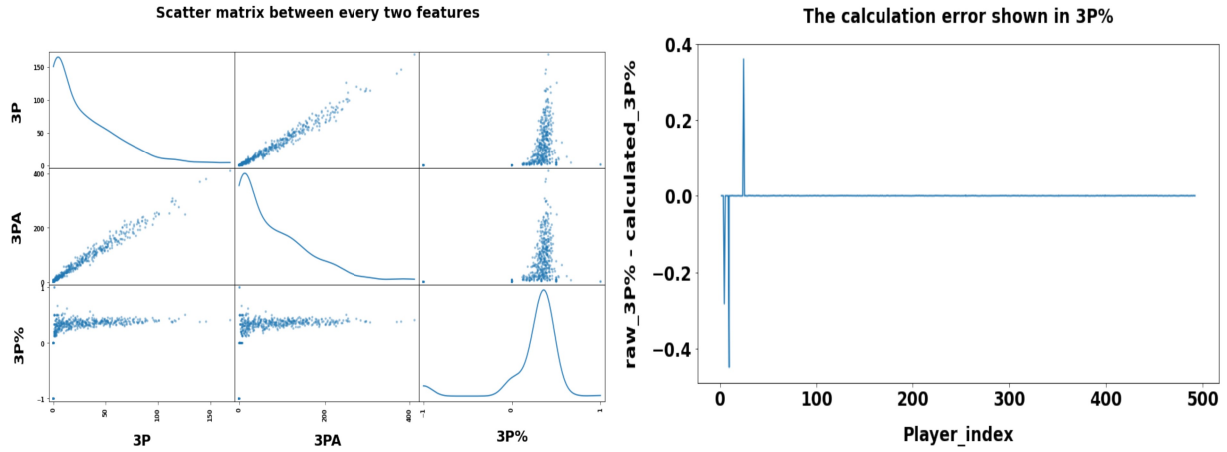
2. (b) shows the total points composition of top 5 players ordered by 3P. Although Stephen Curry is at 2nd rank in PTS, he achieved the highest scores from 3 Points. However, the Top player Bradley Beal's 3 points percentage is only at the 3rd place. Giannis Antetokoumpo's 3 Points is lowest percentage, only at 11.5% of his total points.

3. (c) shows the total points composition of top 5 players ordered by 2P. Giannis Antetokoumpo achieved the most 2 Points scores, followed by Nikola Jokic. Both 2 Points scores are occupied their total points over 65%. Reversely, Stephen Curry and Damian may not prefer getting 2 Points than 3 Points, and their percentages of 2 Points are around 33%.

4. (d) shows the total points composition of top 5 players ordered by FT. Five players have no significant differences in the percentage of FT scores, from 16.7% to 23.6%.

2.2 Task 2.2

Assuming there are data entry errors in Column [3P, 3PA, 3P%], the visualization could be used to identify those potential errors.



(a) Scatter Matrix

(b) Calculation Error

Figure 2: Visualize the assumed data entry error

The Fig 2 (a) scatter matrix shows the relationship between every two features from 3P, 3PA, 3P%, which is a great tool to find a potential error. Firstly, from the scatter plot between 3P and 3PA, we can see a few data point is extremely far away from other data points, indicating they are might be outliers. However, the Scatter plots between 3P% and (3P or 3PA) prove that the 3-Points percentage is falling in a narrowed area, larger than 0.3 but lower than 0.5. Those outliers in 3P and 3PA still shows a reasonable 3P%, sharing a similar probability. It can be concluded that those players probably are great shooters or they just played more games(or MP) to play because the players' total data are investigated.

The Fig 2 (b) simply shows the calculation error in the 3P% where the y axis is calculated by minusing 3P% in the raw data from 3P% in the cleaned data .

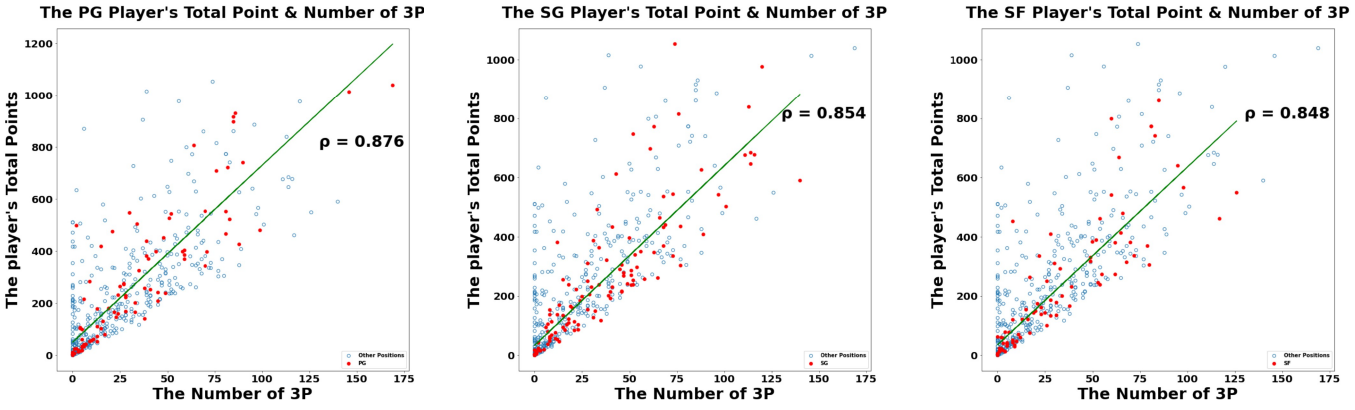
2.3 Task 2.3

To show the relationship between players' total points and other variables, the scatter plot with regression line and Pearson correlation coefficient are used in following analysis.

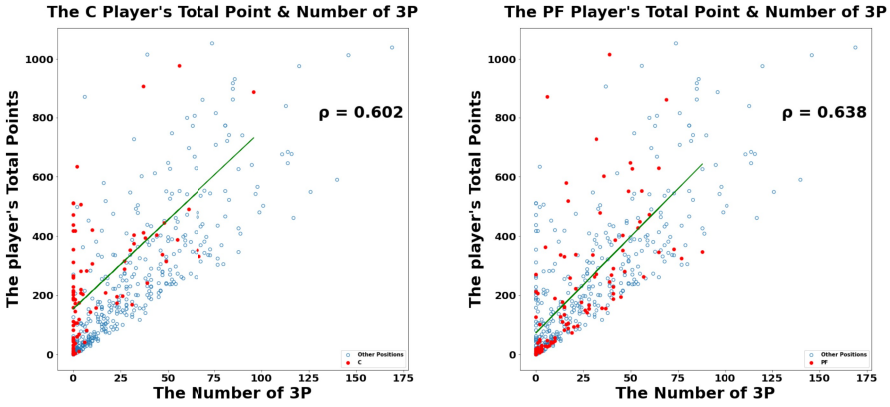
$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

Note that the relationship does not mean causality. Due to some players who played for more than one team, 'TOT' needs to be used and those players' single team data need to be removed because the analysis focuses on the Players' performance.

2.3.1 Task 2.3.1 The relationship between Number of 3-Points and Player's Total Points with different positions



(a) Colored by PG (b) Colored by SG (c) Colored by SF



(d) Colored by C (e) Colored by PF

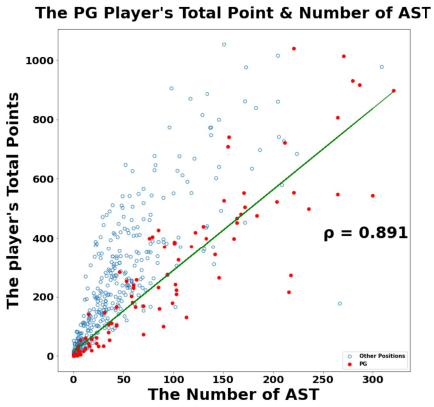
Figure 3: 3P & PTS, Colored by Different Positions

The Fig 3 shows the relationship between Players' Total Points(PTS) and their number of 3-Point Field Goals(3P), considering their positions. Only [PG,SG,SF,C,PF] five positions are considered in this analysis.

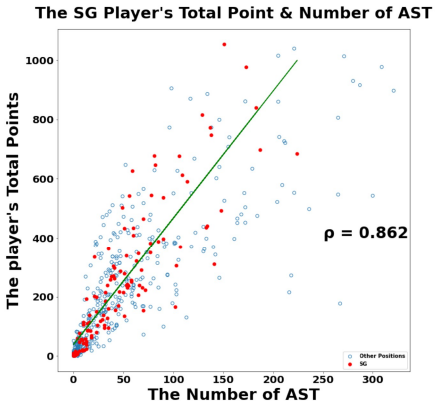
From (a), (b) & (c), they shows the colored points are generally closer to the green regression line. The closer distance shows the stronger correlation. In addition, the ρ marked in (a), (b) and (c) are larger than 0.85, which also indicates a strong positive relationship between 3P and PTS for these positions. On the other hand, Figure (d) & (e) shows there is a weak positive relationship between the PTS and 3P among those players who are in C and PF position, where the ρ are only 0.602 and 0.638, respectively.

It can be concluded that the total points of players who are in PG, SG and SF position are strongly related with their number of 3 Points, while the total points of players who are in C and PF position are weakly related with their number of 3 Points. In another words, as the total points increase, players who are in PG, SG, SF more likely have more 3-Points than another players in C and PF.

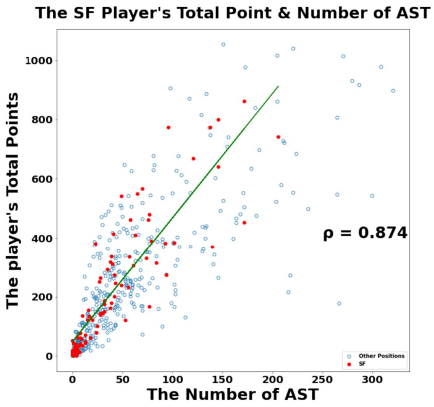
2.3.2 Task 2.3.2 The relationship between Player's AST and Player's Total Points with different positions



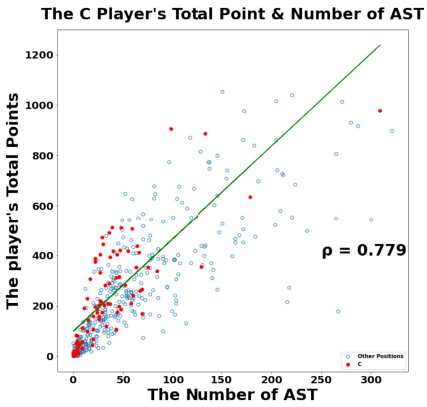
(a) Colored by PG



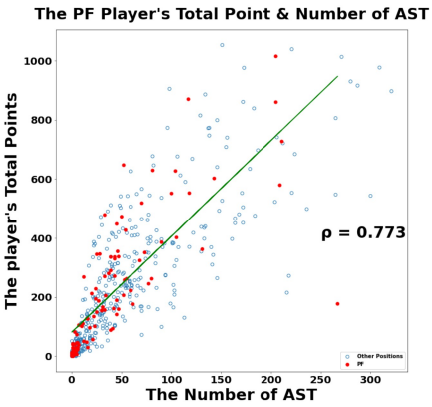
(b) Colored by SG



(c) Colored by SF



(d) Colored by C



(e) Colored by PF

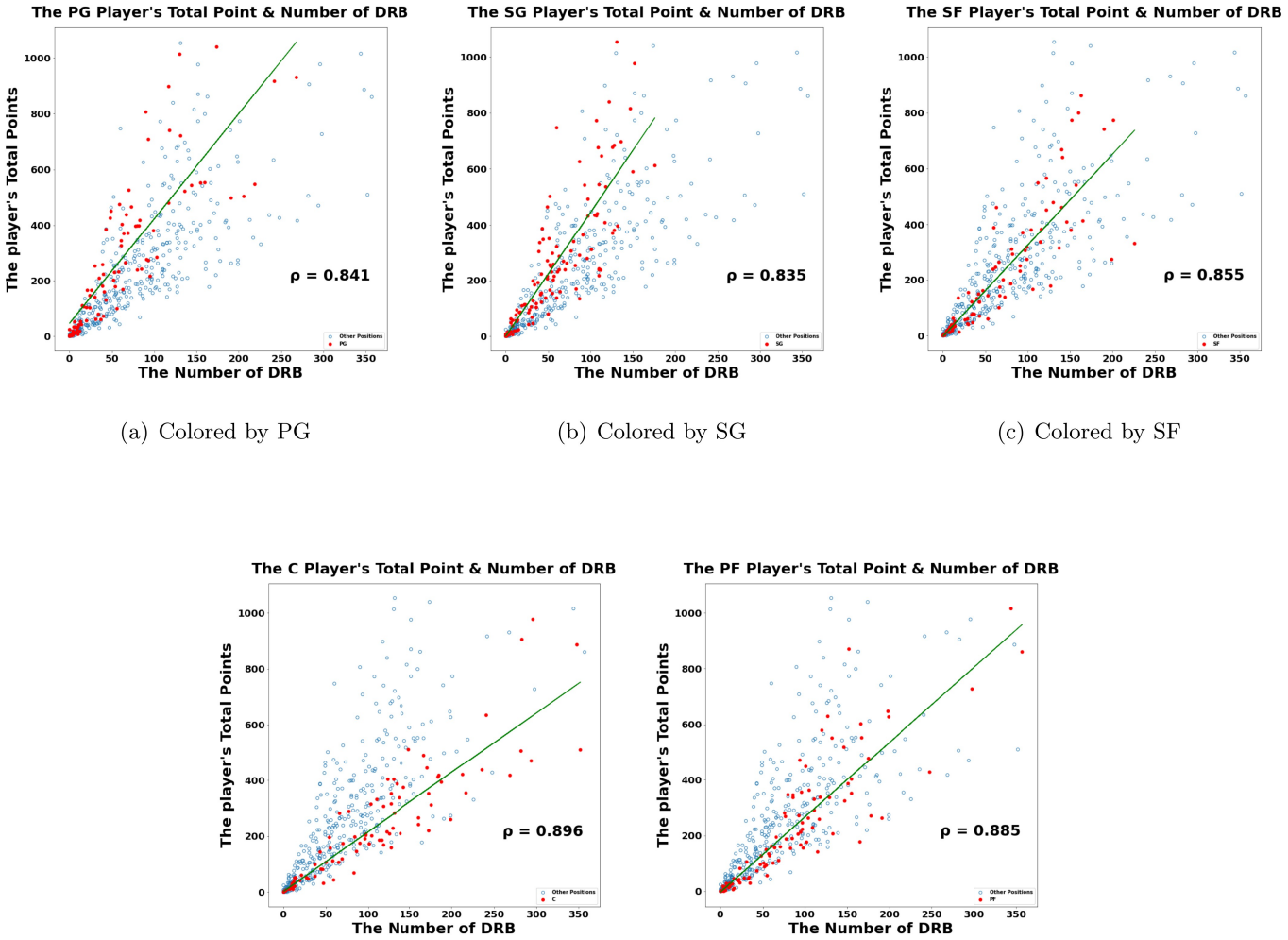
Figure 4: AST & PTS, Colored by Different Positions

The Fig 4 shows the relationship between Players' Total Points and their number of Assists(AST), considering their positions. Only [PG,SG,SF,C,PF] five positions are considered in this analysis.

From (a), (b) & (c), the colored points are closer to make a line, as the green regression line does. It shows there are strong positive relationships between the number of AST and the total points, where the players' positions are PG, SG, SF, respectively. The ρ are all larger than 0.86, where also represent a strong positive correlation mathematically. The players who are in C or PF have also a positive relationship ($\rho > 0.75$) between their number of AST and total points, although this relationships are weaker than those players in PG, SG% SF.

In conclusion, as the total points increase, the PG, SG, SF players more likely have more AST than those C, PF players because the relationship between PTS and AST for PG, SG, SF players are stronger than players in C, PF positions.

2.3.3 Task 2.3.3 The relationship between Player's DRB and Player's Total Points with different positions



(a) Colored by PG

(b) Colored by SG

(c) Colored by SF

(d) Colored by C

(e) Colored by PF

Figure 5: DRB & PTS, Colored by Different Positions

The Fig 5 shows the relationship between Players' Total Points and their number of Defensive Rebounds(DRB), considering their positions. Only [PG,SG,SF,C,PF] five positions are considered in this analysis.

From (a), (b) & (c), the relationship between Players' Total Points and their number of DRB in PG, SG, SF position generally are strong positive relationships, where the ρ are all greater than 0.8 also indicate the strong relationship. However, as the (d) shown, for the players in C position, the relationship between their DRB and their total points are much stronger, where the ρ almost reaches 0.9. Also, (e) shows the relationship between DRB and total points for players in PF position are little weaker than those players in C position, although it is still stronger than same relationship for players in PG, SG, SF.

In conclusion, although it seems that players in these 5 positions are all involved to have Defensive Rebounds, which are all strongly positively related to their total points. However, as the players' total points increase, players who in C position are more likely to have more DRB because of their strongest positive relationship with total points.