CMPN451: Big Data
Credit Hours System – Spring 2024

Team 19

Ammar Nasser	1190543
Anas Sherif	1190375
Marwan Mostafa	1190281
Omar Faheem	1190477

NBA-Shot Logs

Dataset link: https://www.kaggle.com/datasets/dansbecker/nba-shot-logs

Contents

Problem description

The problem under investigation revolves around understanding the shooting patterns and defensive performance in NBA games during the 2014-2015 season. This analysis aims to delve into the intricacies of shot logs, which contain detailed information about each shot attempted during games, including the shooter's position, shot distance, shot result (made or missed), and details about the closest defender.

Key questions we aim to address include:

1. **Shooting Efficiency**: How efficient were players in making their shots during the 2014-2015 NBA season? Were there particular trends or patterns observed in the shooting percentages based on factors such as shot distance, shot clock, or defender proximity?
2. **Defensive Impact**: What was the defensive impact on shooting percentages? Were there correlations between the proximity of the closest defender and the likelihood of a successful shot? Additionally, what insights can be gleaned from analyzing shot logs regarding defensive strategies employed by teams and individual players?
3. **Outlier Detection**: How prevalent were outliers in the dataset, and what impact did they have on the analysis? Are there specific outliers that significantly skew the data, and how should they be addressed to ensure the integrity of our findings?

By addressing these questions, we aim to gain valuable insights into the dynamics of NBA games during the specified season, shedding light on both offensive and defensive aspects of the sport. These insights will not only contribute to a deeper understanding of player performance but also have potential implications for strategic decision-making by coaches and team management.

Project Pipeline

Data loading

We read the data from the csv file “shot_logs.csv” into a pandas dataframe.

Data preprocessing

Then we do some preprocessing on the data to ensure the data has no Nulls and makes sense.

Data visualization

We visualize the dataset with our insights and get the correlation between each column.

Outliers detection

We detect the outliers in each column and plot them using the box plot.

Prediction models

We train four different models (logistic regression, random forest, KNN, gradient boosting).

Map-Reduce

We perform two map and reduce operations to get (the most scoring players in the last 30 seconds of the last quarter, top shot players who scored 3 points with 0 dribbles)

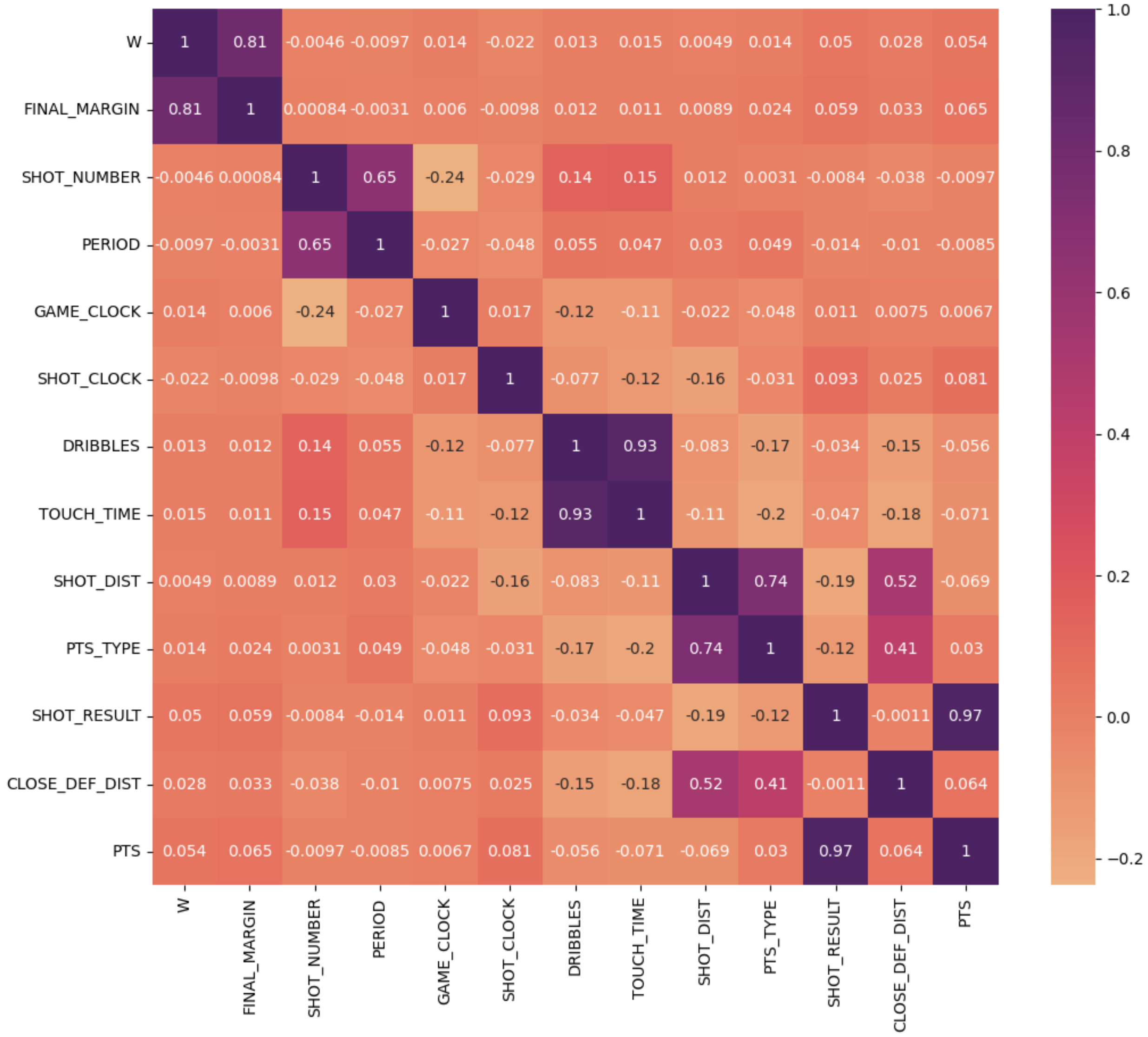
Analysis and solution of the problem

Data preprocessing

- **Handling Null values:** we first extract the columns that contain null values, then we fill them with forward filling.
- **Handling negative touch time:** we replace any negative value in the TOUCH_TIME column with the mean of all the positive values.
- **Removing unnecessary columns:** we remove the FGM column as it corresponds exactly to the column SHOT_RESULT, as the data set is per shot, so the FGM is either 0 if missed and 1 if made.
- **Label encoding:** we change the values from the W column to be 1 or 0 instead of W or L
- **Time conversion:** in the GAME_CLOCK, we calculate the time of the game in seconds instead of the format mm:ss.
- **Handling outliers (optional):** we calculate the lower and upper bound for the outliers in each column then remove that column and return the cleaned data.

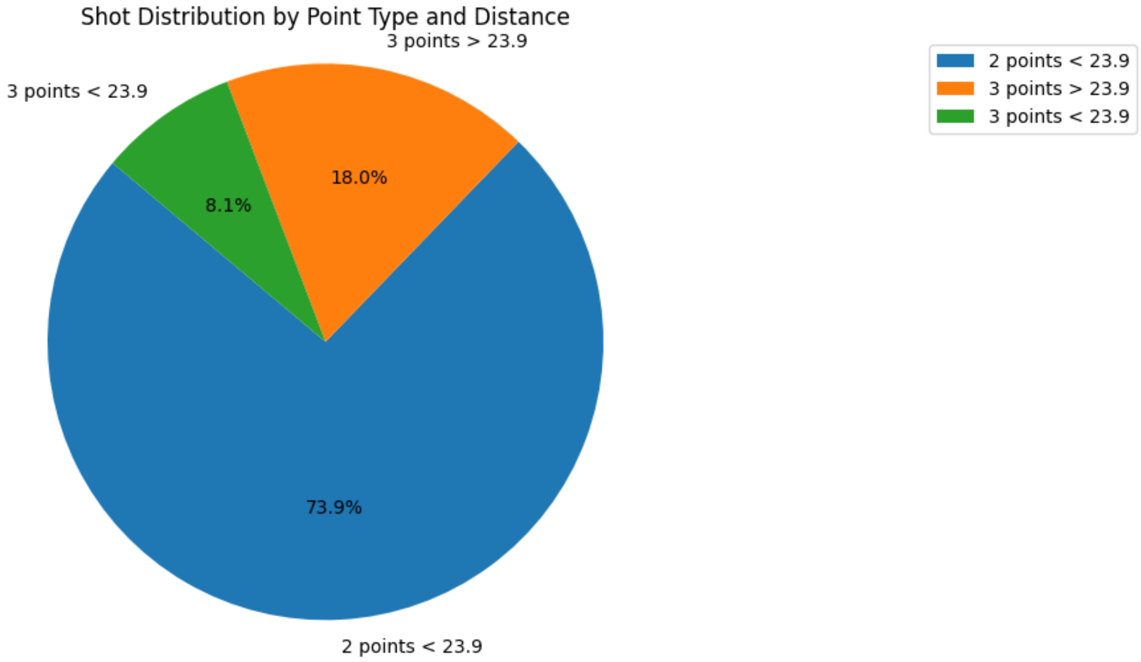
Visualization

Correlation heatmap:



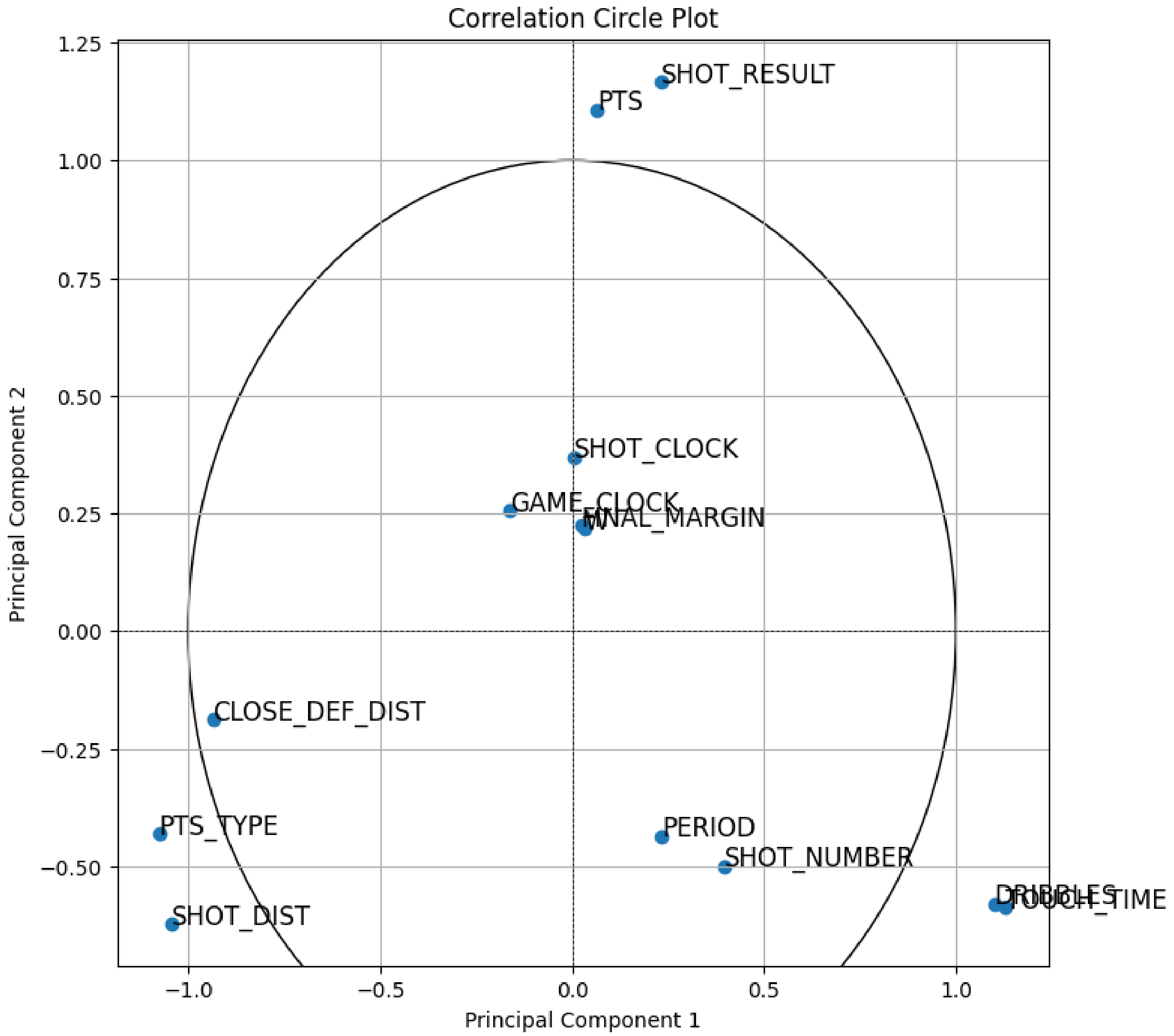
The heatmap uses a color scheme to show the strength of the correlation between two features, where a darker color indicates a weaker correlation.

Interesting insight



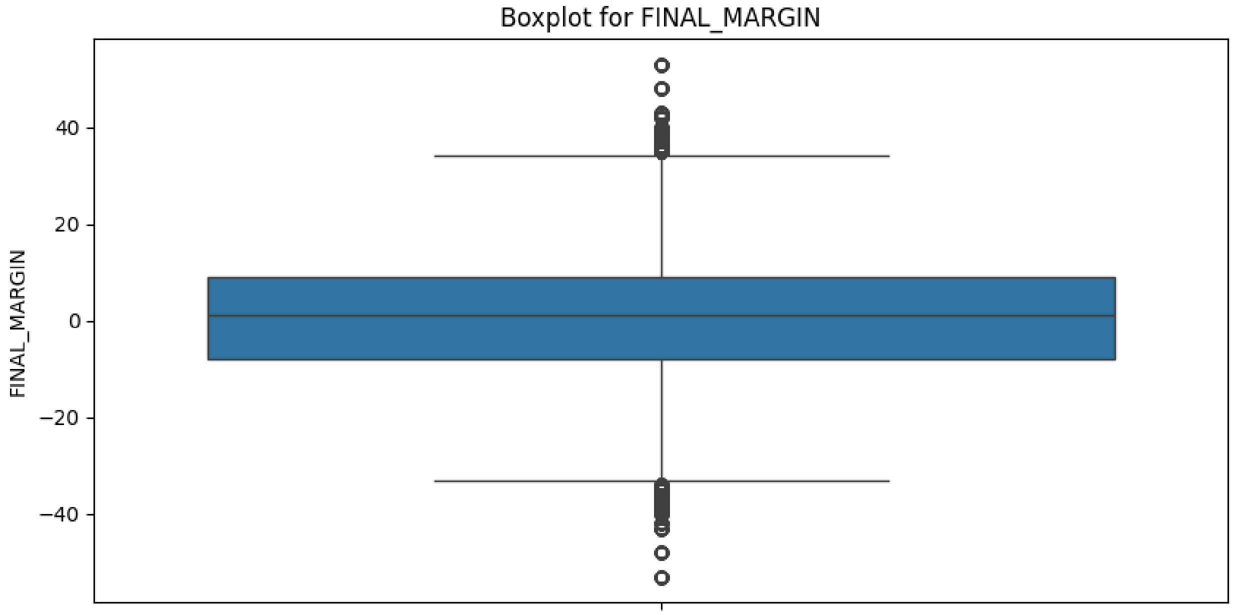
In basketball, shots made from over 23.9 yards count as 3 points. However, if a player is fouled attempting a shot and scores, they get an extra free throw. So, if a player is fouled near the 23.9-yard mark and scores, they earn 3 points (2 points for the basket plus 1 point for the free throw). This can explain why about 8% of shots made from less than 23.9 yards are recorded as 3 points.

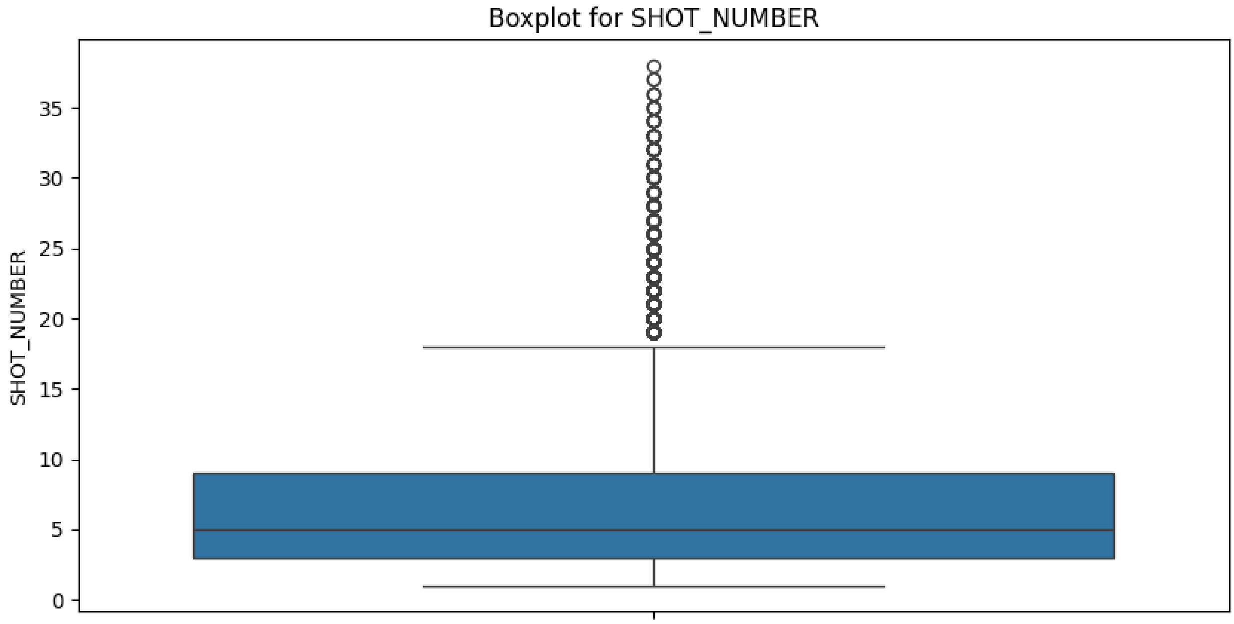
Correlation circle plot

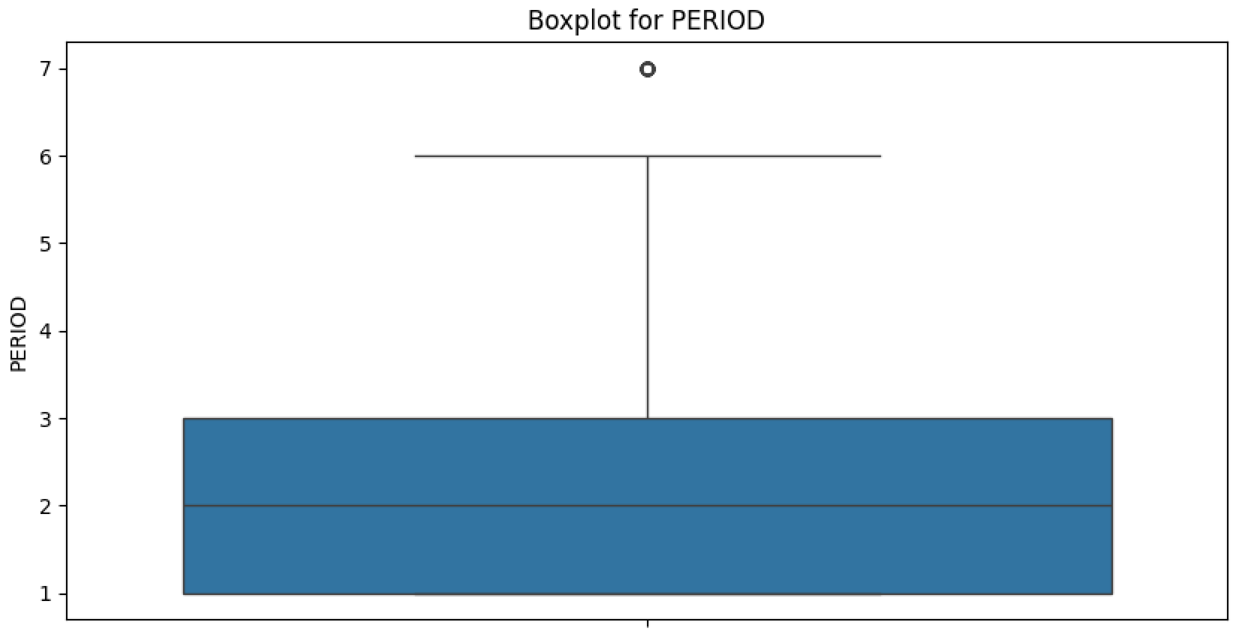


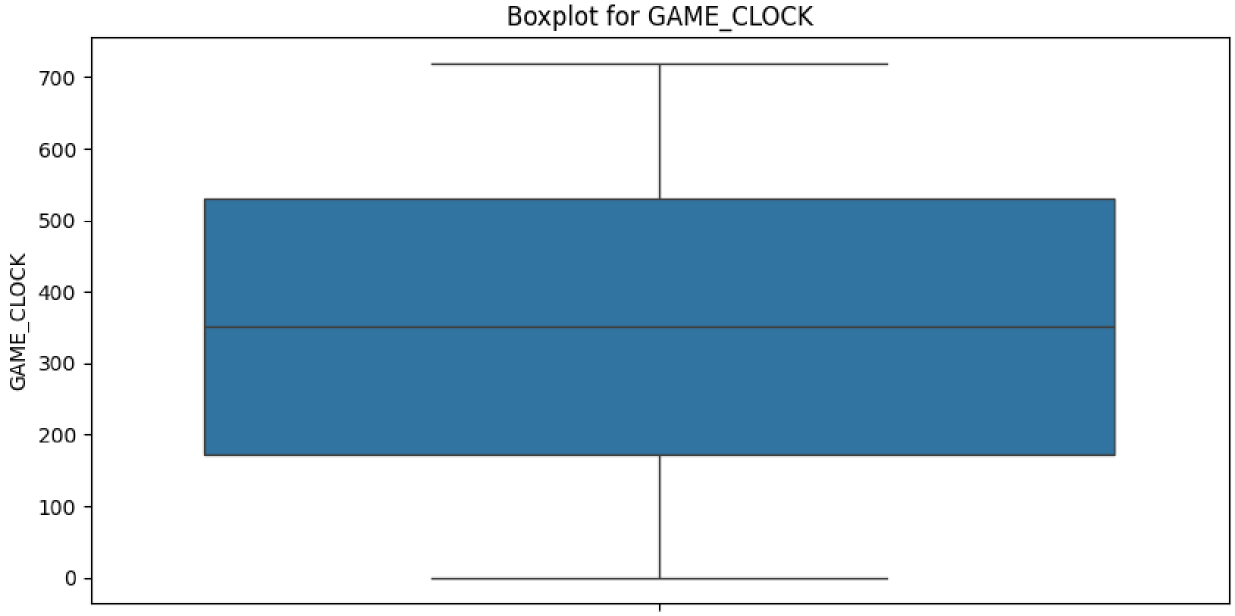
The correlation circle plot visually represents the relationships between multiple features in a dataset. Each feature is represented by a point on the plot, and the proximity of these points to each other indicates the strength and direction of their correlation. The plot uses vectors to illustrate the correlation structure, with the angle between vectors indicating the degree of correlation. A shorter distance between points or vectors implies a stronger correlation, while a longer distance suggests a weaker correlation. Therefore, the plot serves as a visual aid to understand the complex interplay between different features in the dataset.

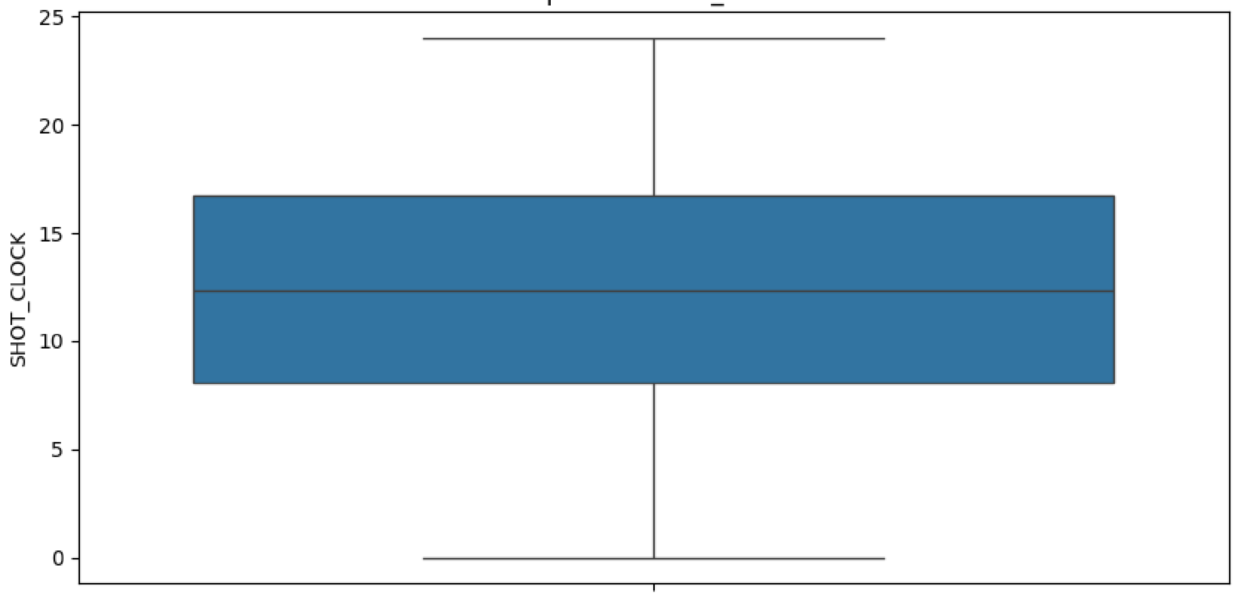
Outlier detection

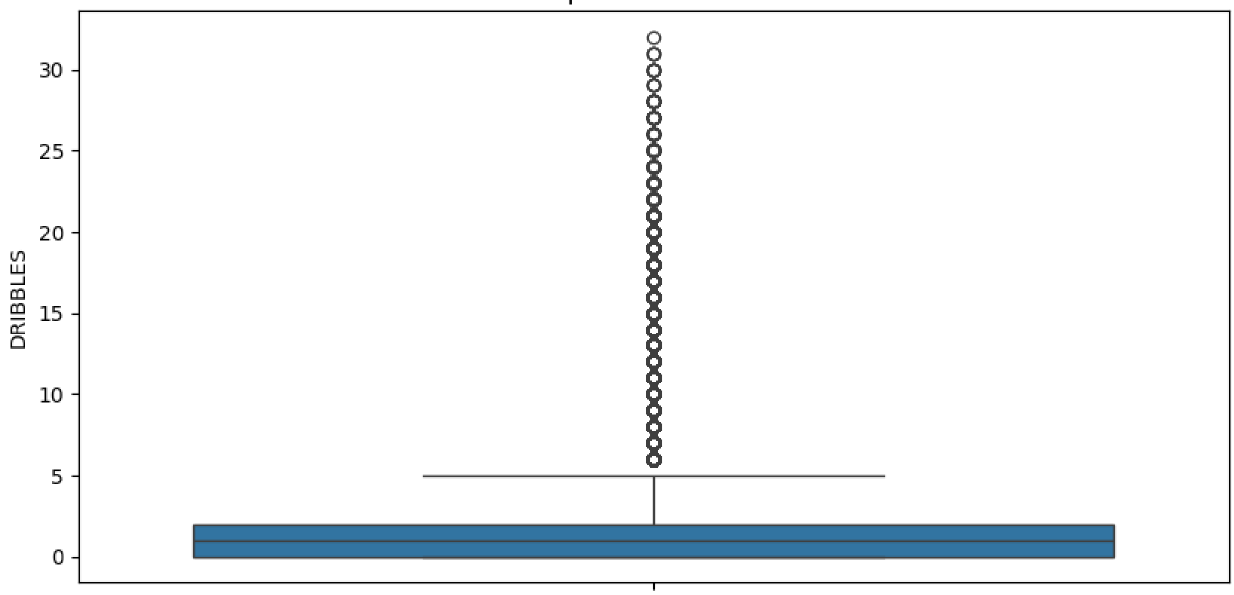


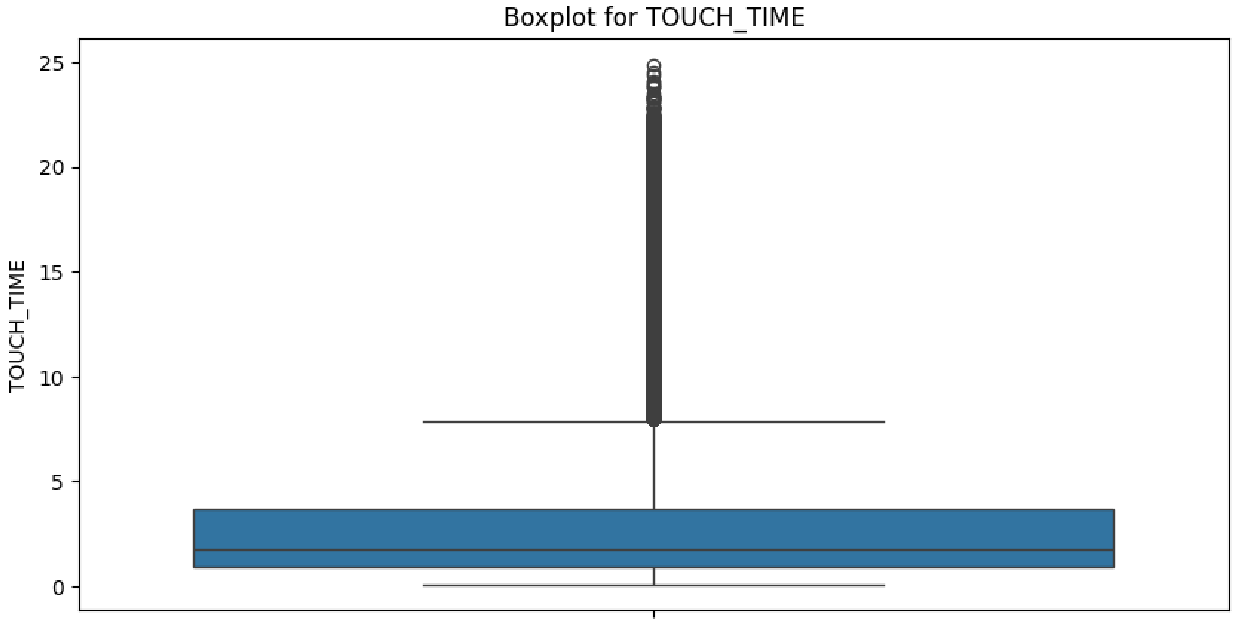


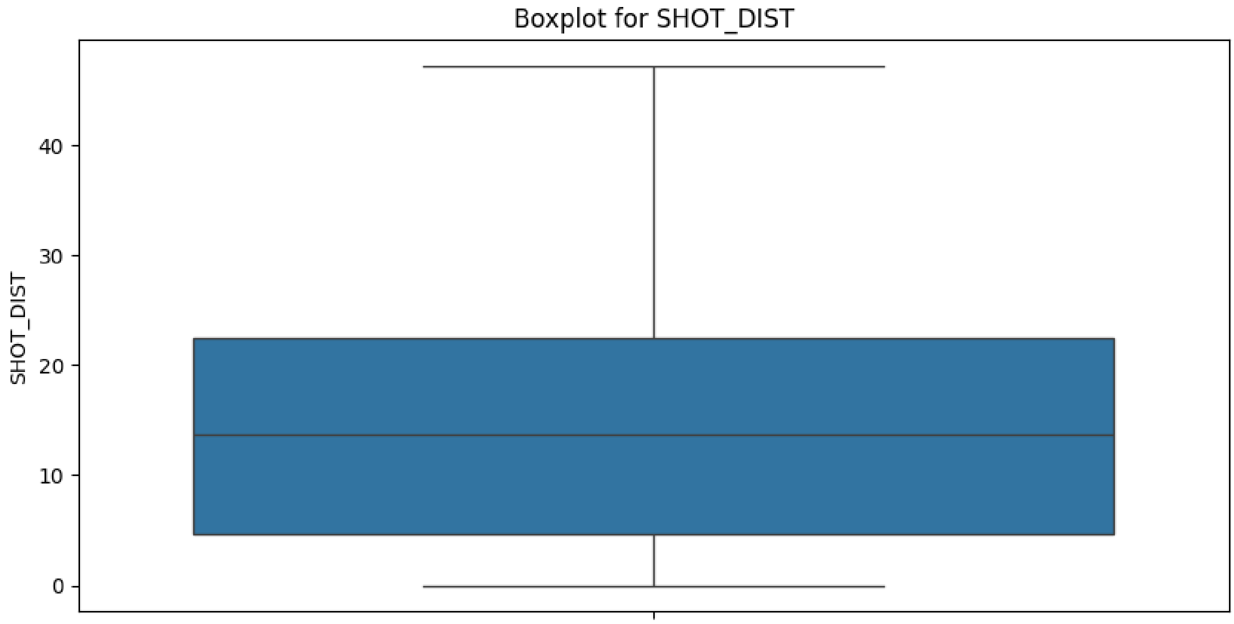

Boxplot for PERIOD

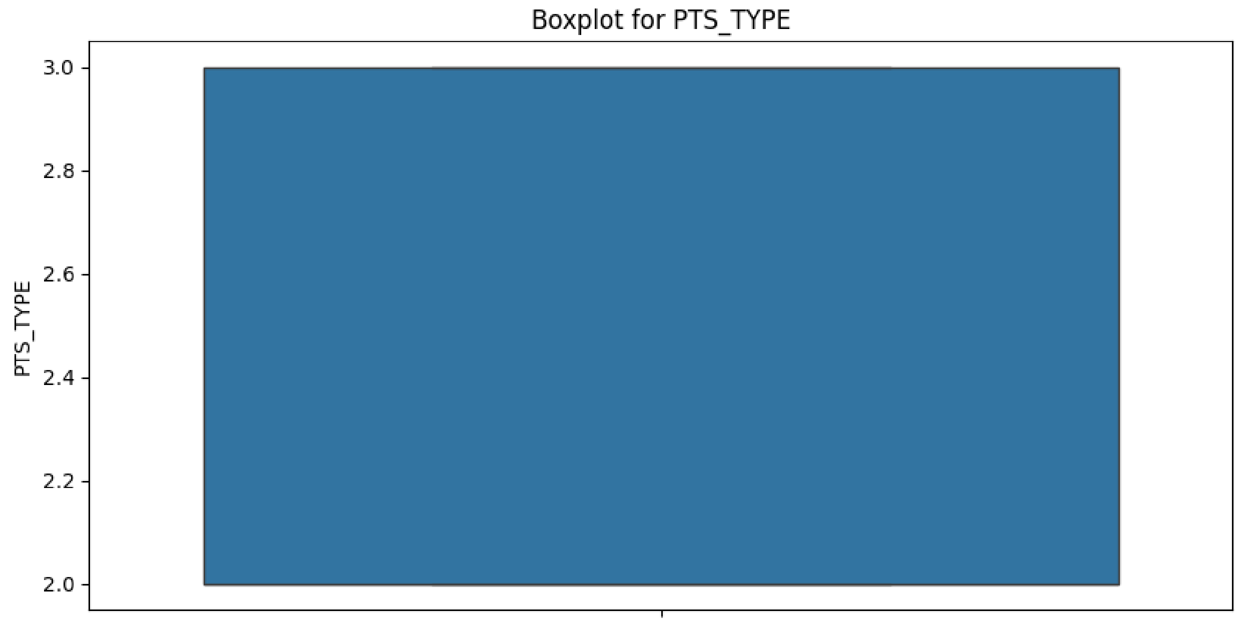

Boxplot for GAME_CLOCK

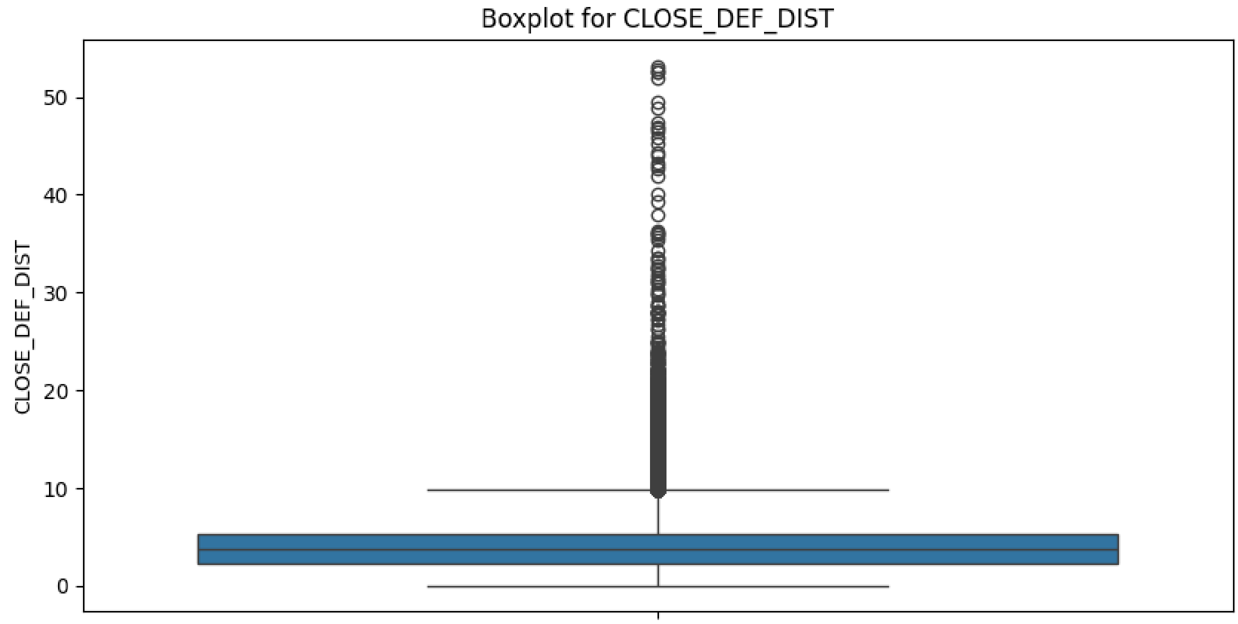

Boxplot for SHOT_CLOCK

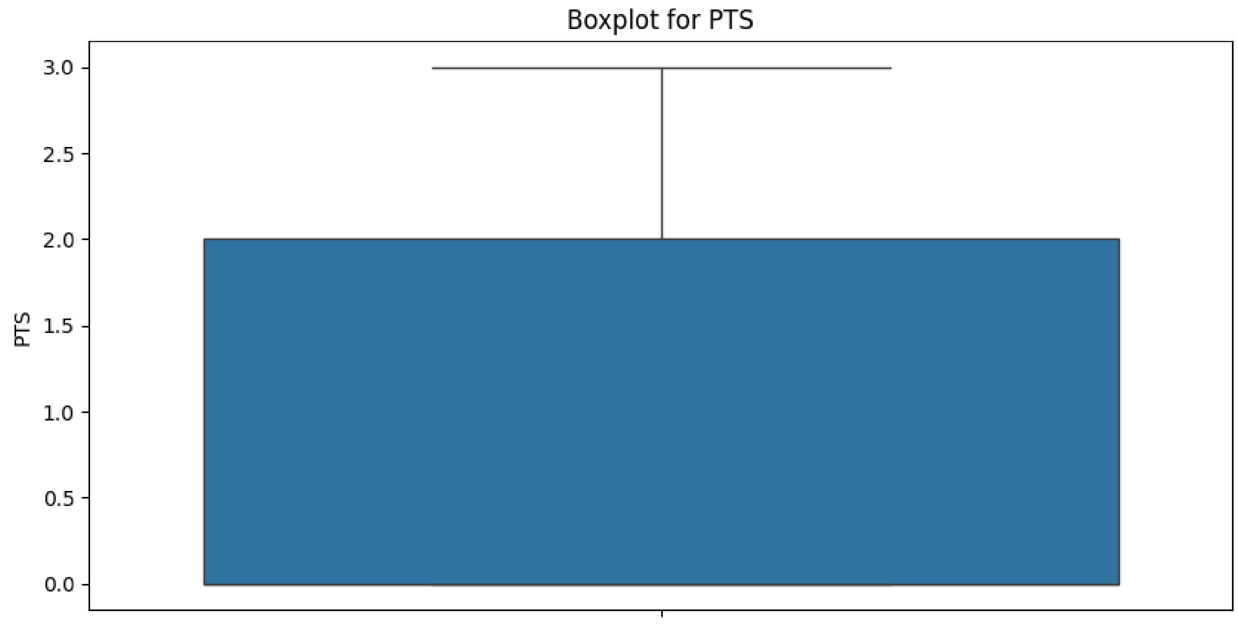

Boxplot for DRIBBLES


Boxplot for TOUCH_TIME
TOUCH_TIME
0
5
10
15
20
25


Boxplot for SHOT_DIST
SHOT_DIST
0
10
20
30
40


Boxplot for PTS_TYPE
PTS_TYPE
3.0
2.8
2.6
2.4
2.2
2.0


Boxplot for CLOSE_DEF_DIST
CLOSE_DEF_DIST
50
40
30
20
10
0

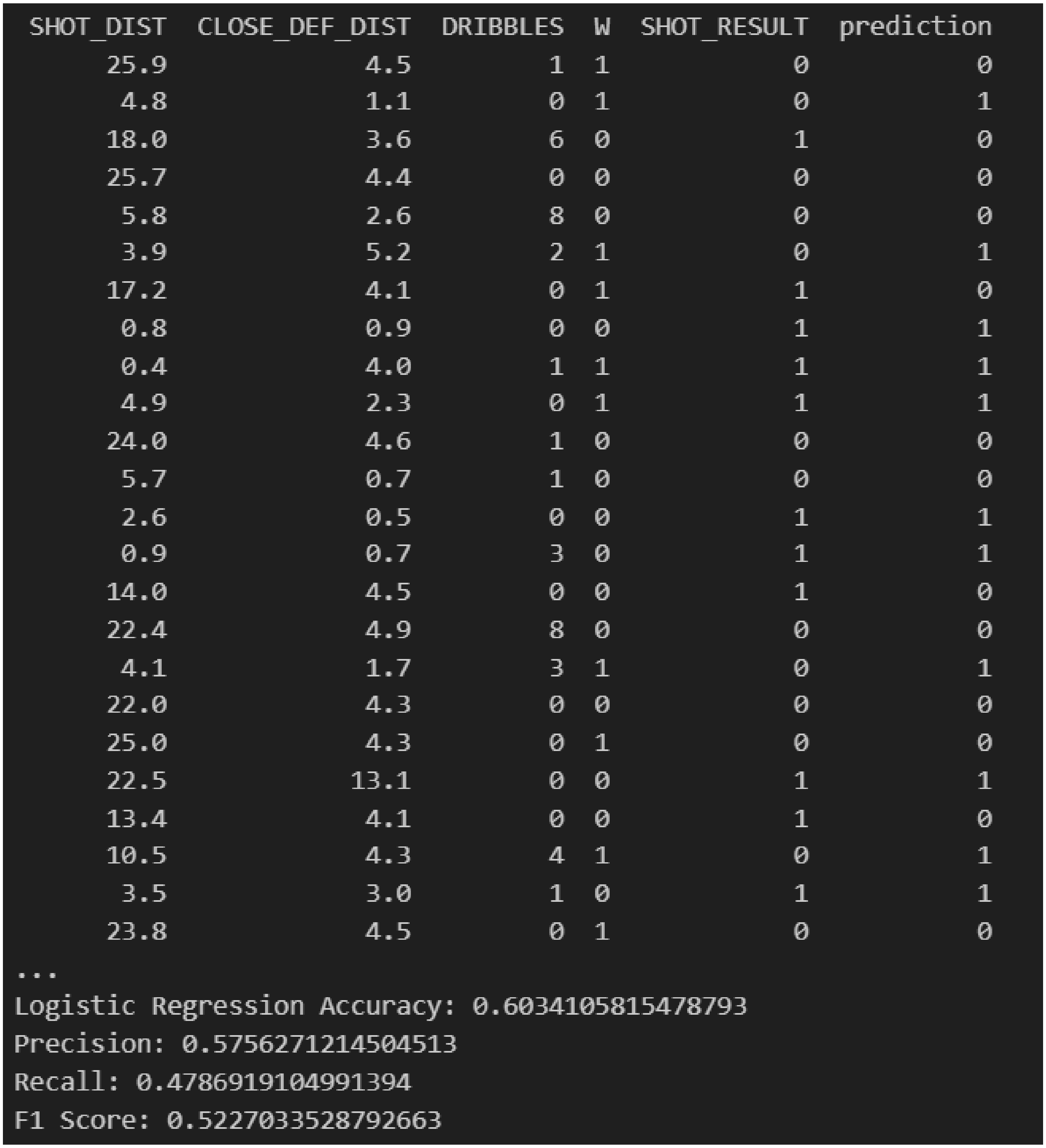

Boxplot for PTS
PTS
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Model/Classifier training

We split the data randomly to 75% for the training and 25% for the evaluation.

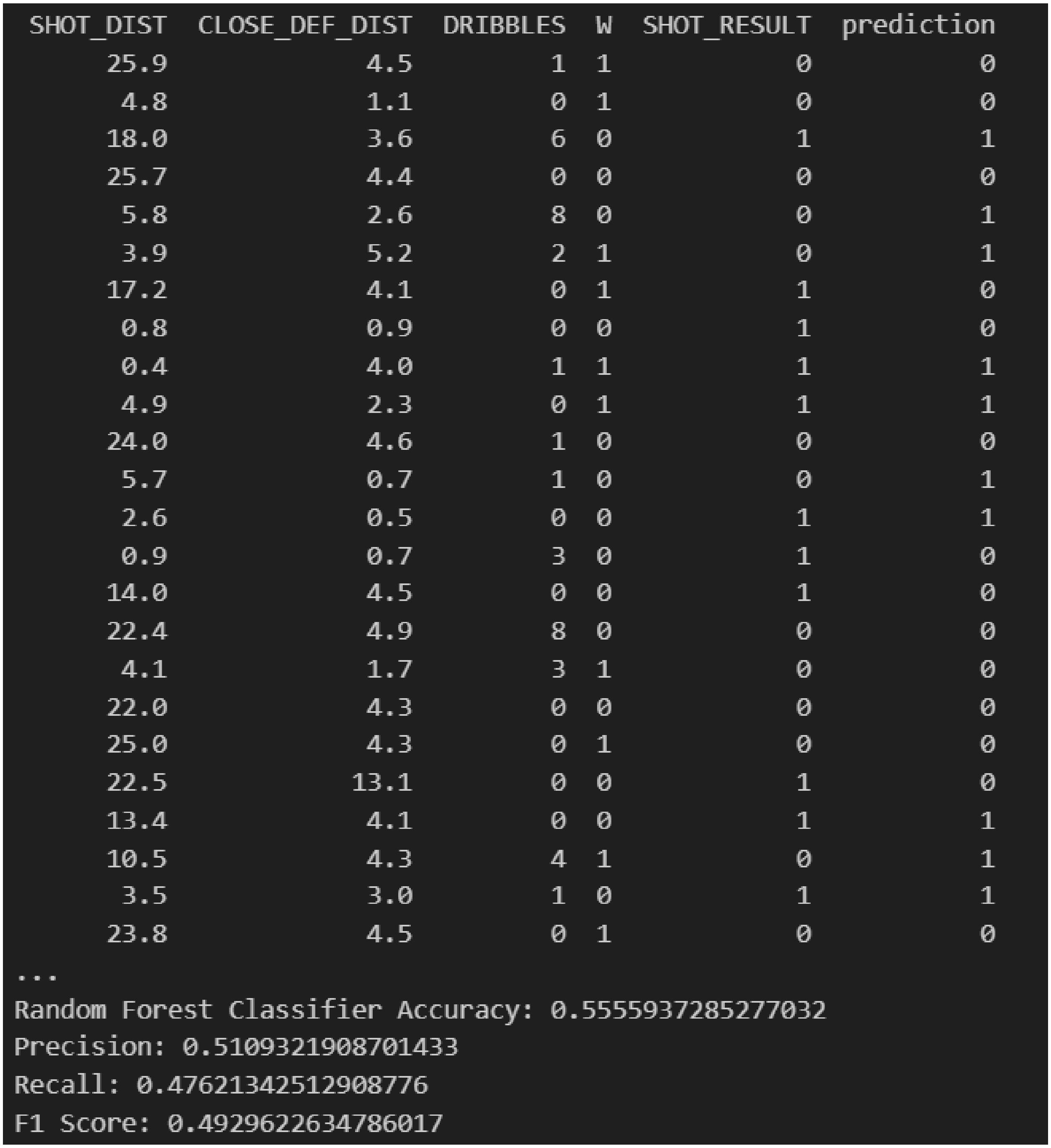
We used 4 different prediction models.

Logistic regression



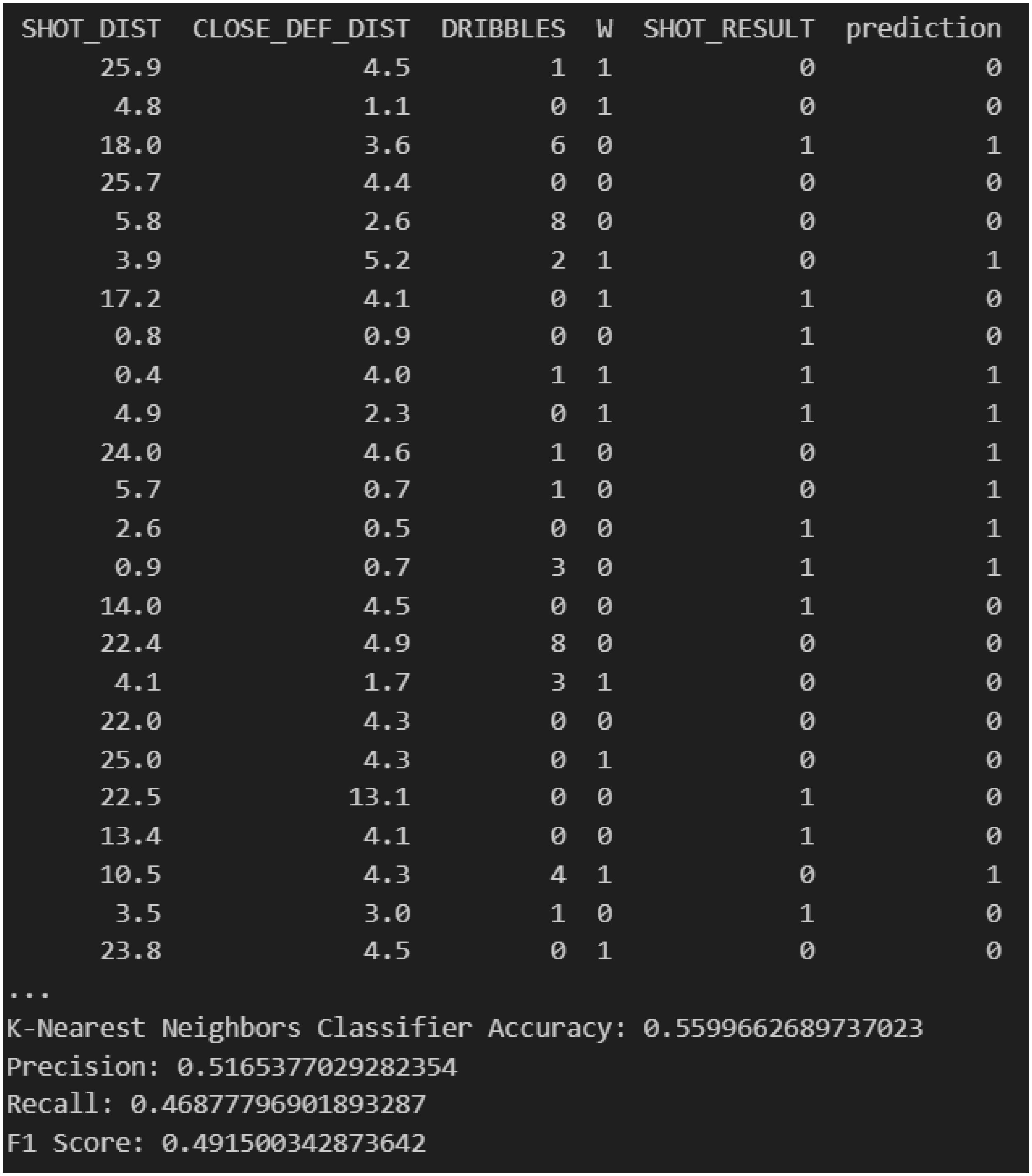
```
 SHOT_DIST  CLOSE_DEF_DIST  DRIBBLES  W  SHOT_RESULT  prediction
      25.9             4.5         1  1            0           0
       4.8             1.1         0  1            0           1
      18.0             3.6         6  0            1           0
      25.7             4.4         0  0            0           0
       5.8             2.6         8  0            0           0
       3.9             5.2         2  1            0           1
      17.2             4.1         0  1            1           0
       0.8             0.9         0  0            1           1
       0.4             4.0         1  1            1           1
       4.9             2.3         0  1            1           1
      24.0             4.6         1  0            0           0
       5.7             0.7         1  0            0           0
       2.6             0.5         0  0            1           1
       0.9             0.7         3  0            1           1
      14.0             4.5         0  0            1           0
      22.4             4.9         8  0            0           0
       4.1             1.7         3  1            0           1
      22.0             4.3         0  0            0           0
      25.0             4.3         0  1            0           0
      22.5            13.1         0  0            1           1
      13.4             4.1         0  0            1           0
      10.5             4.3         4  1            0           1
       3.5             3.0         1  0            1           1
      23.8             4.5         0  1            0           0
...
Logistic Regression Accuracy: 0.6034105815478793
Precision: 0.5756271214504513
Recall: 0.4786919104991394
F1 Score: 0.5227033528792663
```

Random forest



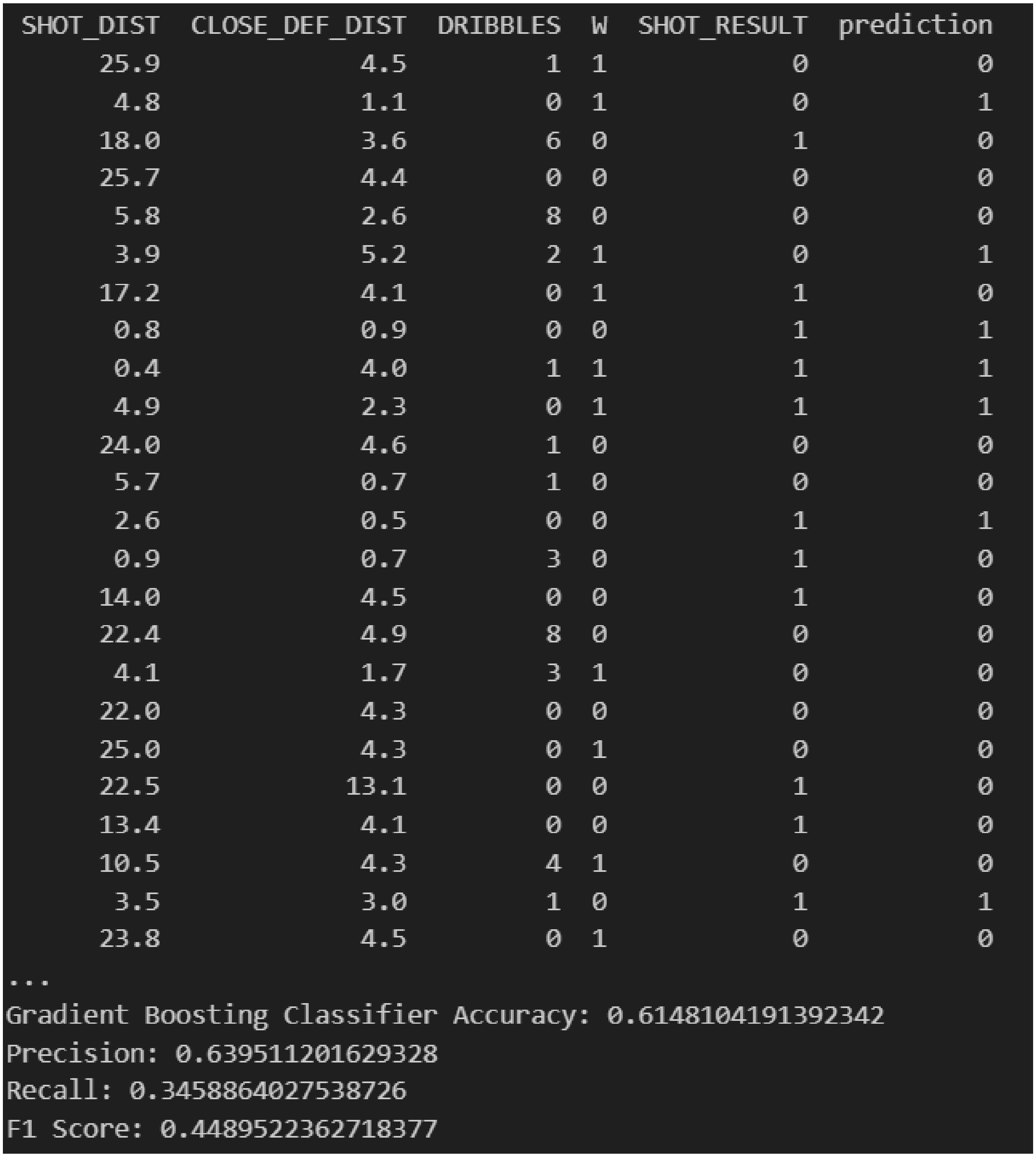
```
 SHOT_DIST  CLOSE_DEF_DIST  DRIBBLES  W  SHOT_RESULT  prediction
      25.9             4.5         1  1            0           0
       4.8             1.1         0  1            0           0
      18.0             3.6         6  0            1           1
      25.7             4.4         0  0            0           0
       5.8             2.6         8  0            0           1
       3.9             5.2         2  1            0           1
      17.2             4.1         0  1            1           0
       0.8             0.9         0  0            1           0
       0.4             4.0         1  1            1           1
       4.9             2.3         0  1            1           1
      24.0             4.6         1  0            0           0
       5.7             0.7         1  0            0           1
       2.6             0.5         0  0            1           1
       0.9             0.7         3  0            1           0
      14.0             4.5         0  0            1           0
      22.4             4.9         8  0            0           0
       4.1             1.7         3  1            0           0
      22.0             4.3         0  0            0           0
      25.0             4.3         0  1            0           0
      22.5            13.1         0  0            1           0
      13.4             4.1         0  0            1           1
      10.5             4.3         4  1            0           1
       3.5             3.0         1  0            1           1
      23.8             4.5         0  1            0           0
...
Random Forest Classifier Accuracy: 0.5555937285277032
Precision: 0.5109321908701433
Recall: 0.47621342512908776
F1 Score: 0.4929622634786017
```

KNN



```
 SHOT_DIST  CLOSE_DEF_DIST  DRIBBLES  W  SHOT_RESULT  prediction
      25.9             4.5         1  1            0           0
       4.8             1.1         0  1            0           0
      18.0             3.6         6  0            1           1
      25.7             4.4         0  0            0           0
       5.8             2.6         8  0            0           0
       3.9             5.2         2  1            0           1
      17.2             4.1         0  1            1           0
       0.8             0.9         0  0            1           0
       0.4             4.0         1  1            1           1
       4.9             2.3         0  1            1           1
      24.0             4.6         1  0            0           1
       5.7             0.7         1  0            0           1
       2.6             0.5         0  0            1           1
       0.9             0.7         3  0            1           1
      14.0             4.5         0  0            1           0
      22.4             4.9         8  0            0           0
       4.1             1.7         3  1            0           0
      22.0             4.3         0  0            0           0
      25.0             4.3         0  1            0           0
      22.5            13.1         0  0            1           0
      13.4             4.1         0  0            1           0
      10.5             4.3         4  1            0           1
       3.5             3.0         1  0            1           0
      23.8             4.5         0  1            0           0
...
K-Nearest Neighbors Classifier Accuracy: 0.5599662689737023
Precision: 0.5165377029282354
Recall: 0.46877796901893287
F1 Score: 0.491500342873642
```

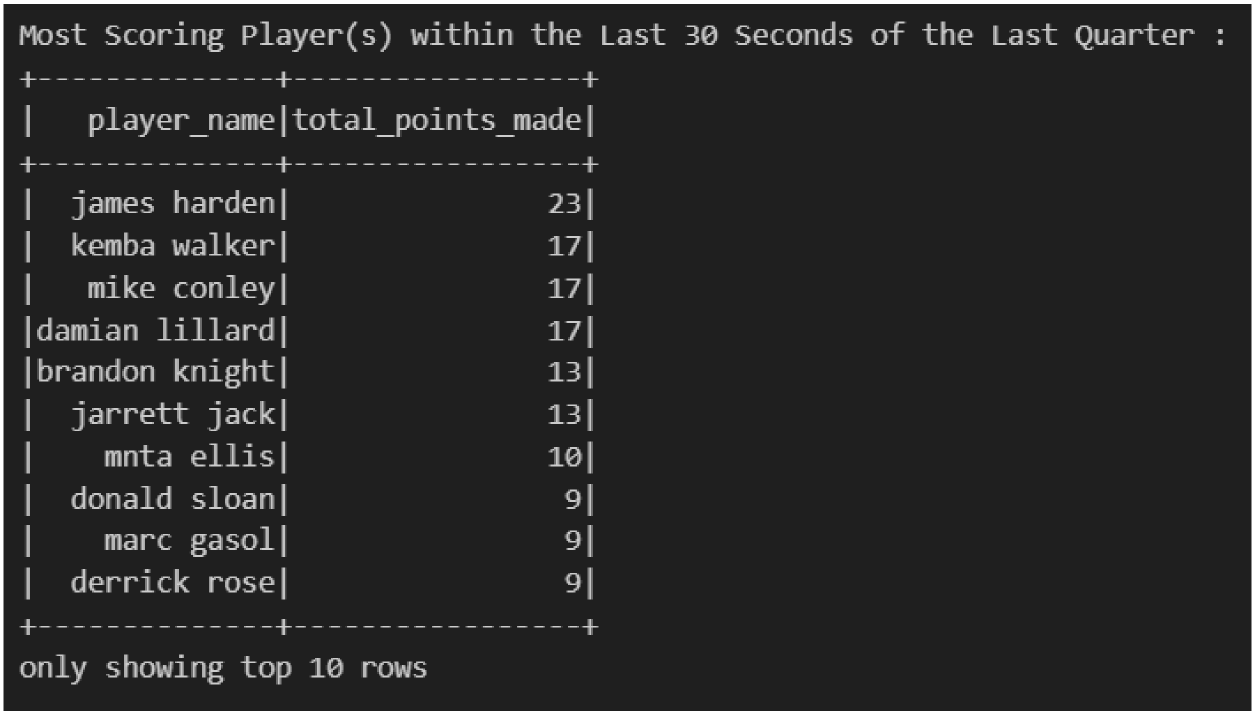
Gradient boosting



```
 SHOT_DIST  CLOSE_DEF_DIST  DRIBBLES  W  SHOT_RESULT  prediction
      25.9             4.5         1  1            0           0
       4.8             1.1         0  1            0           1
      18.0             3.6         6  0            1           0
      25.7             4.4         0  0            0           0
       5.8             2.6         8  0            0           0
       3.9             5.2         2  1            0           1
      17.2             4.1         0  1            1           0
       0.8             0.9         0  0            1           1
       0.4             4.0         1  1            1           1
       4.9             2.3         0  1            1           1
      24.0             4.6         1  0            0           0
       5.7             0.7         1  0            0           0
       2.6             0.5         0  0            1           1
       0.9             0.7         3  0            1           0
      14.0             4.5         0  0            1           0
      22.4             4.9         8  0            0           0
       4.1             1.7         3  1            0           0
      22.0             4.3         0  0            0           0
      25.0             4.3         0  1            0           0
      22.5            13.1         0  0            1           0
      13.4             4.1         0  0            1           0
      10.5             4.3         4  1            0           0
       3.5             3.0         1  0            1           1
      23.8             4.5         0  1            0           0
...
Gradient Boosting Classifier Accuracy: 0.6148104191392342
Precision: 0.639511201629328
Recall: 0.3458864027538726
F1 Score: 0.4489522362718377
```

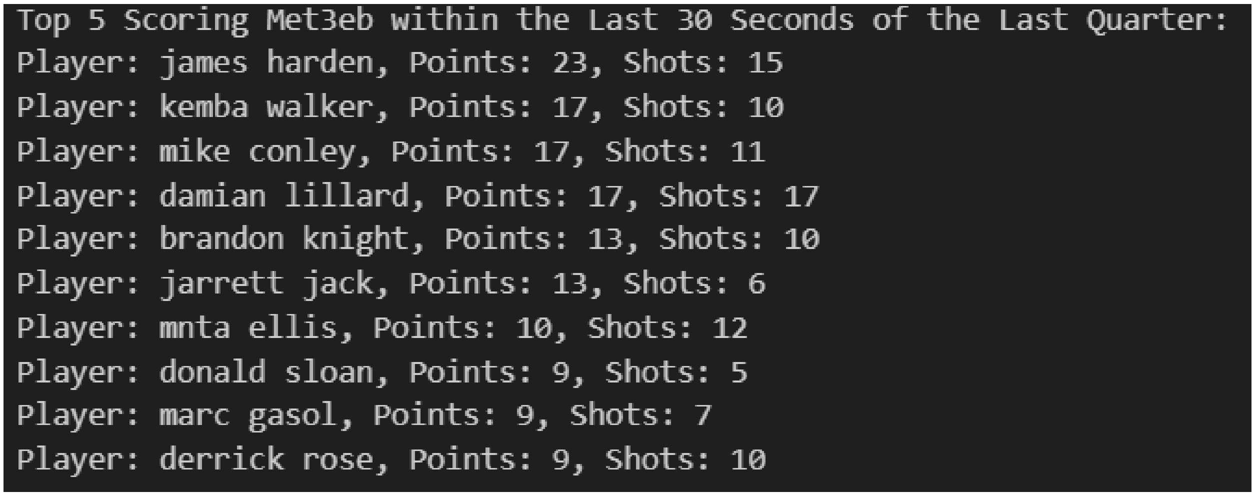
Map reduce

1. **the most scoring players in the last 30 seconds of the last quarter**
   We first made the map reduce function using pySpark.



```
Most Scoring Player(s) within the Last 30 Seconds of the Last Quarter :
+--------------+-----------------+
|   player_name|total_points_made|
+--------------+-----------------+
|  james harden|               23|
|  kemba walker|               17|
|   mike conley|               17|
|damian lillard|               17|
|brandon knight|               13|
|  jarrett jack|               13|
|    mnta ellis|               10|
|  donald sloan|                9|
|    marc gasol|                9|
|  derrick rose|                9|
+--------------+-----------------+
only showing top 10 rows
```

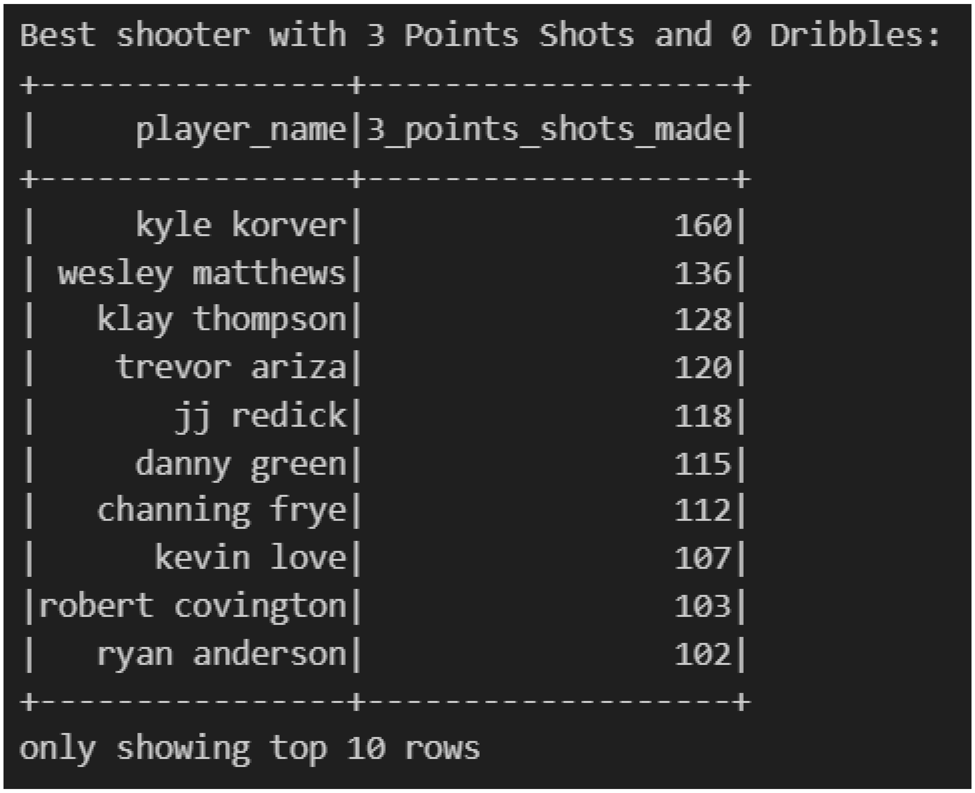
Then we implemented the mapper function and reducer function from scratch.



```
Top 5 Scoring Met3eb within the Last 30 Seconds of the Last Quarter:
Player: james harden, Points: 23, Shots: 15
Player: kemba walker, Points: 17, Shots: 10
Player: mike conley, Points: 17, Shots: 11
Player: damian lillard, Points: 17, Shots: 17
Player: brandon knight, Points: 13, Shots: 10
Player: jarrett jack, Points: 13, Shots: 6
Player: mnta ellis, Points: 10, Shots: 12
Player: donald sloan, Points: 9, Shots: 5
Player: marc gasol, Points: 9, Shots: 7
Player: derrick rose, Points: 9, Shots: 10
```

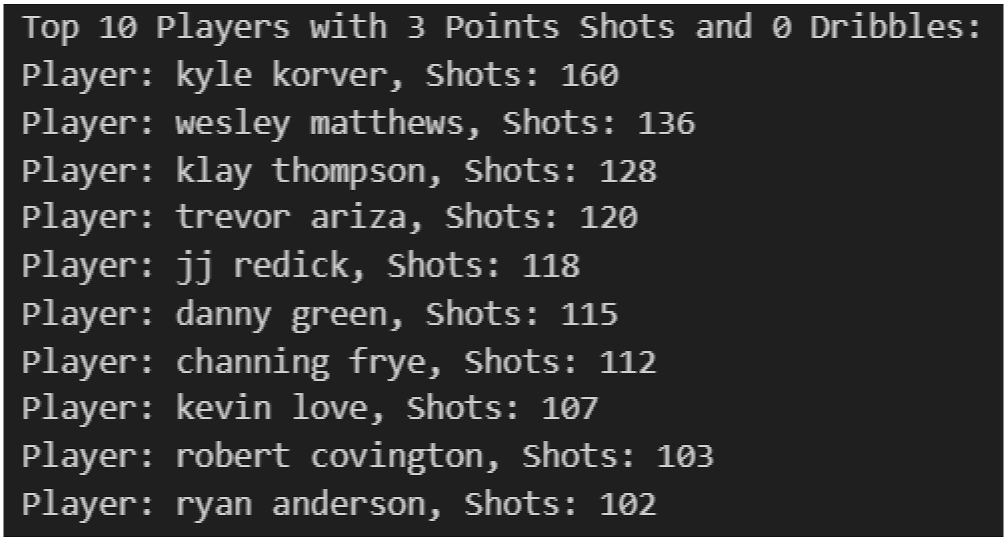
2. **top shot players who scored 3 points with 0 dribbles**

We first made the map reduce function using pySpark.



```
Best shooter with 3 Points Shots and 0 Dribbles:
+----------------+-------------------+
|     player_name|3_points_shots_made|
+----------------+-------------------+
|     kyle korver|                160|
| wesley matthews|                136|
|   klay thompson|                128|
|    trevor ariza|                120|
|       jj redick|                118|
|     danny green|                115|
|   channing frye|                112|
|      kevin love|                107|
|robert covington|                103|
|   ryan anderson|                102|
+----------------+-------------------+
only showing top 10 rows
```

Then we implemented the mapper function and reducer function from scratch.



```
Top 10 Players with 3 Points Shots and 0 Dribbles:
Player: kyle korver, Shots: 160
Player: wesley matthews, Shots: 136
Player: klay thompson, Shots: 128
Player: trevor ariza, Shots: 120
Player: jj redick, Shots: 118
Player: danny green, Shots: 115
Player: channing frye, Shots: 112
Player: kevin love, Shots: 107
Player: robert covington, Shots: 103
Player: ryan anderson, Shots: 102
```