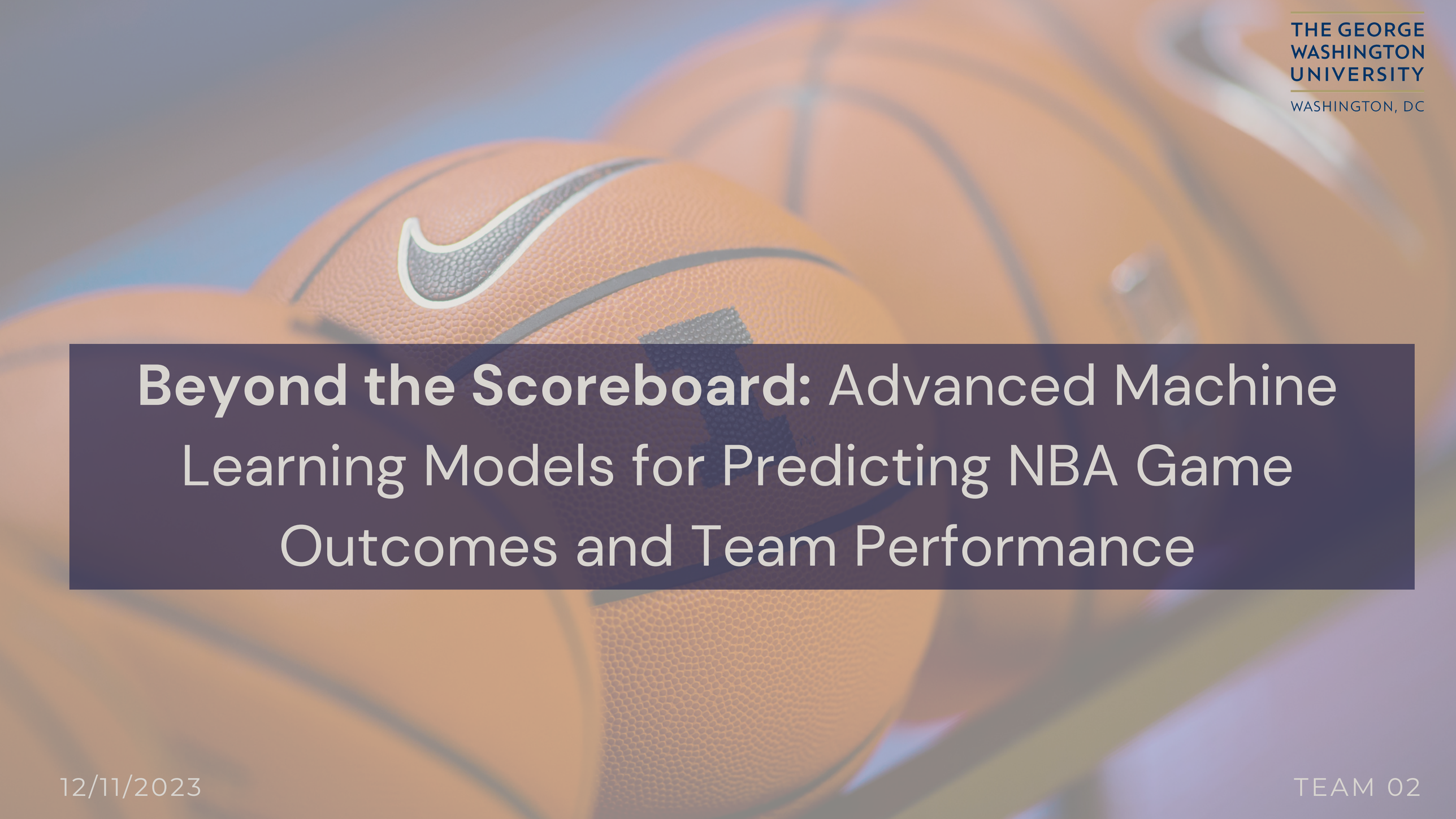


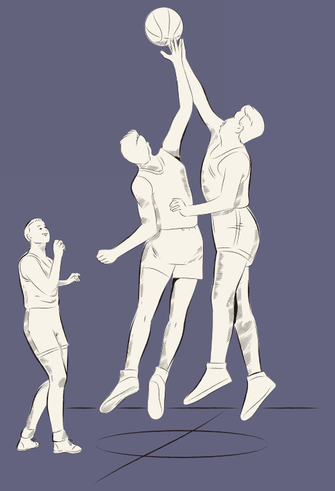
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Beyond the Scoreboard: Advanced Machine Learning Models for Predicting NBA Game Outcomes and Team Performance

12/11/2023

TEAM 02

OUTLINE



INTRODUCTION:

- This project focuses on leveraging NBA dataset to develop predictive models, aimed at classifying home team victories and forecasting home team scores in basketball games.

- The NBA dataset includes a wide array of data points such as player statistics, team performances, game outcomes, and historical trends, providing a robust foundation for analysis.

- We implement advanced statistical methods and machine learning algorithms, notably Polynomial Linear Regression for score prediction and Logistic Regression for win classification.

- By analyzing patterns and correlations in the NBA data, our project contributes to the growing field of basketball analytics, offering insights that could be valuable for teams, coaches, and sports analysts.

DATASET OVERVIEW

- **Source and Collection :**
    - Dataset sourced from Kaggle, originally compiled through web scraping from the official NBA stats website.
    - Primarily aimed at analyzing NBA games data and building predictive and classification models.

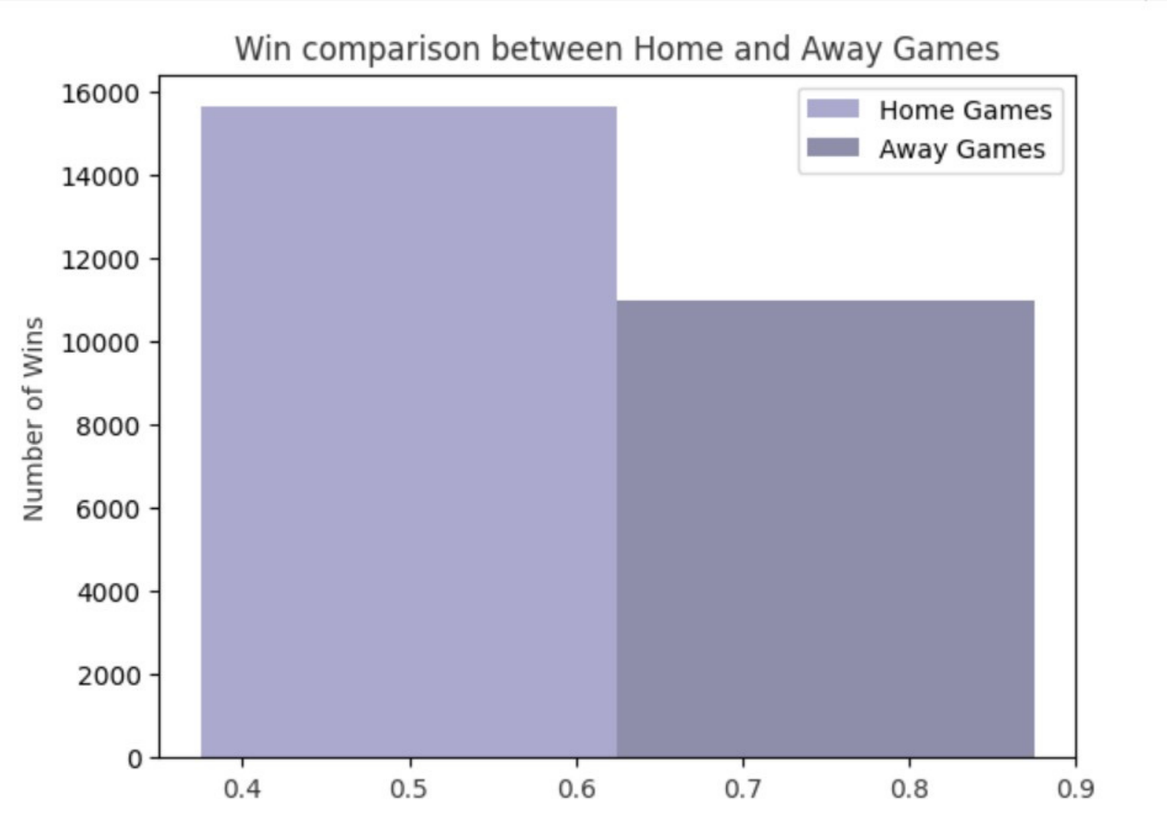
Dataset Composition :

Our dataset comprises five CSV files, with the games.csv file primarily used for building predictive models, while the other four datasets were utilized for extensive Exploratory Data Analysis (EDA).

- **games.csv: Data on NBA games since 2004 (26,000 rows, 21 columns).**
- **games_details.csv: In-depth player stats per game (668,000 rows, 29 columns).**
- **players.csv: Player details (7,000 rows, 4 columns).**
- **ranking.csv: Daily NBA rankings, West and East (210,000 rows, 13 columns).**
- **teams.csv: List of NBA teams (30 rows, 14 columns).**

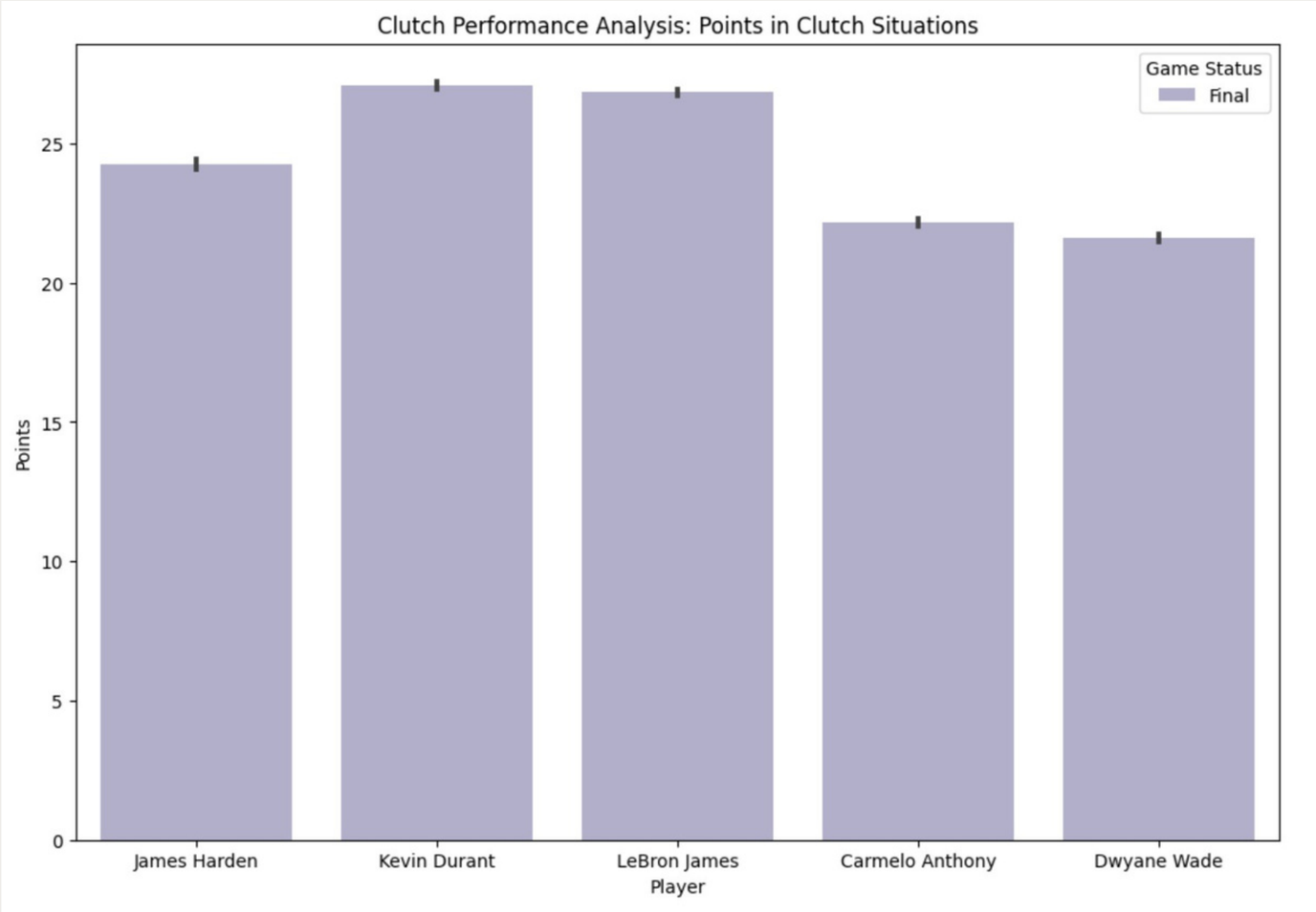
Exploratory Data Analysis

The histogram displays a significant home-court advantage, with a substantially greater number of wins for home games compared to away games, suggesting that teams perform better when playing on their home court.



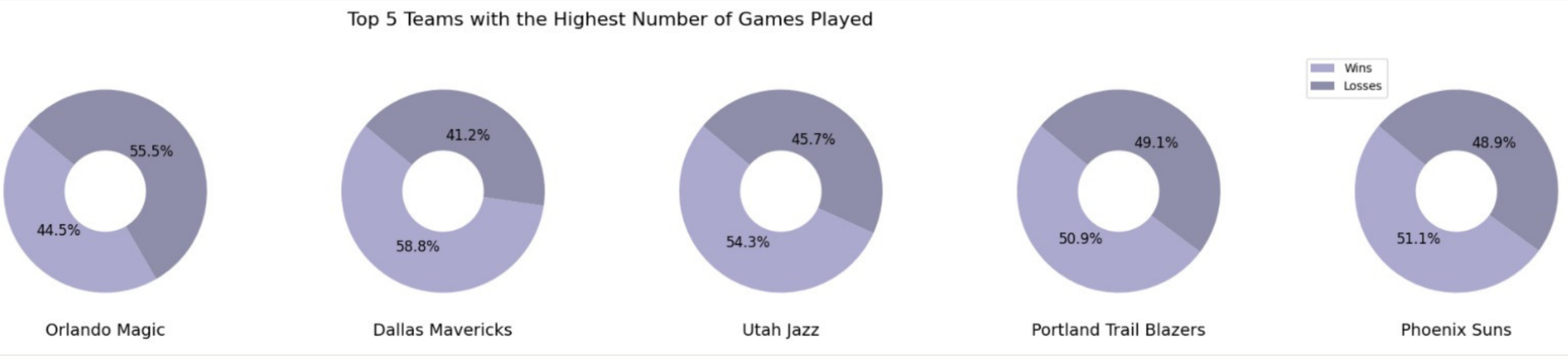
Exploratory Data Analysis

The chart highlights Kevin Durant as the most effective in clutch moments, with Lebron James close behind, and Dwyane Wade with the lowest clutch points among the group.

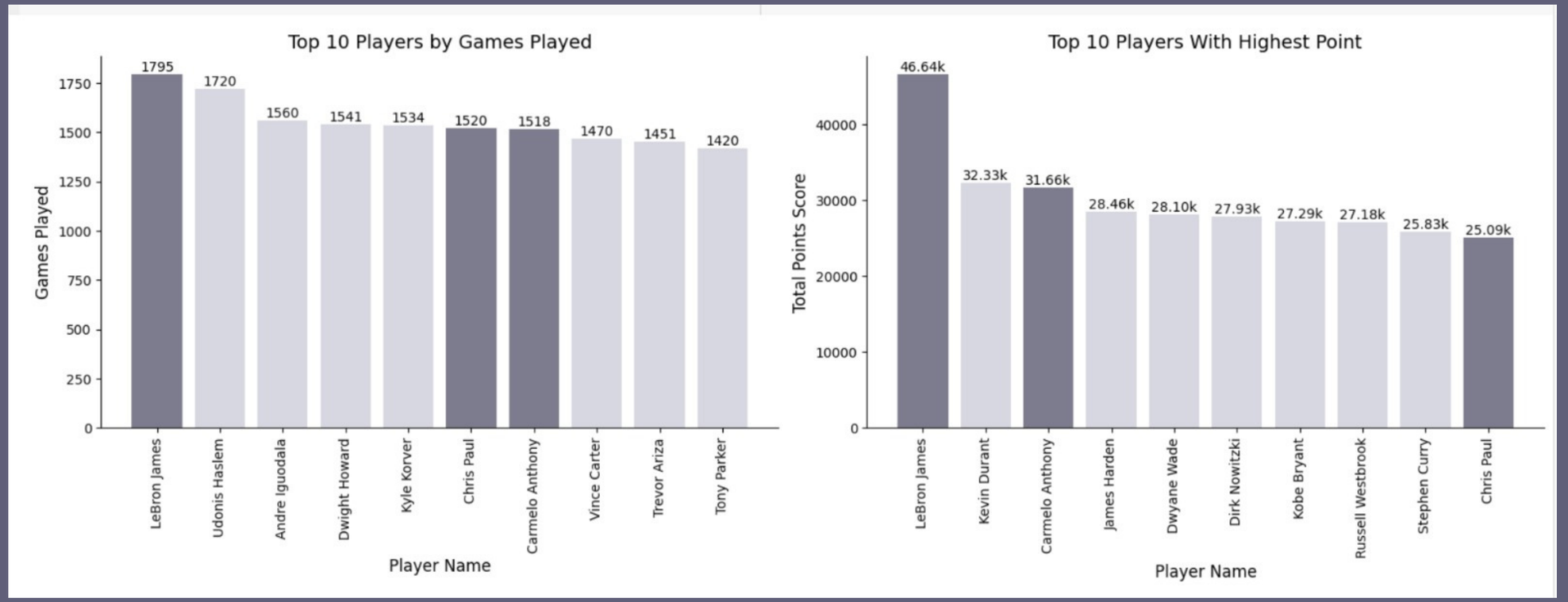


Exploratory Data Analysis

The pie chart displays the top five teams by games played, showing that more games don't guarantee more wins. Notably, Orlando Magic has the highest number of games but also a higher loss percentage than wins.

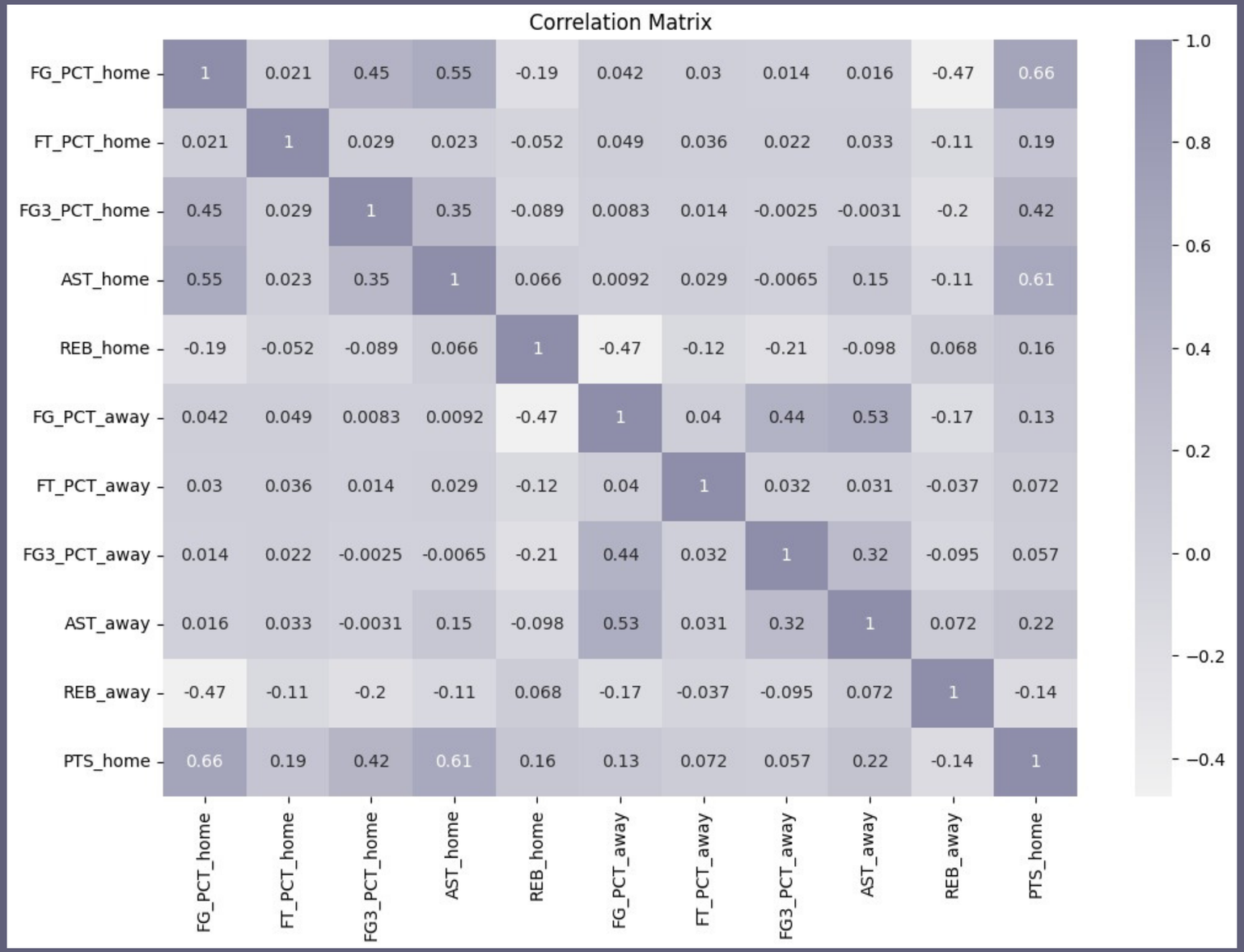


Exploratory Data Analysis

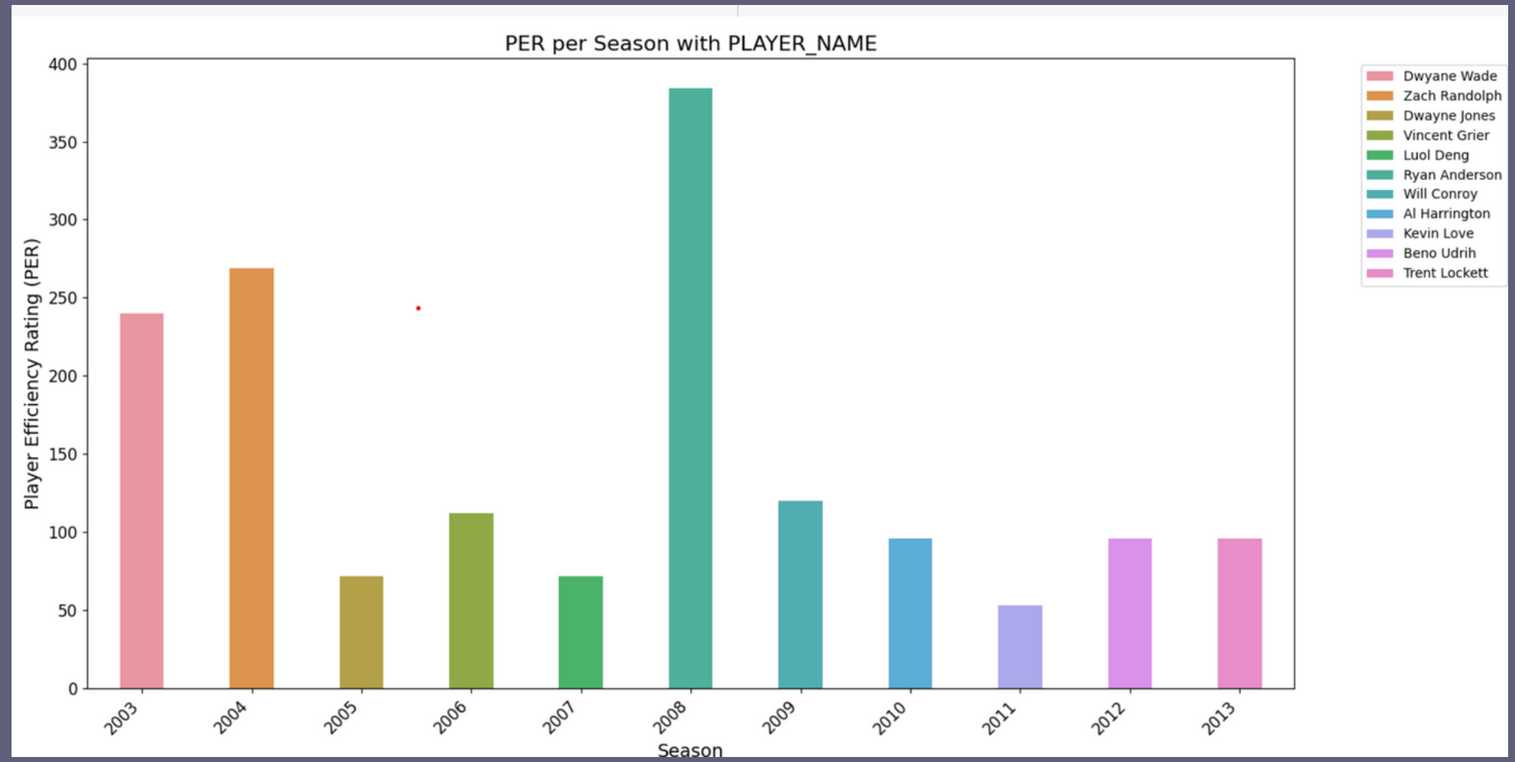


The graph indicates that LeBron James has played the most games and scored the most points. However, not all players with a high number of games have the highest scores. The graph reveals that only three players are featured in both categories.

Exploratory Data Analysis



Exploratory Data Analysis



In the graph we can see that in 2008 Ryan Anderson has highest efficiency among all the seasons. We can see a that efficiency keeps on decreasing through out the seasons.

formula= (merged_data['PTS'] + merged_data['REB'] + merged_data['AST'] + merged_data['STL'] + merged_data['BLK']
- merged_data['FGM'] - merged_data['FTM'] - merged_data['TO'])
/ merged_data['MIN']
) * 48

Data Modeling

Home Team Score Predictor

This model forecasts the total points scored by the home team based on both teams' on-court metrics.

Home Team Win Classifier

This model assesses the likelihood of the home team winning a game using key performance indicators from both the home and away teams.

Algorithmic Framework

Home Score Predictor

Linear Regression

Polynomial Linear Regression

Support Vector Machine Regressor

XGBOOST

Home Win Classifier

Logistic Regression

Decision Tree

Support Vector Machine Classifier

K- Nearest Neighbhor

Results

Home Team Score Predictor

Linear Regression

```
Mean Squared Error: 43.569985782009816
Root Mean Squared Error: 6.600756455286759
Mean Absolute Error: 5.1835799898542065
R-squared: 0.7525821888986025
```

Support Vector Machine Regressor

R-squared value:

```
0.7510757701985349
```

Polynomial Linear Regression

```
Mean Squared Error: 40.435511203163244
Root Mean Squared Error: 6.358892293722488
Mean Absolute Error: 5.015816230465073
R-squared: 0.772501587653443
```

XGBOOST

R-squared value:

```
0.6238871402433316
```

Results

Home Team Win Classifier

Logistic Regression

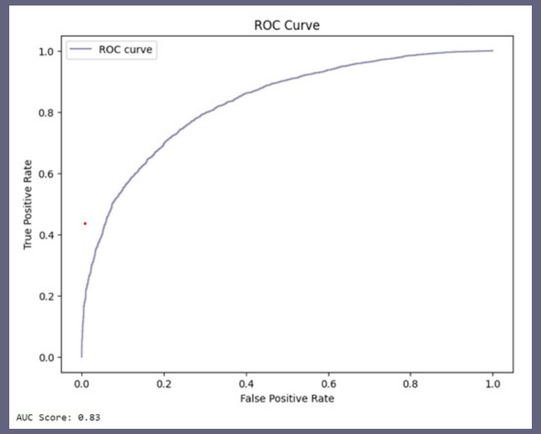
```
Accuracy: 0.76

Confusion Matrix:
[[2202 1075]
 [ 870 3819]]

Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.67      0.69      3277
           1       0.78      0.81      0.80      4689

    accuracy                           0.76      7966
   macro avg       0.75      0.74      0.75      7966
weighted avg       0.75      0.76      0.75      7966
```




```
C=0.001, Accuracy: 0.75
C=0.01, Accuracy: 0.75
C=0.1, Accuracy: 0.76
C=1, Accuracy: 0.76
C=10, Accuracy: 0.76
C=100, Accuracy: 0.76
```

Decision Tree

```
Decision Tree Accuracy: 0.67

Confusion Matrix:
[[1994 1283]
 [1326 3363]]

Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.61      0.60      3277
           1       0.72      0.72      0.72      4689

    accuracy                           0.67      7966
   macro avg       0.66      0.66      0.66      7966
weighted avg       0.67      0.67      0.67      7966
```

Support Vector Machine Classifier

```
SVM Accuracy: 0.75

Confusion Matrix:
[[2111 1166]
 [ 786 3903]]

Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.64      0.68      3277
           1       0.77      0.83      0.80      4689

    accuracy                           0.75      7966
   macro avg       0.75      0.74      0.74      7966
weighted avg       0.75      0.75      0.75      7966
```

K- Nearest Neighbhor

```
KNN Accuracy: 0.72

Confusion Matrix:
[[2081 1196]
 [1034 3655]]

Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.64      0.65      3277
           1       0.75      0.78      0.77      4689

    accuracy                           0.72      7966
   macro avg       0.71      0.71      0.71      7966
weighted avg       0.72      0.72      0.72      7966
```

CONCLUSION:

- Upon evaluating various algorithms, it was concluded that Polynomial Linear Regression offers the highest accuracy for predicting home team points.

- Meanwhile, Logistic Regression is the most accurate in classifying whether the home team will win or lose.

THANK YOU

THANK YOU