STAT 107 Final Project: NBA Team Statistics and Winning

Team 24: Derek de Gracia, Shafin Kazi, Tiffany Huang, Tyler Wong

Due: 2025-12-05

Abstract: Analytics and data-driven metrics have transformed modern NBA strategy, influencing roster decisions, offensive schemes, and defensive priorities. While many statistics are tracked in every game, not all contribute equally to winning. This study examines which team-level offensive and defensive statistics are most strongly associated with winning NBA games during the 2024–25 season.

Using cleaned box score data transformed into team aggregates, we evaluate the effect of shooting efficiency, three-point performance, free throws, defensive pressure (blocks + steals), and total points. We apply hypothesis testing, unsupervised clustering, and logistic regression to measure statistical significance, group team styles, and predict game outcomes. Results show that shooting efficiency (FG%) and defensive pressure are highly predictive of winning, while three-point accuracy contributes strongly but less consistently across teams. Logistic regression confirms that field-goal percentage and total points hold the highest predictive power with above-average classification accuracy.

This analysis provides insights useful to coaches, analysts, and scouting teams seeking to optimize roster construction and during-game strategy in the data-driven era of the NBA.

This report was made in conjunction with Derek De Gracia, Shafin Kazi, Tiffany Huang, Tyler Wong for STAT107 at UCR taught by Jose Sanchez Gomez. The repository is hosted by Github under https://github.com/shafinkazi/STAT107_Team-24_NBA_Analysis.git (https://github.com/shafinkazi/STAT107_Team-24_NBA_Analysis.git)

1. Introduction

The modern NBA is shaped heavily by analytics, efficiency, and data-driven decision-making. Teams increasingly rely on statistical insights to evaluate player performance, optimize game strategies, and gain competitive advantages. Among the many questions that arise in basketball analytics, one of the most fundamental is:

What statistical skills or attributes actually lead to winning games?

Although basketball is a team sport, game outcomes ultimately reflect patterns in offensive and defensive performance. Understanding which team-level statistics correlate most strongly with winning can benefit coaches, analysts, and front offices by highlighting which areas contribute most to success.

In this project, we analyze all team statistics from the 2024–25 NBA season to identify how offensive skills (such as 3-point shooting, assists, and field-goal efficiency) and defensive skills (such as steals, blocks, and turnovers) relate to game outcomes. By transforming player-level game logs into team-level summaries, we

can compare performances across wins and losses and study which metrics are most predictive of success.

Research Question

Which team-level offensive and defensive statistics are most strongly associated with winning NBA games in the 2024–25 season?

Methods Overview

To answer this question, we apply multiple statistical tools, including: - Exploratory Data Analysis - Two-sample t-tests to compare offense vs. defense metrics between wins and losses - K-means clustering to group team performance styles - Logistic regression to predict game outcomes from key features by turning the variables into binary conditions of win or lose

This multi-method approach allows us to evaluate individual statistical differences, build predictive models, and examine broader team-performance patterns.

2. Data

The data used for this analysis was sourced from Kaggle, a popular platform for data sharing that provides publicly available NBA teams' data in their playoffs. The dataset contains player-level box scores for NBA games during the 2024–25 season. Each row represents a single player's performance in one game and includes the offensive and defensive statistics used in our analysis:

Offensive metrics: FG (field goals), FGA (field goal attempts), FG%, 3P (three-pointers made), 3PA (three-point attempts), 3P% (three-point percentage), AST (assists), and PTS (points).

Defensive metrics: STL (steals), BLK (blocks), TOV (turnovers), DRB (defensive rebounds), and TRB (total rebounds).

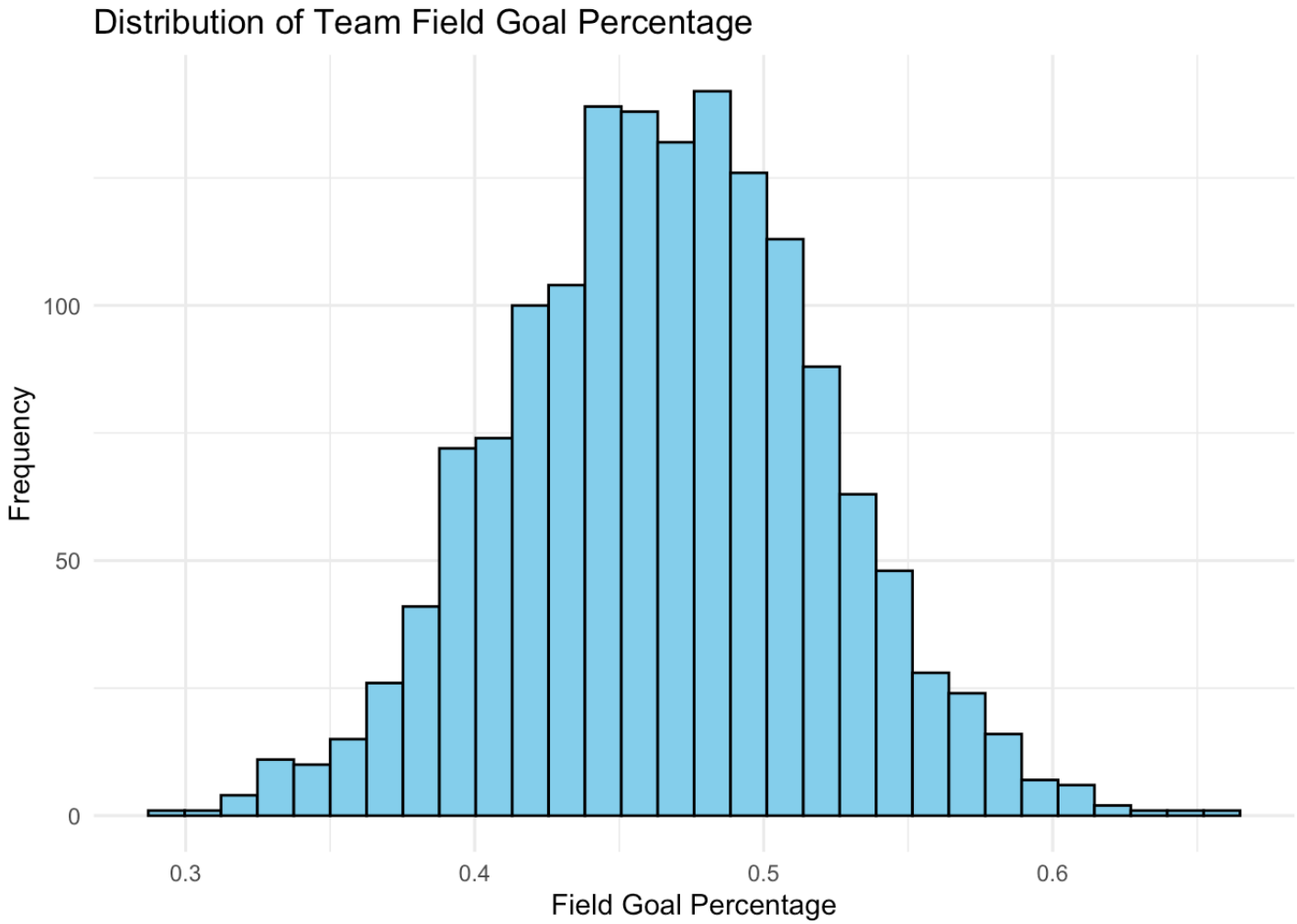
The data-cleaning process begins by importing the raw player box scores from the 2024–25 NBA season using read_csv(), which loads the information into a structured data frame that can be inspected for accuracy with head(). Since each row initially represents a single player's statistics from one game, the next step converts this information into meaningful team summaries. This is accomplished by grouping the data by game date, team, opponent, and game result, then aggregating all player statistics within each group. Using summarise(), field goals, three-pointers, free throws, steals, and blocks are summed with na.rm = TRUE, which removes missing values to prevent statistical distortion. Derived statistics such as field-goal and three-point percentages are calculated, and a new combined defensive metric, "defensive pressure," is created by adding steals and blocks—capturing overall disruption on defense. After creating team-game totals, ungroup() ensures that no unintended grouping carries into later analysis. A similar transformation produces season-level summaries by grouping by team and counting total wins and losses while averaging shooting percentages and summing defensive metrics and free throws. This process not only cleans the dataset by handling missing values and reducing noisy player-level data, but also generates new performance indicators that represent team strength more accurately, preparing the dataset for statistical testing, clustering, and predictive modeling.

```
## # A tibble: 6 × 25
##   Player Tm    Opp   Res      MP    FG   FGA `FG%`  `3P` `3PA` `3P%`    FT   FTA
##   <chr>  <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Jayso… BOS   NYK   W      30.3    14    18 0.778     8    11 0.727     1     2
## 2 Antho… LAL   MIN   W      37.6    11    23 0.478     1     3 0.333    13    15
## 3 Derri… BOS   NYK   W      26.6     8    13 0.615     6    10 0.6       2     2
## 4 Jrue … BOS   NYK   W      30.5     7     9 0.778     4     6 0.667     0     0
## 5 Miles… NYK   BOS   L      25.8     8    10 0.8       4     5 0.8       2     3
## 6 Rui H… LAL   MIN   W      35.1     7    14 0.5       1     4 0.25      3     4
## # ℹ 12 more variables: `FT%` <dbl>, ORB <dbl>, DRB <dbl>, TRB <dbl>, AST <dbl>,
## #   STL <dbl>, BLK <dbl>, TOV <dbl>, PF <dbl>, PTS <dbl>, GmSc <dbl>,
## #   Data <date>
```

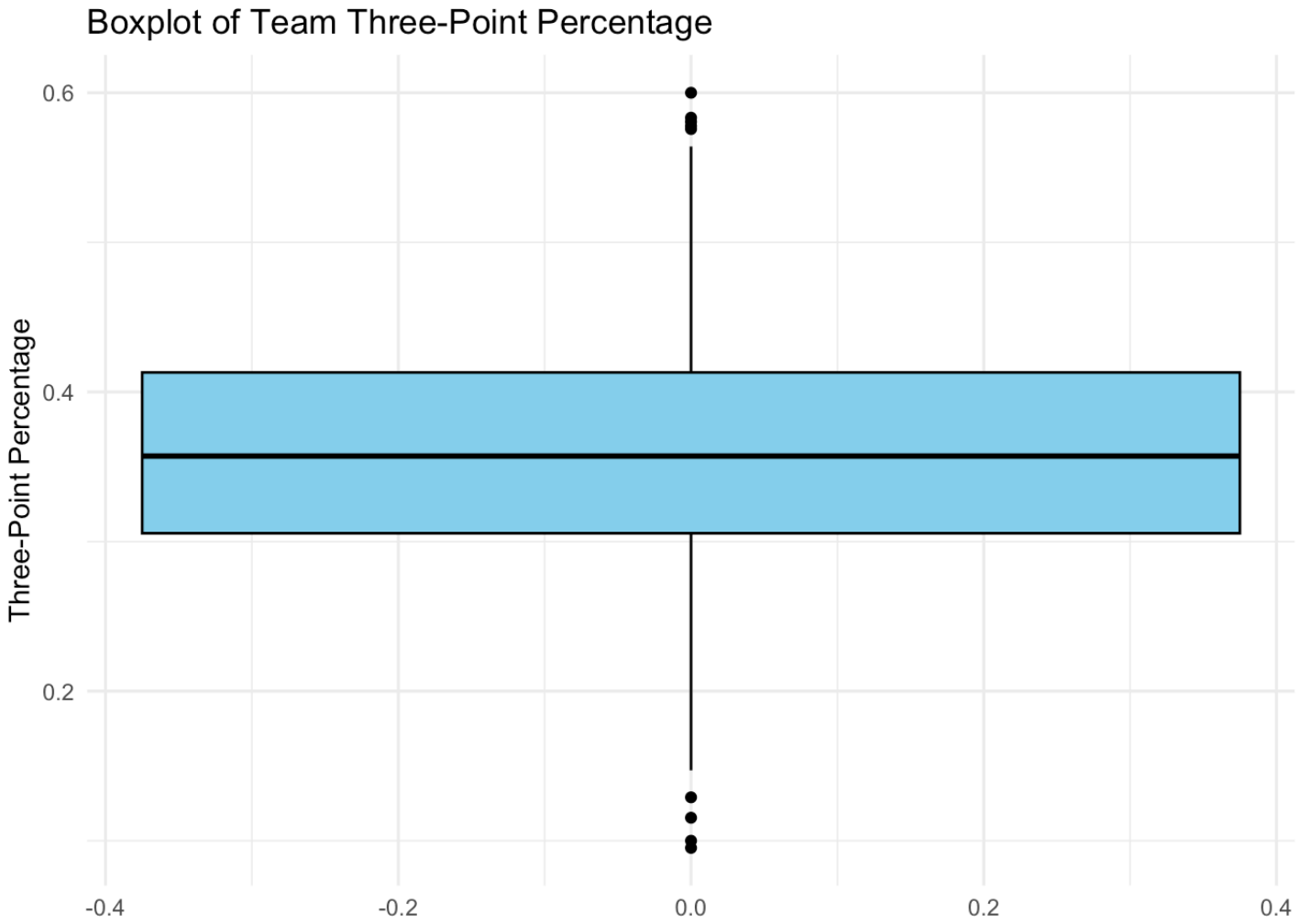
```
## # A tibble: 6 × 15
##   Data       Tm    Opp   Res      FG   FGA FG_pct  `3P` `3PA` `3P_pct`    FT
##   <date>     <chr> <chr> <chr> <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl> <dbl>
## 1 2024-10-22 BOS   NYK   W        48    95  0.505    29    61    0.475     7
## 2 2024-10-22 LAL   MIN   W        42    95  0.442     5    30    0.167    21
## 3 2024-10-22 MIN   LAL   L        35    85  0.412    13    41    0.317    20
## 4 2024-10-22 NYK   BOS   L        43    78  0.551    11    30    0.367    12
## 5 2024-10-23 ATL   BRK   W        39    80  0.488     9    28    0.321    33
## 6 2024-10-23 BRK   ATL   L        40    91  0.440    17    43    0.395    19
## # ℹ 4 more variables: STL <dbl>, BLK <dbl>, defensive_pressure <dbl>,
## #   points <dbl>
```

```
## # A tibble: 6 × 9
##   Tm     wins losses avg_FG_pct avg_3P_pct total_ft total_stl total_blk
##   <chr> <int>  <int>      <dbl>      <dbl>    <dbl>     <dbl>     <dbl>
## 1 CLE      42     10      0.498      0.395      861       430       232
## 2 OKC      41     10      0.474      0.352      826       566       283
## 3 BOS      36     16      0.460      0.368      849       388       297
## 4 MEM      35     16      0.488      0.372      950       472       308
## 5 NYK      34     17      0.495      0.368      868       409       197
## 6 DEN      33     19      0.509      0.380      940       434       252
## # ℹ 1 more variable: avg_def_pressure <dbl>
```

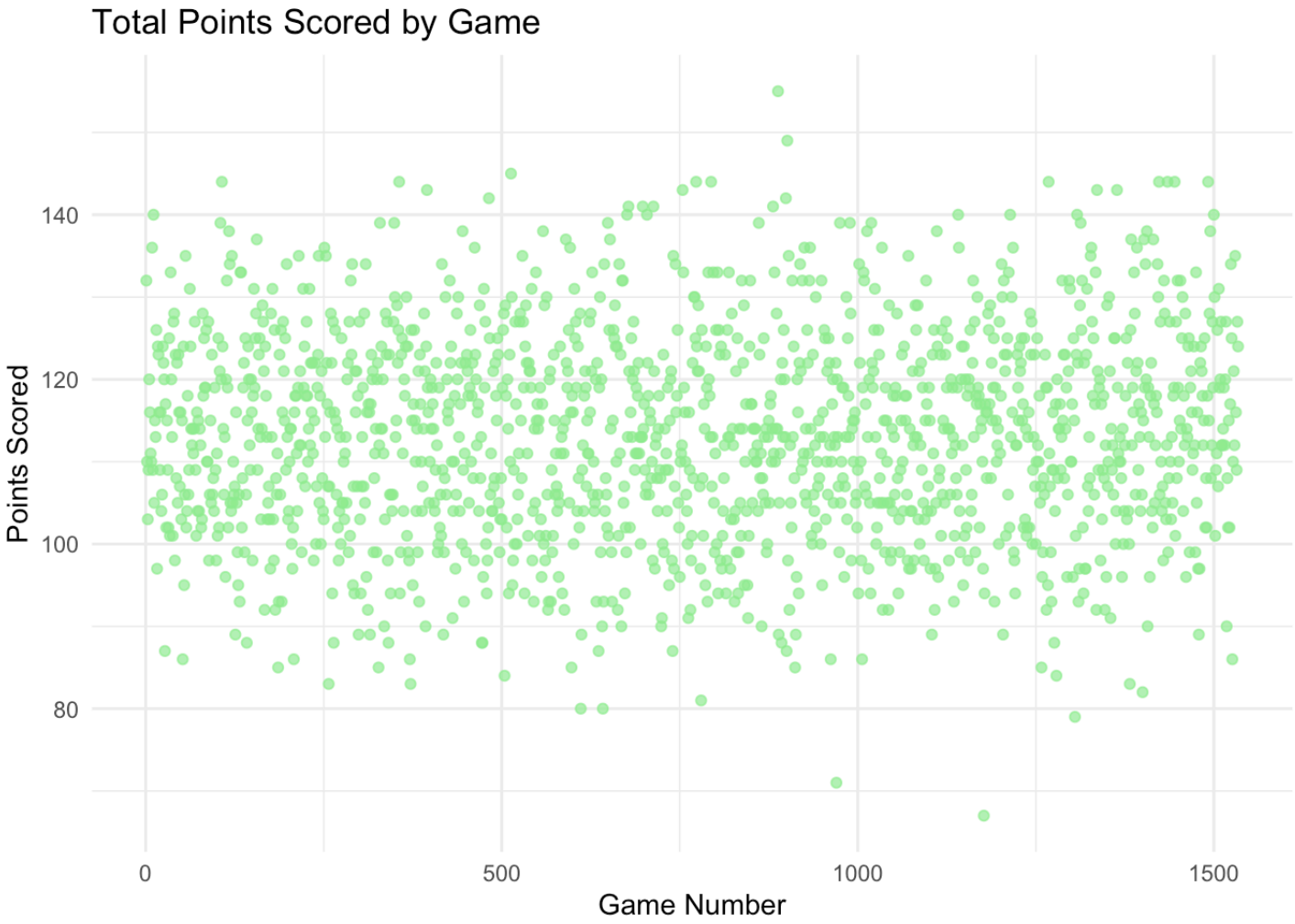
3. Exploratory Data Analysis (Tyler)



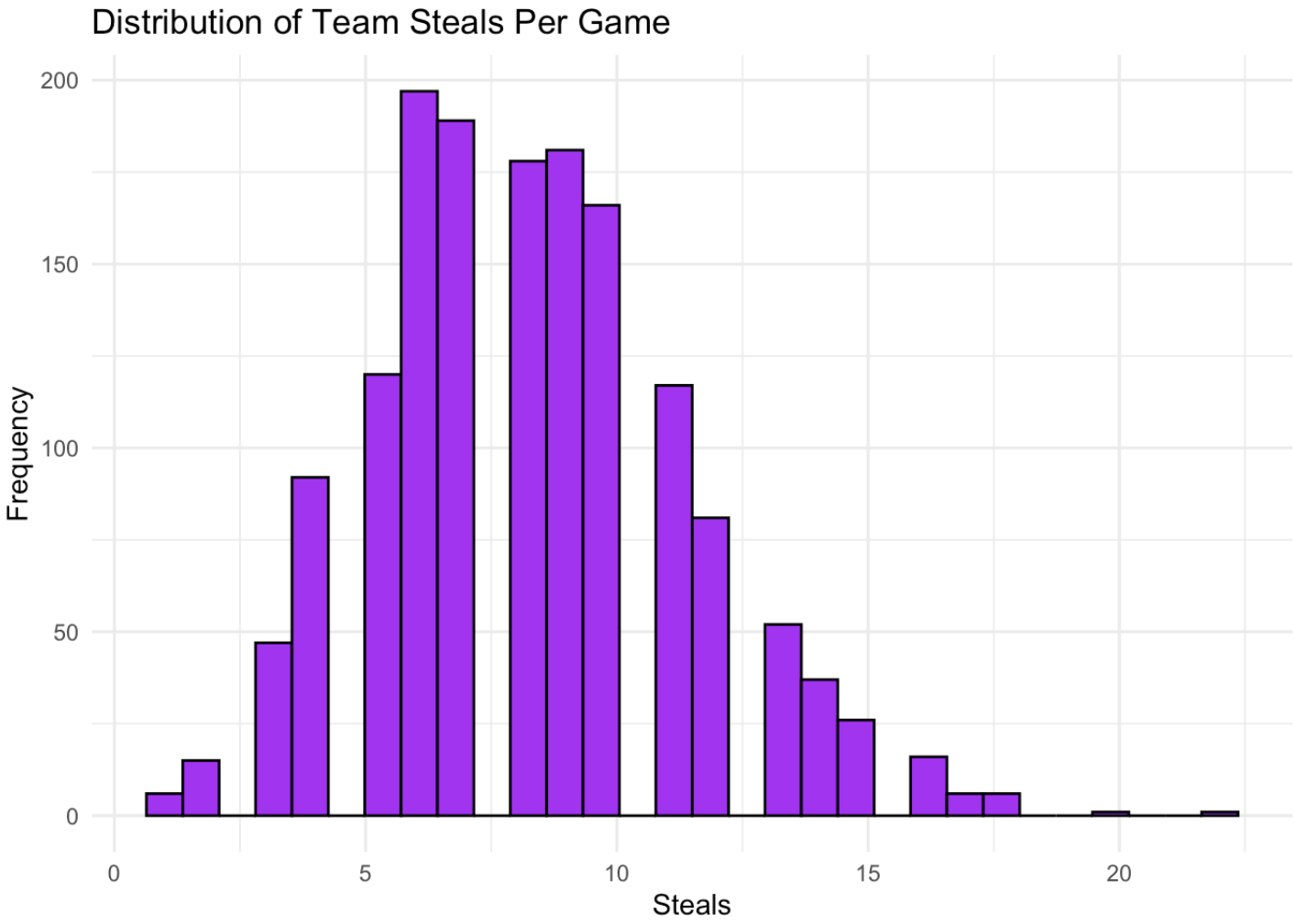
This Histogram plot shows how often different shooting percentages occur across games. Most games fall around a middle FG% range, with fewer games at very low or very high percentages.



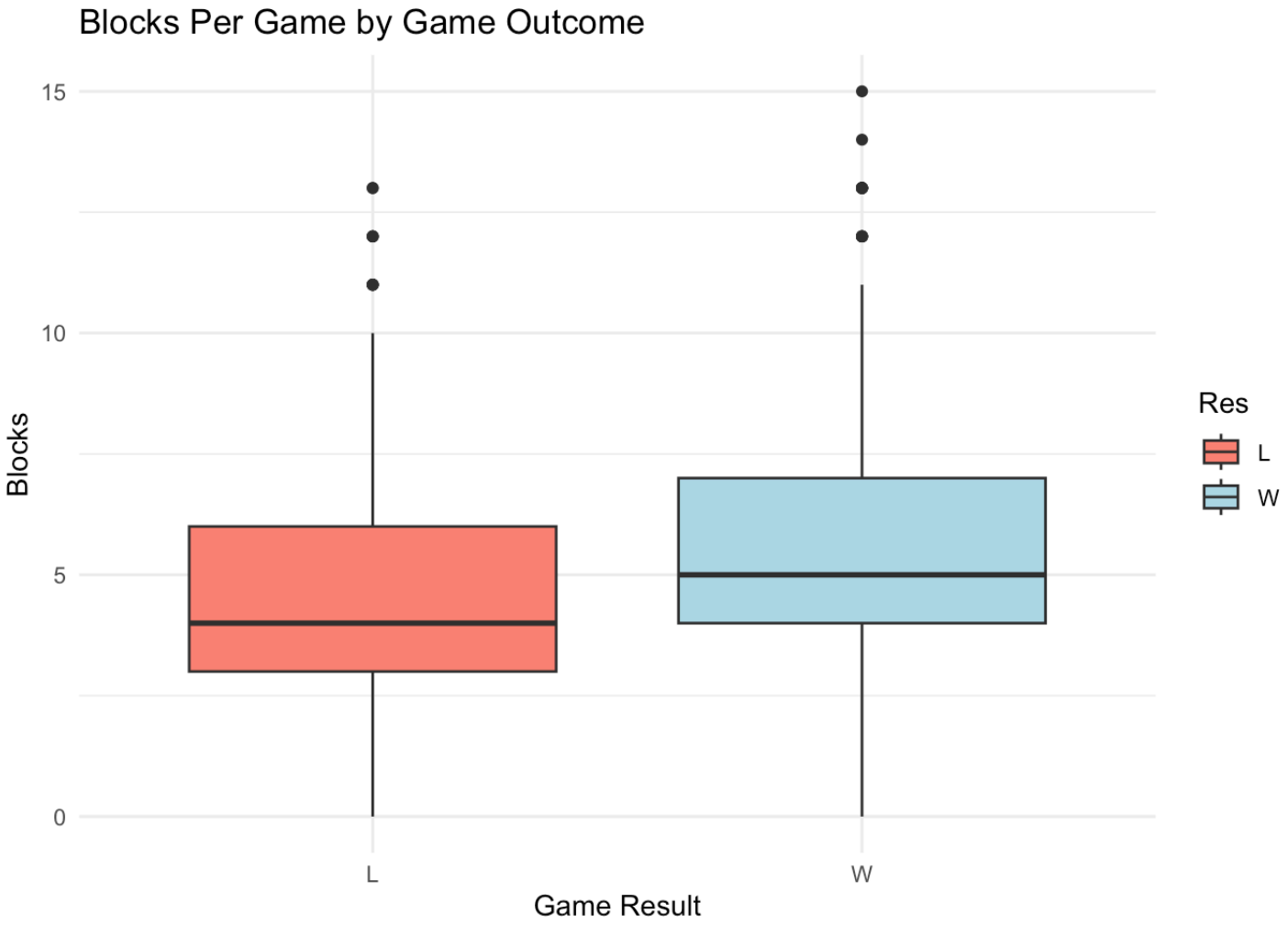
This boxplot shows the spread and variability of three-point shooting. We can see how consistent or inconsistent teams are from beyond the arc, and whether there are games with unusually good or bad three-point shooting.



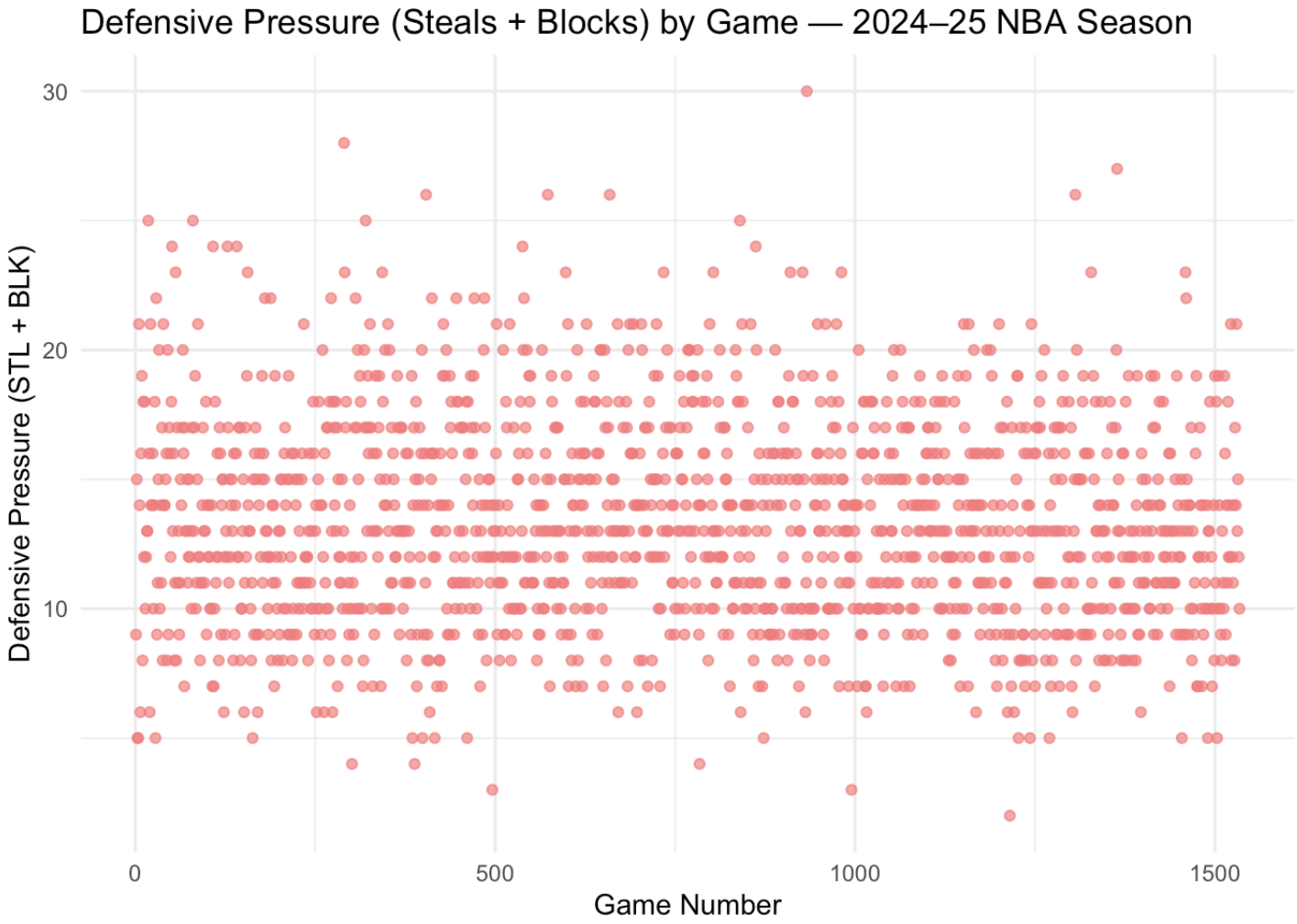
This scatter plot shows how many points teams scored in each game across the season. It helps us see whether scoring is usually clustered in a typical range or if there are many high- or low-scoring outlier games.



This histogram shows how many steals teams generate per game. It highlights whether most games have only a few steals or whether high-steal games are common.



This boxplot compares the distribution of blocks in wins versus losses. It lets us see if winning games tend to have more blocks than losing games.



This scatter plot shows how combined defensive pressure changes from game to game. It helps visualize how often teams produce a lot of steals and blocks in the same game versus games with low defensive disruption.

Exploratory Data Analysis write up

Offensive Statistics

The offensive statistics provide a clear picture of how teams scored throughout the 2024–25 season. The histogram of field-goal percentage shows that most team-game performances cluster around a central range, meaning shooting efficiency tends to stay relatively consistent across games with fewer extreme highs or lows. In contrast, the boxplot of three-point percentage displays a wider spread, indicating that three-point shooting varies more significantly and may play a larger role in separating strong and weak offensive performances. The scatter plot of total points scored by game reveals that while most games fall within a typical scoring range, there are noticeable outlier games where teams score much higher than average. Together, these visuals suggest that offensive performance is driven by both efficiency (FG% and 3P%) and overall scoring output, and that variation in three-point accuracy and occasional high-scoring games may be key factors that influence team success.

Defensive Statistics

The defensive statistics reveal meaningful differences in how teams disrupt their opponents throughout the season. The histogram of steals shows that while many games fall within a moderate range of defensive activity, some games feature significantly higher steal totals, suggesting that certain teams or matchups produce more aggressive turnover-oriented defense. The boxplot of blocks by game result indicates that winning games tend to involve slightly more blocks than losing games, though there is still some overlap between the two outcomes. The scatter plot of overall defensive pressure (steals + blocks) further shows how this combined measure fluctuates across games, with some contests showing consistently high levels of disruption and others showing much lower defensive activity. Taken together, these defensive visuals suggest that disruptive actions—particularly steals and blocks—vary more widely than offensive metrics and may play an important role in differentiating stronger defensive performances from weaker ones.

4. T-Tests (Shafin)

3-Point Percentage (Offensive Efficiency)

H0:There is no difference in the average 3-point shooting percentage (3P%) between games that teams win and games that teams lose.

Ha: There is a difference in the average 3-point shooting percentage (3P%) between games that teams win and games that teams lose.

```
##
##  Welch Two Sample t-test
##
## data:  3P_pct by Res
## t = -14.657, df = 1531.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group L and group W is no
t equal to 0
## 95 percent confidence interval:
##  -0.06374409 -0.04869628
## sample estimates:
## mean in group L mean in group W
##       0.3309174       0.3871376
```

The first t-test compared the average 3-point shooting percentage between games teams won and games they lost. The results show a clear difference: winning teams shot an average of 0.387 from three, while losing teams averaged 0.331. The p-value (< 2.2e-16) shows this difference is statistically significant. This means

teams that win tend to shoot much better from three than teams that lose, suggesting that 3-point efficiency is an important offensive factor connected to winning games.

Defensive Pressure (Steals + Blocks)

H0: There is no difference in the average defensive pressure (steals + blocks) between games that teams win and games that teams lose.

Ha: There is a difference in the average defensive pressure (steals + blocks) between games that teams win and games that teams lose.

```
##
##  Welch Two Sample t-test
##
## data:  defensive_pressure by Res
## t = -8.9814, df = 1511.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group L and group W is no
t equal to 0
## 95 percent confidence interval:
##  -2.190578 -1.405250
## sample estimates:
## mean in group L mean in group W
##        12.42764        14.22555
```

The second t-test compared defensive pressure—defined as steals plus blocks—between wins and losses. Winning teams averaged 14.23 defensive-pressure plays per game, while losing teams averaged 12.43. Again, the p-value (< 2.2e-16) indicates a statistically significant difference. This shows that winning teams typically apply more defensive pressure than losing teams, meaning defensive activity also plays a meaningful role in game outcomes.

Overall Conclusion: Both t-tests show that offensive and defensive metrics differ significantly between wins and losses. Winning teams tend to shoot more efficiently from three and apply stronger defensive pressure. These findings support the research question by showing that specific team-level statistics—such as 3-point percentage and defensive pressure—are clearly associated with winning NBA games. While the t-tests do not reveal which statistic is the most important, they provide strong evidence that both offense and defense contribute to success.

5. K-Clustering (Derek)

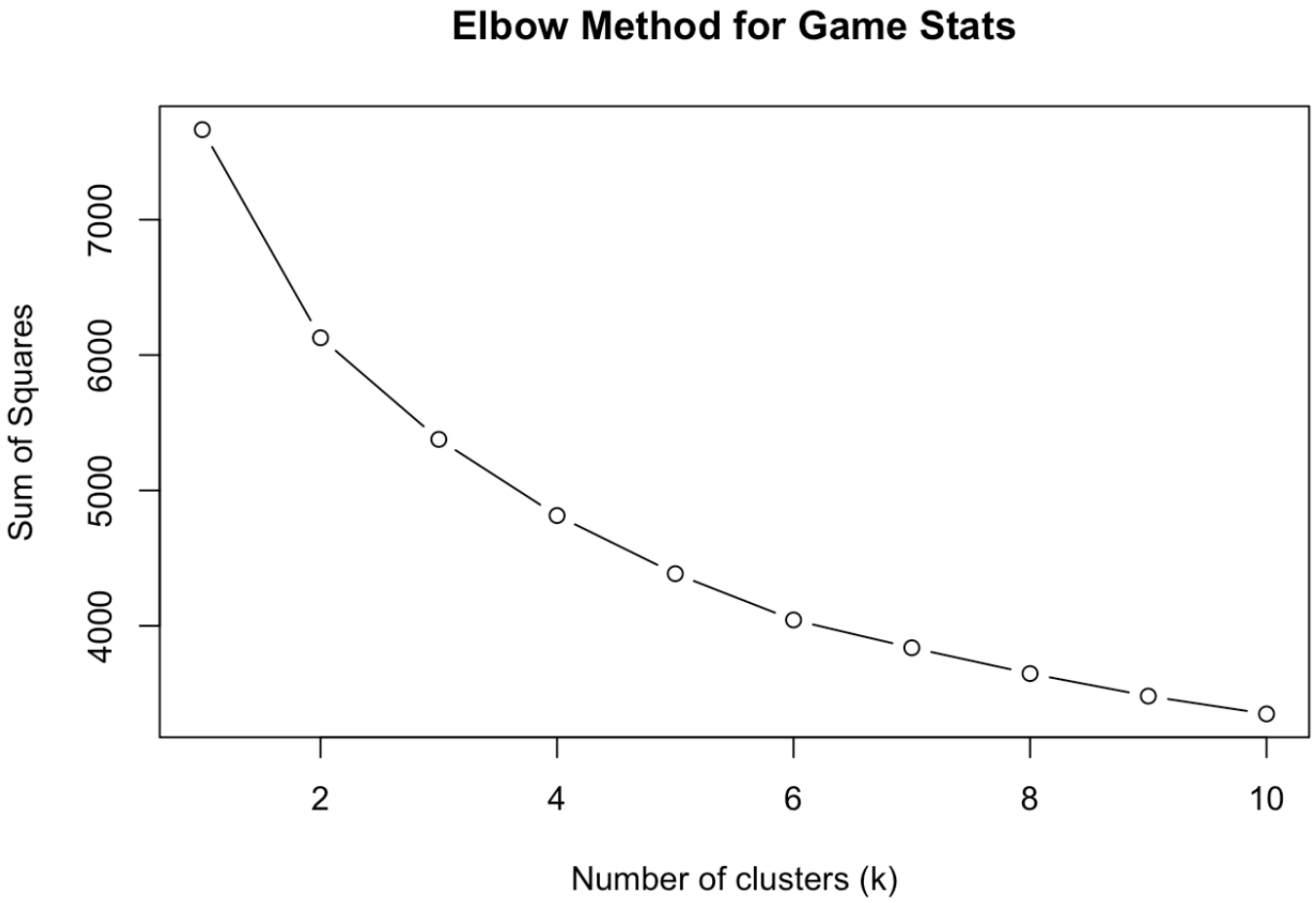
Before clustering, we wanted to see the significance of each variable, so we utilized a multiple linear regression model. The model tries to predict the number of wins a team had based on average 3 point make percentage, average field goal make percentage, total free throws made, total number of steals, and total number of blocks a team had.

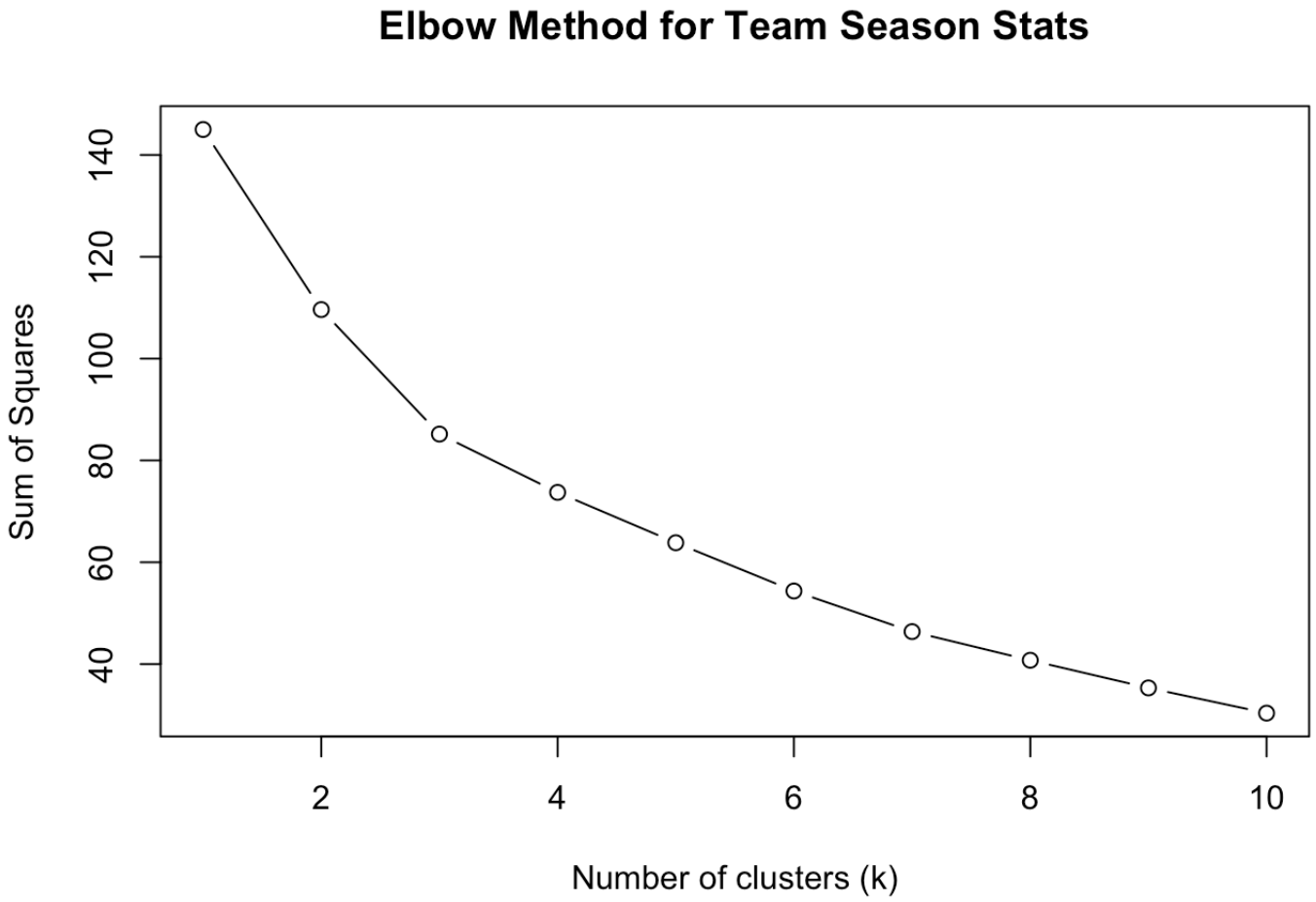
As you can see below, the R^2 score of the model is 0.6003, meaning the model can explain 60.03% of variability in the number of wins a team had in the 2024-2025 NBA season.

```
##
## Call:
## lm(formula = wins ~ avg_3P_pct + avg_FG_pct + total_ft + total_stl +
##     total_blk, data = team_season)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9881  -3.4809  -0.8851   3.8958  11.2911
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.478e+02  3.178e+01  -4.651 0.000101 ***
## avg_3P_pct   1.165e+02  9.123e+01   1.276 0.213993
## avg_FG_pct   2.165e+02  8.791e+01   2.463 0.021321 *
## total_ft    -7.247e-03  2.227e-02  -0.325 0.747639
## total_stl    6.294e-02  2.294e-02   2.745 0.011290 *
## total_blk    3.961e-02  3.290e-02   1.204 0.240325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.695 on 24 degrees of freedom
## Multiple R-squared:  0.6003, Adjusted R-squared:  0.5171
## F-statistic:  7.21 on 5 and 24 DF,  p-value: 0.0003022
```

As shown above, among the five variables, the factors which had the lowest p-values were average field goal make percentage and total number of steals a team had. While the other factors had higher p-values implying that they may not be as significant, removing them reduced the overall R^2 score of the model. With this in mind, teams were clustered based on all five factors.

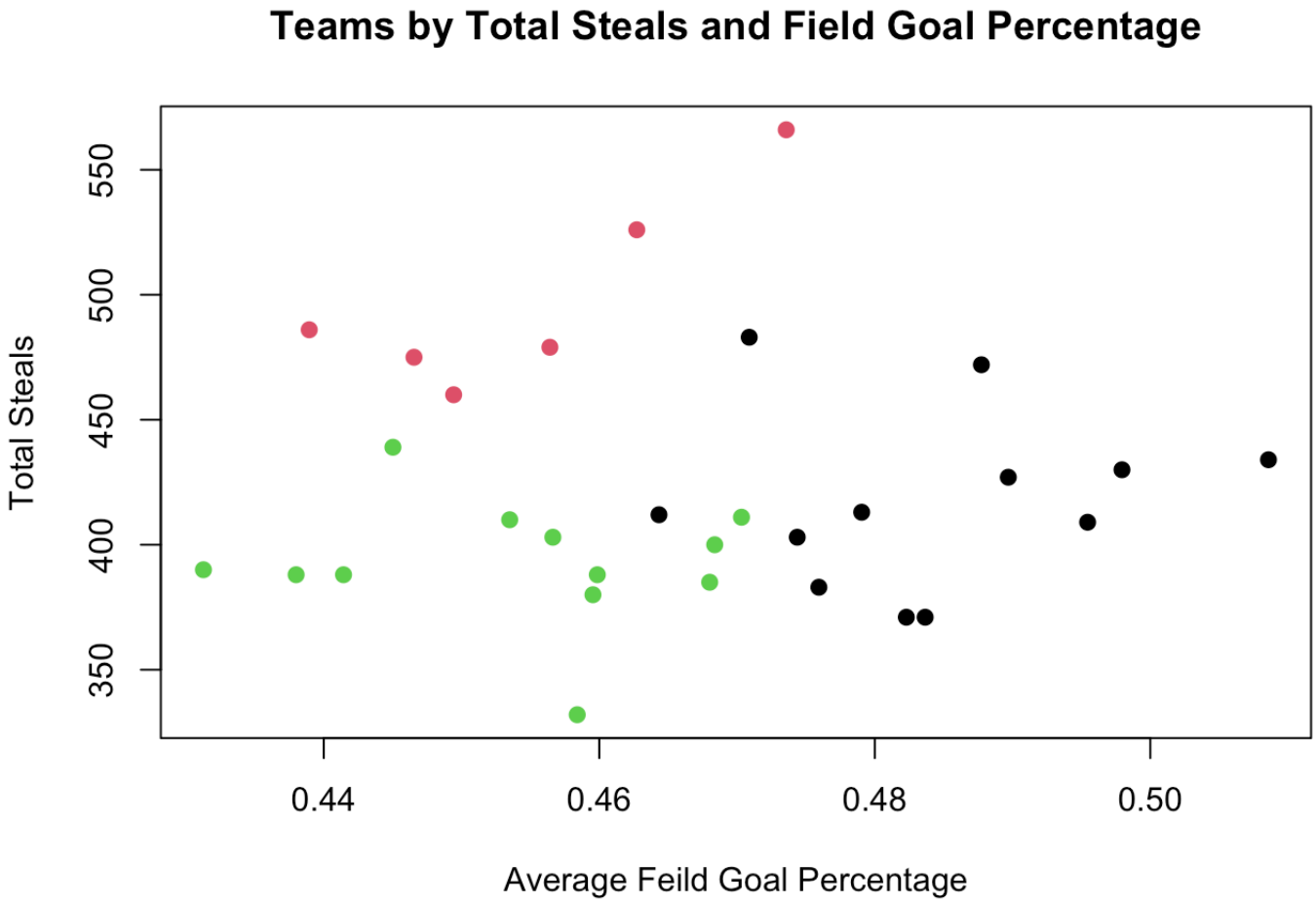
To determine how many clusters to use, an elbow test was conducted for stats by game and stats by each team's season.


Elbow Method for Game Stats
Sum of Squares
4000
5000
6000
7000
2
4
6
8
10
Number of clusters (k)

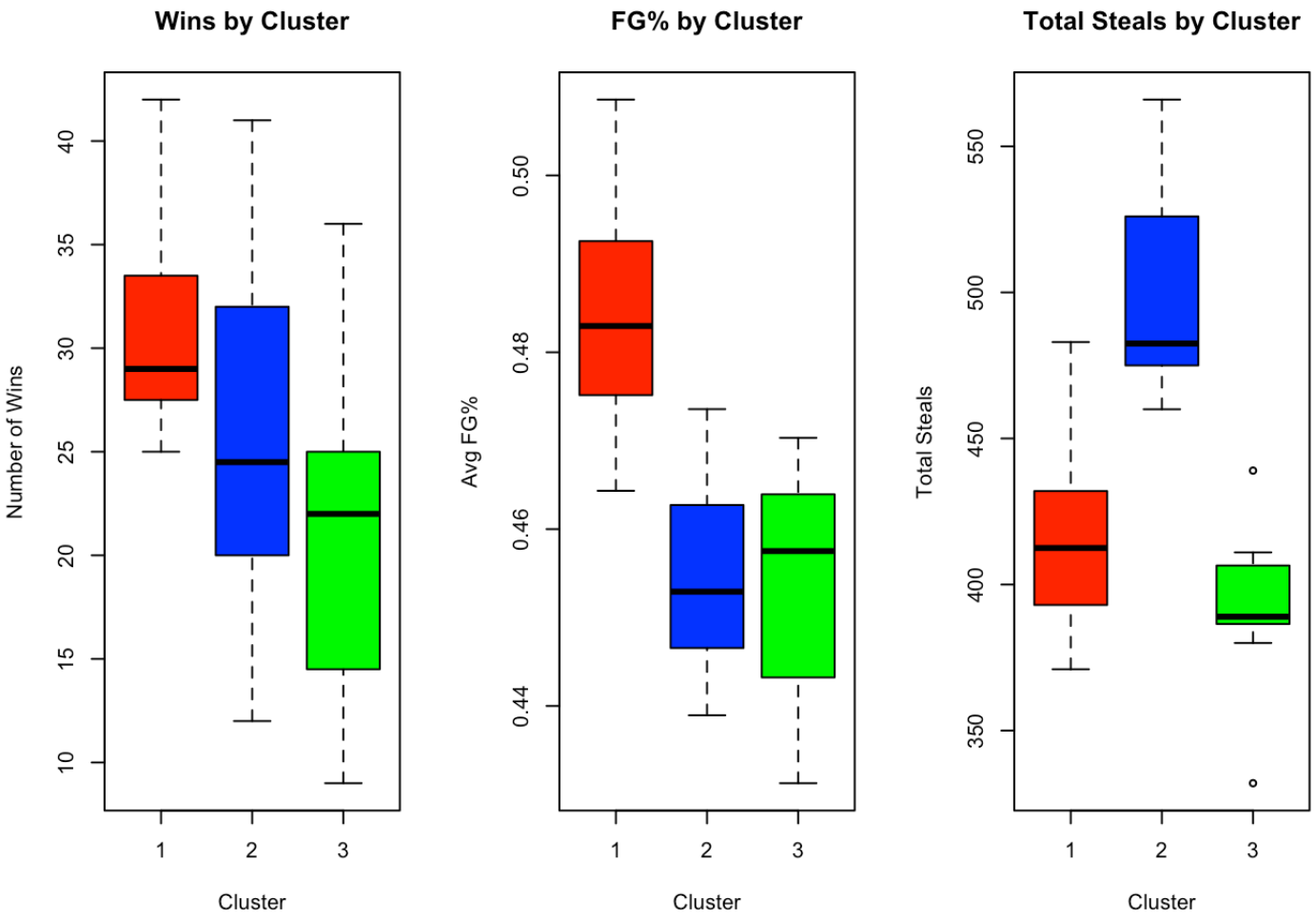


Both of the elbow tests showed that k = 3 would be an appropriate number of clusters.

Below is a scatterplot where each point represents a team. They are plotted based on average field goal percentage as well and total steals in the last season. The color of each point represents the cluster they were grouped into.

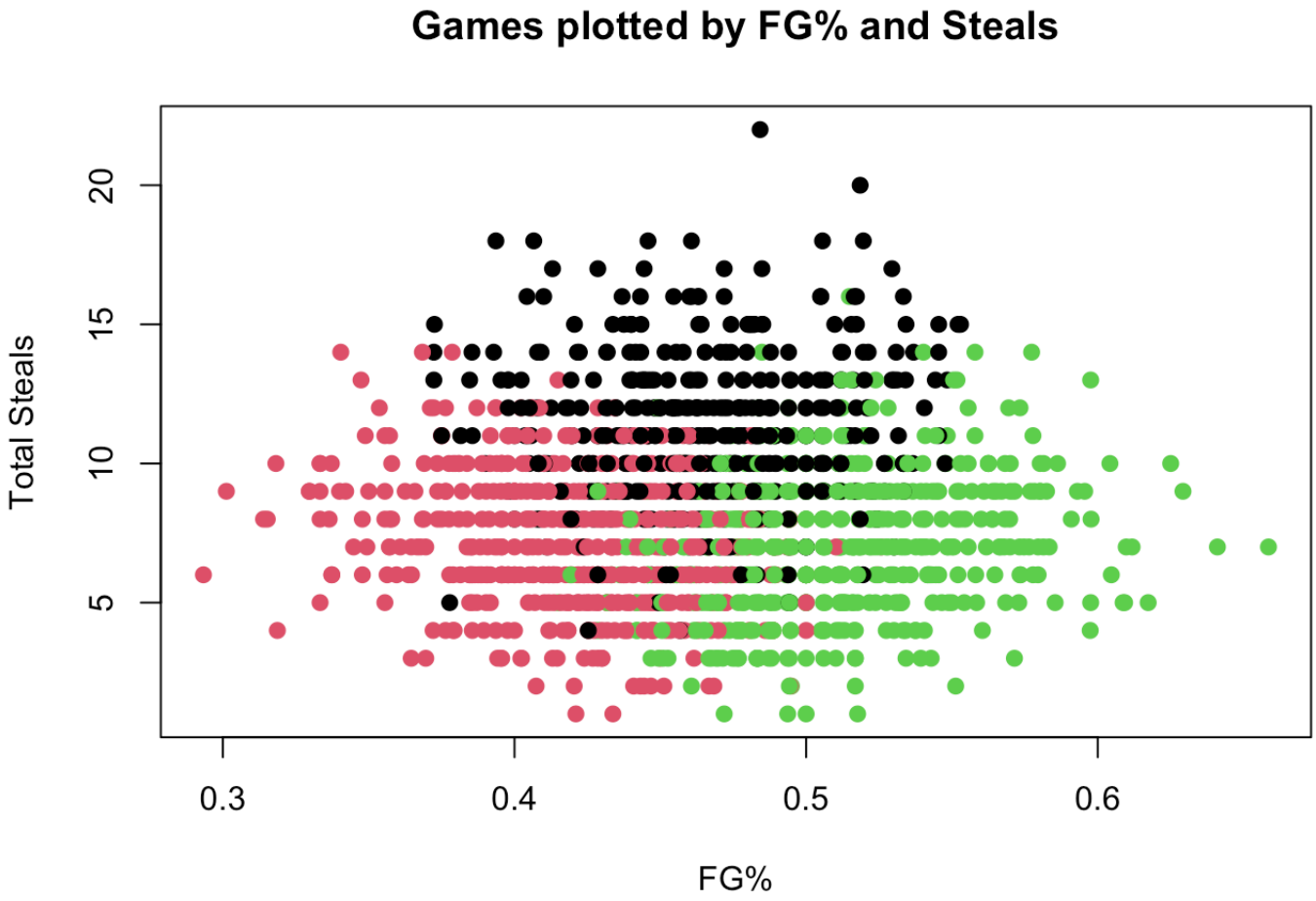


As you can see, each cluster is defined in some way by their number of steals and/or their field goal performance. To dive deeper into this, we created box plots to examine each cluster's win rate, average field goal percentage, and total number of steals.

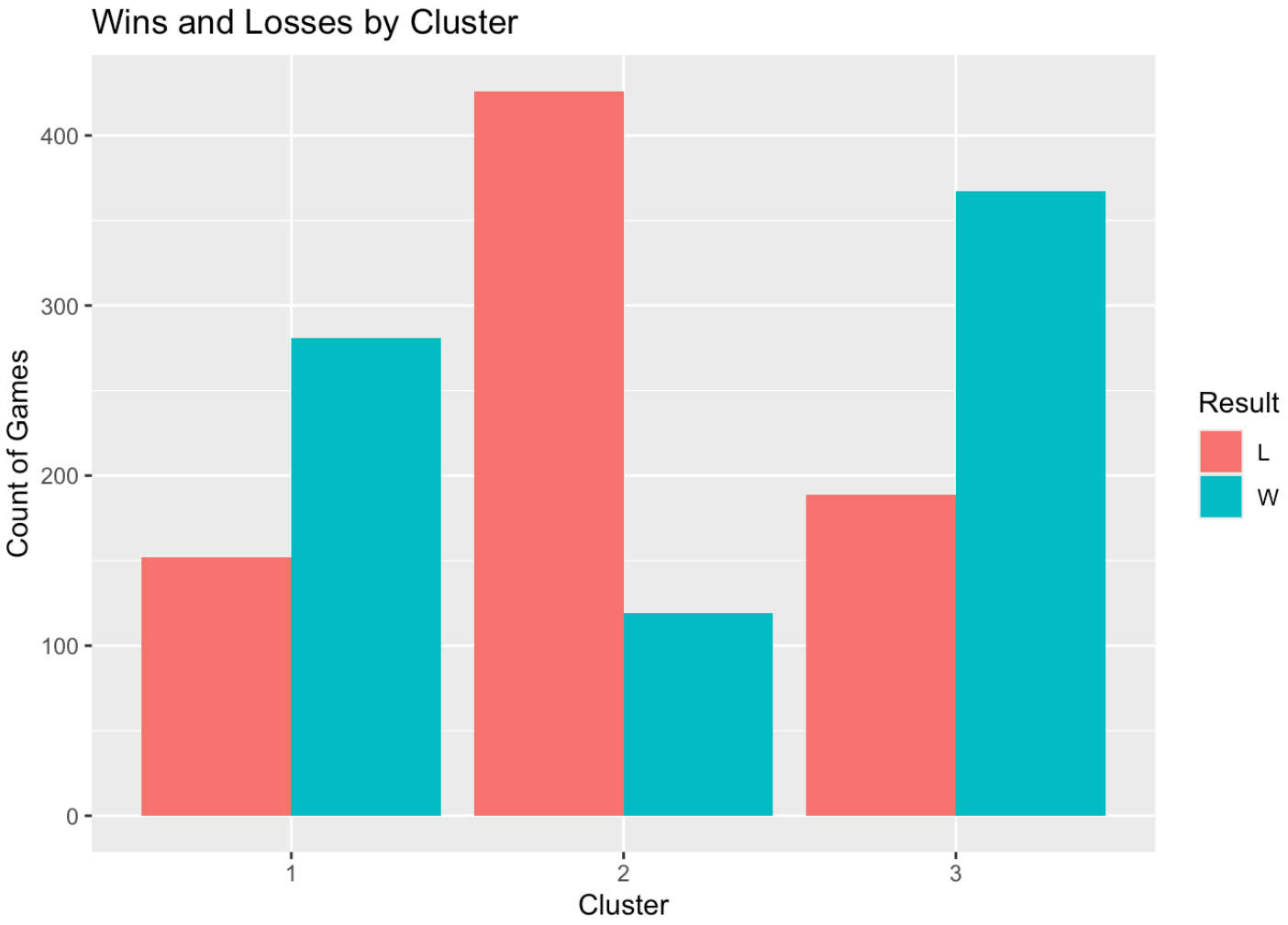


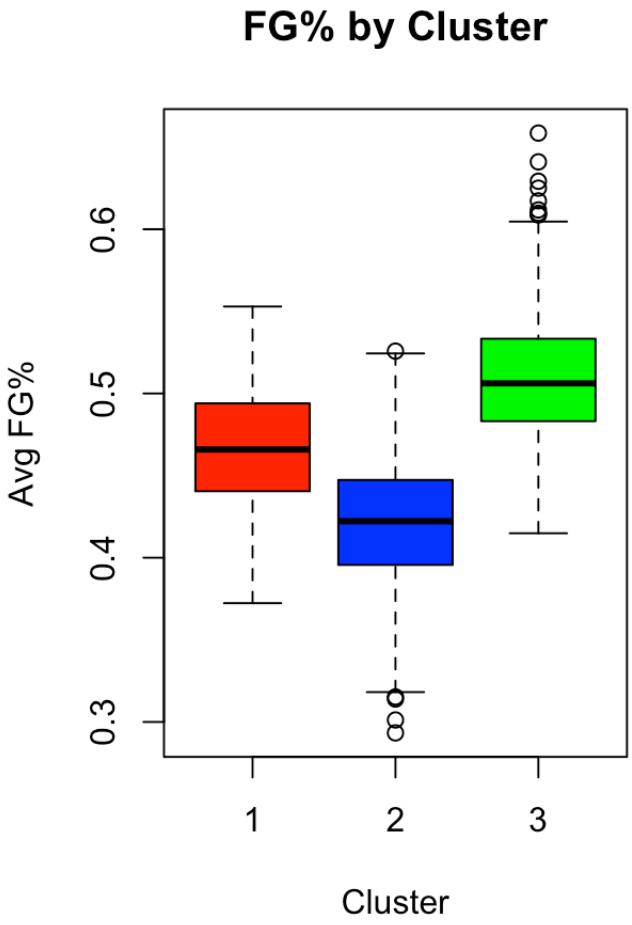
The box plots show that teams that tend to have higher win rates excel in at least one of the stats. Cluster 2 displays high performance in making their shots, while Cluster 1 showed a high number of steals. Cluster 3 did not excel in either of these categories and suffered lower win rates compared to the other two clusters.

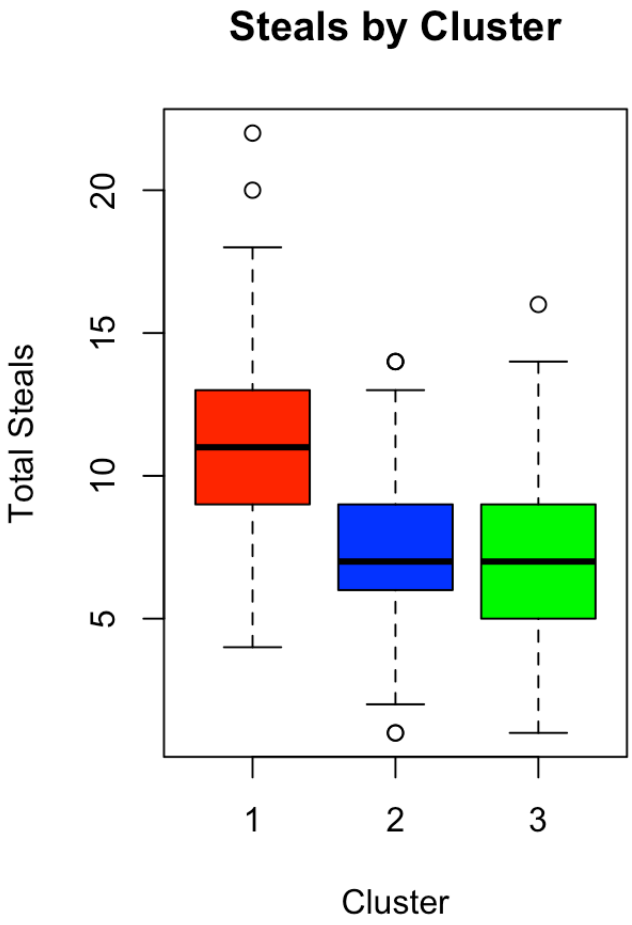
While this finding was very interesting, we were still curious about individual game play, so we clustered and examined individual games as well.



As you can see, there are also some distinctions between individual games that can be highlighted by steals and field goals.







When looking at individual games, the trend is still there. When clustered into three groups, there are typically two groups with higher win rates, and one with a lower rate. The two winning groups either show excellence in shot making or stealing the ball, while the losing group struggles to shine with either.

6. Logistic Regression (Tiffany)

```
##
## Call:
## glm(formula = win_binary ~ FG_pct + `3P_pct` + FT + STL + BLK +
##     defensive_pressure + points, family = binomial, data = team_game)
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -15.740885   0.871078 -18.071  < 2e-16 ***
## FG_pct              11.083830   1.836972   6.034 1.60e-09 ***
## `3P_pct`             3.687497   1.015279   3.632 0.000281 ***
## FT                   0.023832   0.012352   1.929 0.053674 .
## STL                  0.135791   0.020779   6.535 6.36e-11 ***
## BLK                  0.175576   0.026244   6.690 2.23e-11 ***
## defensive_pressure         NA         NA      NA       NA
## points               0.060414   0.008593   7.031 2.05e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2126.6  on 1533  degrees of freedom
## Residual deviance: 1560.8  on 1527  degrees of freedom
## AIC: 1574.8
##
## Number of Fisher Scoring iterations: 5
```

The code begins by creating a binary win/loss variable using `mutate()`, converting `"W"` to 1 and all other results to 0 so that logistic regression can be applied, since it requires a numeric 0/1 outcome. A logistic regression model is then fit using `glm()` with the binomial family, predicting the probability of winning based on field goal percentage, three-point percentage, free throws, steals, blocks, defensive pressure, and points scored. The `summary()` function outputs coefficient estimates, standard errors, z-values, and p-values, which indicate how strongly each predictor influences the odds of winning. To evaluate the model's goodness of fit, a null model is first created with only an intercept, and a likelihood ratio test compares it to the full model via `anova()`, where a significant chi-square p-value suggests the predictors improve model fit. The Hosmer–Lemeshow test is used to assess calibration, a non-significant p-value indicates the predicted probabilities align with observed outcomes. Predicted win probabilities are generated using `predict()`, converted into win/loss predictions at a 0.5 threshold. The ROC curve and AUC are then produced using the pROC package the ROC curve visualizes the model's ability to distinguish wins from losses at varying thresholds, while the AUC shows quality with values closer to 1 indicating stronger performance. The Brier score is computed as the mean squared difference between actual outcomes and predicted probabilities, with lower values indicating more accurate probability estimates.

The logistic regression model estimates how various game statistics influence the probability of winning a game. The intercept is large and negative at −15.74, so without meaningful contributions from shooting efficiency or counting stats, the baseline chance of winning is extremely low. This is expected in logistic

models where the intercept is the model's starting value before adding any performance metrics from percentages and counts.

Field Goal Percentage (FG_pct) has a strong positive coefficient at 11.08 and is highly significant at $p < 1e\text{-}09$. This indicates that even small increases in field goal percentage substantially increase the probability of winning. Because FG% is expressed as a proportion, a change from 0.45 to 0.46 or even 1 percentage point multiplies the odds of winning by exp(0.11) equal to 1.12. FG% is one of the most important predictors where a better shooting efficiency translates into higher win probability.

Three-Point Percentage (3P_pct) also has a positive and significant effect at 3.69, $p < 0.001$. While smaller in magnitude than FG%, each 1 percentage point improvement increases the odds of winning by exp(0.0369) equal to 1.038. This means three-point shooting matters, but not as much as FG%.

Free Throws (FT) shows a small, marginally significant coefficient at 0.0238, p equal to 0.054. This suggests that each additional free throw made improves win probability, but the effect is not as strong. It's possible free throws are partially accounted for by points in the model.

Steals (STL) has a significant positive effect at 0.136, $p < 6e\text{-}11$. Each steal increases the odds of winning by exp(0.136) equal to 1.15, meaning steals are highly impactful. Teams that steals often stop opponents' possessions of the ball.

Blocks (BLK) also have a notable positive effect at 0.176, $p < 2e\text{-}11$. Each block increases the odds of winning by exp(0.176) equal to 1.19. This reinforces that defensive pressure at the rim and preventing high-value shots contributes significantly.

Points has a small significant coefficient at 0.0604, $p < 2e\text{-}12$. Each additional point increases the odds of winning by exp(0.0604) equal to 1.062. Although getting more points is obviously related to winning, its effect is reduced because shooting efficiency and other metrics already explain the scoring variety.

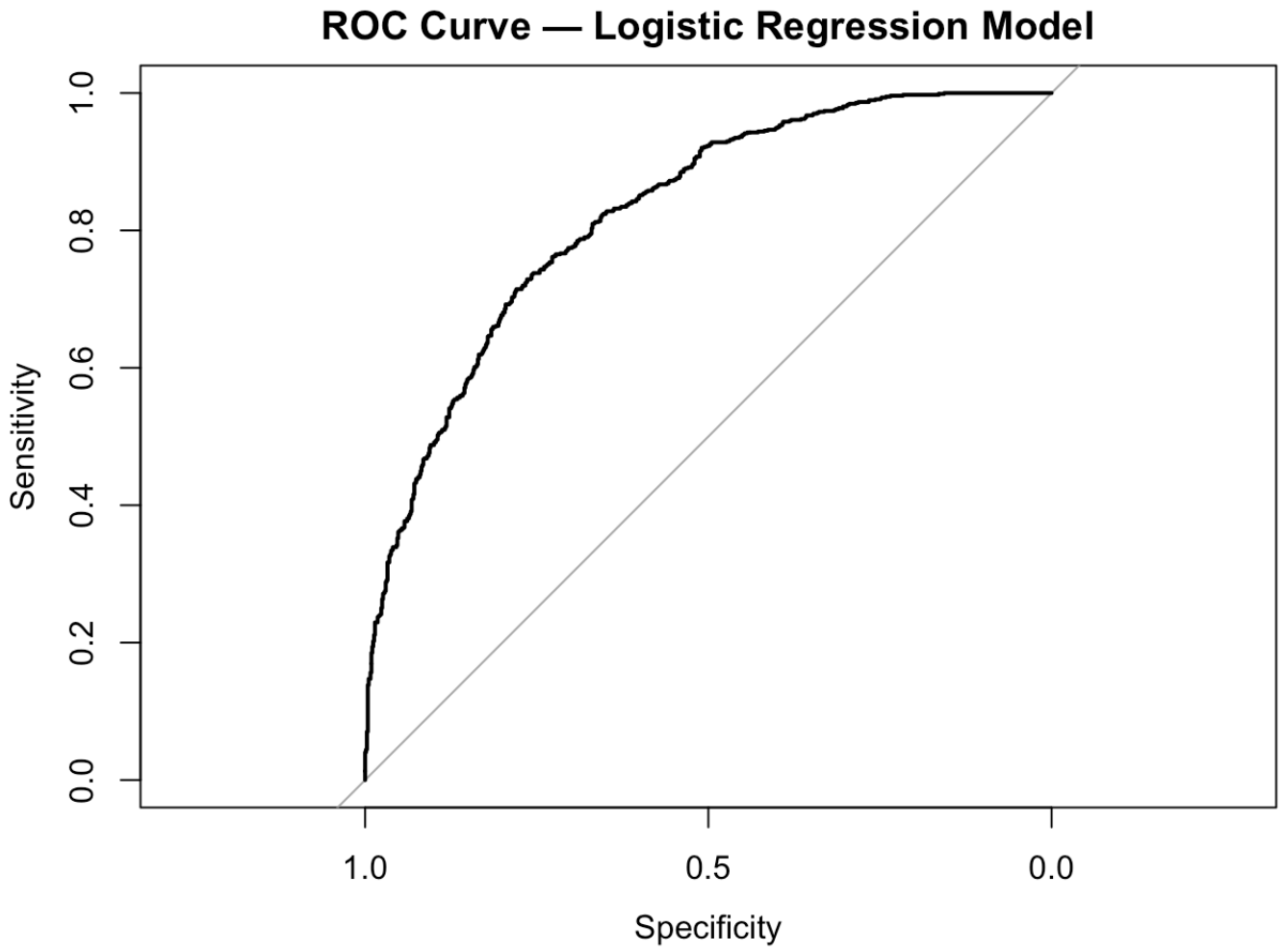
7. Goodness of Fit Tests (Tiffany)

```
## Analysis of Deviance Table
##
## Model 1: win_binary ~ 1
## Model 2: win_binary ~ FG_pct + `3P_pct` + FT + STL + BLK + defensive_pressure +
##     points
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1533     2126.6
## 2      1527     1560.8  6   565.73 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  team_game$win_binary, fitted(logit_model)
## X-squared = 12.196, df = 8, p-value = 0.1427
```

```
## fitting null model for pseudo-r2
```

```
##           llh       llhNull            G2      McFadden          r2ML
##  -780.4206661 -1063.2877750   565.7342177     0.2660306     0.3084340
##          r2CU
##     0.4112454
```




```
## Area under the curve: 0.8248
```

```
## [1] 0.1709783
```

The Likelihood Ratio Test comparing the fitted model to the null model shows a reduction in deviance at 565.73 with a p-value far below 0.001, demonstrating that the set of predictors—including shooting percentages, defensive actions, and total points provides more explanatory power than with no predictors. The chosen variables meaningfully improve the model's ability to distinguish wins from losses.

Calibration quality was assessed using the Hosmer–Lemeshow test, which evaluates how well the predicted probabilities align with observed outcomes across deciles of predicted risk. The test result of X-squared = 12.196, p = 0.1427 does not indicate a statistically significant lack of fit, meaning there is no evidence that the model over- or under-predicts win probability for any group of games. The model's predicted probabilities match the real distribution of outcomes. The pseudo R-squared values each suggest a strong fit for regression because they account for variation in win probability.

8. Conclusion

The results of our analysis show that both offensive and defensive performance play meaningful roles in determining whether an NBA team wins a game, but some statistics stand out as especially influential. On offense, field goal percentage emerged as the strongest predictor of winning. The logistic regression model showed that even small increases in shooting efficiency substantially raise a team's probability of winning, and this aligns with our exploratory data analysis, which showed that teams with more consistent shooting tend to score within a higher and more stable range. Three point percentage was also important, as supported by our t test, which found a significant difference in three point percentage between wins and losses. Although more variable than field goal percentage, strong three point shooting appears to separate top offensive performances from average ones.

On the defensive side, steals were the most impactful statistic across our models. The logistic regression showed that each additional steal significantly increases the odds of winning, and teams in our clustering analysis that generated more steals tended to fall into higher performing clusters. Blocks also contributed positively, and our defensive exploratory data analysis showed that teams with more blocks per game often performed better and created stronger defensive disruption. Taken together, these findings indicate that teams that apply consistent defensive pressure, especially through steals and blocks, gain meaningful advantages during games.

The t tests reinforced these conclusions by showing that winning teams consistently had higher three point percentages and higher defensive pressure, measured through combined steals and blocks, compared to losing teams. Meanwhile, the k means clustering revealed that teams do not need to dominate every category to find success. Many winning teams specialized either in strong shooting or in disruptive defense, and both strategies were linked to higher win totals.

Overall, our multi method approach demonstrates that shooting efficiency, especially field goal percentage, and defensive disruption through steals are the most influential factors in predicting wins during the 2024 25 NBA season. This suggests that NBA teams can improve their chances of winning by focusing on efficient shot selection and cultivating a defensive style that creates turnovers and limits opponent opportunities.

Team Contributions by individual: - Derek: Data cleaning, K means clustering, linear regression model, clustering analysis, Conclusion - Shafin: File Template, Introduction, Source file setup, Data cleaning, T-Test and analysis, Readme.rmd, Conclusion - Tiffany: Abstract, Introduction, Logistic regression and analysis, Goodness of Fit tests and analysis, Data cleaning, Readme.rmd - Tyler: Offensive and Defensive EDA/EDA Analysis