Student Name: Syed Nazmul Kabir Student ID: 3874060

Data Preparation

1.1: Data Import:

Here a detailed explanation is provided about the steps that are followed while dealing with the potential issues/errors.

- At first the required libraries are imported. Three libraries are used: pandas, matplotlib and string. ‘NBA_palyers_starts.csv’ is imported.
- ‘Rk’ is a random number and cannot be considered as a feature. So, to keep separated, ‘Rk’ column is converted to index column.
- The size of the imported dataset is checked, it is 512. All columns are thoroughly checked and is ensured that the dataset is imported correctly.

1.2: Data Cleaning:

1.2.1: Error 1: Handling Missing Values/NaN:

- A thorough analysis on NaN values reveals that all NaN values are in numeric data.
- Those NaN are generated from zero division error.
- That means those NaN are created from the calculation of ‘zero divide by zero’ cases.
- So, those data can easily be considered as zero.
- Hence all NaN values are converted to zero value using fillna() method.

1.2.2: Error 2: Checking Name Column:

- At first an investigation is done on all values to find out any unnecessary white-space available in data using str.isspace() method from string library. There is none.
- The name structure is in Title format in the data. However, there are some discrepancies. To convert all name in Title format, str.title() is applied.
- –′ . and few other abnormal alphabets are being replaced appropriately.

1.2.3: Error 3: Sanity Check on ‘Age’ Variable:

- Applying value_counts () on ‘Age’ feature finds two impossible values: 280 and -19.
- Those are from typos.
- The data are corrected to 28 and 19 respectively and placed in dataset.
- Later, further checking is done for the confirmation of replacement of wrong values.
- Other numbers are all in between 22 to 37. So, those are valid numbers.

1.2.4: Error 4: ‘Pos’ Feature Contains Some Invalid Data:

- After removing white spaces and converting to upper cases, the value of ‘Pos’ variable is checked whether those are present in approved position list or not.
- Approved ‘Pos’ list is: ['PF', 'PG', 'C', 'SG', 'SF', 'PG-SG', 'SF-PF'].
- A for loop on unique values of ‘Pos’ feature reveals that three values are not in the list. These are: ‘SGA’, ‘SF.’,’PFA’.
- Again these are generated from typos. Corresponding appropriate values should be ‘SG’, ‘SF’,’PF’. The corrected values are placed in appropriate place of the data set.
- Further checking is done for confirming the replacement of wrong data.

1.2.5: Error 5: Abnormal Data in ‘Tm’ Feature:

- The following code reveals that some values have white spaces in the names.
  [ateam for ateam in clean_nbaPlayers['Tm'].unique() if ateam not in team_list]
- Here team_list is the approved team list provided in business statement.
- Value ‘HOU’ is corrected with corresponding data from team_list.
- Some values are not capitalized whereas all data of team_list is in capital letter.
- So, all values are capitalized.
- Final checking is done to confirm that wrong data is replaced by the corrected data.

1.2.6: Checking on data duplication finds no duplication in the whole data-set.
1.2.7: Sanity test on ‘G’ column does not find any abnormalities. All values are in 0 to 82
1.2.8: An investigation on all numeric data is done to reveal whether any non-numeric data is there or not. The experiment finds that all numeric feature contains only numeric data.

1.2.9: Error 6: ‘PTS’ Feature Contains Some Errors:

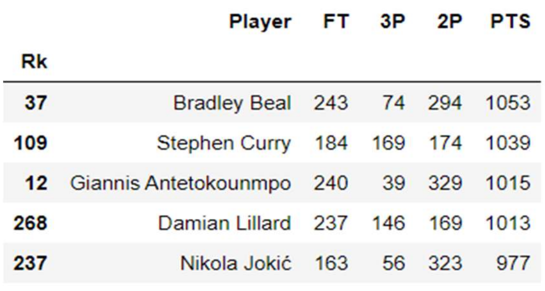
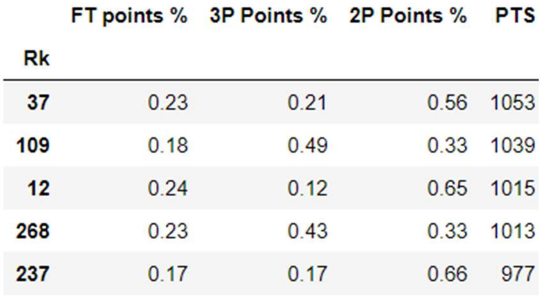
- According to business statement a player’s point can not be more than 2000
- Sanity check on ‘PTS’ finds that Rk-2 and Rk-5 contains very high value for PTS.
- Again these are type errors.
- The errors are handled with the solution generated from task 2.1 where the total points composition is identified.
- Using the composition, the values for Rk-2 and Rk-5 is corrected in the data set.

Data Exploration

Task 2.1

According to game rules, each 3-point field goal contributes 3 points, 2-point field goal contributes 2 points and a free through contributes 1 point to the total points. By adding the values of these three columns with corresponding ratios gives output of the values exactly like the PTS column. So, the composition of 3P, 2P and FT is 3:2:1.

Table1: Number of Scores and Percentages of FT,2P, 3P that makes Players Total points.

Scores of FT,3P,2P and Total Points	Fraction of FT,2P, 3P in Total Points
	

Rk	Player	FT	3P	2P	PTS
37	Bradley Beal	243	74	294	1053
109	Stephen Curry	184	169	174	1039
12	Giannis Antetokounmpo	240	39	329	1015
268	Damian Lillard	237	146	169	1013
237	Nikola Jokić	163	56	323	977

Rk	FT points %	3P Points %	2P Points %	PTS
37	0.23	0.21	0.56	1053
109	0.18	0.49	0.33	1039
12	0.24	0.12	0.65	1015
268	0.23	0.43	0.33	1013
237	0.17	0.17	0.66	977

The topmost point holder, Bradley Beal scores the greatest number of free throws among the top five points-holder. But he holds maximum points (56%) from 2P goals. However, Bradley scores less than half of the Stephen’s 3P goal numbers. Stephen scores maximum points from 3P goal (49%) in his total points. Though Stephen is in second position, he shows most al rounding capabilities as he scores almost same number of FT,3P and 2P. On the contrary, Giannis is adept in scoring one type of goal. He is the champion in scoring 2P goals (65% of his total points), but he has shown real weakness in scoring 3P goals (only 39) among the top five point-holders.

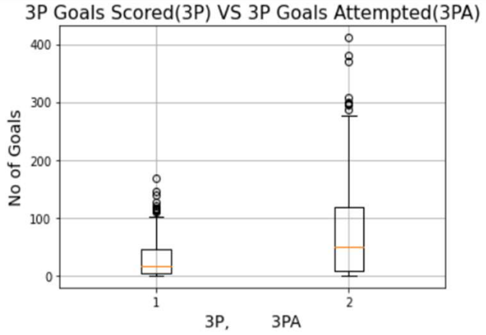
Damian is the fourth highest point holder who was lagging by only two points of Giannis total points. His strength in making free throws and scoring 3Ps is significant. But he is the lowest scorer of 2P points. Nikola scores second highest 2P goals which comprises 66% of his total points. He is the second lowest scorer of 3Ps goals as well. His capacity in making free throughs is lowest (163) among top five.

Task 2.2

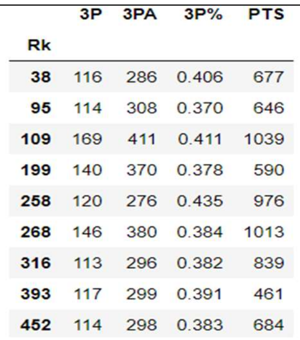
As there is possibility of errors in 3P,3PA and 3P% column, a thorough analysis is performed based on evaluating outliers, searching impossible values:

- During data cleaning steps, it is confirmed that all numeric columns have only numeric data. So, data type variation is not the problem.
- Initially there were some NaN values. Those are also handled in data cleaning stages.
- Availability of outliers in 3P and 3PA can be the source of impossible values. So, outliers are detected in steps 2.2.1 and 2.2.2. However, these values are real performances. How? Some players attempt more times to score goals and some players score more goals compared to others. So those outliers should remain in data for further analysis.

BoxPlot showing Outliers



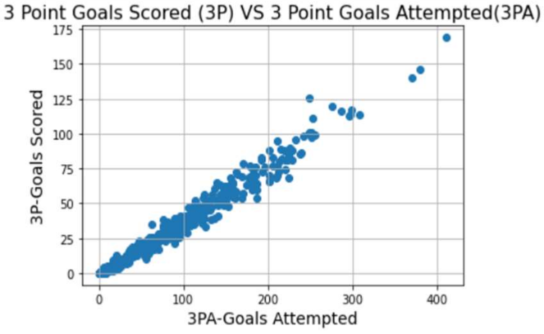
The Outlier Data



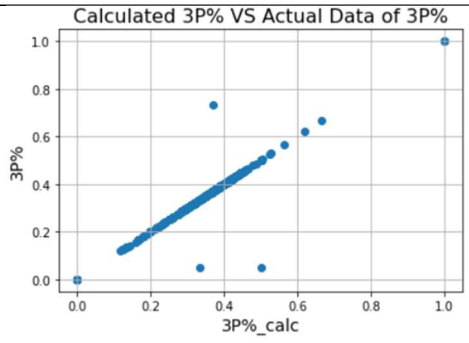
Rk	3P	3PA	3P%	PTS
38	116	286	0.406	677
95	114	308	0.370	646
109	169	411	0.411	1039
199	140	370	0.378	590
258	120	276	0.435	976
268	146	380	0.384	1013
316	113	296	0.382	839
393	117	299	0.391	461
452	114	298	0.383	684

- Another important point is that 3P column values cannot be more than 3PA column values. In 2.2.3 of ipynb file, a scatter plot confirms that values in those columns are complying with this rule. So, error is not found in 3P and 3PA columns.

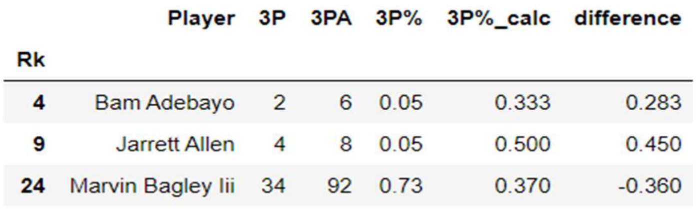
Figure showing no abnormalities



Three points are abnormal



- Let’s check now the values of 3P% columns. These are the compound values. Division of values of 3P by the values of 3PA results 3P%. So, a calculation is being performed and a new feature ‘3P%_calc’ is created.
- Now the values of 3P%_calc and 3P% should be exactly same. A scatter plot in 2.2.4 among those features discovers that three data points are not matching with rest of the data positions. So those three data are troublesome.
-

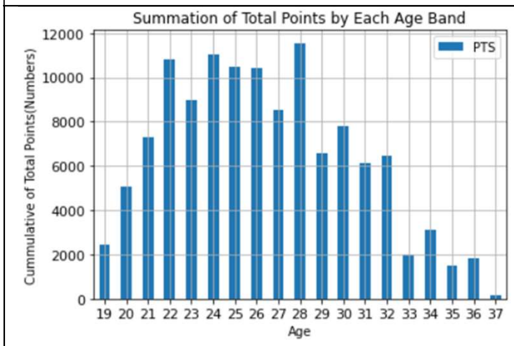
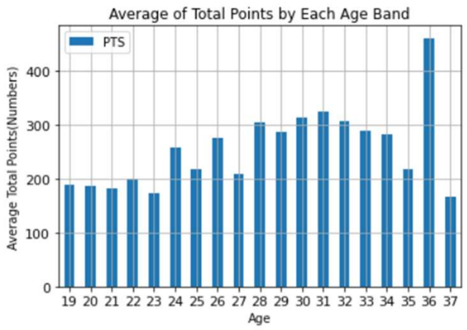


Rk	Player	3P	3PA	3P%	3P%_calc	difference
4	Bam Adebayo	2	6	0.05	0.333	0.283
9	Jarrett Allen	4	8	0.05	0.500	0.450
24	Marvin Bagley Iii	34	92	0.73	0.370	-0.360

- Finally, the three data is revealed in step 2.2.5 of ipynb file and they are the data of the player named: ‘Bam Adebayo’, ’Jarrett Allen’ and ‘Marvin Bagley’. Those data are corrected with the calculated values. This is the way the errors in 3P% column is identified and handled.

Task 2.3

2.3.1: Age vs Players Cumulative Total Points: We know age is an important factor in performance of any sports. So, to make insightful analysis, two bar plots are created to find out real impact of age on PTS. The first bar-plot between total cumulative points and age clearly shows that only the players between the age of 22 to 28 score more than 8000 cumulative points. One of the reasons is that maximum players are in the age range of 22 to 28 (see section 1.2.3 of IPYNB file). Not too many players are scoring who are under 20 or over 32.

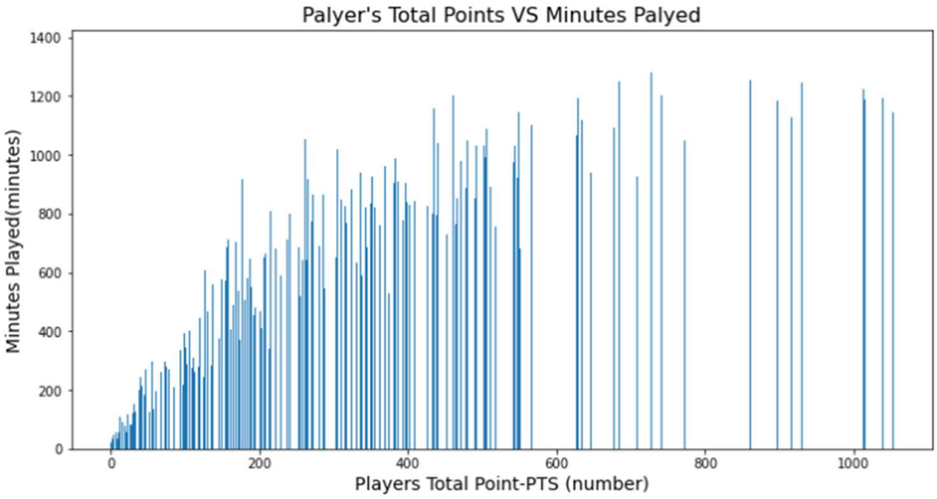
Cummulative Total Points(PTS) by Age	**Average Ttoal Points(PTS) by Age**
	

Age vs Players Average Total Points: However, when considering the effectiveness, the average scores of the players in age range of 28 to 34 is higher compared to that of other age ranges with exception of 36 years old-aged-players. This is caused because, only four players are playing at that old-age and among them one is high scorer.

Section 2.1.1 of IPYNB file also shows that top five scorers are all aged >= 25 years. So certain aged players performance is better compared to the others and 28 years of old players are most effective considering both cases (highest total numbers, higher average scores.)

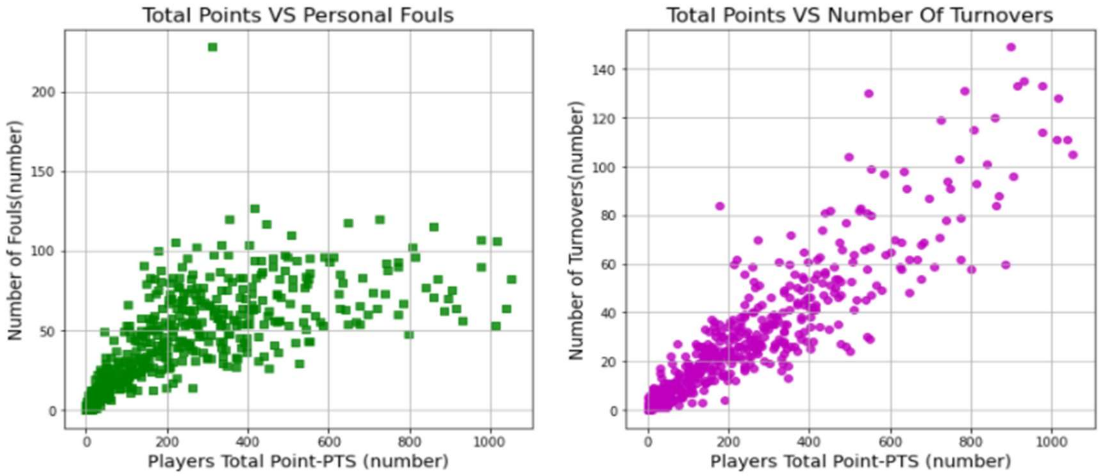
2.3.2: Players Total Points (PTS) by Minutes Played (MP):

A bar-plot of PTS by MP feature shows that players points are directly proportional to the minutes they played up to the range of 400 points. Beyond that, player's performance does not vary with minutes and is almost dependent on individual skill rather than playing more minutes.



However, every high scorer (who scores more that 600 points) plays more than 800 minutes at least. So player should have minimum level of stamina to score more than 600 points.

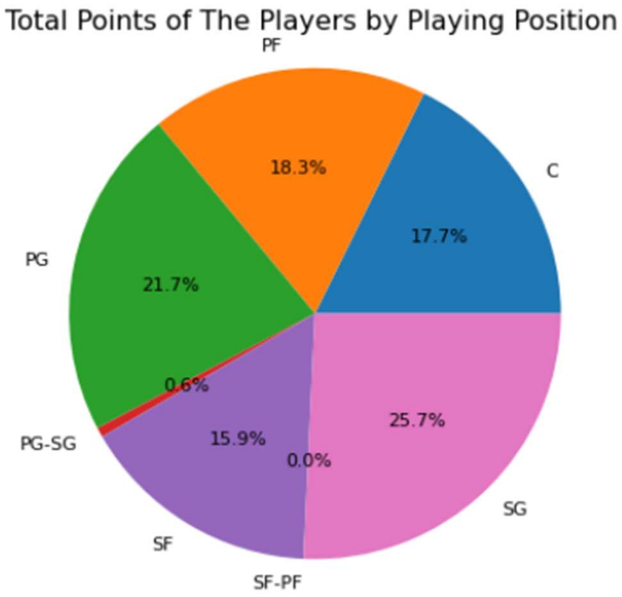
2.3.3: Analysing player's total points (PTS) with personal fouls(PF) and number of turnovers (TOV):



The above figure clearly shows that all players foul less than 125 except one case where a player commits more than 220 fouls. This is an outlier. The plot is showing that all higher point holders (>600) commit minimum fifty number of fouls and players who score less than hundred commit less than 50 Fouls.

On the other hand, the relationship of number of turnovers with player's point is almost linear. Every player who scores above 1000 points make 100 turnovers in a season. So having capability to make higher number of turnovers is a characteristics of top point holders in this game.

2.3.4: Analysing player's total points (PTS) with players position (Pos):



The pie-diagram is showing it clearly that players who play in the position of 'PG-SG' and 'SF-PF' can hardly score goals. 'SG' positioned players score the greatest number of goals (25.7%) followed by 'PG' (21.7%) position holders. Player's total points do not vary that much (less than 2.5%) by the position of 'PF','C','SF'.

REFERENCES:

Boschetti, A. & Massaron, L. (2016) *Python Data Science Essentials* (2nd ed.). Packt Publishing.

matplotlib. (2021). *Pyplot tutorial: Intro to pyplot*
https://matplotlib.org/stable/tutorials/introductory/pyplot.html#sphx-glr-tutorials-introductory-pyplot-py

RMIT Canvas. (2021). *Practical Data Science with Python COSC2670: Modules: Week1,2,3*
https://rmit.instructure.com/courses/79792/modules