Student name: NGOC BAO VY LE Student number: s3828276

ASSIGNMENT 1

Task 1. DATA PREPARATION

Pandas and matplotlib.pyplot library were imported for reading csv file and'NBA_players_stats.csv' file was opened to get a glance on data structure. As the data set already has header and its features are separated by coma, therefore, a few parameters of read_csv() function were used such as sep=',' and decimal='.'. Data loaded properly, and its features got separated in columns.

By using info() funtion, it showed that the data set had 512 entries and 29 features with missing values in FG%, 3P%, 2P% and FT %. It also displayed information about data type for each column.

The isna(), value_counts() and plot.barh() functions were used to find missing values in FG%, 3P%, 2P% and FT%; and fillna() to fill replace missing values. For the columns that had data type as object, using value_counts() counting the occurrence of each unique values to find typos, str.strip() to strip white spaces and loc[] to locate error's rows. For columns that has data type as integer or float, using plot.hist() to find the outlier and Boolean logic to check for data consistency across columns, then loc[] to locate error's row. Finally, after all errors were cleaned and fixed, the data got saved as cleaned_NBA_players_stats in a csv file.

1.1 Missing value

- FG%

There were 3 missing values. They got filled by results of the division between FG and FGA.

- 3P%

There were 33 missing values. Filling missing values by the results of the division between 3P and 3PA.

- 2P%

There were 7 missing values. Filling them by results of the division between 2P and 2PA.

- FT%

There were 32 missing values. Missing values filled by results of the division between FT and FTA.

1.2 Extra white spaces

There were 6 white space errors in Pos and 22 white space errors in Tm. To fix them, using str.strip() function, to convert data type to string, then strip white spaces.

1.3 Typos

There were 7 typos in Pos and 2 typos in Tm. To fix these typos, using upper() converting the characters to upper case, masking the typos, applying loc[] function to find their positions and changed them to the correct values accordingly.

1.4 Impossible values

- *Age*

There were 2 impossible values in Age. The unique() function list unique values in Age, and in the values –19 and 280 were out of range. As age cannot be less than 0 and more than 117, they were typos mistakes by human, therefore, I used replace() function to change –19 to 19 and 280 to 28 after checking these player's birth.

- PTS

There were 2 impossible values in PTS at 20000 and 28800, as PTS was less than 2000. Using loc[] function to find their position and replace their value by sum (2P*2 + 3P*3 + FT).

1.5 Calculation errors

- *3P%*

When comparing values in 3P% to the values of 3P divided by 3PA, there was a lot of errors. However, most of these errors were either the roundup of number of decimals or having divisor as 0 in value. Therefore, they were not considered as errors. After filtering these values out, there were 3 calculation errors in 3P% as these values were not equal to the corresponding values of 3P divided by 3PA. Filling these errors by results of the division between 3P and 3PA.

Task 2 DATA EXPLORATION

2.1 Analyzing the composition of PTS of top five players with the most points.

As Players were able to play in multiple teams in one season, the total points of players were the sum of their total points of teams they had played in one season.

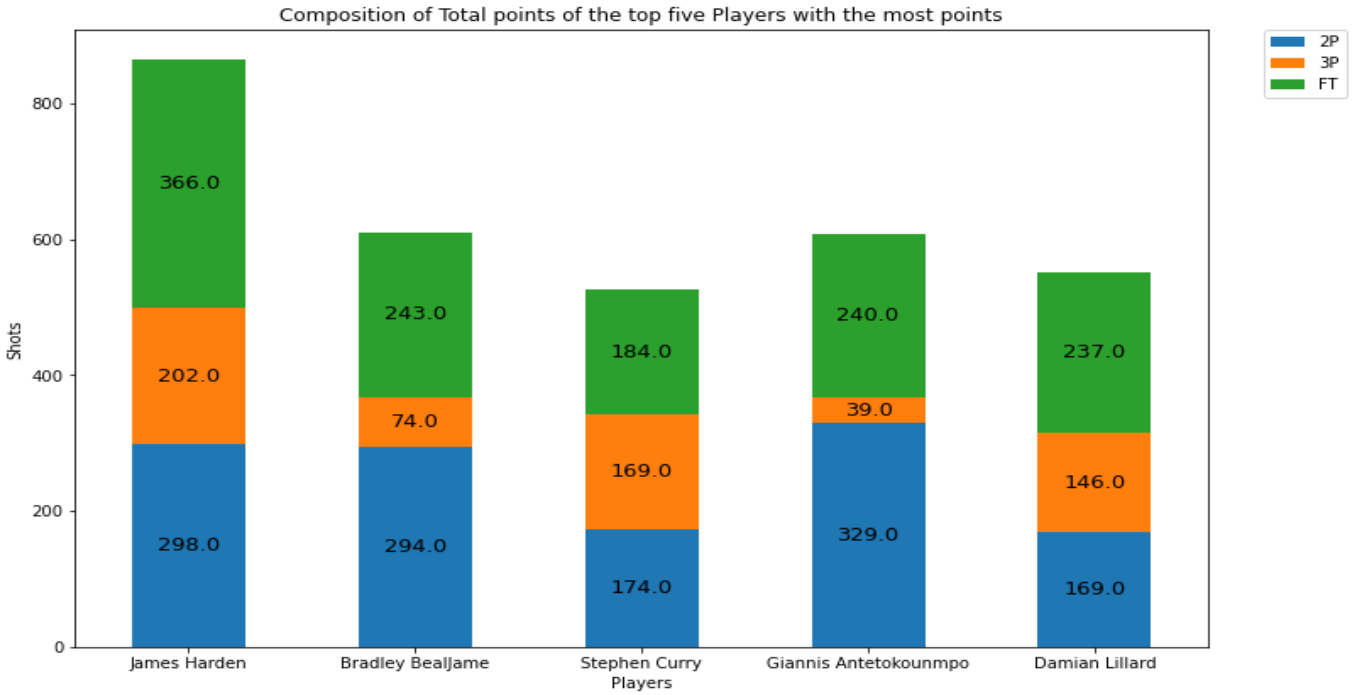

Figure 1: Stacked bar chart describes the number of shot's type of top five player with the most total points.

	2P	3P	FT
mean	252.8	126.0	254.0
min	169.0	39.0	184.0
max	329.0	202.0	366.0

Figure 2: Statistic summary of composition of Total Points of top five players with the most points.

The above graph and statistic table showed that players in the top five made less 3P shot compared to 2P and FT shots. The Free Throw and 2-point dominated shot category with the mean of 254.0 and 252.8 respectively and 3-point has the mean of 126.0.

Players who had high 2-point shots, were solid 2-point shooters (David, Donald & Vince 2018, p.340). Leading in this category were Giannis with 329 shots, following by James with 298 shots and Bradley with 292 shots. They were often near the rim. Therefore, they were less likely to have many 3-point shots (Fadi, Li, & Neda 2019, p.108). Bradley and Giannis's 3P were at 74 and 39 shots respectively which were below the 3P's mean of 126.0 shots. However, that was not the case for James as he had the highest 3-point at 202 shots. Due to often stay near the rim, they were most likely to get offended by the opposition compared to others. As being the top three in Free Throw shots, they were particularly good at creating offensive opportunities which James lead at 366 shots, followed by Bradley at 243 shots and Giannis at 240 shots. That made James the most effective Free Throw shooter in the 2020-2021 season.

James, Stephen and Damian had a high volume of shots across all categories (2P, 3P and FT), meaning they would often have the ball in their hands while in the court (David, Donald & Vince 2018, p.339). James, Stephen and Damian also were effective 3-point shooters with 202, 169 and 146 shots, respectively. That made James the most effective 3-point shooter of 2020-2021 season. Stephen and Damian were getting the Free Throw line often and skilled at creating offensive opportunities with 237 shots for Damian and 184 shots for Stephen. Whereas James was leading in Free Throw category with 366 shots. Even though, leading second and third in 3P, Stephen and Damian's 2-point shots were below the mean of 254.0 with 174 shots for Stephen and 169 shots for Damian. While James came second in 2-point shot with 298 shots. In short, these three players often moved around the court and made effectively shoots across all categories (2P, 3P and FT).

James was an effective shooter across all categories (2P, 3P and FT). He came first in 3-point and Free Throw shots; and came second in 2-point shot. Because of playing for multiple teams in this season, it would help him being more familiar with others player's tactic. Therefore, he would be able to tackle them effectively to improve his total points. With the highest total point, James was the best NBA player in 2020-2021 season.

2.2 Exploring errors in 3P, 3PA and 3P% by visualization and solution.

2.2.1 Missing values and calculation errors

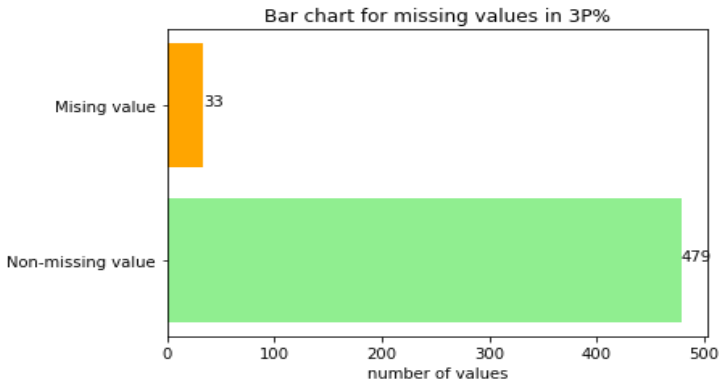

Figure 3: Bar chart for missing values in 3P%. Yellow and green bar represents for missing values and non-missing values, respectively.

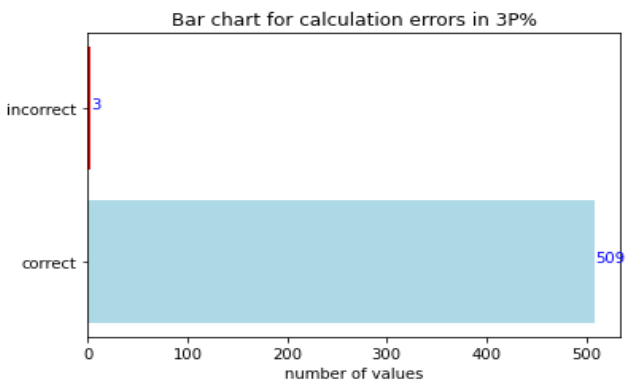

Figure 4: Bar chart for calculating errors for 3P%. Red and blue bar represents. for incorrect calculated values and correct calculated values, respectively.

The bars charts showed 3P% had 33 missing values and 3 calculation errors. Missing values were filled out by 0 and calculation errors were filled out by corresponding results of the division of 3P and 3PA.

2.2.2 Outliers

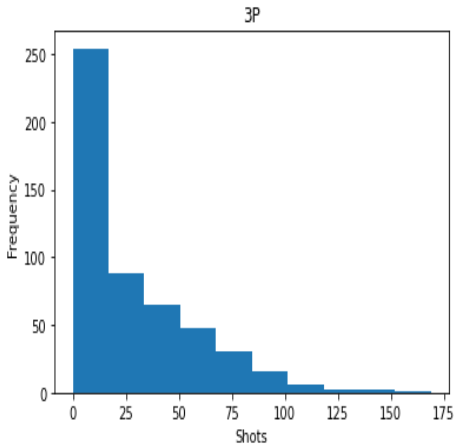

Figure 5: Frequency histogram of 3P showing the number of 3-Point Field Goals

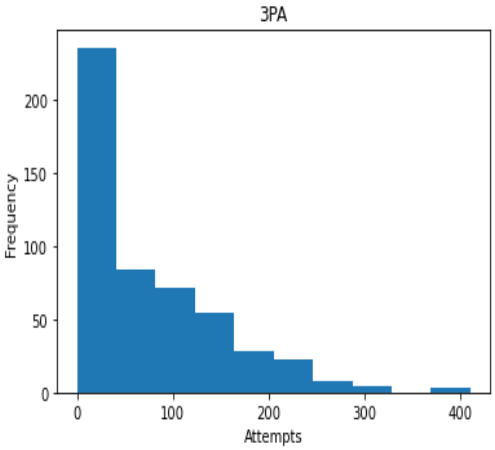

Figure 5: Frequency histogram of 3PA showing the number of 3-Point Field Goal Attempts

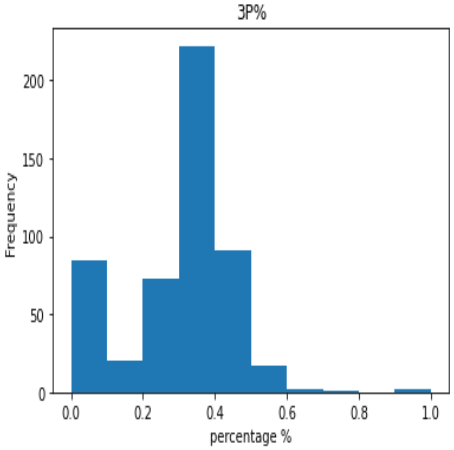

Figure 6: Frequency histogram of 3P% showing the number of3-Point Field Goal Percentage

The above charts showed no outlier in 3P, 3PA and 3P%.

2.3 Analyzing the relationship between STL, MP and TOV to PTS

2.3.1 Analyzing the association between STL and PTS.

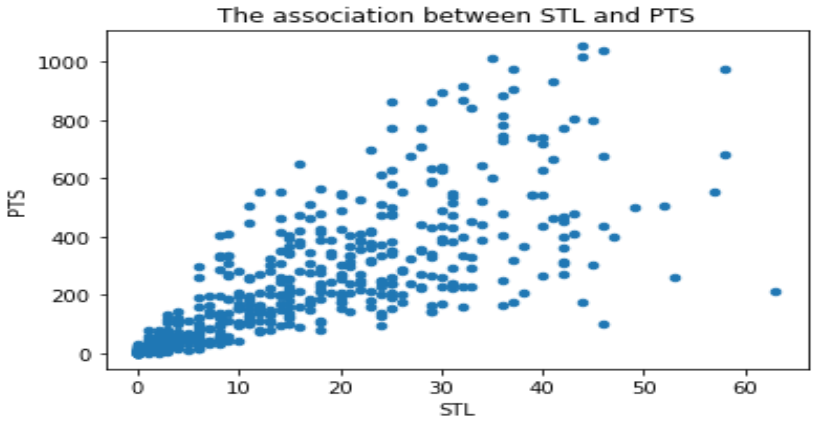

Figure 8: Scatter plot showing the effects of number of Steals on Total Points

The above graph showed a positive, weak linear relationship between numbers of Steal and Total Points with the correlation point at 0.76. The players had a greater number of Steals, tent to have high Total of Points. However, not all players followed this trend. There were outliers where players had the same number of Steals but having different Total Points. In short, Total Points increased when number of Steals increased, but not.

2.3.2 Analyzing the association between MP and PTS.

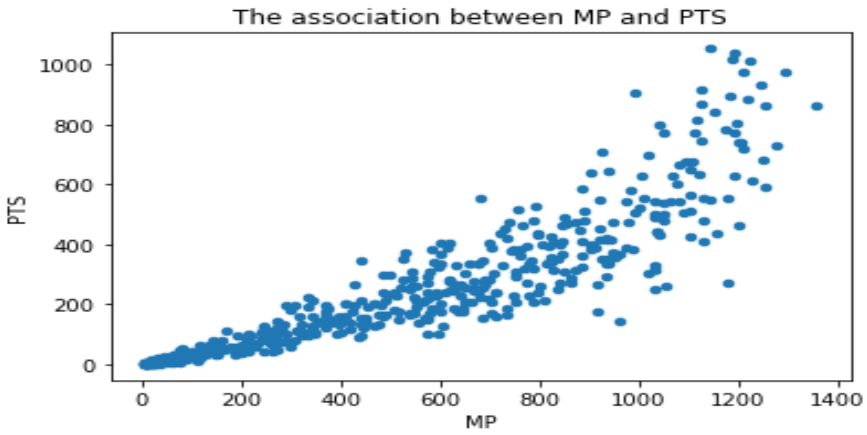

Figure 8: Scatter plot showing the effects of Minutes played on Total Points.

The scatter graph described a positive strong nonlinear relationship between Minus Played and Total Points of players with the correlation point at 0.9. The Total Points increased when Minute Played increased. Players played less than 1150 minutes, followed this trend closely. Whereas Players with numbers of minutes played above 1150 minutes, their Total Points did not always increase when number of minutes played increased. There were some outliers in the range 900 – 1200 minutes where players had high number of minutes played, had low Total Points. In short, Total Points increased when number of Minutes Played increased.

2.3.3 Analyzing the association between TOV and PTS.

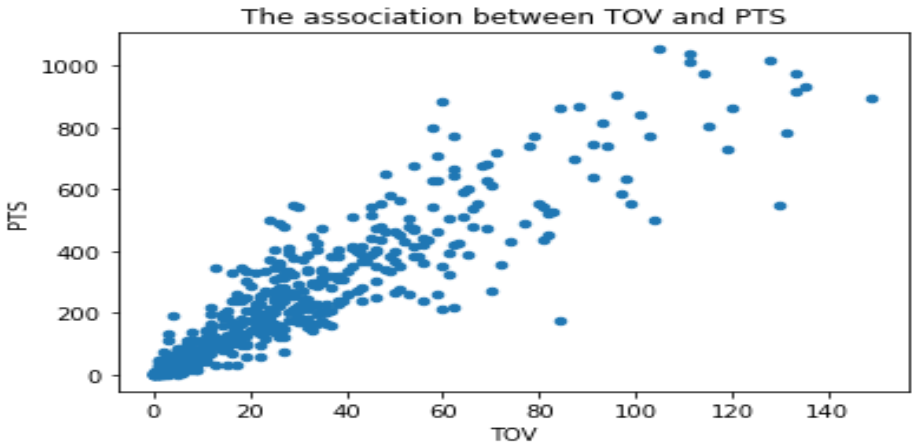

Figure 8: Scatter plot showing the effects of number of Turnovers on Total Points

The graph described a positive, strong linear relationship between TOV and PTS with correlation point at 0.91. More than half of player's PTS and TOV distributed in the bottom left of the graph. Players had TOV below 60, their PTS increased when TOV increased. Players had TOV beyond 60, their PTS did not always increase when TOV increased. In short, PTS of players increased when their TOV increased.

Reference:

Fadi, T, Li, Z, and Neda, A, 2019, NBA Game Result Prediction Using Feature Analysis and Machine Learning, Anal of Data Science, Vol.6, pp.103-110.

David, W, Donald, s and Vincent, B, 2018, Power, performance, and expectations in the dismissal of NBA coaches: A survival analysis study, Sport Management Review, Vol.21, Issue.4, pp.333-346.