1. Defining the problem

One of the most long-standing debates in basketball analytics is whether defensive or offensive capabilities matter more for winning games. Conventionally, basketball analysis has prioritized offensive metrics as the key to winning games. Understanding these relationships can highlight how different play styles can contribute to success and offer insight into how performance can be optimized. Our project addresses this by examining performance data across multiple seasons of NCAA Division I men's basketball.

2, 3 & 4. Data Collection, Preparation, & Exploration

The dataset was scraped from barttorvik.com and includes all NCAA Division I men's basketball programs competing during the 2013-2023 seasons, excluding the 2020 season due to its cancellation. Rather than working with a sample, we have data for the complete population of Division I teams during this period.

Sample Characteristics:

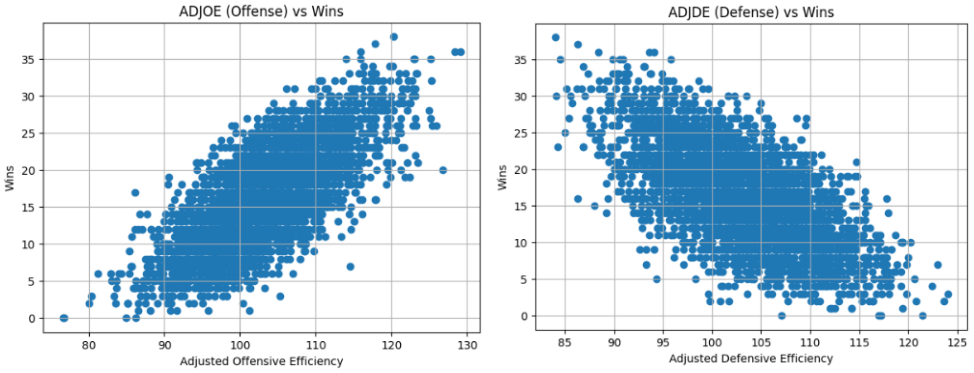
- Number of Samples: Approximately 3,600 team-season records (roughly 350 teams across 10 seasons)
- Number of features: Multiple performance metrics organized into several categories

The dataset includes the following types of variables that were relevant to our data analysis:

- **Efficiency Metrics:** Adjusted offensive efficiency (ADJOE), adjusted defensive efficiency (ADJDE).
- **Tempo Variables:** Possessions per game and pace-adjusted statistics (ADJ_T).
- **Outcome Variables:** Win (W) and number of games played (G) records; we later added a column for win percentage (calculated as the number of wins divided by the number of games played).
- **Boolean Variables:** Engineered own columns by comparing a team's ADJOE and ADJDE with the mean of the entire population of teams. New engineered features include High_Defense and High_Offense; True if the team had higher-than-average efficiency and False if it had lower-than-average efficiency. We ended up creating another variable called "Style," which helped narrow down the sample to teams that had either strong defense or strong offense, but not both.

The analysis primarily focuses on the adjusted efficiency metrics (ADJOE, ADJDE, ADJ_T) and some of our own engineered features (Win_PCT, Style, High_Defense, High_Offense), as these provide the most comprehensive view of what drives winning while controlling for schedule strength and tempo differences between teams.

The purpose of these two visualizations is to recognize that ADJOE and the number of wins a team has a strong positive linear correlation. Meanwhile, ADJDE and number of wins has a strong negative linear correlation. All this means is that as the value of offensive efficiency, that is, the more points that a team scores on offense increases, the higher the number of wins they get. On the other hand, as defensive efficiency increases, that is the number of points that a team allows to be scored on them, the lower the number of wins they get.



POSTSEASON, SEED, and other contextual variables were either removed or ignored, as some represent outcomes that occur after the regular-season performance being analyzed.
Rows with missing values in critical variables (W, ADJOE, ADJDE, ADJ_T, ORB) were removed, as these variables are essential to the analysis, and the missing efficiency metrics suggest the season data may be incomplete.

The dataset has several limitations that should be acknowledged. First, it only includes Division I men's basketball, so the findings cannot be generalized to women's basketball or lower and upper divisions, where playing styles differ. Additionally, because the data are summarized at the team-season level, they don't include individual player performance or other factors such as injuries. Teams with fewer total games may have less stable efficiency averages, potentially skewing their metrics.

Still, the dataset has several significant advantages: it covers 3,523 team-seasons over 10 years, includes every Division I program, and accounts for detailed adjustments to opponent quality and tempo. Altogether, it provides a strong, reliable foundation for examining how offensive and defensive efficiency relate to team success in college basketball.

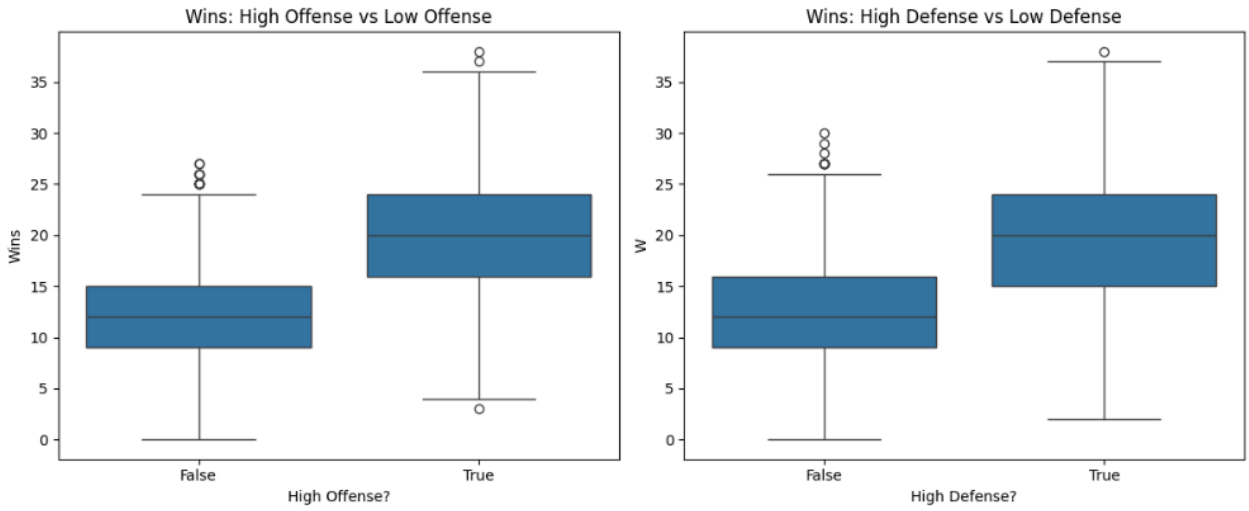
The data exploration phase aims to gain insights into the datasets' characteristics and to inform the feature model and the eventual model choice. The primary step was to create summary statistics and visualize the distributions of important variables, especially the target variable. To comprehend the data's central tendency, spread, and outliers. In our project, we used box plots and scatter plots to examine the distribution of key metrics. This step helps us determine whether certain features exhibit significant variance, indicating they should be included and encoded. A preliminary model was built as a crucial step in the modeling process. By fitting a

linear regression model to the scaled training data and making predictions on the test set, we gained important insights. Below is a table of summary statistics using the .describe() method.

	ADJOE	ADJDE	ADJ_T	ORB	W
count	3523.000000	3523.000000	3523.000000	3523.000000	3523.000000
mean	103.151320	103.153250	67.735339	29.308544	15.990633
std	7.264859	6.511989	3.091703	4.214131	6.572893
min	76.600000	84.000000	57.200000	14.400000	0.000000
25%	98.200000	98.400000	65.700000	26.500000	11.000000
50%	102.800000	103.200000	67.700000	29.400000	16.000000
75%	107.900000	107.800000	69.700000	32.100000	21.000000
max	129.100000	124.000000	83.400000	43.600000	38.000000

To be more specific, we calculated summary statistics for Adjusted Offensive Efficiency (ADJOE) and Adjusted Defensive Efficiency (ADJDE) across all Division I men's basketball teams. The average ADJOE was 103.15, with a range of 52.5, while the average ADJDE was 103.15, with a range of 40. These results suggest that offensive and defensive efficiencies are relatively balanced on average, but with considerable variation across teams. The wide range in both metrics indicates meaningful differences in team performance, supporting our goal of comparing which—stronger offense or defense — contributes more to consistent winning over time.

Additionally, we spent time looking through the data and experimenting with different plots (boxplots, scatterplots) and tables, as well as feature engineering.



Based on the above box plots, it looks like better offense has a slightly higher average for win percentage than better defense.

We created the following variables using the averages and the following computation:

```
#for offense, higher number is better efficiency, so use > symbol
df_temp["High_Offense"] = df_temp["ADJOE"] > df_temp["ADJOE"].mean()
#for defense, lower number is better efficiency, so use < symbol
df_temp["High_Defense"] = df_temp["ADJDE"] < df_temp["ADJDE"].mean()
```

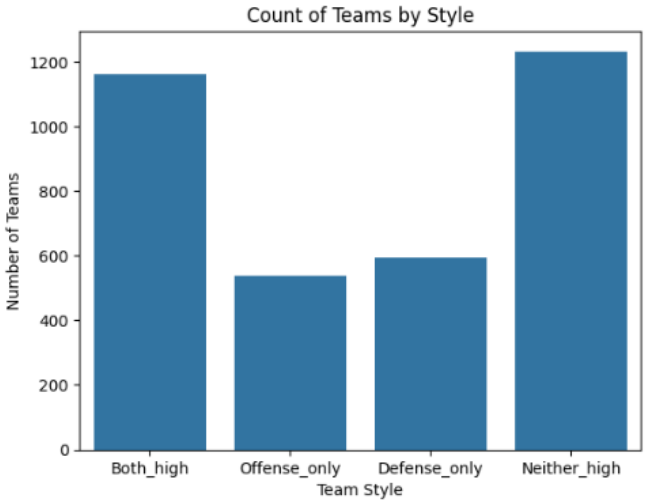
Later compiling these variables into a new column called “Style.”

```
df_temp["Style"] = np.select(
    [
        (df_temp["High_Offense"] & ~df_temp["High_Defense"]),
        (~df_temp["High_Offense"] & df_temp["High_Defense"]),
        (df_temp["High_Offense"] & df_temp["High_Defense"]),
    ],
    ["Offense_only", "Defense_only", "Both_high"],
    default="Neither_high"
)
```

And finally, generating the following data aggregations for each:

```
group_stats = df_temp.groupby("Style")["Win_PCT"].agg(["mean", "std", "count"])
print(group_stats)
```

Style	mean	std	count
Both_high	0.658695	0.129850	1160
Defense_only	0.519808	0.146818	593
Neither_high	0.369266	0.139634	1233
Offense_only	0.545387	0.131712	537



The above visualization was to look at the distribution of teams and what their playing style is. Based on the above plot, our dataset shows the majority of teams are not one-sided (only good at defense or only good at offense), they are either strong at both or not that strong at both. Also, the sample of offense_only and defense_only are around the same (537 to 593 respectively), so computing their averages or other metrics will not be heavily skewed.

5. Model Building

We chose a linear regression model to predict a team's win percentage (Win_PCT) using three features (ADJOE, ADJDE, ADJ_T). We selected these variables based on domain knowledge, observed patterns during exploratory data analysis, and their direct perceived links to winning basketball games. We went with a simple linear regression because our goal was to measure the strength of each variable's contribution to winning, and because exploratory data analysis indicated fairly linear relationships between the predictors and win percentage. Linear regression also provides interpretable coefficients that quantify how changes in each variable are associated with changes in win percentage. Since the predictors in our dataset have different scales, we applied MinMax scaling to the features before fitting the model to ensure that no single variable dominated the others solely due to scale differences.

ADJOE represents offensive efficiency; higher values indicate better offense. Scatterplots and the linear regression model showed a strong positive association between offensive efficiency and win percentage.

ADJDE represents defensive efficiency; lower values indicate stronger defense. Scatterplots and the regression model suggest that defensive efficiency is also closely related to winning, and domain knowledge supports this. Because defense is known to be a significant factor in team success, we included ADJDE to examine how strongly it contributes relative to offensive efficiency.

ADJ_T measures tempo/pace (number of possessions per 40 minutes). Two teams with identical offensive and defensive efficiencies may play at very different speeds, which can influence game outcomes. Including tempo allows us to test whether playing faster or slower is associated with higher winning percentages. Although the correlation between tempo and win percentage was relatively weak, we included this variable to test whether pace adds predictive value beyond efficiency metrics.

We focused on these three variables because they capture distinct components of team performance: offense, defense, and play style. Together, they allow us to investigate how multiple aspects of a team's approach relate to its overall success. While we could have incorporated more features such as two-point percentage, three-point percentage, offensive rebounds, two-point percentage allowed, three-point permitted percentage, and offensive rebounds allowed, we decided to keep the model simple and sufficient for our analysis by adding too many features and making it complex.

Data splitting was also performed using train_test_split, yielding X_train, X_test, y_train, and y_test, to enable proper model evaluation on unseen data. We also applied Min-Max scaling to the numerical features to normalize their ranges. That step was crucial for the performance of the linear regression model. The MinMaxScaler was used only for the training data, then applied to both the training and test sets, preventing data leakage from the test set.

In addition to the linear regression model, we also conducted a permutation hypothesis test with the same goal as before: to test whether offense heavy teams win more than defense heavy teams. The reason we went with a permutation test was because it not only cross checks the regression story, but it serves as a simulation based hypothesis test of our big picture problem (offense-only versus defense-only win%). The other advantage is that permutations tests (unlike parametric tests such as t-tests or z-tests or models like linear regression) have no assumptions about distributions or linearity.

We had the following null and alternative hypotheses.

H0: Offense-heavy and defense-heavy teams have the same mean Win_PCT
H1: Offense-heavy teams have higher mean Win_PCT than defense-heavy teams

Our test statistic was the observed difference in mean win percentage between offense heavy teams and defense heavy teams. This value was calculated by performing the following calculations.

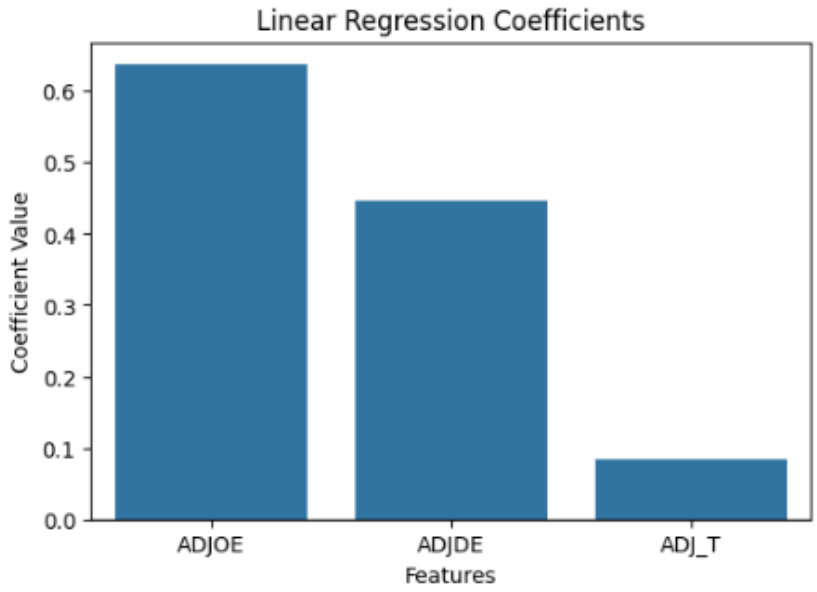
```
# 2. Observed difference in means (offense - defense)
obs_diff = off.mean() - defn.mean()
print("Observed mean win% (offense):", off.mean())
print("Observed mean win% (defense):", defn.mean())
print("Observed diff (off - def):", obs_diff)
```

```
Observed mean win% (offense): 0.545387111202113
Observed mean win% (defense): 0.5198075433179106
Observed diff (off - def): 0.02557956788420246
```

Once we had our observed p value, the rest was essentially conducting a one sided test (offense > defense). We pooled all of our win percentage values from both offense-heavy and defense-heavy teams, randomly shuffled them into groups of equal size, and then ran a permutation test simulated under the null condition. This was repeated a total of 10,000 times.

6. Model Evaluation

B) The hypothesis we tested was “To what extent do offensive efficiency, defensive efficiency, and tempo predict a Division I men’s basketball team’s win percentage, and which factor has the strongest effect?” Our model showed that the coefficients for ADJOE and ADJDE were significant, showing that both offense and defense have strong linear relationships with winning. The coefficient for ADJ_T was close to zero, showing that tempo matters way less than efficiency. The following data visualization reinforces this fact.



The above plot is a bar plot of the coefficients. We flipped the sign of ADJDE for the sake of this plot and easily visualized the three features on the same scale. We also flipped the sign to avoid potentially misleading one from seeing ADJDE being negative could imply that strong defense in basketball can not lead to winning games. However, we had to keep in mind that in our linear regression, the beta (coefficient) for ADJDE was -0.45.

C) For evaluation, we used the R^2 score, which measures the proportion of variation in win percentage explained by the model. Our model achieved an R^2 of approximately 0.60, meaning that the three predictors together explain about 60% of the variability in win percentage across teams. This level of accuracy shows that efficiency-based statistics contribute most to team success in this dataset.

D)
i) Linear regression makes decisions by assigning a coefficient to each predictor, representing how much that variable contributes to the predicted win percentage. We think these coefficients are straightforward to interpret and align with basketball domain knowledge.

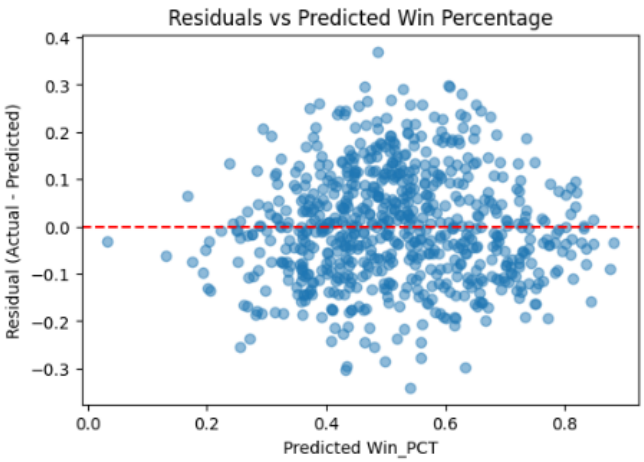
ii) To check for overfitting or underfitting, we compared the model's performance on a single train/test split and examined the coefficient patterns and scatter distributions. Linear regression is a low-complexity model with mostly linear relationships in the data; it is unlikely to overfit. The close alignment between the regression line and actual data points also suggests the model is neither overfitting nor underfitting.

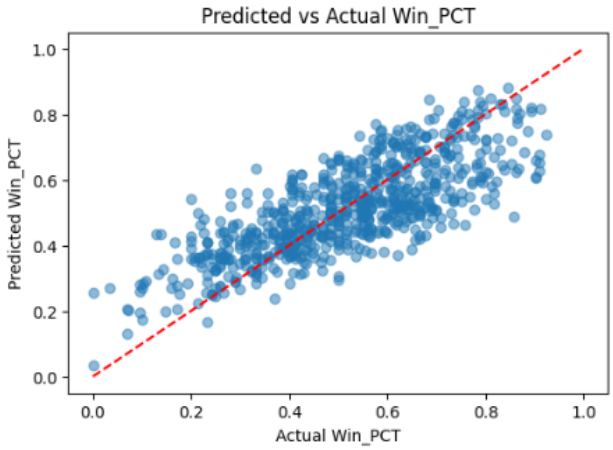
iii) Our model does not suffer from class imbalance because our target variable (win percentage) is continuous, not categorical.

iv) We normalized all our features using MinMax scaling because the variables are measured on very different scales, and scaling prevents any single variable from dominating the model only

due to its units. Effects include making coefficient magnitudes directly comparable, helping us interpret which predictors matter most.

Below are two of the data visualizations we made after the model was trained.

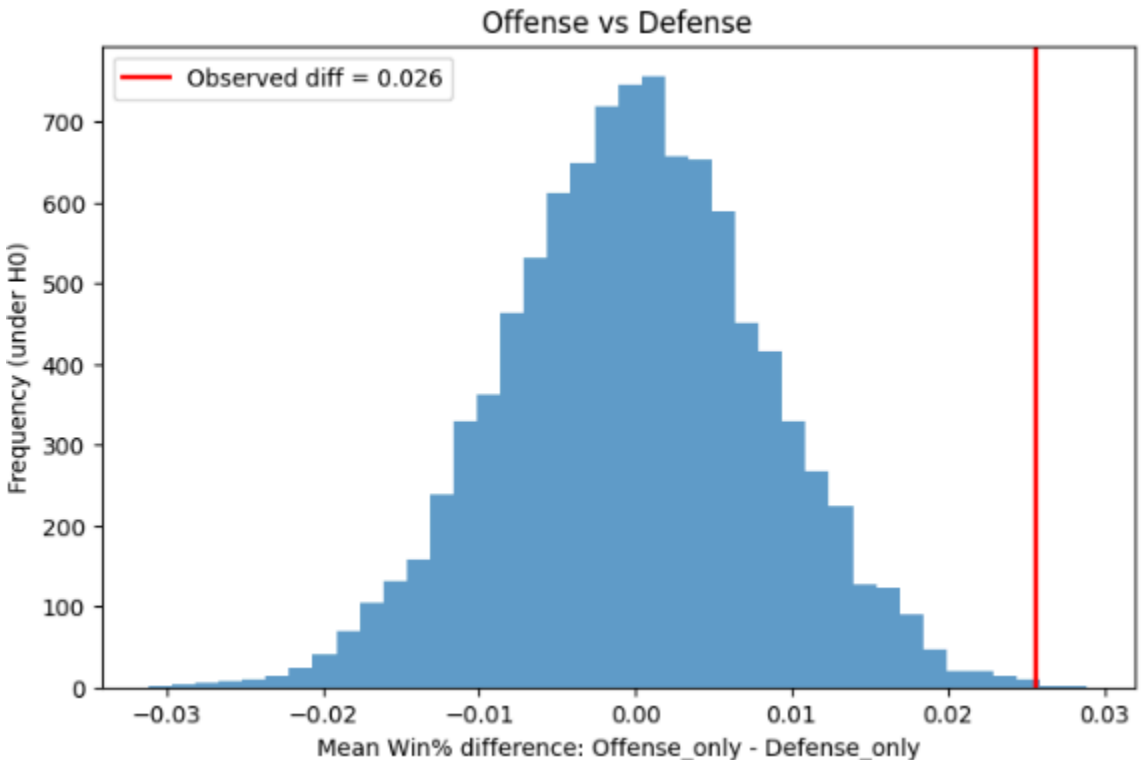




Residual plot - if scattered about 0 and not resembling any pattern or shape, then linear regression is reasonable. By "reasonable" in our case, the model was able to capture a linear relationship between efficiency and win% based on the residual plot check. The residual plot shows that we are good for linear regression.

Predicted vs. actual Win_PCT plot - the points on the red line corresponds to a perfect prediction from the model.

The below visualization shows the results of our permutation test.



Our observed difference of 0.026 (2.6%) is plotted to the far right. Our p-value ended up being 0.0005. This is also portrayed in the above visualization by looking at the little bit of data that was to the right of the red vertical line. This alpha value essentially means fewer than 0.5% of shuffles generated a difference as extreme or more extreme than our observed difference. Therefore, we reject the null hypothesis and find statistically significant evidence for the alternative hypothesis.

7. Model Deployment

If we deployed this model in a real basketball analytics setting, it could serve as a practical tool for coaches and performance staff to better understand the factors most strongly associated with winning games. Because the model predicts a team's win percentage from just three key variables (offensive efficiency, defensive efficiency, and tempo), it could be used before or during a season to estimate how well a team is likely to perform based on its underlying performance indicators. For example, when deciding whether to prioritize improving offense or defense, staff can examine the model's coefficients, which reveal that offense correlates slightly more strongly with winning. The insights the model reveals are therefore beneficial for teams in planning how to allocate practice time, which lineups maximize efficiency, and how their current strengths or weaknesses will translate into outcomes over the course of a season.

From a broader analytics perspective, recruiters and scouts could use the model to estimate how a new player or lineup might influence a team's offensive or defensive efficiency and, by

extension, its win potential. Even media outlets or fans could use a simplified version of the model to visualize team performance beyond raw win–loss records. Because offensive and defensive efficiencies are already widely used in sports analytics, integrating this model into existing analytics dashboards would be technically straightforward.

A potential risk we face when deploying is the model's simplicity. Since the model only considers three features, it might not generalize well to complex real-world scenarios. This limited feature set (three: ADJDE, ADJOE, ADJ_T) might limit its performance once deployed. One way to address this limitation before deployment is to expand the model to include additional features. Doing so could make the model more representative of the real world and improve its overall usability.

Despite these challenges, our model remains a strong starting point for understanding how offensive and defensive quality translates into winning games, and how coaches, critics, and fans can interpret those insights to their advantage.