Student Name: Rahul Kumar Student ID: S3802846

Data Preparation

I read dataset from csv file into data frame and then analysed each column for potential errors as per description provided in assignment details.
For e.g. I checked that FG value for any player should not be greater than FGA since FG represents goals scored and FGA represents field goals attempts and actual goals cannot be greater than attempts.
I performed similar checks for each column as per column description and found few errors, details of which are provided below:

Error 1: Whitespaces:

There were extra whitespaces in the end in column values 'Pos' and 'Tm'. Removed these white spaces since extra spaces can result same values into 2 distinct records.

Error 2: Invalid values/Typos:

In column description, 'Pos' and 'Tm' columns are provided with list of valid values which means values in these columns should be within respective list. There were few records in these 2 columns which were not matching with provided list and looks like typo since it includes error like 'SF' entered with one dot in the end as 'SF.' in Pos column or 'NYK' entered with y in small as 'NyK' in Tm column.

Steps followed to correct these invalid values:

i) Returned all distinct values present in these columns.
ii) Entered all valid values in a variable.
iii) Fetched all invalid records by comparing column values with variable.
iv) Replace all invalid values obtained in step (iii) with valid values.

Error 3: Impossible values:

Found 2 impossible values, -19 and 280 in column Age by selecting unique values. These values are impossible since age cannot be negative and 280 is not a practically possible age.

Steps followed to correct these impossible values:

i) Replaced -19 with 19 since negative sign seems to be a typo and 19 is a valid age.
ii) Created a NaN (not a number value) and assigned it to player whose age is 280.
iii) Calculated mean of Age for all players.
iv) Replaced NaN value with mean since actual age for this player is not known.

Error 4: Missing/NaN values:

Searched NaN/missing values in all columns and found null values in 4 columns – FG%, 3P%, 2P% and FT%.
Replaced these NaN values with 0 since these columns are calculated as FG/FGA, 3P/3PA, 2P/2PA and FT/FTA respectively and as per data, these columns are populated with 0 for rows having Nan values in percentage column.

Error 5: Incorrect Values:

In column description, it is mentioned that PTS value must less than 2000.
Found PTS value greater than 2000 for 2 players with rank 2 and 5.

Steps followed to correct these incorrect values:

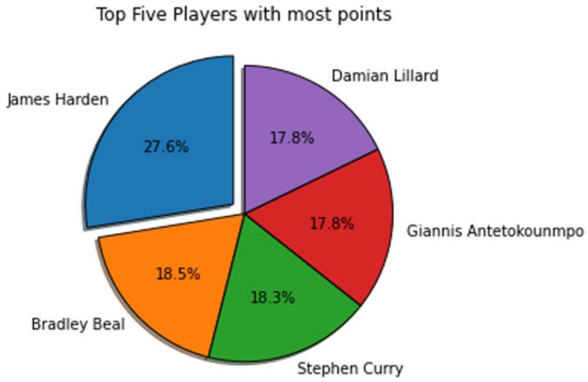
i) It is mentioned that PTS is the sum of 2P, 3P and FT fields where 2P is worth 2 points, 3P is worth 3 points and FT is worth 1 point.
ii) Calculated PTS using formula (2P*2 + 3P*3 + FT) and stored in a variable 'total_points'
iii) Populated PTS for Rank 2 and 5 players for which PTS value was incorrect with calculated value.

Data Exploration

Task 2.1

I have analysed the composition of the total points of the top five players with the most points using table and pie chart.

Player	Points
James Harden	1568
Bradley Beal	1053
Stephen Curry	1039
Giannis Antetokounmpo	1015
Damian Lillard	1013
Total	**5688**



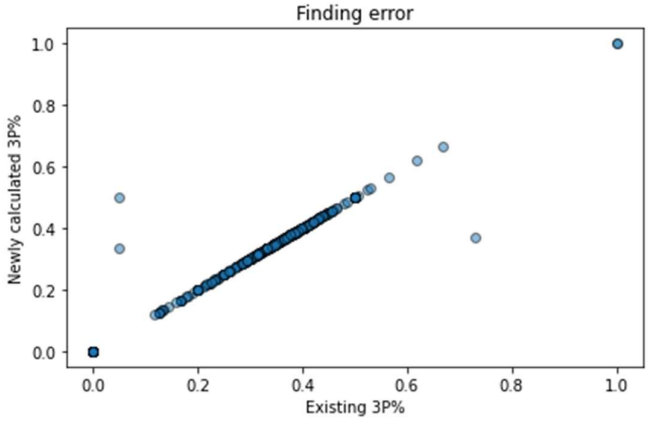
The above table shows the top five players with the most points which is calculated by adding PTS for each player for all teams they have played for and the pie chart explains composition of these points at player level.
Looking at the pie chart, it is clear that James Harden is the top player which has scored 27.6% of the total points. Furthermore, there is a difference of only 0.2% between second and third position occupied by Bradley Beal with 18.5% and Stephen Curry with 18.3% respectively.
Surprisingly, Damilian Lillard and Giannis Antetokounmpo have same contribution which is 17.8% of the total points scored by these five players.
Overall, it is clear that James Harden which plays for multiple teams is the top scorer among all the players.

Task 2.2

To explore errors in 3P, 3PA and 3P% columns, I have used scatter plot.

Steps followed to correct to visualize and correct errors:

i) First, I created a new column 'new_3P%' for calculating correct values for 3P% which is calculated as 3P/3PA.
ii) Then I plotted a graph between existing column '3P%' and new column 'new_3P% to compare and find if any value in existing column is incorrectly calculated.

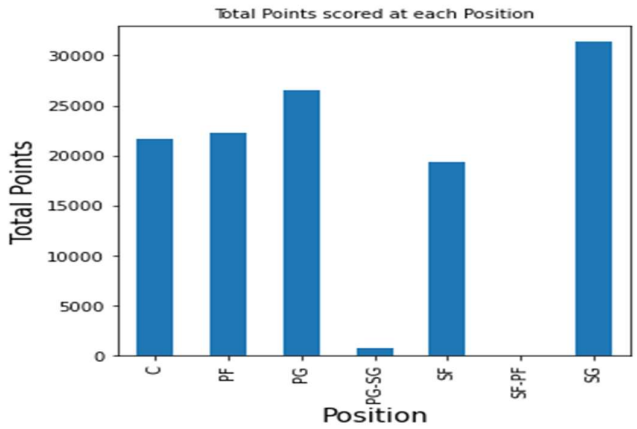


iii) By looking at the graph, it can be deduced that there are errors in existing 3P% column which is represented by dots which are not in a straight line (dots not in a straight line means that values in 2 columns are not matching).
iv) Compared all values in 2 columns using code and found 43 mismatches or errors.
v) Corrected these errors by replacing existing 3P% column values with newly calculated column new_3P%.

Task 2.3

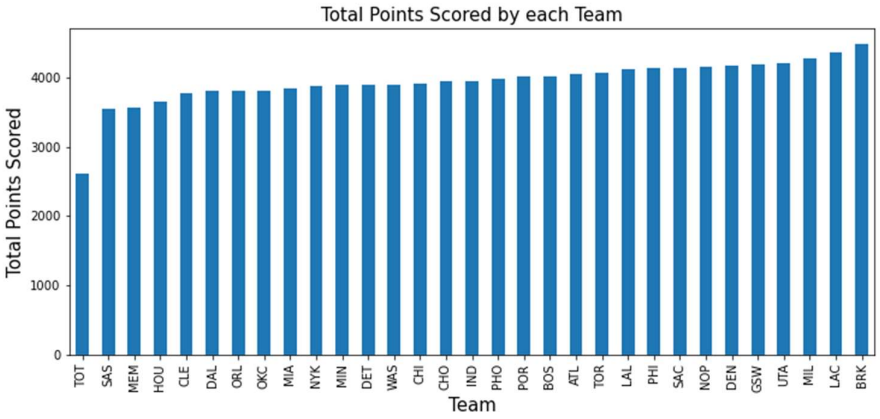
For this task, I have used Position (POS), Team (Tm) and Player attributes.

Total Points Vs Position



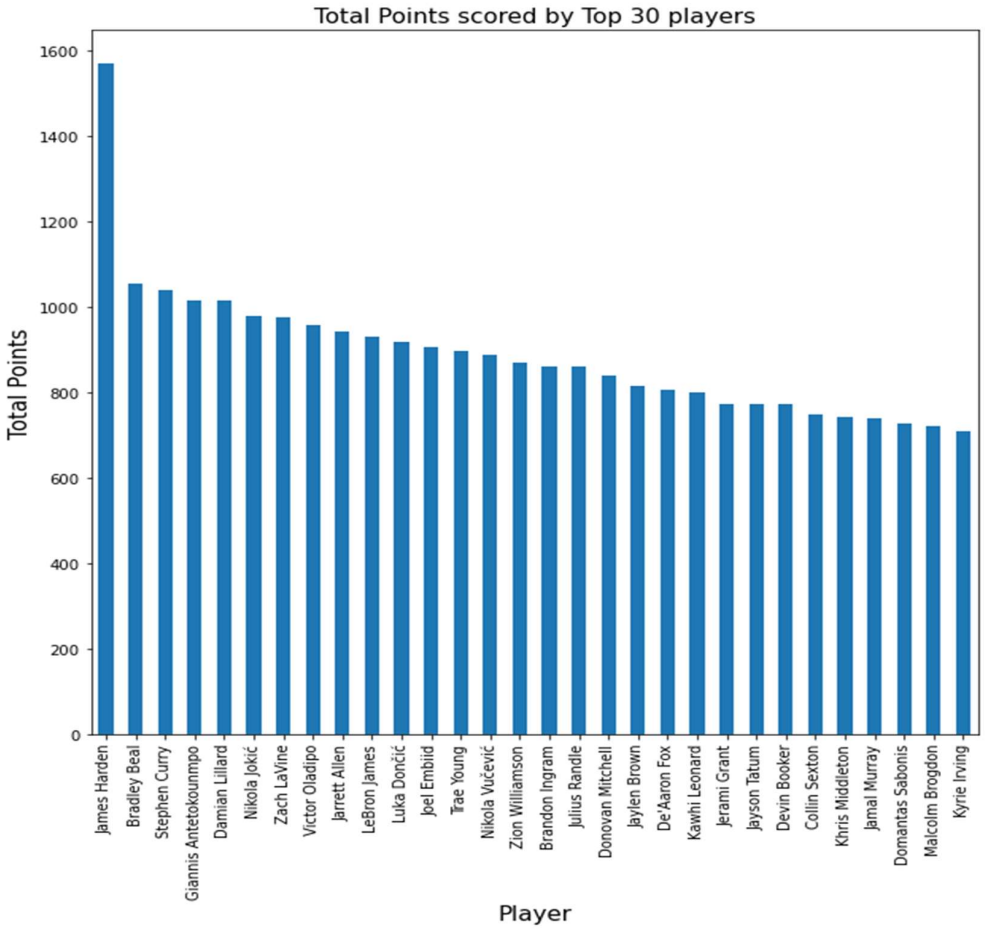
The above graph shows the total points scored at different positions.
It can be seen that most of the points are scored at position SG while none of the points are scored at position SF-PF.
More than 25000 points are scored at position PG which is second highest while C and PF have almost same number of points scored at these positions. Furthermore, points scored at SF position is nearly 20000 while PG-SG is standing at nearly 1000 points.
In conclusion, SG is the position where most of the points are scored.

Total Points Vs Team



The bar graph shows the total points scored by all teams.
It is clearly evident from the graph that BRK team has scored highest points which is more than 4000 points. MIL and LAC at second and third positions have scored almost same points while other teams have minor difference in the points scored by them. TOT stands at last with only 2500+ points.
Overall, top team BRK has scored almost double points as compared to least scorer team TOT.

Total Points Vs Players



The above bar graph shows total points scored by top 30 players.
As observed, James Harden is the top scorer while Kyle Irving is the least scorer.
It can be deduced from the graph that James Harden, top scorer player has significant difference when compared to points scored by other players while other players have very little difference in points scored by them.

References:

- Boschettiand L. Massaron, **Python Data Science Essentials**, Chapters 2 and 6
- Pandas read_csv: http://pandas.pydata.org/pandas-docs/
- https://stackoverflow.com/questions/28706968/how-to-correct-typos-in-pandas-dataframe
- https://stackoverflow.com/questions/43256402/pandas-analogue-of-sqls-not-in-operator
- https://stackoverflow.com/questions/37366717/pandas-print-column-name-with-missing-values
- https://stackoverflow.com/questions/27842613/pandas-groupby-sort-within-groups
- https://matplotlib.org/