Assignment 1

1 Overview

This report shows the examination of the “NBA_players stats.csv” dataset, which contains the stats of 492 NBA players during the 2020 – 2021 season, by presenting the results of the analysis using the first two steps in data science which are data cleaning and data exploring. To support the process of examination, three Python libraries were imported including pandas, NumPy, and Matplotlib. All codes and functions mentioned in this report are written inside the < > bracket as followed <code>.

2 Data Preparation

In this section, audiences will be guided through certain steps of the data cleaning process, that have been performed, from importing data to identifying and handling issues/errors. As the data will be cleaned accordingly to the description of each column, an assumption that the descriptions are accurate has been made.

2.1 Data Retrieving (Program: Section 1.1)

In this first step, data was imported from the <.csv> file into the dataframe named <NBA_players_stats> using the <read.csv()> function. As the first column of the Excel file named “Rk” is not based on a player’s performance and it contains progressive numbering, it was specified as an index column using <index_col = 0> to prevent leakage in modelling stage. To ensure that all data has been loaded accurately, <.head()> (showing the first 5 lines of the data frame), <.tail()> (showing the last 5 lines of the data frame), and <.shape()> (getting the shape of the array) functions were used. The dataframe has 512 rows and 28 columns (29 if including the index column).

2.2 Initial Data Inspection (Program: Section 1.2)

Following Data Retrieving is Initial Data Inspection. This step was conducted to get an overview of the dataframe, to check if all the data type is correct and if there are any missing values. Firstly, the function <.info()> was used. It provides a summary of the dataframe and the number of data entries in each column. The following conclusions were made because of this function:

- All data types are correct – what is categorized as object is object, int64 is in64, and float64 is float64.
- There are <null> data located in FG%, 3P%, 2P%, FT% columns as the number of <non – null> data is less than the number of rows.

To cross-check the conclusions above, <.describe()> and <.isnull()> functions were used showing the numbers of data entries in each column and number of <null> values in each column, respectively. It was confirmed that FG% column has 3 <null> values, 3P% has 33 <null> values, 2P% has 7 <null> values and FT% has 32 <null> values.

2.3 Errors Handling

Overview Observation

Typos and spelling mismatches were checked for all categorical data. This includes “Player” column, “Pos” column, and “Tm” column. Functions <.value_counts()> and <.unique()> were used at the beginning of the analysis, for each column, to identify the unique values and their number of appearances. These two functions were reused, at the end of an analysis, to make sure that all errors have been corrected.

2.3.1 Data Entry Error (Program: Section 1.3.1)

After being spotted out, spelling errors and inconsistencies were fixed using the <.replace()> function, which replace the current data with the assumingly accurate one.

2.3.2 Redundant Whitespace (Program: Section 1.3.2)

The same method of detection was carried out for redundant whitespace. The string function <.str.strip()> was utilized to eliminate whitespace on the left and right of the data.

2.3.3 Capital Letter Mismatches (Program: Section 1.3.3)

Capital letter mismatches often occurred in data with initials. In this case, it occurred in "Pos" and "Tm" columns. All values were turned into upper case to ensure consistency using <.str.upper()>.

Narrow Observation

The <.describe()> function was used to observe the statistics of all categorical and numerical data to determine whether there are impossible values, missing values, outliers.

2.3.4 Impossible Values and Sanity Checks (Program: Section 1.3.5)

The data gathered from <.describe()> function was used to compare with the description for each column from the assignment brief. The following conclusions were made for impossible values:

- There are significant low (-19) and high (280) value within the "Age" column.
- There is significant high value in "Personal Foul" (PF) (228). As when this number is divided by the highest number in "Games" (G) (38), the result is 6 (higher if he played less games) which is equal to the "disqualified" criteria number. In this case, this number is unlikely possible because a player cannot foul six times in every game he played.
- The "Total Points" or "PTS" need to be less than 2000, however, its maximum value is 28800.

The program continues by deeply checking all the columns and compare it with the requirement for them in the assignment brief. Only the relevant ones with impossible values will be mentioned in this report.

2.3.4.1 *"Age"*

Through research, the oldest players currently playing in the NBA is Udonis Haslem at the age of 40 (Green 2021). And the minimum age for a NBA basketball player is 19 (Mccann 2019). Using these information, two impossible values (Anthony Gill – 280 years old – and Killian Hayes – 19 years old) were found. Their actual ages were determined and changed to be 28 and 19 years old, respectively.

2.3.4.2 *"Personal Fouls"*

According to the project brief, an NBA player can have six personal fouls per game. However, the data of Ivica Zubac shows that he has six personal fouls every game he played, which is very unlikely.

This data was cross-checked with the information provided by the NBA official stats (Ivica Zubac committed 2.6 fouls/game in 2020 – 2021) (NBA Stats 2021), which results in this data being fixed to 99 (2.6 * 38 games). The <.loc()> function was used to locate the data of Ivica Zubac and replace it.

2.3.4.3 *"PTS"*

The "Total Points" of a player must be less than 2000, however, there are two data (Jaylen Adams – 20000 points and LaMarcus Aldridge – 28800 points) which are higher than this. These two data were addressed using the <.iloc()> function, and were re-calculated using the formula:

$$PTS = FT + 2P * 2 + 3P * 3$$

The new data for these two players are 2 points for Jaylen Adams and 288 points for LaMarcus Aldridge.

2.3.5 Missing Values (Program: Section 1.3.6)

In previous section, missing values have been identified in columns: "FG%", "3P%", "2P%", and "FT%". To clearly observe which rows, do the missing values appeared on, the program used the <.isnull()> function. It was concluded that the missing values, in these columns, were caused due to inputter attempt to divide a number by 0 using the formula:

$$Percentage = \frac{Succeeds}{Attemps}$$

The approach to fixing these missing values were to impute a static value of 0. The reason behind this decision was that the numbers of successful attempt, for players with attempts lower than 3, are very low, and often equal to 0. It is acknowledged that, using such a static value could lead to false estimations in future modelling process, however, in this case, they are very minimal.

2.3.6 Outliers (Program: Section 1.3.7)

Multiple histograms were plotted to observe the outliers in the numerical data. From observation, there are certain data which can be considered outliers in columns: “2P%, “3P”, “3P%”, “3PA”, “BLK” and “FG%”. However, as not all outliers are bad data and some of them can even be possible values (Ex: a player who is very confident in attempting 3 points can lead to him being an outlier in “3PA” column, but he is also a very important data which should be included), close observation into these data was conducted and proved that these data are possible. It was come to the decision that the program would move forward without dropping these.

Further investigation and analysis are recommended to double-checked on the outliers in future stages.

2.3.7 Saving Cleaned File

The Dataframe was double-checked using the <.info()> function, before, being saved into the new csv file named “cleaned_NBA_players_stats.csv”.

3 Data Exploration

3.1 Explore the composition of the total points of the top 5 players with the most points.

3.1.1 Identifying the top 5 players with the most points (Program: Section 2.1.1)

Before exploring and visualizing the total points of the top 5 players with the most points, these five players must first be identified.

The dataset that was provided, however, contains data of players who have played for more than one team in the season, so it is crucial to summarize the total points of these players. The total points (“PTS”) of these players were calculated and were compared with the score of the players who only play for one team in the whole season. The result shows that the top 5 scoring players and their PTS are:

Table 1: Total Points of the top 5 players with the most points.

Players	PTS (from high to low)
James Harden	1568
Bradley Beal	1053
Stephen Curry	1039
Giannis Antetokounmpo	1015
Damian Lillard	1013

These are the players that will be analysed in this section.

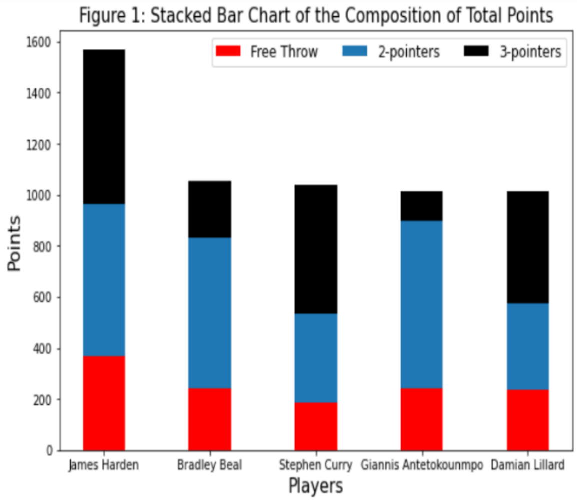


Figure 1: Stacked Bar Chart of the Composition of Total Points

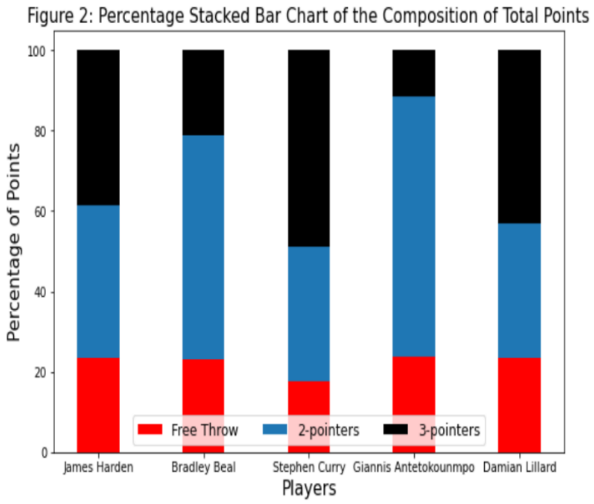


Figure 2: Percentage Stacked Bar Chart of the Composition of Total Points

3.1.2 Visualising (Program: Section 2.1.3)

Stacked bar charts were used to analyse the composition of the total points of the 5 players.

Figure 1 is a stacked bar chart, that presents the total points of the top 5 players with the most points during NBA season 2020 – 2021, across two variables: the total points and the composition of the total points. The primary variable, which is the total points, shows that James Harden is the player that scored the most and Damian Lillard is the one who scored the least. Figure 1 also depicts that James Harden scored the most out of the 5 players from three points and free throw, whereas Giannis Antetokounmpo scored the most from two points.

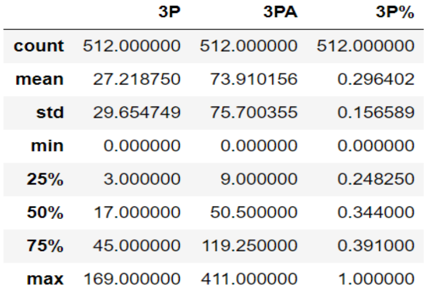
To further explored these compositions, a percentage stacked bar chart of the total points was plotted in figure 2. Percentage stacked bar chart allows better analysis on the second variable, which is the composition of the total points (Yi 2019). From figure 2, it was deduced that James Harden, who scored the most, is also the one having the most well-rounded composition– 23% of his score comes from free throw, 38% comes from 2 points and 39% from 3 points. The other 4 players were divided into two categories, those that scored more from 2 points and those that scored more from 3 points. Bradley Beal and Giannis Antetokounmpo lie in the first category; their percentages scoring from 2 points are 56% and 65%. These numbers are superior to both their 3 points and free throw. On the other hand, Stephen Curry and Damian Lillard lies in the second category; they scored more from 3 points with 49% from Stephen Curry and 43.2% from Damian Lillard. The figure also reveals that free throw contributes the least to the composition of the total points – around 20% for these 5 players – which is understandable because free throw in basketball only accounted for 1 point.

3.2 Data visualization to identify the errors occurred in “3P”, “3PA”, and “3P%” columns.

3.2.1 Exploring Descriptive Statistics (Program: Section 2.2.1)

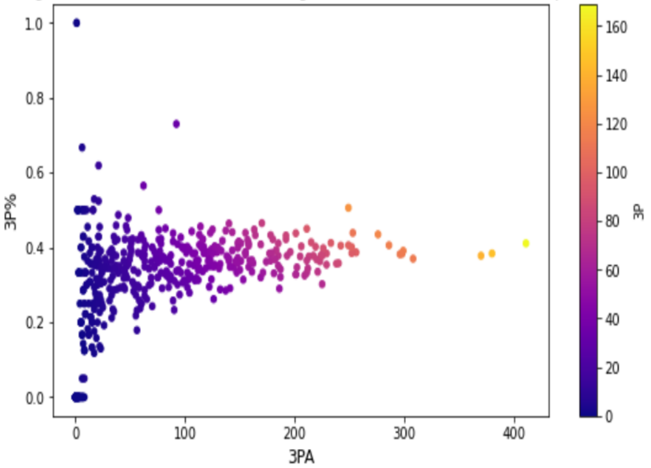
Before getting into the visualization process and identify the errors, the descriptive statistics of the 3 columns were re-summarized using the <.describe()> function. The results are showed in the table 2.

Table 2: Descriptive Statistics of columns: "3P", "3PA", "3P%".



	3P	3PA	3P%
count	512.000000	512.000000	512.000000
mean	27.218750	73.910156	0.296402
std	29.654749	75.700355	0.156589
min	0.000000	0.000000	0.000000
25%	3.000000	9.000000	0.248250
50%	17.000000	50.500000	0.344000
75%	45.000000	119.250000	0.391000
max	169.000000	411.000000	1.000000

Figure 3: 3-Point Field Goals Percentage versus 3-Point Field Goal Attempts



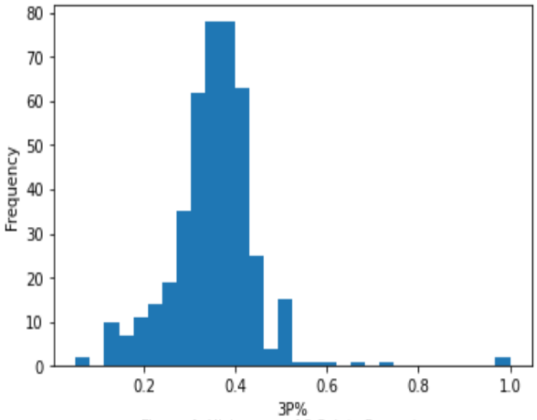


Figure 4: Histogram of 3 Points Percentage

The descriptive statistics, summarized in table 2, provide a measure of the central tendency and dispersion of the data in the three columns; it can also be used to show where outliers might be located. These numbers will be used in later section.

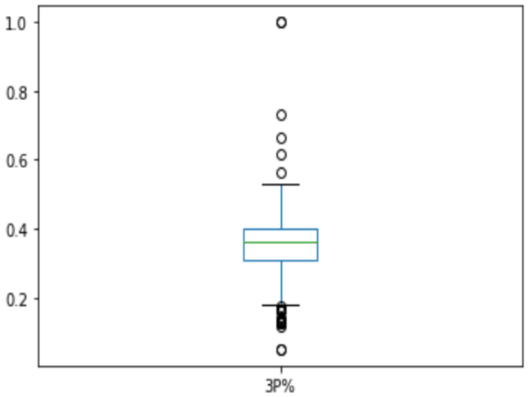


Figure 5: Boxplot of 3P% higher than 0%

3.2.2 Visualizing (Program: Section 2.2.2 and 2.2.3)

To identify the errors, it is also very crucial to understand the correlation between these three variables. For this reason, figure 3, showing the relationship between the percentage and the attempts, is used; the graph also included the number of succeeded points as the 3rd variables. The chart demonstrates that the relationship between 3PA and 3P% has a horizontally bimodal shape with two peaks, one at 0% and the other one at around 40% this is compatible with data in table 2. When analysing a figure with two peaks like this, it is crucial to separate the data into two systems and observe bot (PQsystems n.d) , however, in this case, it is decided that data with 0% can be ignored – we have gone through analysing these data when filling in missing values in section 2.3.5. The data of 3P and 3PA also have been checked in previous section testing with FG and FGA as well as PTS.

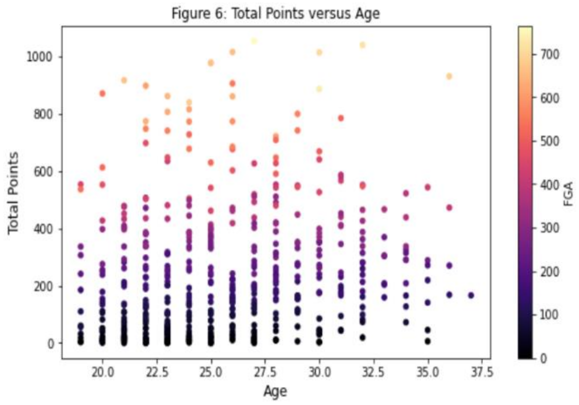
For this reason, figures 4 and 5 were plotted for only data in 3P% column, whom percentage is higher than 0%. From these 2 graphs and table 2 above, the data located at 5% level and higher than 50% level were put under investigation as these were considered outliers. The result of the investigation identified 3 errors occurred with the data of Marvin Bagley III, who has 73% percentage succeed rate instead of 37%, in the data of Bam Adebayo and Jarrett Allen, who has 5% percentage succeed rate instead of 33%. These values were fixed in section 2.2.4 of the program.

3.3 Analyse the relationship between the player's total points and the rest features.

Including in this report is the relationship between the total points with three other features, please refer to the actual program – section 2.3 – for more.

3.3.1 Hypothesis 1

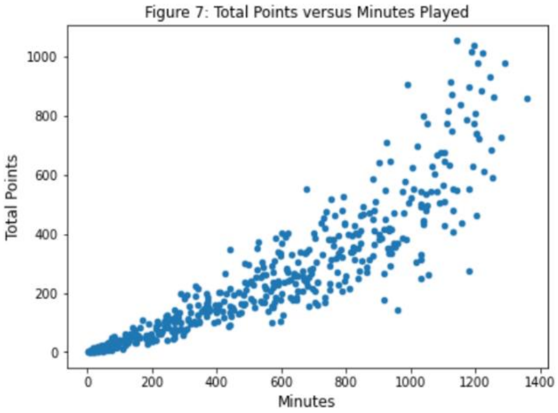
H0 – Total points depend on the ages of the player. H1 – Total points do not depend on the ages of the player.



Conclusion: Figure 6 shows that there is barely any relationship between the total points a player can scored during one season and the age of the player. Although, we can see a downhill trend of the total points as the age keep on going up, there are many outliers, and the data does not provide much information. *These reasons lead to the rejection of the null hypothesis and that total points do not depend on the ages of the player.*

3.3.2 Hypothesis 2

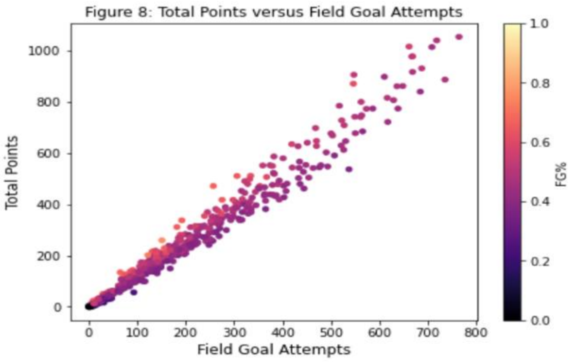
H0 – Total points depend on the Minutes Played. H1 – Total points do not depend on the Minutes Played.



Conclusion: Figure 7 shows a positive strong relationship between the total points and the minutes a player played on the field. The relationship gets weaker as the minutes going up, but we can still see a very strong uphill trend. *For these reasons, it was concluded that the null hypothesis is accepted and that total points depend on the minutes a player played on the field.*

3.3.3 Hypothesis 3

H0 – Total points depend on the Field Goal Attempts. H1 – Total points do not depend on the Field Goal Attempts.



Conclusion: Figure 8 shows an even stronger positive relationship, between the total points and the number of field goal attempts of a player played on the field. This can be understood as the player who attempt to score the most are the one who scored the most. *The null hypothesis is accepted in this case.*

4 Reference List

Jack Green 2021, 'The oldest NBA players of 2021 and all-time', blog post, 25 January, viewed 17 April 2021, https://blog.betway.com/basketball/the-oldest-nba-players-current-and-all-time/#:~:text=Udonis%20Haslem%20is%20the%20oldest,in%20the%202020%2D21%20season.

Mccann, M 2019, 'Examining What a Change to the NBA's One-and-done Rule Could Mean for All Involved', Sports Illustrated, viewed 17 April 2021, https://www.si.com/nba/2019/03/03/legal-analysis-change-age-eligibility-rule-one-and-done

NBA News 2021., Ivica Zubac stat, NBA, viewed 17 April 2021, <(https://www.nba.com/stats/player/1627826/>

PQsystems n.d., *Histogram: study the shape*, PQsystems, viewed 17 April 2021, <https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/interpretation/histogram_shape.php>

Yi, M 2019, *A Complete Guide to Stacked Bar Chart*, Chartio, viewed 17 April 2021, <https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/>