Python Project - Data Analytics and Machine Learning

1. Abstract

This project analyses NBA player performance data using data analytics and machine learning techniques. The dataset spans the 2021–2024 NBA seasons and focuses on variables such as points scored, minutes played, and assists. Using Linear Regression and Random Forest models, the project predicts players' scoring averages and identifies key factors influencing performance. I chose Linear Regression and Random Forest as my dependent variable points scored is continuous. The project employs hyperparameter tuning and data visualization to uncover patterns.

The project also includes unsupervised learning techniques such as PCA (Principal Component Analysis) and K means clustering to group players based on performance metrics. Results reveal that minutes played and field goals attempted are strong predictors of scoring, with Random Forest achieving better predictive accuracy based on the Mean Absolute Error metric. I have some domain knowledge in this sport which was what drove me to analyse certain metrics

---

2. Introduction

Project Topic and Significance

In professional sports, data analytics is critical for evaluating performance and shaping strategies. This project explores NBA player statistics to predict scoring performance and determine the key factors contributing to success. Insights from this analysis can support team managers, analysts, and fans in making data-driven decisions.

Dataset Overview

The dataset includes NBA player statistics across multiple seasons (2021–2024), containing metrics such as:

- **PTS (Points Scored):** Target variable.
- **MP (Minutes Played):** Time spent on the court.
- **TRB (Total Rebounds):** Defensive and offensive impact.
- **AST (Assists):** Team playmaking contributions.
- **FGA (Field Goals Attempted):** Offensive activity.
- **STL (Steals):** Defensive performance.
- **3P (Three-Point Field Goals Made)**: Successful three-point shots.
- **3PA (Three-Point Field Goals Attempted)**: Three-point shot attempts.
- **3P% (Three-Point Percentage)**: Success rate of three-point shots.
- **DRB (Defensive Rebounds)**: Rebounds collected on the defensive end.

The dataset is suited for predictive modelling, enabling analysis of player contributions and performance patterns.

---

3. Data

Dataset Description

The dataset consists of season-on-season NBA player statistics. It includes:

- **Numerical features** (e.g., points, rebounds, assists, steals).
- **Categorical features** (e.g., player names and teams).
- **Target variable:** Points scored (PTS).

Source

The dataset was sourced from publicly available NBA statistics repositories. You can find the dataset here. https://www.basketball-reference.com/

---

4. Importing

Method

The data was imported into Python using the pandas library:

1. The dataset was loaded from a flat CSV file using read_csv().
2. Missing values were identified and handled using Python functions.
3. Season-on-season statistics were aggregated to create a comprehensive view of each player's performance.

---

5. Data Preparation

Steps Taken

1. **Handling Duplicates:**
    - Players appeared multiple times in the dataset, representing individual seasons.
    - Data was aggregated using the groupby() function to calculate averages for key statistics (e.g., PTS, AST, MP).

    - I also removed any line that contained 2TM as this included already aggregated data and needed to be removed. 2TM represents players who have played for 2 teams in the course of a season
2. **Handling Missing Values:**
    - I checked for missing values in the important PTS column with the fillna() function in case there were values that were missing. These would have been filled with the mean of all points scored however it did not appear that there was missing data for the points column
3. **Filtering Data:**
    - Players with fewer than 10 games played were excluded to focus on consistent contributors.
4. **Feature Engineering:**
    - Efficiency metrics, such as PTS per MP, AST-to-PTS ratio, and FGA efficiency, were calculated to provide deeper insights into player performance.
5. **DataFrames and Grouping:**
    - Data was grouped by player names to compute average statistics for analysis.

---

6. Data Visualization

Key Visualizations

- **Distribution of Assists-to-Points Ratio (Seaborn)**

A histogram showing that most players rely less on assists for scoring, with an AST-to-PTS ratio below 0.3. *(Graph 1)*

- **Correlation Heatmap (Seaborn)**

Highlights strong correlations between minutes played (MP) and points scored (PTS), and between field goals attempted (FGA) and PTS. *(Graph 2)*

- **Scoring Trends for Top 10 Players (Matplotlib)**

Line charts showing the scoring trends of the top 10 players, illustrating consistency in performance. *(Graph 3)*

- **Feature Importance (Seaborn)**

A bar plot from the Random Forest model highlights PTS per MP and FGA as the most critical features for predicting scoring. *(Graph 4)*

- **Clusters Visualized Using PCA (Seaborn)**

A scatterplot showing clusters of players based on their performance metrics, reduced to two principal components (PCA1 and PCA2). Clusters illustrate distinct player groups *(Graph 5)*

- **Minutes Played Distribution (Matplotlib)**

A boxplot shows the range of minutes played, emphasizing variations in court time. *(Graph 6)*

- **Scoring Efficiency vs. Shooting Efficiency (Seaborn)**

A scatterplot demonstrates the relationship between PTS per MP and FGA efficiency, color-coded by assists. *(Graph 7)*

- **Points Distribution (Matplotlib)**

A scatterplot shows the distribution of scoring averages, with most players scoring fewer than 20 points per game. *(Graph 8)*

---

7. Machine Learning

Objective

The goal was to predict players' average points scored (PTS) using supervised and unsupervised learning algorithms. Two models were implemented, I chose these two models as the dependent variable of points scored is continuous:

Unsupervised Learning:

K-Means Clustering with PCA
K-Means clustering was implemented to group players into performance-based clusters. Using PCA, the dataset's dimensions were reduced to two principal components, capturing the most important variance. This dimensionality reduction allowed for effective visualization and clustering of players into three groups:

1. **Cluster 0**: Players with lower contributions across metrics.
2. **Cluster 1**: Balanced players contributing across multiple categories.
3. **Cluster 2**: High-performing players with strong scoring and playmaking contributions.

The cluster labels were also included as a feature in the supervised learning models to enhance predictions of scoring performance.

Supervised Learning:

1. **Linear Regression:** Captures linear relationships between features.
2. **Random Forest Regressor:** Models non-linear interactions such as the impact of defensive stats on points scored and provides feature importance.

Implementation Steps

1. **Data Splitting:**
    - Data was divided into training (80%) and testing (20%) sets to evaluate the performance of machine learning models.
    - The training set contains 80% of the data and is used to teach the model to learn relationships between independent variables and the dependent variable
    - The testing set contains 20% of the data and is kept separate during training. This is used to evaluate how the model performs on data it has not seen
2. **Model Training:**
    - Both models were trained on features like MP, AST, TRB, STL, G, 3P, 3PA, 3P%, DRB and FGA. After running the model and viewing the relationship certain variables had on PTS, I removed defensive stats DRB and STL
    - Linear Regression model was trained to find a linear relationship between independent variables and the target variable of PTS
3. **Hyperparameter Tuning:**

- Random Forest hyperparameters were optimized using RandomizedSearchCV. This searches through a large number of hyperparameter combinations and chooses the best performing

4. **Evaluation Metrics:**
    - Models were evaluated using:
        - **Mean Absolute Error (MAE):** Measures prediction accuracy.
        - **R^2 Score:** Indicates variance explained by the model.

Results

The below results include features such as defensive rebounds and steals. Please see graph 10. After viewing this graph, I decided to tune my model further by removing these stats – steals (STL) and Defensive Rebounds (DRB)

Metric	Linear Regression	Random Forest
Mean Absolute Error	0.442	0.388
R^2 Score	0.991	0.987

Metric	Linear Regression	Random Forest
Mean Absolute Error	0.436	0.386
R^2 Score	0.990	0.990

Both models performed very well in predicting points for the R^2 predictor. Anywhere close to 1 indicates a good model fit. Similarly, both models achieve a good mean absolute error score with Random Forest outperforming Linear Regression by producing a lower number indicating a better fit

---

8. Insights

1. **Scoring Distribution:**
    - Most players score fewer than 20 points per game, with only a few outliers exceeding 25 points. Players who score more than 25 points have a high Field Goals Attempted, Minutes Played and Assists. See graph 6 which is colour coded for assists. *(Graph 7)*
2. **Feature Importance:**
    - **Minutes Played (MP)** and **Field Goals Attempted (FGA)** are the most significant predictors of scoring *(Graph 4)*

3. **Assists Contribution:**
    - High-assist players generally score more, emphasizing the role of playmaking in offensive success. This might encourage coaches to emphasize the importance of passing and ball movement *(Graph 7)*
4. **Court Time Impact:**
    - Players scoring the most tend to have higher average minutes played. This is an expected finding from my relevant domain knowledge as translates to more opportunities to score through field goal attempts and more chances to generate assists. Coaches rely on scorers to play more minutes to take advantage of their scoring potential *(Graph 2)*
5. **High-Scoring Efficiency Outliers**:
    A subset of players achieves high scoring (PTS) with limited minutes played or field goal attempts, showcasing their efficiency in converting attempts into points. These players can be critical assets in rotations. *(Graph 12)*
6. **Defensive Impact:**
    - Defensive contributions like Steals (STL) had a relatively small impact on scoring performance. This would indicate defensive specialization ie. players who are focused on defence focus less on scoring, see graph 4 which illustrates a weak correlation between steals and points scored *(Graph 4)*
7. **Top Scorers:**
    - A few players stand out as consistent high scorers across games. Top scorers are also very efficient 3 point shooters and shoot at a high volume *(Graph 10)*
8. **Team Dynamics:**
    - Players on teams with higher assist averages tend to score more, reflecting strong team play. As seen in the correlation heat map – see graph 2 the ratio of assists to points scored is .75 which would indicate that players on teams with a passing emphasis score more *(Graph 2)*

---

9. Results and Conclusion

Summary

This analysis revealed key insights into NBA player performance:

1. **Field goals attempted and Points per minutes played** are the strongest predictors of scoring. Field Goals Attempted are the aggregate of 2 point and 3 point attempts.
2. **Random Forest** outperformed Linear Regression slightly in predictive accuracy (lower MAE).
3. **Visualization** highlighted correlations between scoring, playing time, field goals attempted (FGA) and assists.

Conclusion

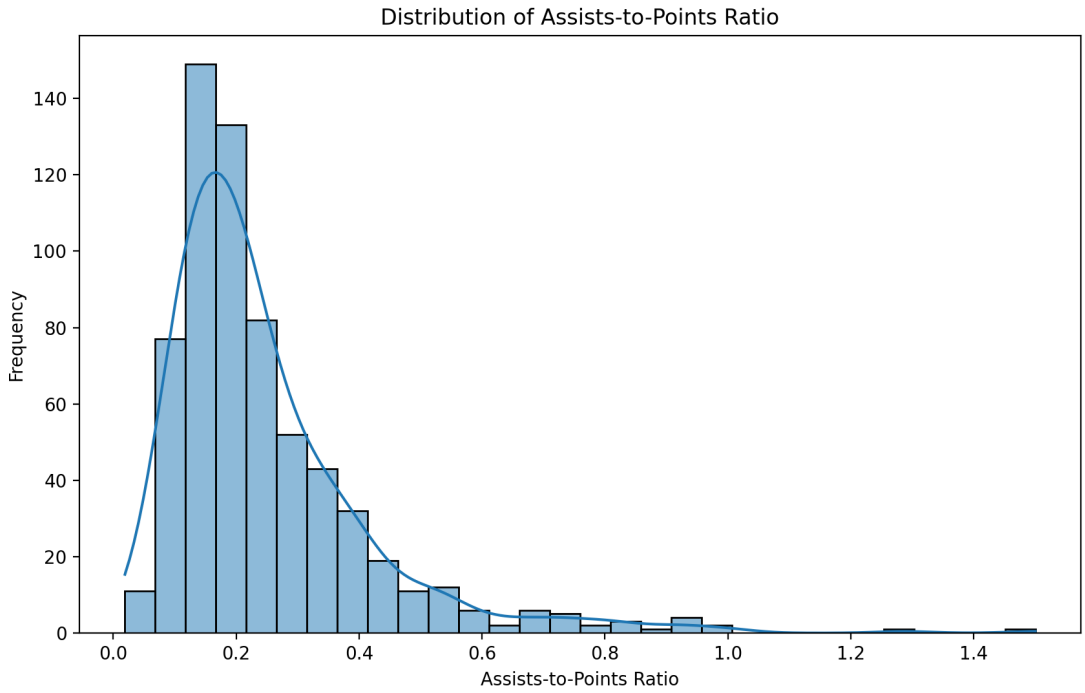
This project revealed that supervised learning models can effectively predict player scoring with Random Forrest slightly outperforming Linear Regression. Additionally, unsupervised

learning using K-means clustering identified three distinct player groups based on performance metrics supporting insights into player roles and contributions.

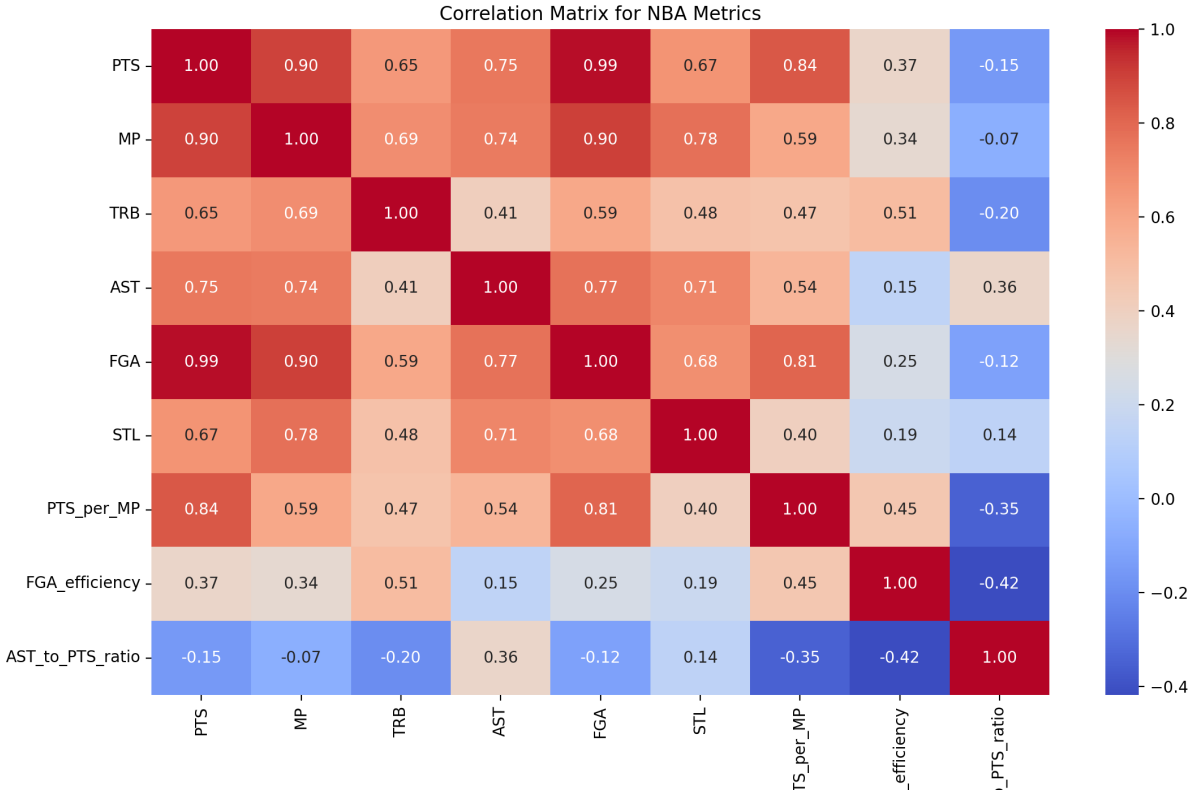
This project showcases the value of data analytics and machine learning in sports analysis. The insights can guide teams in identifying high-impact players and refining strategies. Future work could integrate more advanced models, such as neural networks, or analyse team-level statistics for deeper insights.

10. Graphs

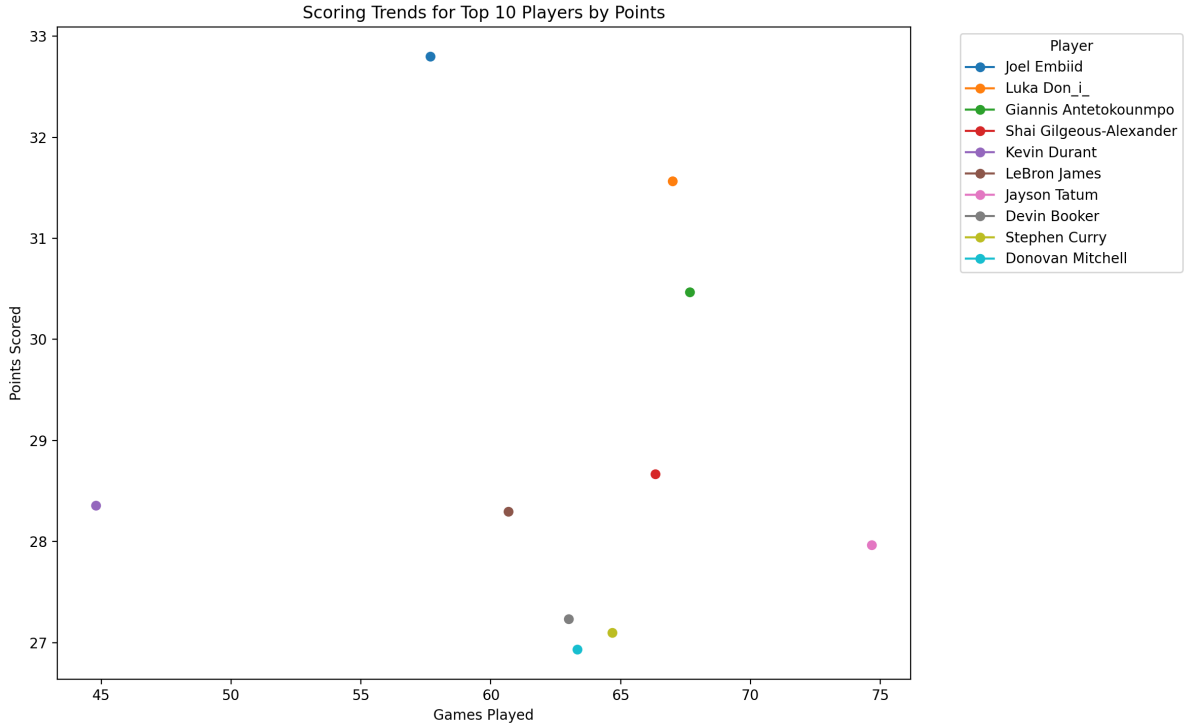
Graph 1



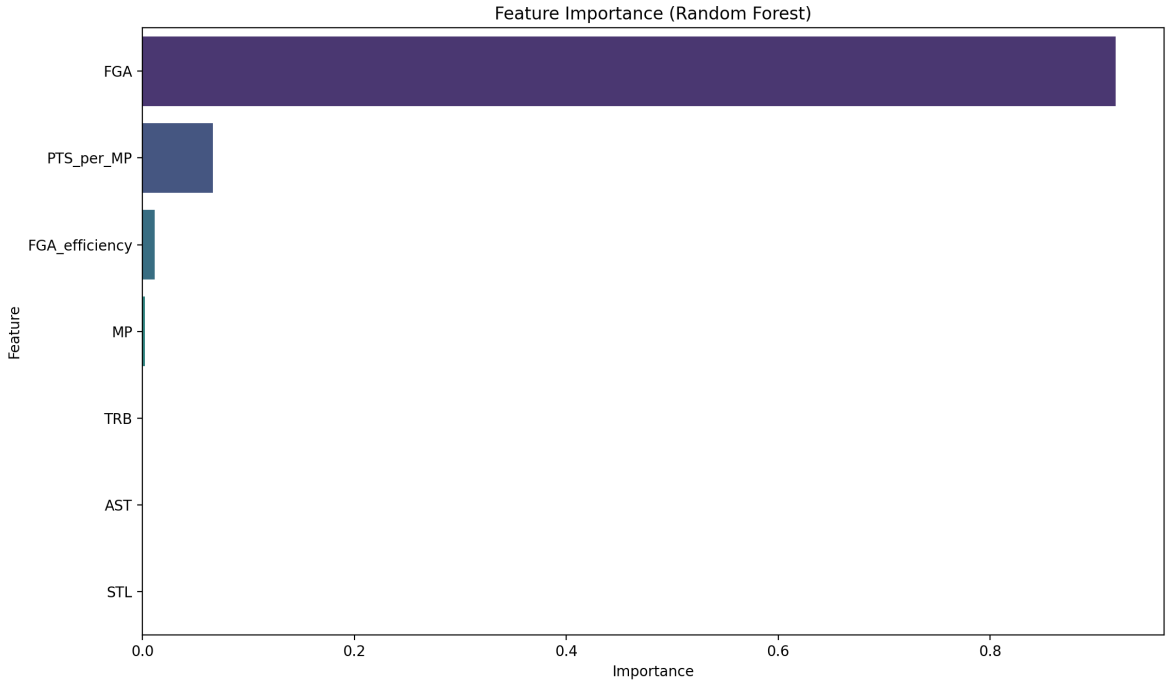
Graph 2


Correlation Matrix for NBA Metrics
PTS
MP
TRB
AST
FGA
STL
PTS_per_MP
FGA_efficiency
AST_to_PTS_ratio
1.00 0.90 0.65 0.75 0.99 0.67 0.84 0.37 -0.15
0.90 1.00 0.69 0.74 0.90 0.78 0.59 0.34 -0.07
0.65 0.69 1.00 0.41 0.59 0.48 0.47 0.51 -0.20
0.75 0.74 0.41 1.00 0.77 0.71 0.54 0.15 0.36
0.99 0.90 0.59 0.77 1.00 0.68 0.81 0.25 -0.12
0.67 0.78 0.48 0.71 0.68 1.00 0.40 0.19 0.14
0.84 0.59 0.47 0.54 0.81 0.40 1.00 0.45 -0.35
0.37 0.34 0.51 0.15 0.25 0.19 0.45 1.00 -0.42
-0.15 -0.07 -0.20 0.36 -0.12 0.14 -0.35 -0.42 1.00
PTS
MP
TRB
AST
FGA
STL
TS_per_MP
_efficiency
_PTS_ratio
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

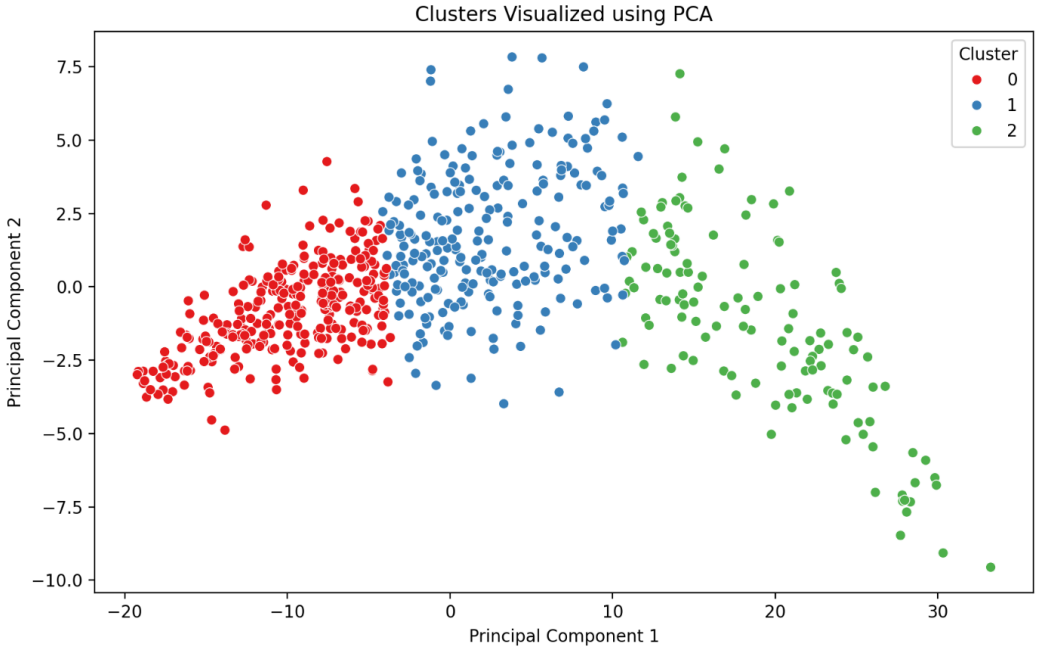
Graph 3


Scoring Trends for Top 10 Players by Points
Points Scored
Games Played
33
32
31
30
29
28
27
45
50
55
60
65
70
75
Player
Joel Embiid
Luka Don_i_
Giannis Antetokounmpo
Shai Gilgeous-Alexander
Kevin Durant
LeBron James
Jayson Tatum
Devin Booker
Stephen Curry
Donovan Mitchell

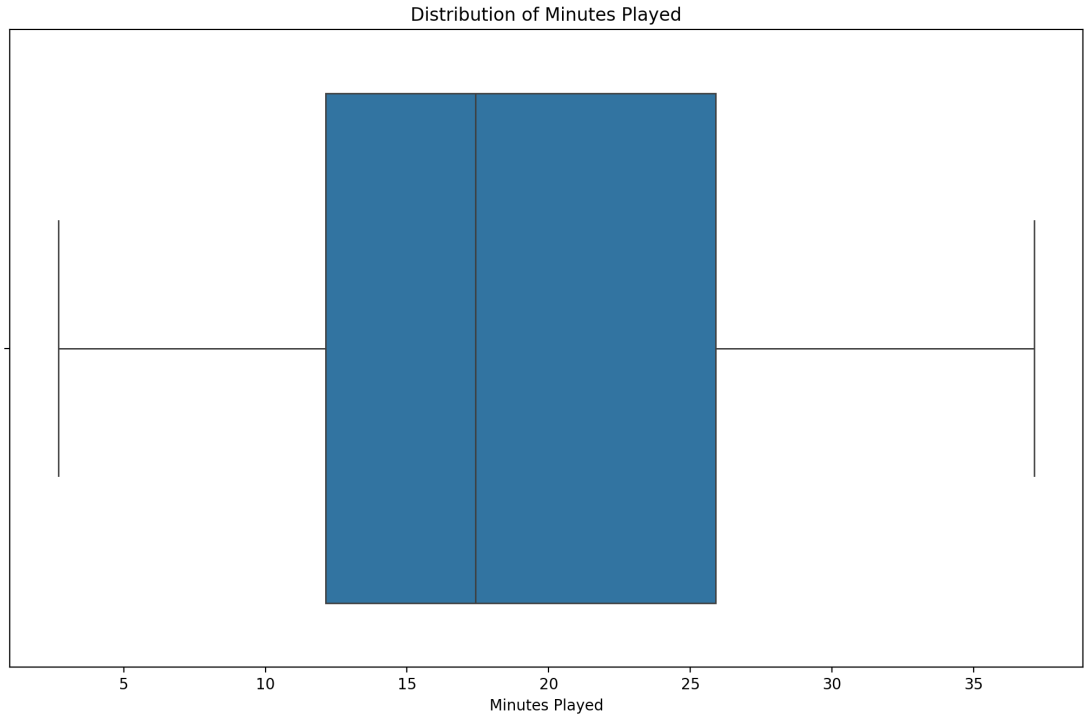
Graph 4


Feature Importance (Random Forest)
FGA
PTS_per_MP
FGA_efficiency
MP
TRB
AST
STL
Feature
0.0
0.2
0.4
0.6
0.8
Importance

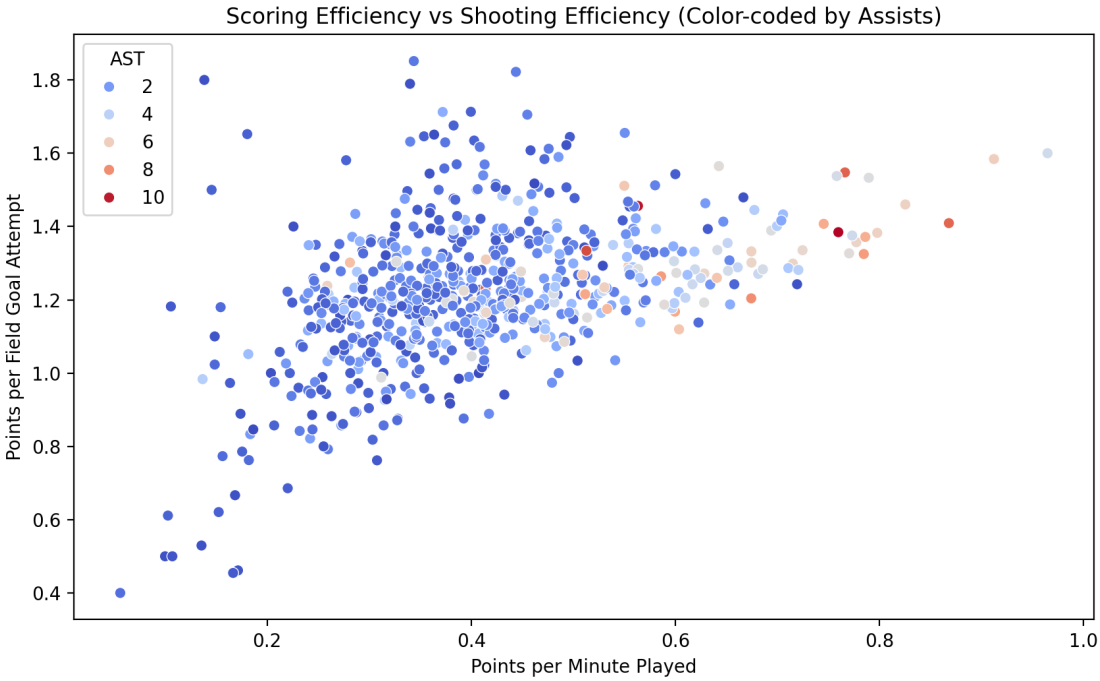
Graph 5


Clusters Visualized using PCA
Cluster
0
1
2
Principal Component 2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−20
−10
0
10
20
30
Principal Component 1

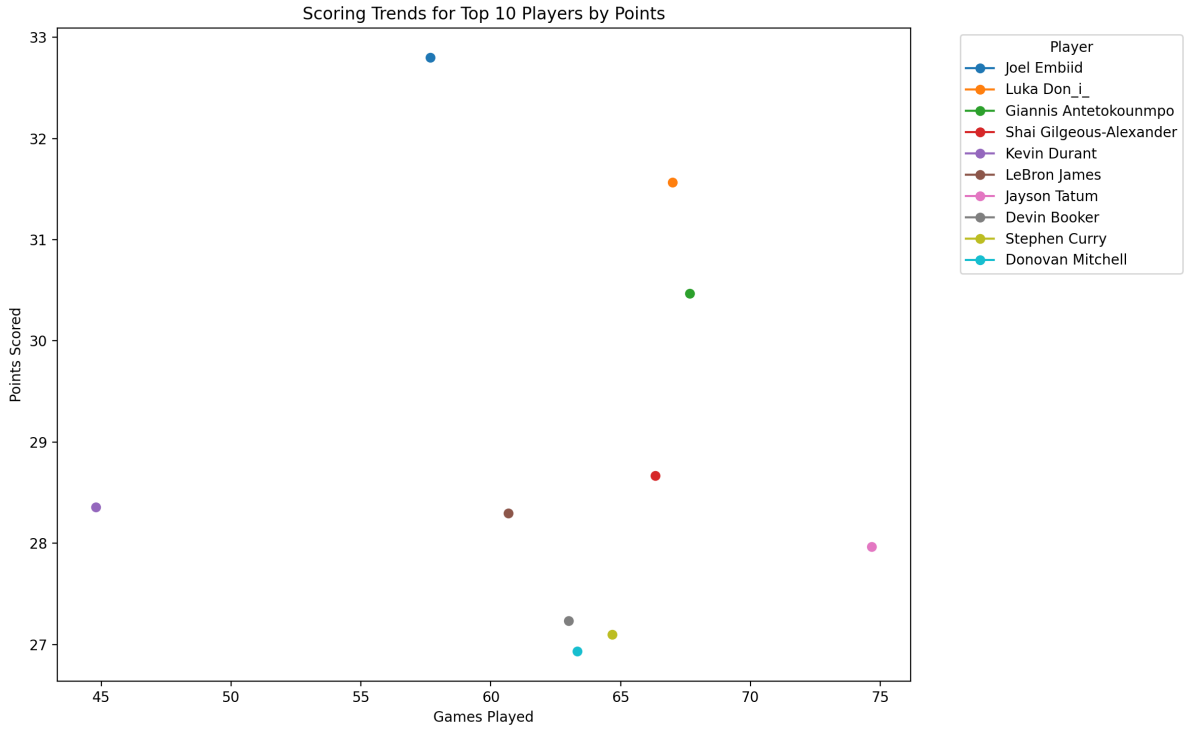
Graph 6


Distribution of Minutes Played
5
10
15
20
25
30
35
Minutes Played

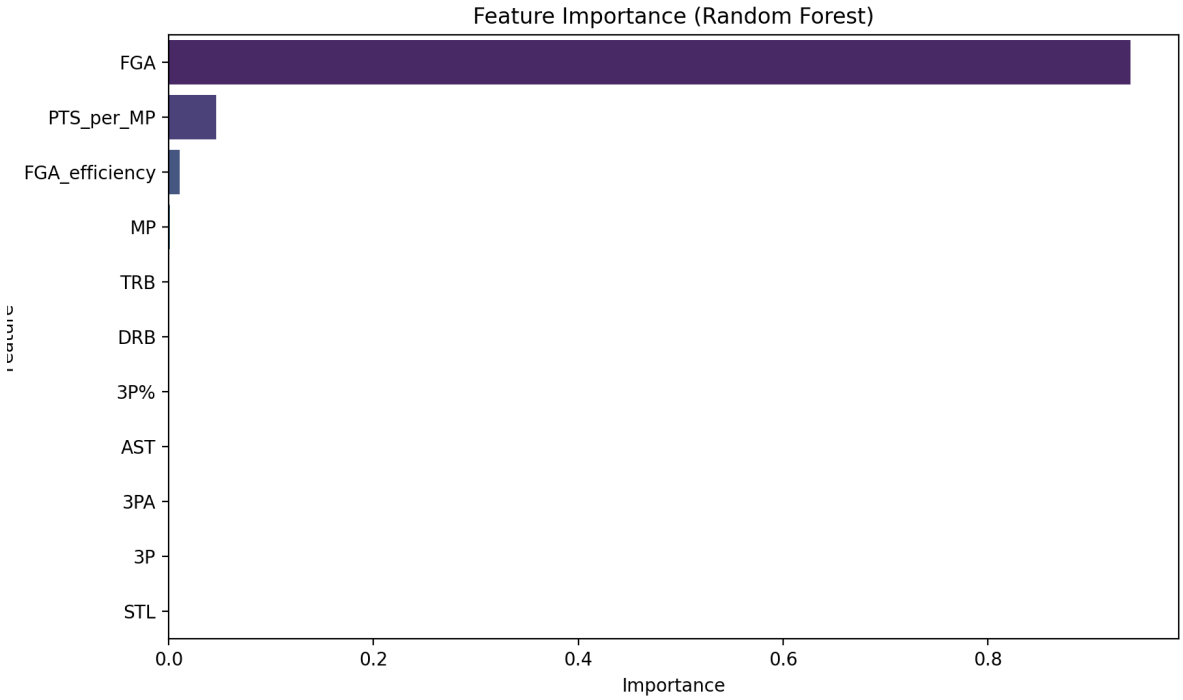
Graph 7


Scoring Efficiency vs Shooting Efficiency (Color-coded by Assists)
AST
2
4
6
8
10
Points per Field Goal Attempt
Points per Minute Played
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0.2
0.4
0.6
0.8
1.0

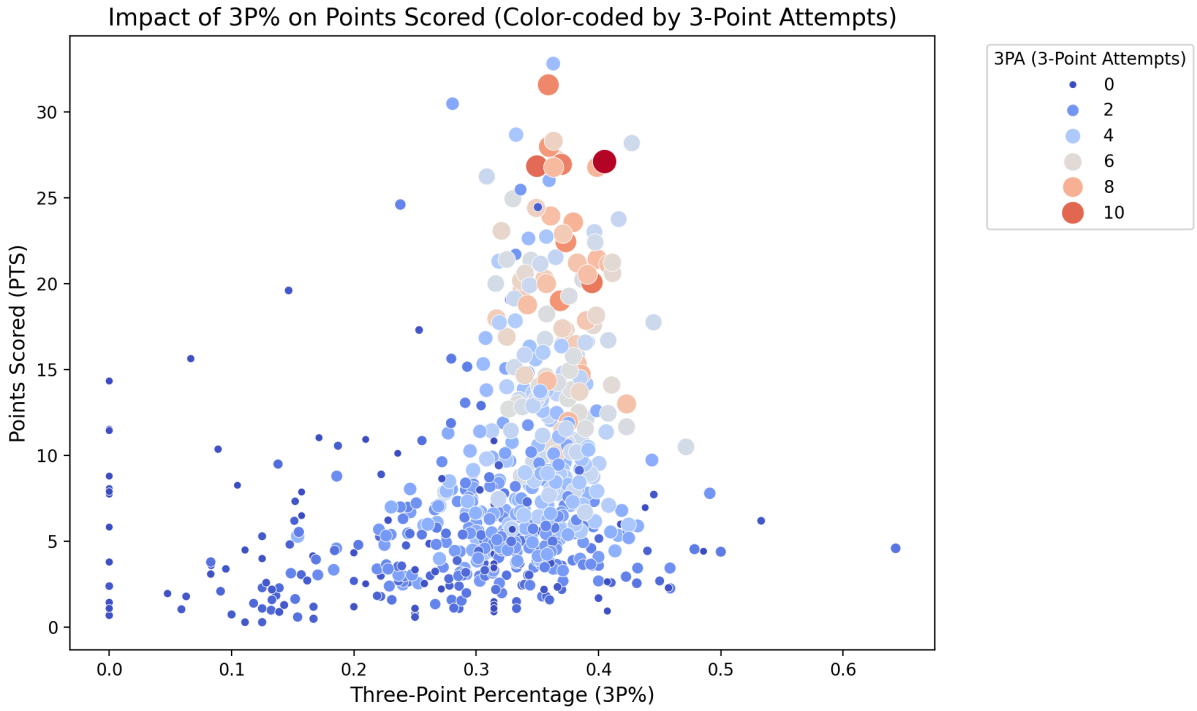
Graph 8


Scoring Trends for Top 10 Players by Points
Points Scored
Games Played
33
32
31
30
29
28
27
45
50
55
60
65
70
75
Player
Joel Embiid
Luka Don_i_
Giannis Antetokounmpo
Shai Gilgeous-Alexander
Kevin Durant
LeBron James
Jayson Tatum
Devin Booker
Stephen Curry
Donovan Mitchell

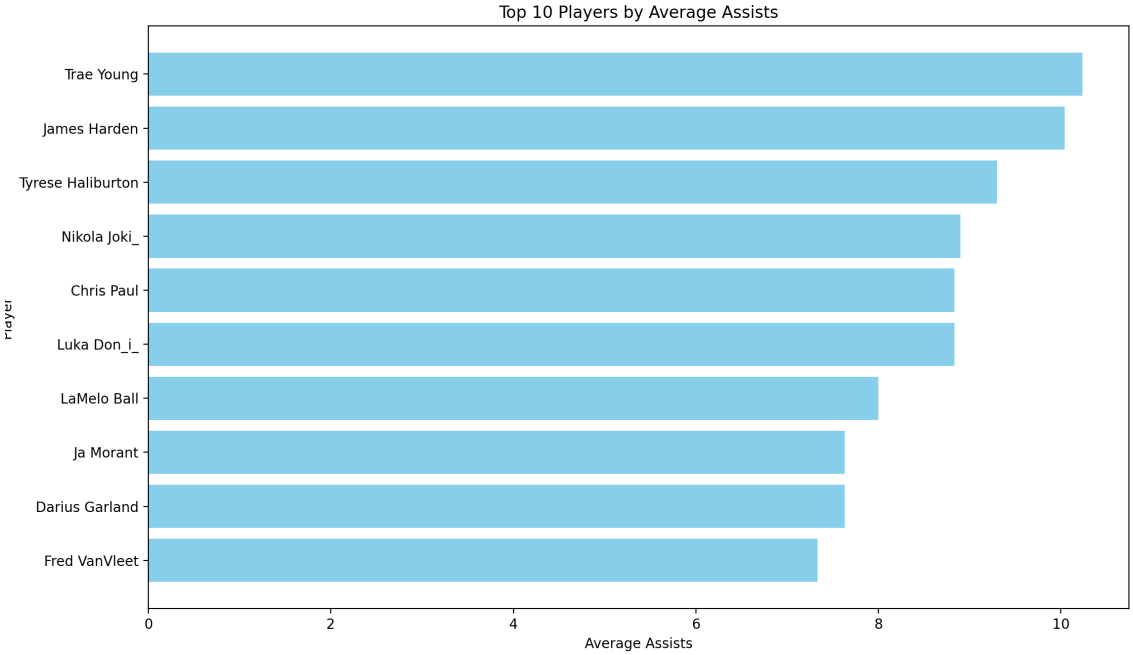
Graph 9
Visualisations from initial model:
Feature importance including Defensive stats such as Steals (STL) and Defensive Rebounds (DRB). Also, including 3 point statistics 3PA, 3P% and 3P



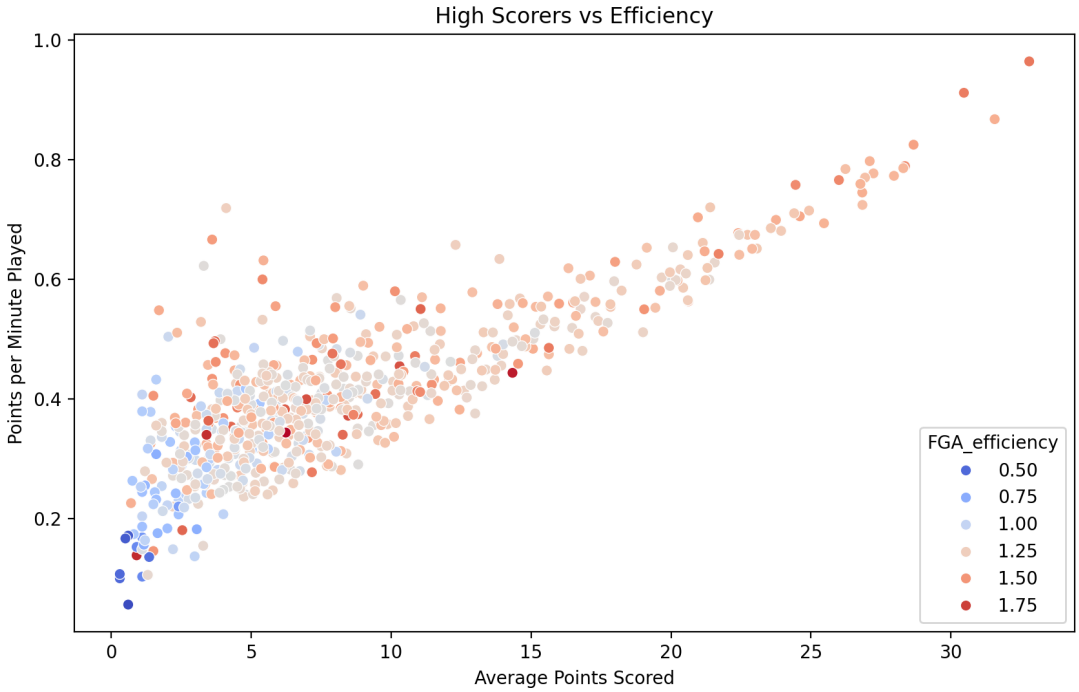
Graph 10



Graph 11



Graph 12



Graph 13