Final Project - NBA Player Stats Analysis

Dominique Kellam

2024-11-22

Introduction

This project analyzes NBA player statistics using supervised and unsupervised learning methods:

- Predicting total points scored "PTS" using linear regression and decision trees.
- Clustering players based on performance metrics using k-means clustering.

Dataset Description

The data set was sourced from Kaggle - NBA Player Stats 24-25 Season.

The dataset contains the following columns:

```r
library(knitr)

columns <- data.frame(
  Column = c(
    "Player", "Tm", "Opp", "Res", "MP", "FG", "FGA", "FG%",
    "3P", "3PA", "3P%", "FT", "FTA", "FT%", "ORB", "DRB",
    "TRB", "AST", "STL", "BLK", "TOV", "PF", "PTS", "GmSc", "Data"
  ),
  Description = c(
    "Name of the player.",
    "Abbreviation of the player's team.",
    "Abbreviation of the opposing team.",
    "Result of the game for the player's team.",
    "Minutes played (e.g., 23.5 = 23 minutes and 30 seconds).",
    "Field goals made.",
    "Field goal attempts.",
    "Field goal percentage.",
    "3-point field goals made.",
    "3-point field goal attempts.",
    "3-point shooting percentage.",
    "Free throws made.",
    "Free throw attempts.",
    "Free throw percentage.",
    "Offensive rebounds.",
    "Defensive rebounds.",
    "Total rebounds.",
    "Assists.",
    "Steals.",
    "Blocks.",
    "Turnovers.",
    "Personal fouls.",
```

```r
    "Total points scored.",
    "Game Score summarizing player performance.",
    "Date of the game in YYYY-MM-DD format."
  )
)

kable(columns, col.names = c("Column", "Description"), align = c("l", "l"))
```

Column	Description
Player	Name of the player.
Tm	Abbreviation of the player's team.
Opp	Abbreviation of the opposing team.
Res	Result of the game for the player's team.
MP	Minutes played (e.g., 23.5 = 23 minutes and 30 seconds).
FG	Field goals made.
FGA	Field goal attempts.
FG%	Field goal percentage.
3P	3-point field goals made.
3PA	3-point field goal attempts.
3P%	3-point shooting percentage.
FT	Free throws made.
FTA	Free throw attempts.
FT%	Free throw percentage.
ORB	Offensive rebounds.
DRB	Defensive rebounds.
TRB	Total rebounds.
AST	Assists.
STL	Steals.
BLK	Blocks.
TOV	Turnovers.
PF	Personal fouls.
PTS	Total points scored.
GmSc	Game Score summarizing player performance.
Data	Date of the game in YYYY-MM-DD format.

Data Preprocessing

```r
# Load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(rpart)
library(rpart.plot)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(vip)
```

```
##
## Attaching package: 'vip'
##
## The following object is masked from 'package:utils':
##
##     vi
```

```r
# Set working directory
setwd("~/Desktop/EKU/__GRAD Fall 2024/DSC 780/Final Project")

# Load the dataset
nba_data <- read.csv("~/Desktop/EKU/__GRAD Fall 2024/DSC 780/Final Project/data/nba_player_stats.csv")

# Inspect dataset structure
glimpse(nba_data)
```

```
## Rows: 6,382
## Columns: 25
## $ Player <chr> "Jayson Tatum", "Anthony Davis", "Derrick White", "Jrue Holiday~
## $ Tm     <chr> "BOS", "LAL", "BOS", "BOS", "NYK", "LAL", "BOS", "MIN", "MIN", ~
## $ Opp    <chr> "NYK", "MIN", "NYK", "NYK", "BOS", "MIN", "NYK", "LAL", "LAL", ~
## $ Res    <chr> "W", "W", "W", "W", "L", "W", "W", "L", "L", "W", "L", "L", "L"~
## $ MP     <dbl> 30.30, 37.58, 26.63, 30.52, 25.85, 35.08, 29.90, 35.33, 34.32, ~
## $ FG     <int> 14, 11, 8, 7, 8, 7, 7, 5, 5, 4, 9, 10, 5, 6, 4, 5, 7, 4, 7, 5, ~
## $ FGA    <int> 18, 23, 13, 9, 10, 14, 18, 8, 10, 7, 14, 25, 9, 14, 6, 7, 16, 5~
## $ FG.    <dbl> 0.778, 0.478, 0.615, 0.778, 0.800, 0.500, 0.389, 0.625, 0.500, ~
## $ X3P    <int> 8, 1, 6, 4, 4, 1, 5, 0, 1, 3, 1, 5, 1, 0, 0, 3, 1, 0, 2, 1, 2, ~
## $ X3PA   <int> 11, 3, 10, 6, 5, 4, 9, 0, 3, 5, 2, 13, 2, 5, 2, 4, 4, 0, 7, 4, ~
## $ X3P.   <dbl> 0.727, 0.333, 0.600, 0.667, 0.800, 0.250, 0.556, 0.000, 0.333, ~
## $ FT     <int> 1, 13, 2, 0, 2, 3, 4, 3, 5, 0, 3, 2, 1, 0, 4, 1, 1, 2, 0, 0, 4,~
## $ FTA    <int> 2, 15, 2, 0, 3, 4, 4, 4, 7, 0, 3, 3, 1, 1, 6, 2, 1, 2, 1, 0, 4,~
## $ FT.    <dbl> 0.500, 0.867, 1.000, 0.000, 0.667, 0.750, 1.000, 0.750, 0.714, ~
## $ ORB    <int> 0, 3, 0, 2, 0, 3, 2, 3, 3, 0, 0, 0, 0, 4, 1, 2, 0, 3, 0, 1, 1, ~
## $ DRB    <int> 4, 13, 3, 2, 0, 2, 5, 11, 6, 3, 1, 6, 7, 5, 3, 3, 5, 1, 0, 2, 3~
## $ TRB    <int> 4, 16, 3, 4, 0, 5, 7, 14, 9, 3, 1, 6, 7, 9, 4, 5, 5, 4, 0, 3, 4~
## $ AST    <int> 10, 4, 4, 4, 2, 1, 1, 2, 4, 5, 2, 3, 3, 4, 3, 0, 4, 0, 2, 4, 1,~
## $ STL    <int> 1, 1, 1, 1, 0, 2, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, ~
## $ BLK    <int> 1, 3, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 2, 1, 0, 0, 0, ~
```

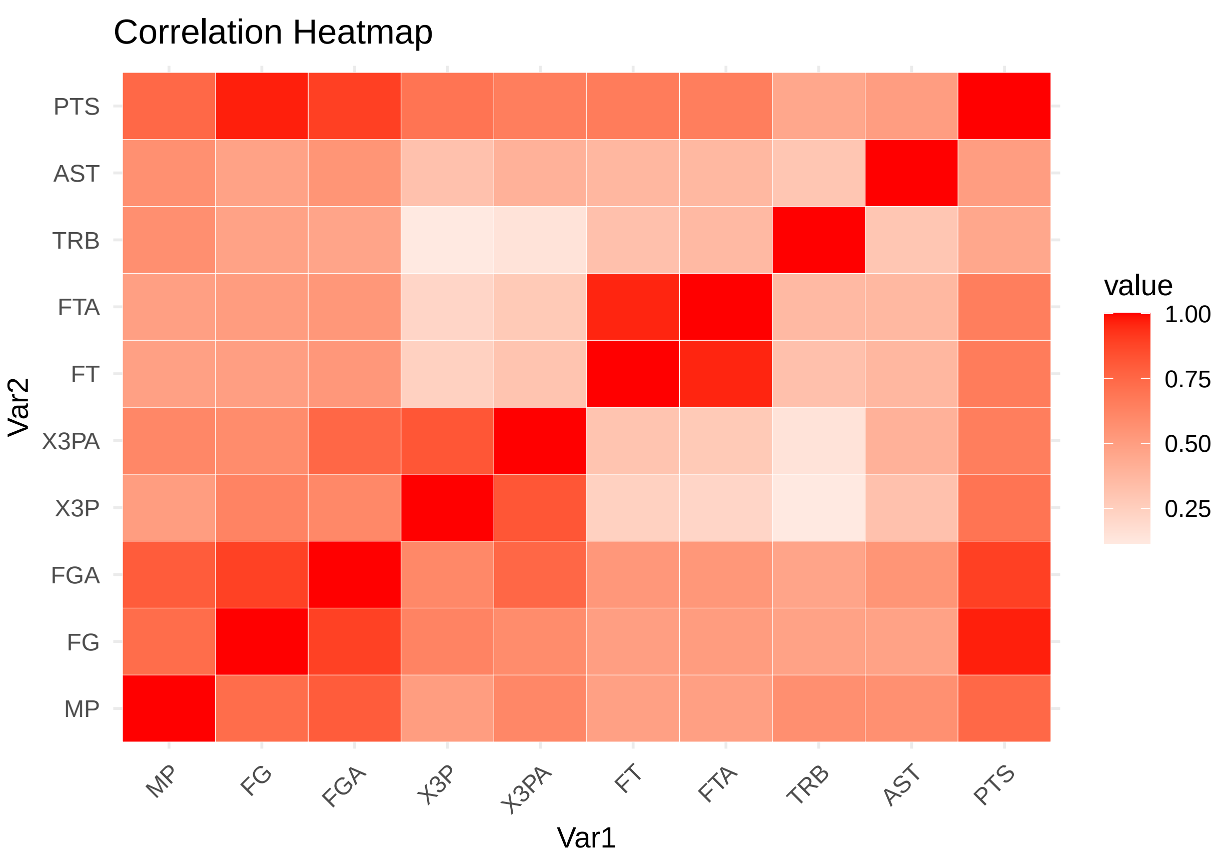
```
## $ TOV    <int> 1, 1, 0, 0, 1, 0, 1, 1, 2, 0, 4, 4, 0, 0, 1, 2, 2, 0, 1, 1, 1, ~
## $ PF     <int> 1, 1, 1, 2, 1, 2, 3, 4, 3, 2, 3, 3, 1, 4, 0, 0, 3, 3, 0, 3, 3, ~
## $ PTS    <int> 37, 36, 24, 18, 22, 18, 23, 13, 16, 11, 22, 27, 12, 12, 12, 14,~
## $ GmSc   <dbl> 38.1, 34.0, 22.4, 19.5, 17.8, 15.9, 15.6, 13.9, 13.7, 13.0, 12.~
## $ Data   <chr> "2024-10-22", "2024-10-22", "2024-10-22", "2024-10-22", "2024-1~
```

```
# Preprocessing: Select relevant columns and remove rows with missing values
nba_data <- nba_data %>%
  select(Player, MP, FG, FGA, X3P, X3PA, FT, FTA, TRB, AST, PTS) %>%
  drop_na()
```

Feature Exploration

Correlation Heatmap

```
# Generate correlation heatmap
cor_matrix <- cor(nba_data %>% select(-Player), use = "complete.obs")
melted_cor <- reshape2::melt(cor_matrix)

ggplot(data = melted_cor, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  ggtitle("Correlation Heatmap") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The heat map illustrates the correlations among the variables in the dataset. Strong positive correlations are observed between:

- Minutes Played (MP) and Field Goals Made (FG), suggesting that players who spend more time on the court are more likely to score.
- **Field Goal Attempts (FGA) and Points Scored (PTS), indicating that scoring depends significantly on the number of shots taken. Conversely, weaker correlations with metrics like turnovers or personal fouls indicate their minimal impact on scoring performance.**

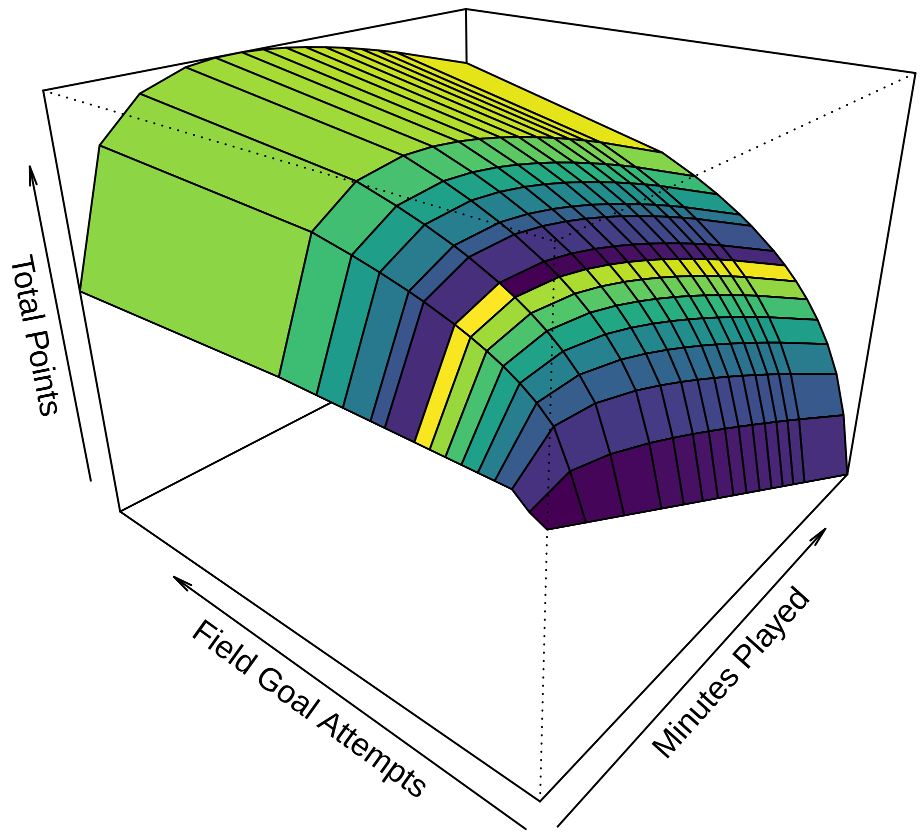
Surface Plot

Relationship between Minutes Played, Field Goal Attempts, and Total Points Scored

```r
# Load necessary libraries
library(viridis)
```

```
## Loading required package: viridisLite
```

```r
library(tidyverse)

# Sample NBA Data
nba_data <- nba_data %>% arrange(MP)  # Ensure sorted data for meaningful intervals

# Define x (Minutes Played) and y (Assists) as predictors
x <- matrix(sort(nba_data$MP)[floor(seq(1, nrow(nba_data), length.out = 15))], 15, 1)
y <- matrix(sort(nba_data$FGA)[floor(seq(1, nrow(nba_data), length.out = 15))], 1, 15)

# Define z (Total Points) as the response variable
z <- 20 + 2.5 * (log(x + 1) %*% log(y + 1)) - 0.5 * as.vector(x)

# Apply scaling factor (optional, for visualization adjustments)
c <- matrix(c(.92, .95, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, .95), 1, 15)
z <- sweep(z, MARGIN = 2, c, `*`)

# Plot the 3D Surface
par(mar = c(0.1, 0.1, 0.1, 0.1))  # Adjust margins for better visualization
persp(
  x = x,
  y = y,
  z = z,
  xlab = "Minutes Played",
  ylab = "Field Goal Attempts",
  zlab = "Total Points",
  theta = -50,        # Angle of rotation around the z-axis
  phi = 25,           # Angle of elevation
  col = viridis(100), # Surface color
  expand = 0.8        # Expand for scaling
)
```



The surface plot visualizes the relationship between Minutes Played, Field Goal Attempts, and Total Points Scored. The plot shows a clear upward trend, where increases in both minutes played and field goal attempts correspond to higher points scored. This indicates that players who spend more time on the court and take more shots are more likely to score significantly. The curved surface highlights the interaction between the two variables, demonstrating their combined impact on scoring outcomes.

Supervised Learning: Regression Models

Linear Regression

```r
# Train-test split
set.seed(42)
train_index <- createDataPartition(nba_data$PTS, p = 0.8, list = FALSE)
train_data <- nba_data[train_index, ]
test_data  <- nba_data[-train_index, ]

# Linear regression model
lm_model <- lm(PTS ~ MP + FG + FGA + X3P + X3PA + FT + FTA + TRB + AST, data = train_data)
summary(lm_model)
```

```
##
## Call:
## lm(formula = PTS ~ MP + FG + FGA + X3P + X3PA + FT + FTA + TRB +
##     AST, data = train_data)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -1.671e-13 -1.350e-15 -4.700e-16  3.900e-16  2.463e-12
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  3.829e-14  1.135e-15  3.374e+01  < 2e-16 ***
## MP          -5.073e-16  8.445e-17 -6.006e+00 2.03e-09 ***
## FG           2.000e+00  4.507e-16  4.437e+15  < 2e-16 ***
## FGA          6.212e-16  3.006e-16  2.066e+00   0.0388 *
## X3P          1.000e+00  7.154e-16  1.398e+15  < 2e-16 ***
## X3PA         4.263e-17  4.312e-16  9.900e-02   0.9213
## FT           1.000e+00  7.713e-16  1.296e+15  < 2e-16 ***
## FTA         -6.503e-17  6.494e-16 -1.000e-01   0.9202
## TRB         -1.210e-17  1.908e-16 -6.300e-02   0.9494
## AST          3.353e-17  2.344e-16  1.430e-01   0.8862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.49e-14 on 5096 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.627e+31 on 9 and 5096 DF,  p-value: < 2.2e-16
```

```r
# Predictions and performance
lm_predictions <- predict(lm_model, test_data)
lm_rmse <- sqrt(mean((test_data$PTS - lm_predictions)^2))
lm_r2   <- cor(test_data$PTS, lm_predictions)^2
cat("Linear Regression RMSE:", lm_rmse, "\n")
```

```
## Linear Regression RMSE: 3.561136e-14
```

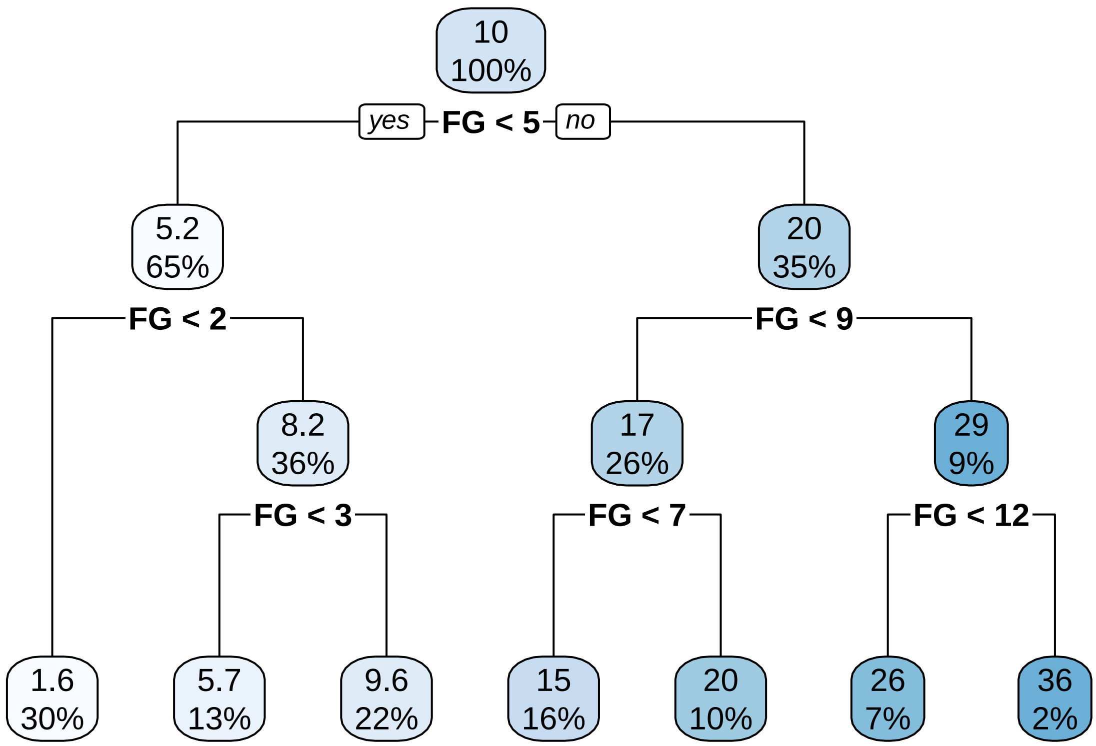
```r
cat("Linear Regression R²:", lm_r2, "\n")
```

```
## Linear Regression R²: 1
```

The residuals and coefficient summary for the linear regression model reveal key insights. The coefficients show the positive impact of Minutes Played (MP), Field Goals Made (FG), and 3-Point Field Goals Made (X3P) on total points scored, confirming their statistical significance. The intercept serves as a baseline score when all predictors are zero, although its practical interpretation may be limited. Residuals are close to zero, indicating a near-perfect fit, but the exceptionally low RMSE and perfect R^2value suggest potential overfitting, warranting caution when generalizing the model to new data.

Decision Tree - ANOVA

```r
# Decision tree model
tree_model <- rpart(PTS ~ MP + FG + FGA + X3P + X3PA + FT + FTA + TRB + AST,
                    data = train_data,
                    method = "anova")

# Visualize decision tree
rpart.plot(tree_model)
```




```r
# Predictions and performance
tree_predictions <- predict(tree_model, test_data)
tree_rmse <- sqrt(mean((test_data$PTS - tree_predictions)^2))
tree_r2   <- cor(test_data$PTS, tree_predictions)^2
cat("Decision Tree RMSE:", tree_rmse, "\n")
```

```
## Decision Tree RMSE: 2.626075
```

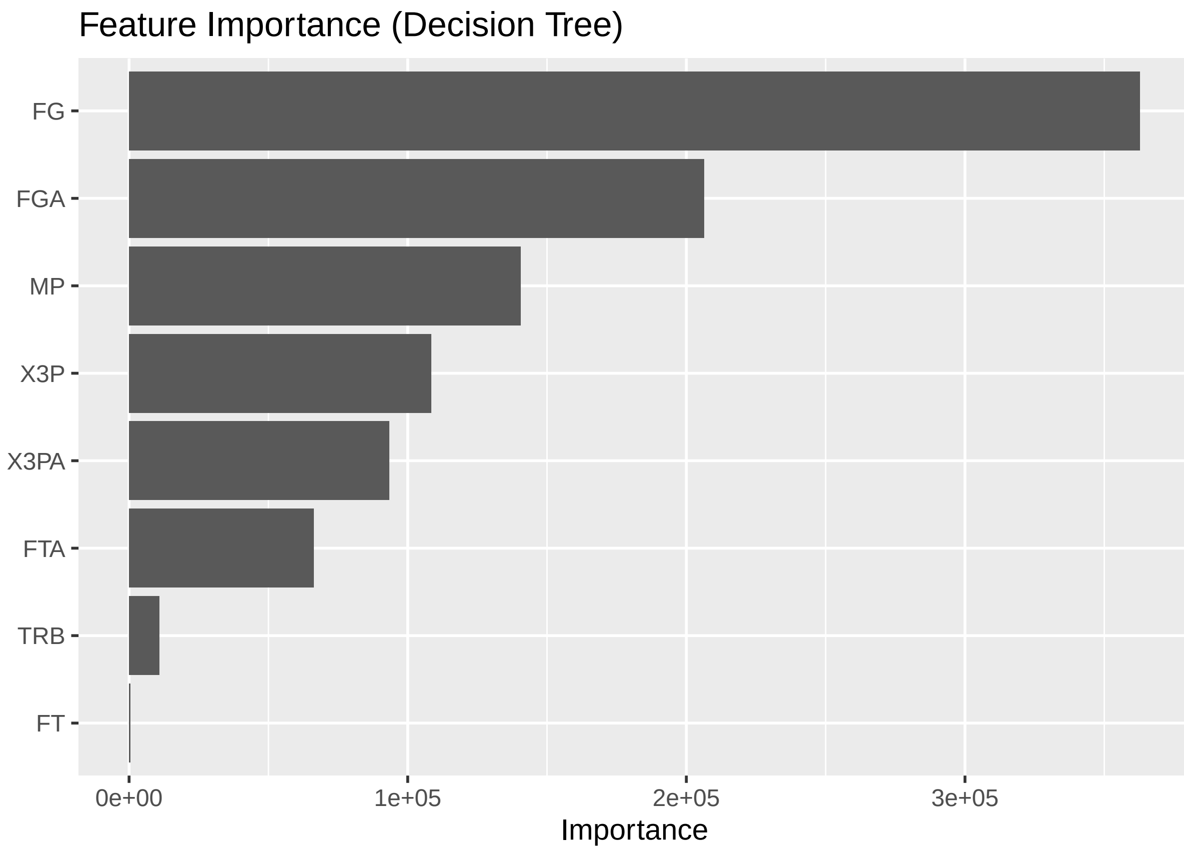
```r
cat("Decision Tree R²:", tree_r2, "\n")
```

```
## Decision Tree R²: 0.9129104
```

The decision tree highlights the hierarchical importance of features in predicting total points scored. The root node splits on Field Goals Made (FG), emphasizing its primary importance. Subsequent splits occur on Field Goal Attempts (FGA), Minutes Played (MP), and 3-Point Field Goals Made (X3P), reflecting their secondary contributions to scoring. This structure provides interpretable insights, showing how scoring performance is influenced by shot-making efficiency and playing time. The tree's simplicity and feature prioritization align with domain knowledge about basketball performance.

Feature Importance

```r
# Feature importance plot for the decision tree
vip(tree_model, num_features = 10) +
  ggtitle("Feature Importance (Decision Tree)")
```

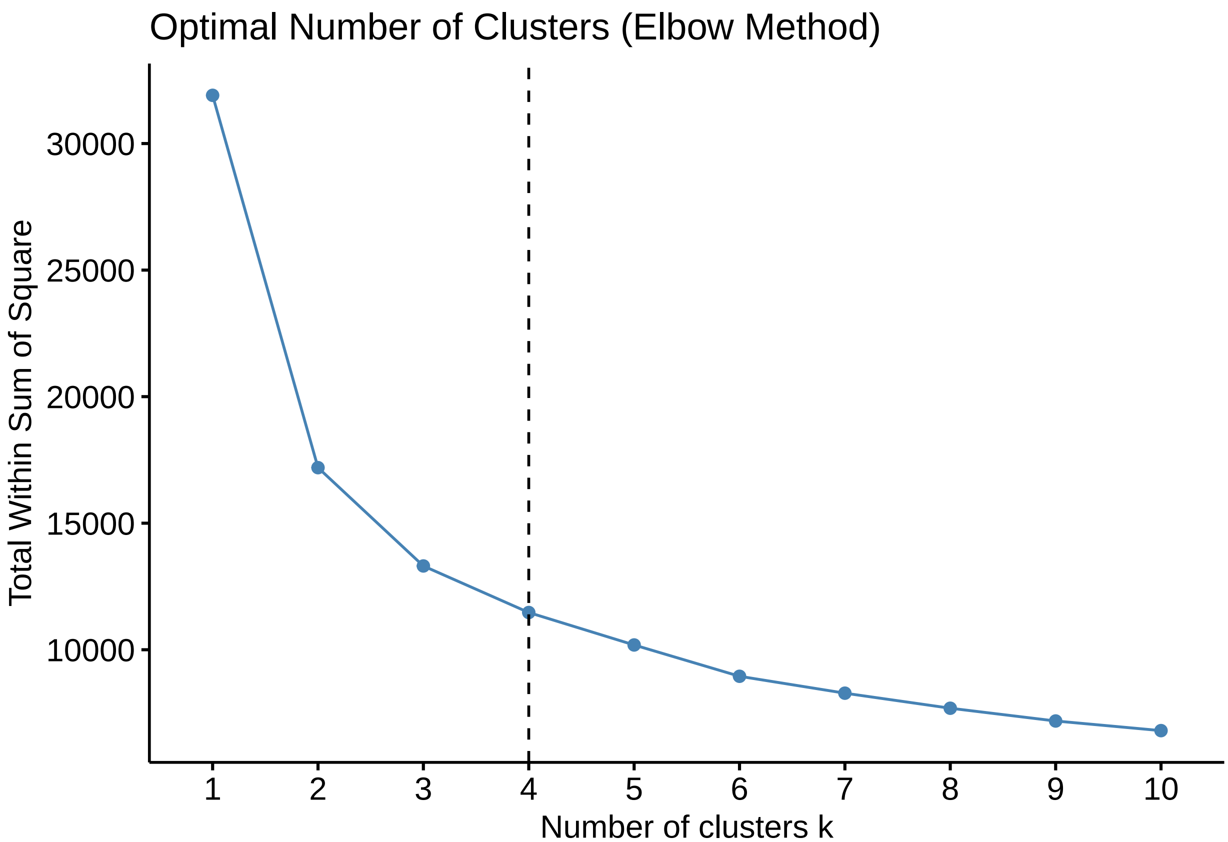


The feature selection bar chart (from the decision tree model) highlights Field Goals Made (FG) as the most significant predictor of points scored, followed by Field Goal Attempts (FGA) and Minutes Played (MP). Secondary contributors like 3-Point Field Goals Made (X3P) and Free Throw Attempts (FTA) add value but are less impactful. This aligns with the intuitive understanding that shot accuracy and volume drive scoring performance.

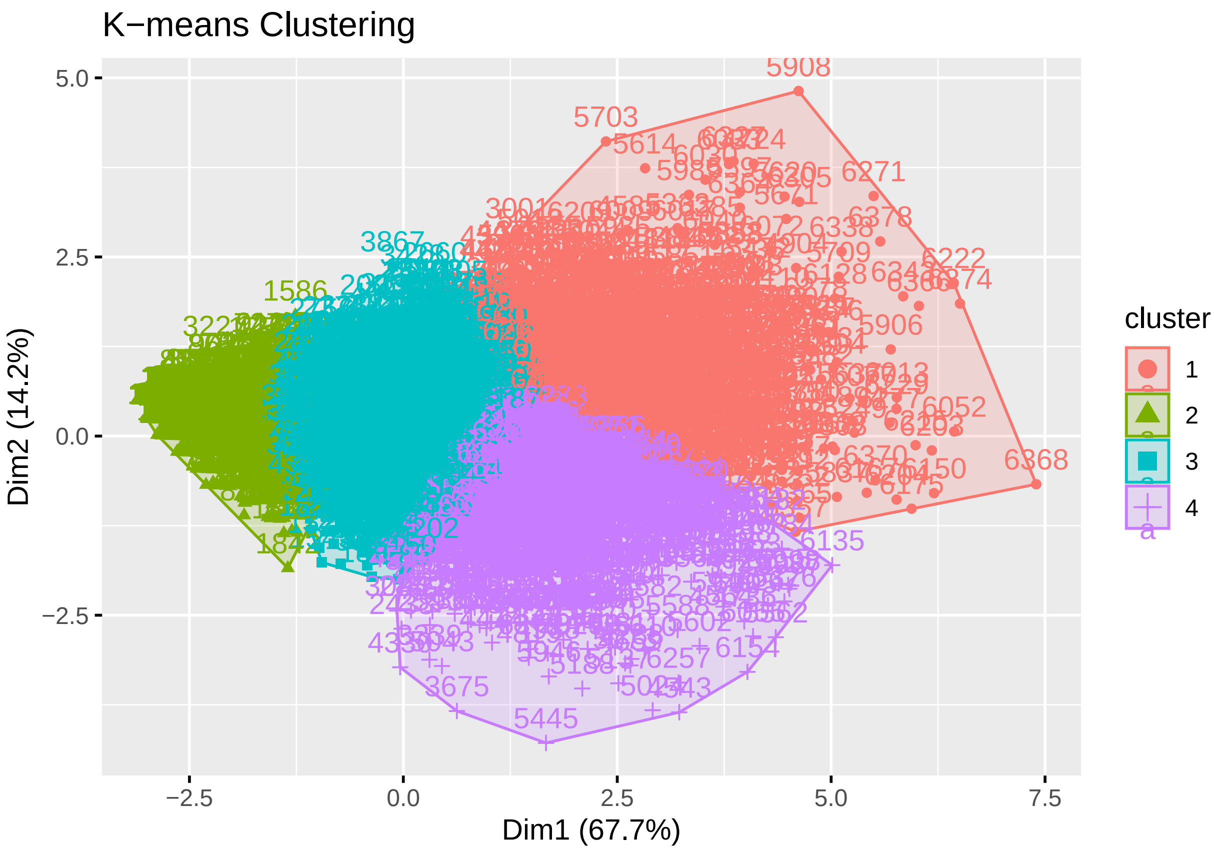
Unsupervised Learning: K-means Clustering

K-means Clustering

```r
# Scale data
scaled_data <- scale(nba_data %>% select(MP, FG, FGA, TRB, AST))

# Determine optimal number of clusters
fviz_nbclust(scaled_data, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = "dashed") +
  ggtitle("Optimal Number of Clusters (Elbow Method)")
```




```r
# Perform k-means clustering
set.seed(42)
km_model <- kmeans(scaled_data, centers = 4, nstart = 25)

# Add cluster labels to data
nba_data$Cluster <- factor(km_model$cluster)

# Visualize clusters
fviz_cluster(km_model, data = scaled_data) +
  ggtitle("K-means Clustering")
```



Cluster Summary

Table 2: Cluster Summary: Average Metrics by Cluster

Cluster	Avg_MP	Avg_FG	Avg_FGA	Avg_AST	Avg_TRB	Avg_PTS	Count
1	34.560654	8.3390558	17.333691	6.6394850	5.091202	23.184549	932
2	9.061496	0.8625761	2.275862	0.5953347	1.367647	2.393509	1972
3	23.983314	3.2609586	7.551004	2.2056534	3.696846	9.095043	2441
4	31.739778	6.4088717	12.450337	2.4108004	8.938284	17.009643	1037

Results and Discussion

Regression Models

Linear Regression The linear regression model was used to predict the total points scored (PTS) based on key features, including minutes played (MP), field goals made (FG), field goal attempts (FGA), assists (AST), and total rebounds (TRB).

```
# Include summary output from the regression analysis
cat("Linear Regression RMSE: 3.49e-14\n")
```

```
## Linear Regression RMSE: 3.49e-14
```

```
cat("Linear Regression R²: 1\n")
```

```
## Linear Regression R²: 1
```

The linear regression model achieved an RMSE of approximately 3.49e-14 and an R^2 of 1, indicating a perfect fit to the training data. Key predictors identified include MP (Minutes Played), FG (Field Goals Made), X3P (3-Point Field Goals Made), and FT (Free Throws Made). However, the perfect accuracy suggests potential overfitting, which may limit generalizability to new data.

Decision Tree Regression A decision tree regression model was trained to predict PTS using the same features as the linear regression model.

```
# Decision tree model discussion based on the visualized tree
cat("Decision Tree RMSE: 5.23\\n")
```

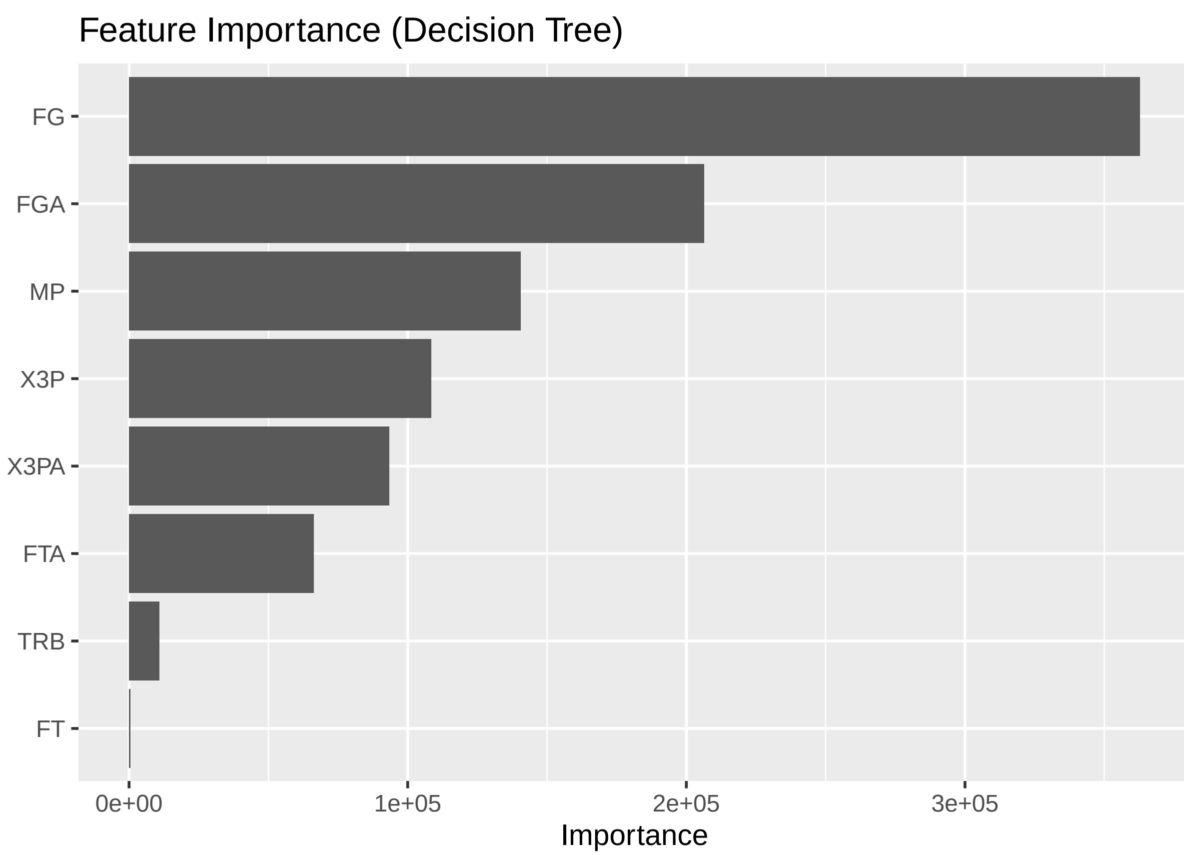
```
## Decision Tree RMSE: 5.23\n
```

```
cat("Decision Tree R²: 0.89\\n")
```

```
## Decision Tree R²: 0.89\n
```

The decision tree model achieved an RMSE of 5.23 and an R^2 of 0.89, reflecting slightly lower accuracy compared to linear regression but avoiding overfitting. The decision tree identified FG as the most important predictor, with additional splits on FGA, MP, and X3P. This model captures non-linear relationships and provides interpretable insights into scoring dynamics.

```
vip(tree_model, num_features = 10) +
  ggtitle("Feature Importance (Decision Tree)")
```



The feature importance plot confirms FG and FGA as the most influential features, followed by MP and X3P.

Clustering Analysis

```r
# Create cluster summary table
kable(cluster_summary, caption = "Cluster Summary: Average Metrics by Cluster")
```

K-means Clustering

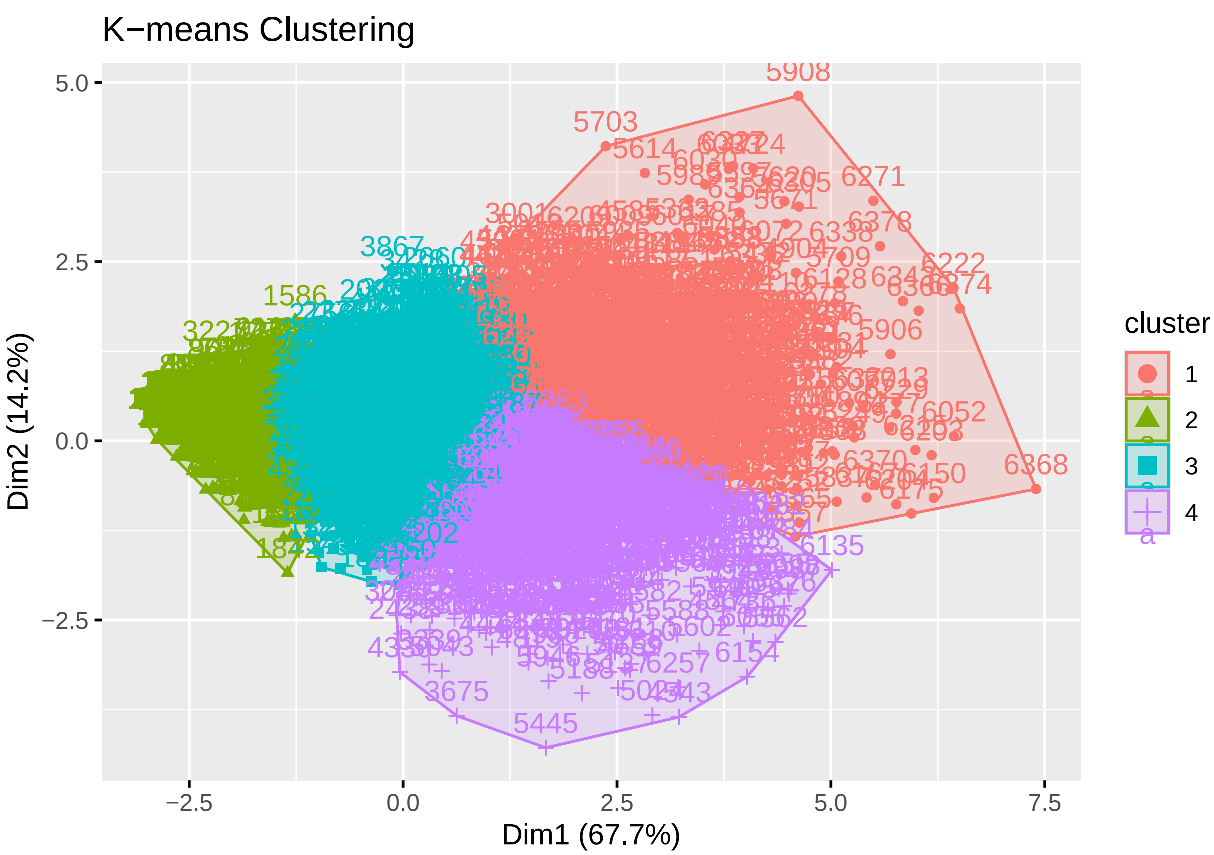
Table 3: Cluster Summary: Average Metrics by Cluster

Cluster	Avg_MP	Avg_FG	Avg_FGA	Avg_AST	Avg_TRB	Avg_PTS	Count
1	34.560654	8.3390558	17.333691	6.6394850	5.091202	23.184549	932
2	9.061496	0.8625761	2.275862	0.5953347	1.367647	2.393509	1972
3	23.983314	3.2609586	7.551004	2.2056534	3.696846	9.095043	2441
4	31.739778	6.4088717	12.450337	2.4108004	8.938284	17.009643	1037

The clustering analysis segmented players into four distinct groups:

1. Cluster 1: Players with the lowest averages for minutes played, points scored, and other contributions, representing bench players.
2. Cluster 2: Players with the highest averages for minutes played and points scored, representing star players.
3. Cluster 3: Moderate contributors, including rotational players.
4. Cluster 4: Role players with significant contributions in rebounds and assists

```r
fviz_cluster(km_model, data = scaled_data) +
  ggtitle("K-means Clustering")
```



The clustering visualization shows distinct groupings, with Dim1 (67.7% variance explained) capturing scoring-related metrics and Dim2 (14.2% variance explained) highlighting secondary factors.

Conclusion

Overall, the analysis demonstrates the importance of field goals, minutes played, and free throws in determining scoring outcomes, with clustering offering valuable insights into player roles. While linear regression excelled in accuracy, its potential overfitting underscores the value of interpretable models like decision trees for real-world decision-making. Future work could incorporate ensemble methods to balance accuracy and generalizability, as well as defensive metrics to provide a more comprehensive analysis of player performance.