Assaf Bitton

Analyzing Team Performance and Winning Trends in Baseball

Abstract:

This project was largely focused on looking through different variables in the game of baseball. I focused on different variables separated into offense and defense and how they affected wins. I used the Lahman package which I found by researching the different datasets that RStudio came with. The main focus was to figure out which baseball statistics mattered most for winning games. In order to accomplish this, I created multiple regression models, plotted the statistics using R-script, and performed statistical tests. After running tests to compare offensive and defensive metrics, I found that runs scored had the strongest correlation with wins, while walks allowed was the strongest defensive metric correlated with wins. A surprising finding was that hitting home runs or triples was negatively correlated with wins, indicating that raw power is not the only strategy for winning (I described how this was the case).

My other focus was on how teams changed over time. I looked at the difference in the means of the significant predictors and what that meant. Interestingly, when looking at change in predictors over time, I saw an outlier in the 2020 season. This was likely a result of the COVID-19 pandemic which shut down the entire world for almost an entire year. It was a reminder that when analyzing statistics, you should be wary of anomalous environmental changes. Finally, I focused on how win percentages differed among different teams (franchises). I found that I could make indirect inferences on which teams might win in a match off by looking at their overall performance over a multi-decade period (1990-2023).

Introduction:

Baseball has long been one of the most popular sports, with team performance depending on a combination of factors such as offense and defense. Understanding the relationships between these factors can help predict team success and inform strategies. This project will analyze the “Teams” dataset from the Lahman package to explore how variables such as defensive and offensive stats impact team performance and winning percentages. The “Teams” dataset is a collection of variables which describe performance metrics of baseball players across multiple decades, multiple teams, and multiple leagues. The research question is: "What factors most significantly influence team success in baseball?" To answer this, **I** will use inferential statistics and regression models to identify key drivers of performance.

Data Collection and Description:

This project will use the “Teams” dataset from the Lahman package. The dataset includes information on leagues from 1871 to the present. Key variables to look at include: offensive stats (Runs Scored (R), Home Runs (HR), and Stolen Bases (SB), etc) and defensive stats (Base on Balls (BB), Strikeouts (SO), and HRA (Home Runs Allowed), etc). The dataset also has data on the outcomes like Wins (W), making it ideal for analyzing team performance across different leagues and eras.

Although players change over time, franchises often carry a legacy that influences long-term trends. Thus, for this project, I decided to focus on data starting from 1990. This allows me to use the most current and relevant information to reflect modern information regarding baseball and statistical patterns. So, by narrowing the dataset to this period, I can avoid challenges like incomplete or inconsistent historical data. This is why focusing on recent years and selecting key variables ensures a meaningful and reliable analysis of how offensive and defensive metrics impact team success.

The methodology used for Offensive and Defensive metrics:

1. **Data Preparation**: The analysis aimed to understand how offense and defense influence wins using data from the Lahman package. The dataset focused on teams from 1990 onward and have important key metrics for both offensive and defensive performance. Before I start to analyze, to ensure accuracy, missing data was removed before proceeding.

2. **Data Visualization and Exploration:** Scatter plots with regression lines were created to visualize how each metric relates to wins, making it easier to identify the strongest correlations. Additionally, I created a bar plot of the regression coefficients that showed which metrics had the greatest impact on wins.
3. **Regression Analysis:** Two regression models were created: one for defense and one for offense. Offensive metrics (Runs Scored (R), Home Runs (HR), etc.) were analyzed alongside defensive metrics like (Strikeouts (SO), Home Runs Allowed (HRA), etc.).

Analysis of offensive metrics:

To start off, the offensive metrics include Runs Scored (R), Home Runs (HR), Stolen Bases (SB), Hits (H), At-Bats (AB), Doubles (X2B), and Triples (X3B), with the dependent variable being team wins (W).

First, I will perform a multiple linear regression to evaluate the impact of various offensive metrics on wins. The regression coefficients reveal both the strength and direction of each metric's relationship with wins. As seen in Offensive Figures 1 and 2, Runs Scored (R) stands out as the most significant predictor, with a coefficient of 0.1076. This means that for every additional run scored, the team's expected wins increase by approximately 0.11. By looking at the p values, the result is highly significant ($p < 2e-16$), which means that this metric is important for increasing wins. Taking a closer look at Offensive Figure 3, the scatter plot shows a strong positive relationship. As the number of runs scored increases, there is a clear upward trend in the number of wins, which can be seen by the red regression line. Overall, with Runs Scored being the most important offensive metric, it's fitting to think of the dots on the scatterplot like baseballs.

Furthermore, as seen in Offensive Figures 1 and 2, Stolen Bases (SB), Hits (H), and At-Bats (AB) show positive and important effects on wins, with those metrics having p-values less than 0.05 and a coefficient of 0.023. This means that for every additional stolen base, a team's expected wins increase by approximately 0.023. Similarly, for every additional hit, a team's expected wins increase by approximately 0.0280. At-Bats (AB), however, has a relatively smaller impact, with a coefficient of 0.0101. Overall, these metrics (SB, H, and AB) emphasize the importance of maintaining offensive opportunities. It is important to note that since Doubles (X2B) have a p-value greater than 0.05, they are not statistically significant and are therefore ignored.

Interestingly, as seen in Offensive Figures 1 and 2, Triples (X3B), H (Hits), and Home Runs (HR) have a negative relationship and have important effects on wins, with those metrics having p-values less than 0.05. When you think about stats like home runs and triples, they are usually seen as key to winning games, big, exciting moments that fans love. But here, the data tells a different story. The regression shows that home runs have a coefficient of -0.0315, meaning each additional home run slightly decreases wins. Triples have an even stronger negative coefficient of -0.1244, and hits also show a small negative effect at -0.0280. This doesn't mean hitting home runs or triples makes teams worse, it is more about how these numbers interact with other parts of the game. Metrics like runs, hits, and home runs often overlap, and that overlap can skew the results. Teams that primarily focus on offense might miss out on other important strategies, like baserunning tactics and strategies. While these results seem surprising, they reflect how complex winning in baseball really is and how it's about balancing offense and not just making big hits. It's a psychological factor. When a big hitter makes a significant impact, other players on the team might feel pressured to get the same result (hit a homerun). The pressure of following up a homerun is immense. Note that home runs differ from runs in that a home run is a run scored by a single player hitting the ball out of play and reaching home plate in one at-bat, while runs can be scored by advancing to home plate from any base.

Lastly, as seen in Offensive Figure 2, the residual standard error (RSE) of 9.54 reflects the average difference between the predicted and actual number of wins. This means that, on average, the model's predictions are off by about 10 wins. Considering that teams play around 162 games each season, this level of error is relatively small and highlights the model's reliability. While a lower RSE would indicate an even better fit, this value suggests the model does a solid job of predicting team win totals within a reasonable margin. To summarize, as seen in Offensive Figure 2, the model as a whole is highly statistically significant, as evidenced by the F-statistic of 207.1 and a p-value less than 2.2e-16, which means that the offensive metrics collectively contribute to explaining team wins. Notably, Runs Scored (R) has the strongest positive impact, followed by Stolen Bases (SB) and Hits (H), while Triples (X3B) and Home Runs (HR) show significant negative relationships when controlling for other variables. This analysis shows the importance of consistent run production and base running over power-hitting metrics in predicting team success.

Analysis of defensive metrics:

First I am going to conduct a descriptive analysis for the sample that I have chosen.

- SO(Strikeouts) - total number of batters the team's pitchers strike out
- RA(Runs Allowed) - total number of runs allowed allowed by team
- E(Errors) - total number mistakes made by fielders that allow opposing team to advance or score
- HRA(Home Runs Allowed) - total number of home runs allowed by a team's pitchers
- BB(Base on Balls) - total number of times a pitcher throws 4 balls outside the strikeout zone, resulting in a walk-on-base

Now let's look at the scatter plot of defensive metrics vs. wins. The scatterplot for SO vs. Wins as seen in Defensive Figure 4 shows a weak positive correlation, as indicated by the upward-sloping regression line with a correlation coefficient of 0.31. The upward regression line shows a weak relationship but indicates that teams with more strikeouts tend to win more games. In baseball, strikeouts are one of the key defensive metrics a team's pitcher uses to prevent the ball from being hit into play. However, this weak correlation suggests that while strikeouts are helpful, other predictors may have stronger correlations with wins.

The scatterplot for RA vs. Wins as seen in Defensive Figure 5 shows a weak positive correlation, with a correlation coefficient of 0.075. The positive regression line indicates that teams that allow fewer runs tend to win more games. The goal of the team is to decrease the total number of runs allowed by the opponent.

The scatterplot for E vs. Wins as seen in Defensive Figure 6 shows a weak positive correlation, with a correlation coefficient of 0.151. The positive regression line indicates that teams with fewer errors tend to win more games. The scatterplot for HRA vs. Wins as seen Defensive Figure 7 shows a weak positive correlation, with a correlation coefficient of 0.1046, showing that teams who allow fewer home runs may have a slight advantage.

The scatterplot for BB vs. Wins as seen in Defensive Figure 3 shows a moderate positive correlation, with a correlation coefficient of 0.671. The upward-sloping regression line indicates that teams that issue more walks tend to win more games. This may seem counterintuitive, but it is based on the idea that these walks are part of a team's defensive strategy, not solely negative.

Overall, this shows that defensive metrics like Base on Balls(BB) tend to have a stronger correlation to wins, showing their importance in a team's success. On the other hand, metrics such as RA, HRA, and E show a weaker correlation, indicating that while they do play a factor in contributing to a team's success, they are not the sole predictor.

Next, I will do a regression model to look into how different defensive metrics influence the number of wins for a baseball team. The analysis includes Strikeouts (SO), Runs Allowed (RA), Errors (E), Home Runs Allowed (HRA), and Base on Balls/Walks Allowed (BB), with the dependent variable being team wins (W). Defensive Figures 1 and 2 shows that Base on Balls (BB) is the most significant predictor, with a coefficient of 0.1189. This means that for every one increase in walks allowed, the team's wins increase by approximately 0.12. BB has a p-value of 2e-16 < 0.05, showing that it is highly significant and important to manage the number of walks allowed in order to win games. The scatter plot in Defensive Figure 3 supports this positive correlation, showing a positive slope with a correlation of 0.671. The scatterplot shows that as the number of walks increases, there is an increase in the number of wins, as shown by the red regression line.

In addition, Defensive Figures 1 and 2 also show that Errors (E) and Strikeouts (SO) lead to more wins. Both predictors each have p-values less than 0.05. Errors (E) have a coefficient of 0.069 which shows that for every additional increase of error, a team's chance of winning increases by 0.069. Strikeouts (SO) also have a positive relationship but with a smaller coefficient of 0.013 which means that for every additional strikeout, wins increase by 0.013. These results show the importance and significance of pitching and defense during a game to increase wins.

Interestingly, as seen in Defensive Figures 1 and 2, Runs Allowed (RA) has a significant negative relationship with wins, with a coefficient of -0.0499. This means that for every additional run allowed, the team expects their wins to decrease by 0.05. This supports the idea that limiting the number of runs is crucial for success in baseball. On the other hand, Home Runs Allowed (HRA) has a weaker positive relationship with a coefficient of 0.0224. However, it has a p-value greater than 0.05, meaning it is not as significant and does not strongly contribute to explaining wins.

As seen in Defensive Figure 2, the residual standard error of 10.14 shows the average difference between the predicted and actual wins. This means that, on average, the model's predictions are off by about 10 wins. This makes sense because a team plays more than 100 games a season, which can lead to errors, showing that this model is realistic and reliable. Having a lower RSE would result in a better fit and provide more reliability in predicting wins.

Overall, the multiple regression model is significant, with an F-statistic of 233.38 and a p-value of 2.2e-16, which is less than 0.05. This shows that it is useful in assessing which defensive metrics contribute more to wins. Walks Allowed (BB) is the best predictor and has the strongest impact on wins, followed by Errors (E) and Strikeouts (SO). Runs Allowed (RA) has the most significant negative relationship with wins. Therefore, defense is important for wins, as minimizing runs allowed and having good pitching can lead to team success.

Comparing model fit between offense and defense:

After analyzing both offensive and defensive metrics, **I** compared their respective models to see which one is the better predictor for predicting wins. So, **I** looked at which model produces a more significant and larger r-squared value. Based on the adjusted R-squared values, the offensive model has an Adjusted R-squared of 0.59, while the defensive model has an Adjusted R-squared of 0.53. The offensive model explains approximately 59.1% of the variability in team wins, while the defensive model explains approximately 53.9% of the variability in team wins.

Although the offensive metrics used in **my** model produced a higher r-squared value, the difference between the two models is relatively small. This suggests that both offensive and defensive metrics play a significant role in determining the outcome of a game, and both should be considered when analyzing team performance.

Franchise Comparison Section:

Methodology:

In this part of my study, I changed my focus from looking at predictors for wins to looking at franchise teams as a whole over a chosen time period (1990-2023). In order to do this without knowing if the groups have similar variances I chose Welch's two-sample t-test (Franchise Figure 3). My main question was whether the average win percentage (WinPct) of one franchise—like the New York Yankees (NYY)—was higher than that of another franchise, such as the Boston Red Sox (BOS). The null hypothesis stated there would be no difference, while the alternative hypothesis suggested that one team had a statistically significant greater win percentage.

To get a clearer picture, I also created visual summaries using the ggplot2 package in R. A boxplot (Franchise Figure 1) allowed me to see how each team's WinPct was distributed, and another plot (Franchise Figure 2) showed the mean WinPct with 95% confidence intervals. These intervals give me a sense of how precise my estimates are, providing a range in which the "true" average likely lies.

Analysis and Results:

Figure 1 shows a boxplot of WinPct for several franchises (ATL, BOS, CHC, LAD, NYY) from 1990 to 2023. By looking at where the median line falls inside each box and how wide the boxes are, I can tell if a team's performance was stable or varied a lot. A higher median WinPct suggests a stronger team overall, and a tighter box indicates more consistent results year after year.

Figure 2 takes this further by plotting the mean WinPct for each franchise and including confidence intervals. If a team's mean WinPct is comfortably above 0.55 and the interval is fairly narrow, it suggests that they enjoyed reliably strong seasons. On the other hand, if two teams' intervals overlap a lot, it's harder to claim one was truly better.

Our Welch's tests helped confirm what we saw visually (Franchise Figure 3):

- **NYY vs. ATL:** With a p-value around 0.28, I did not see a statistically significant difference. NYY's average (0.57) was a bit higher than ATL's (0.56), but not by enough to be sure it wasn't just chance.
- **NYY vs. BOS:** A p-value near 0.01067 showed a significant difference, suggesting NYY's mean WinPct (about 0.57) was truly higher than BOS's (about 0.53). This aligns with the idea that NYY consistently outperformed BOS during this period.
- **NYY vs. CHC:** With a p-value of roughly 1.83e-06, the difference was very strong. NYY's mean (~0.57) topped CHC's (~0.49) by a large margin, and the confidence interval fully supported this result.

By combining these tests with my plots, I gained both statistical and visual insights into how different franchises compare to each other. The sudden drop in all of the predictors around 2020, likely due to the COVID-19 pandemic, shows how environmental change can cause outliers in the data. Overall, these methods worked together to give me a richer understanding of how teams compared over time.

Timespan Analysis Section:

Methodology:

To see how key baseball performance metrics evolved from 1990 to 2023, I used the statistically significant predictors discussed in the offense and defense metrics section. The predictors are Runs (R), Earned Run Average (ERA), Runs Allowed (RA), and Errors (E). I calculated the averages for these predictors across my timespan. Instead of zeroing in on specific teams, this approach lets me see bigger trends and shifts in the game's overall offensive and defensive environment.

After finding the annual averages, I reorganized the data so I could plot all four metrics together. Using ggplot2, I made a set of line graphs (Timespan Figure 1). I applied LOESS smoothing (shown as dashed lines) to highlight long-term patterns and make reading the plot less susceptible to misinterpretations due to short term changes. Environmental changes like new training methods, strategies, or regulations are highlighted better as they represent slower and more gradual change.

Analysis and Results:

Figure 4 breaks down each each predictor in a separate plot, making readability easy:

- **Errors (E):** The red line shows a steady drop in errors over time, suggesting that fielding has gotten better as coaches, players, and equipment have improved.
- **Earned Run Average (ERA):** The green line shows that ERA went through ups and downs. High ERA periods might reflect stronger hitting or less effective pitching, while low ERA phases hint at better pitching and tighter defense.
- **Runs (R):** The blue line shows fluctuations in scoring, sometimes rising and sometimes dipping. These patterns could be tied to a wide assortment of changes such as player talent, ballpark conditions, or team strategies.
- **Runs Allowed (RA):** The purple line for RA looks very similar to ERA, since runs allowed depend heavily on pitching and defense (since ERA the number of earned runs a pitcher allows). When RA is higher it means that the opposing team had a strong offense; when RA is lower, the team being observed likely had good performances in pitching and defense.

Around 2020, **I** notice a sudden drop in these metrics, which is probably related to the COVID-19 pandemic. That season was shorter and less traditional, so the data from that year might not follow the usual patterns.

Though **I** did not run formal tests on these trends, the smoothed lines help **me** spot important shifts. The pronounced decline in errors is especially noteworthy, showing a clear improvement in defensive play. Meanwhile, the changing values for ERA, R, and RA suggest that many factors, such as league rules, player development, and environmental conditions, have shaped the game's style over the years.

These observations set the stage for further questions. For example, **I** might look into how team spending or coaching techniques influence these long-term patterns. In summary, these graphs and **my** analysis provide a valuable overview of how key aspects of the game have changed since 1990, with 2020 standing out as an unusual year due to the pandemic's impact.

Conclusion:

My project aimed to figure out what really influences a baseball team's success by looking closely at offensive and defensive stats from the Lahman "Teams" dataset, covering the years 1990 to 2023. I used a combination of statistical models, visual charts, and tests to see which factors matter most for winning games, and also to compare how well offensive and defensive numbers predict success. On top of that, I looked at how these important stats vary across different teams and over time.

The results showed a few clear takeaways. First, offensive numbers stood out, with runs scored being the most important predictor of wins. In my model, about 59% of the difference in teams' win totals could be explained by offensive metrics. Defensive factors also mattered, though slightly less. For example, walks (BB) ended up being a key defensive measure that helped teams succeed. About 54% of the difference in wins could be explained by these defensive stats. While this gives a bit of an edge to offense, my findings show that both sides of the game, hitting and pitching/fielding, matter for winning.

I also found a few surprises. For instance, hitting a lot of home runs or triples did not always mean more wins. This suggests that focusing too much on power hitting might not help if it's not part of a well-rounded approach. My initial guess was that offensive stats might be a bit stronger than defensive ones and that turned out to be true. Even though both sets of numbers were fairly close, it's clear you should not ignore one side of the game.

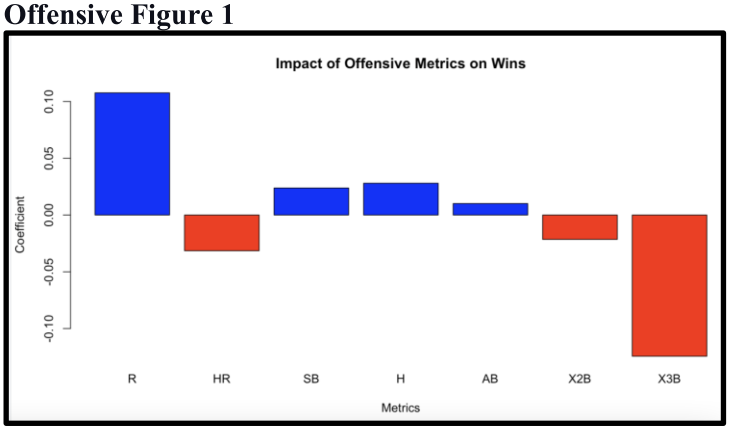
There were a few limitations. I only looked at one dataset and did not factor in things like injuries, spending, or weather, all of which can influence results. Also, while my models explained a good chunk of the differences in wins, there were still other unexplained factors. The unusual 2020 season, affected by the COVID-19 pandemic, also introduced some odd data points.

In the future, it would be interesting to dig deeper into how certain strategies, team finances, or coaching styles might shape performance. Adding more data, like information from other leagues or different eras could help me understand how the game is changing. Overall, I found that both offense and defense count a lot toward winning in baseball, with offense having a slight upper hand. My findings remind managers and coaches that a balanced, well-thought-out approach is key to achieving lasting success in the sport.

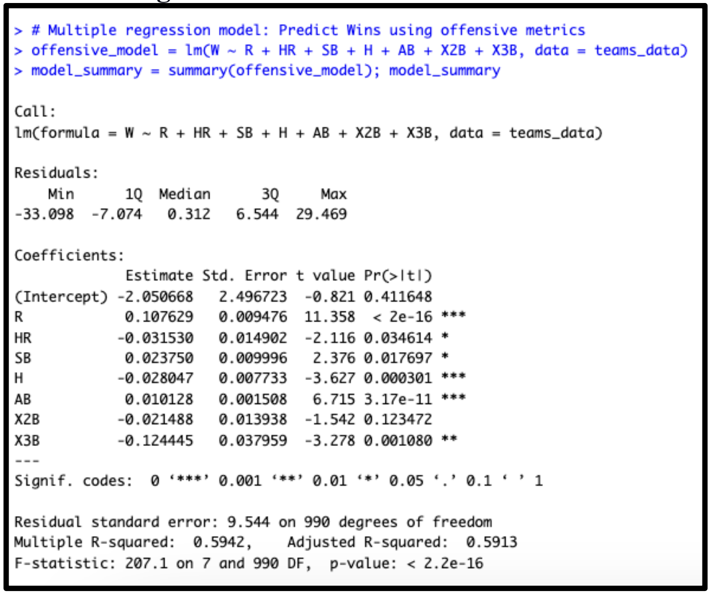
Appendices

Data Exploration/Visualization and Interpretation for offensive metrics:

Offensive Figure 1

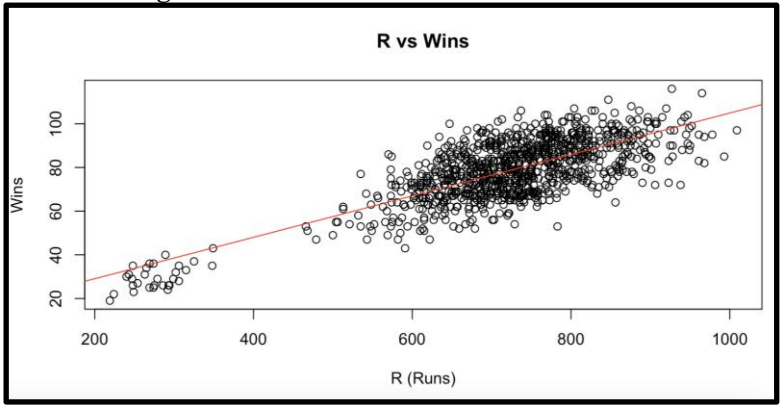


Offensive Figure 2



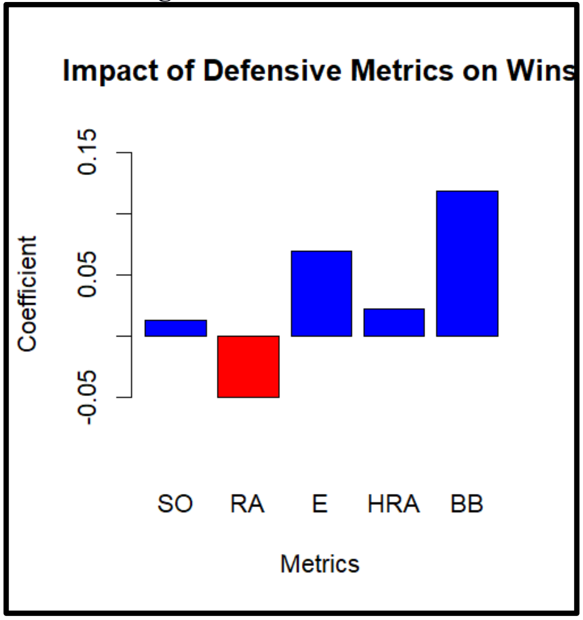
```
> # Multiple regression model: Predict Wins using offensive metrics
> offensive_model = lm(W ~ R + HR + SB + H + AB + X2B + X3B, data = teams_data)
> model_summary = summary(offensive_model); model_summary

Call:
lm(formula = W ~ R + HR + SB + H + AB + X2B + X3B, data = teams_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.098  -7.074   0.312   6.544  29.469 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.050668   2.496723  -0.821 0.411648    
R            0.107629   0.009476  11.358  < 2e-16 ***
HR          -0.031530   0.014902  -2.116 0.034614 *  
SB           0.023750   0.009996   2.376 0.017697 *  
H           -0.028047   0.007733  -3.627 0.000301 ***
AB           0.010128   0.001508   6.715 3.17e-11 ***
X2B         -0.021488   0.013938  -1.542 0.123472    
X3B         -0.124445   0.037959  -3.278 0.001080 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.544 on 990 degrees of freedom
Multiple R-squared:  0.5942,	Adjusted R-squared:  0.5913 
F-statistic: 207.1 on 7 and 990 DF,  p-value: < 2.2e-16
```

Offensive Figure 3

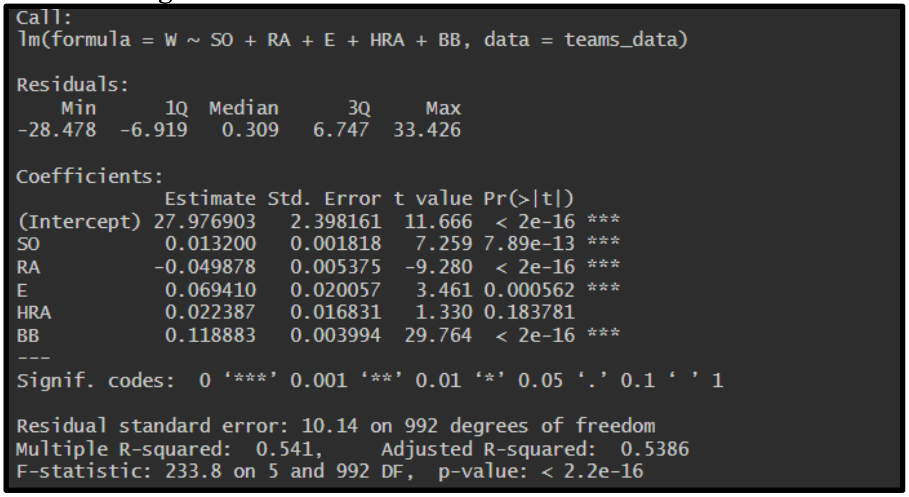


Data Exploration/Visualization and Interpretation for defensive metrics:

Defensive Figure 1

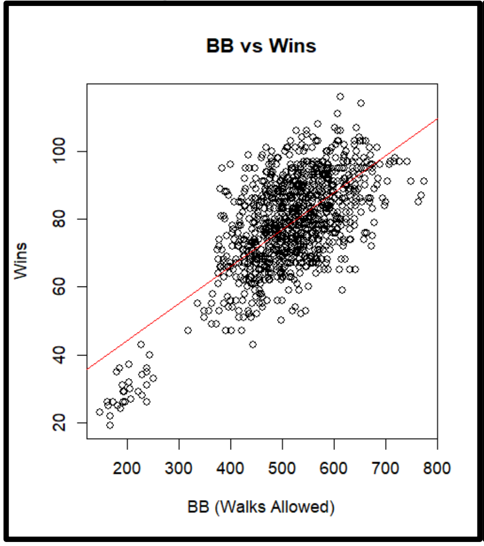


Defensive Figure 2

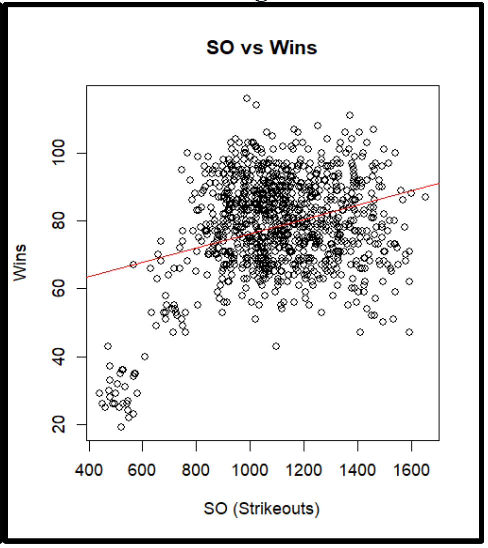


```
Call:
lm(formula = W ~ SO + RA + E + HRA + BB, data = teams_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.478  -6.919   0.309   6.747  33.426 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 27.976903   2.398161  11.666  < 2e-16 ***
SO           0.013200   0.001818   7.259 7.89e-13 ***
RA          -0.049878   0.005375  -9.280  < 2e-16 ***
E            0.069410   0.020057   3.461 0.000562 ***
HRA          0.022387   0.016831   1.330 0.183781    
BB           0.118883   0.003994  29.764  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.14 on 992 degrees of freedom
Multiple R-squared:  0.541,	Adjusted R-squared:  0.5386 
F-statistic: 233.8 on 5 and 992 DF,  p-value: < 2.2e-16
```

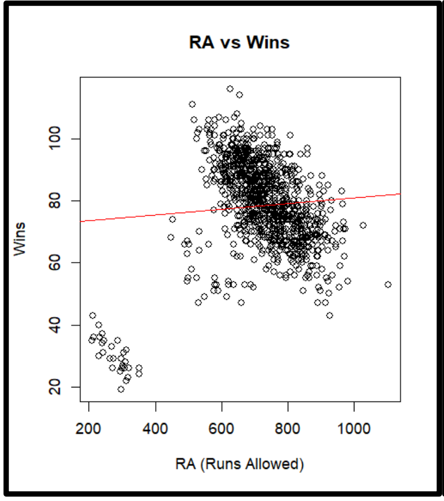
Defensive Figure 3



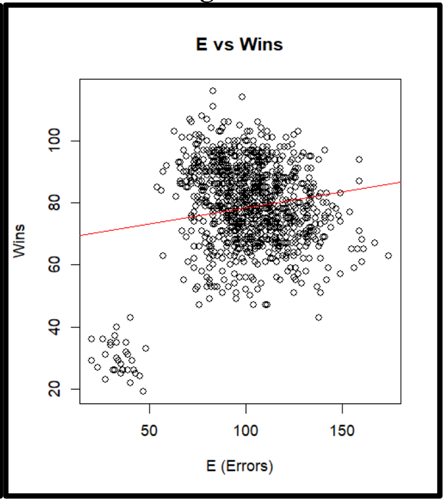
Defensive Figure 4



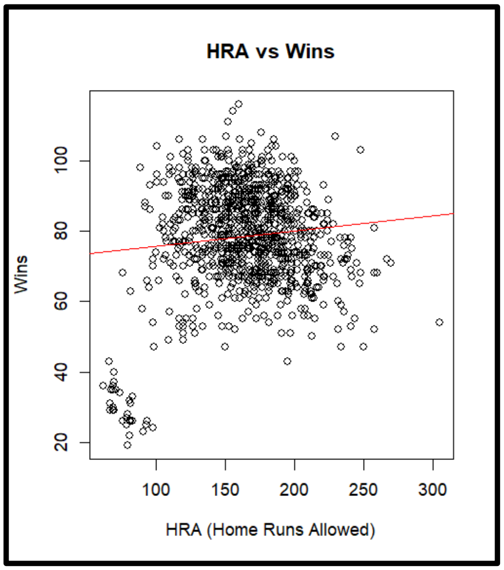
Defensive Figure 5



Defensive Figure 6



Defensive Figure 7



Franchise Figures and Output:

Figure 1:

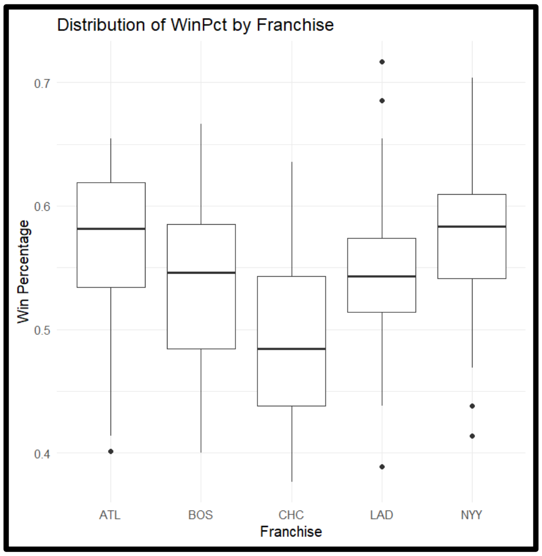


Figure 2:

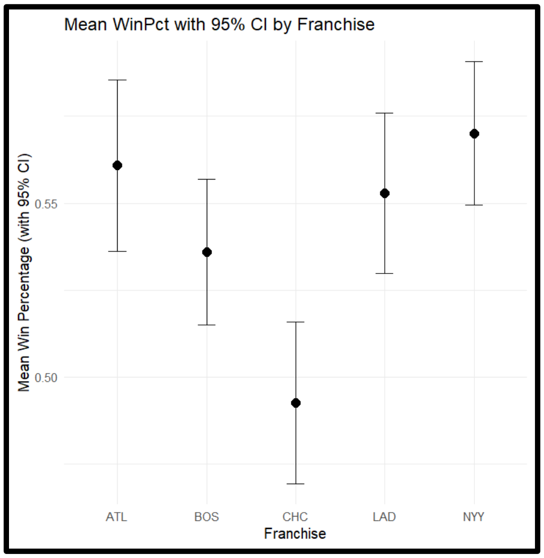
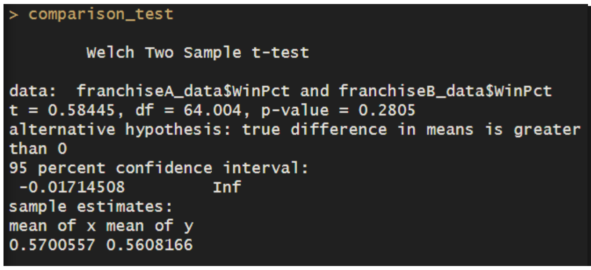


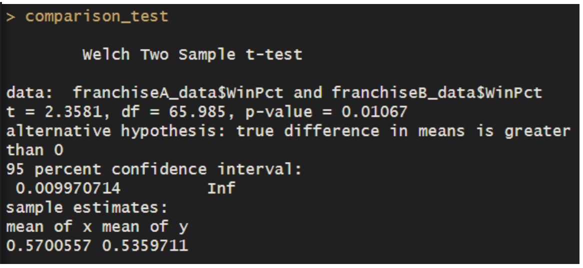
Figure 3:

NYY vs ATL:



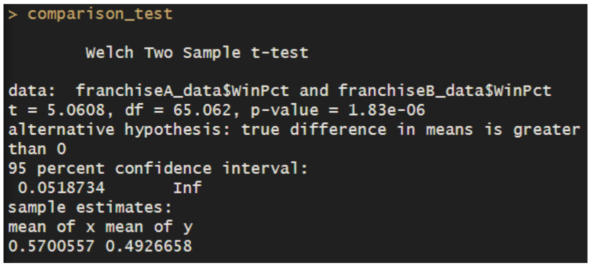
```
> comparison_test

	Welch Two Sample t-test

data:  franchiseA_data$WinPct and franchiseB_data$WinPct
t = 0.58445, df = 64.004, p-value = 0.2805
alternative hypothesis: true difference in means is greater
than 0
95 percent confidence interval:
 -0.01714508         Inf
sample estimates:
mean of x mean of y 
0.5700557 0.5608166 
```

NYY vs BOS:



```
> comparison_test

	Welch Two Sample t-test

data:  franchiseA_data$WinPct and franchiseB_data$WinPct
t = 2.3581, df = 65.985, p-value = 0.01067
alternative hypothesis: true difference in means is greater
than 0
95 percent confidence interval:
 0.009970714         Inf
sample estimates:
mean of x mean of y 
0.5700557 0.5359711 
```

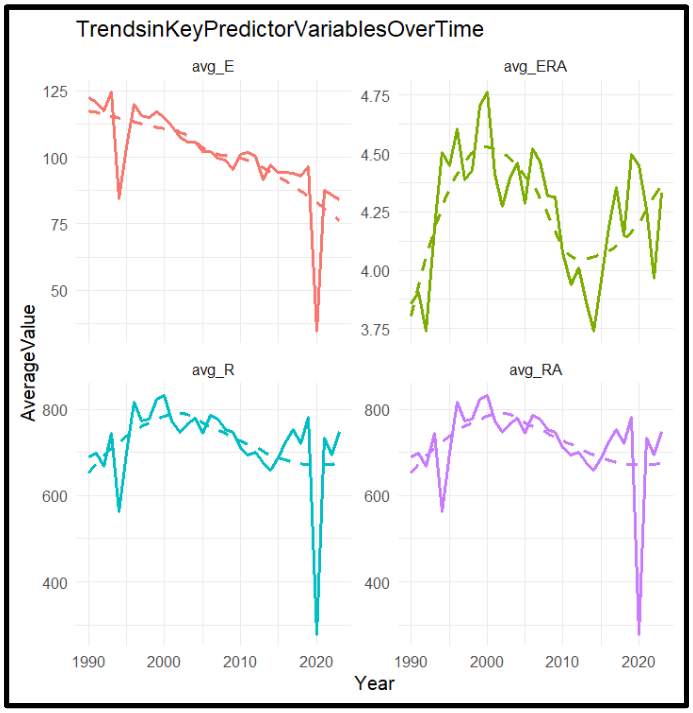
NYY vs CHC:



```
> comparison_test

	Welch Two Sample t-test

data:  franchiseA_data$WinPct and franchiseB_data$WinPct
t = 5.0608, df = 65.062, p-value = 1.83e-06
alternative hypothesis: true difference in means is greater
than 0
95 percent confidence interval:
 0.0518734       Inf
sample estimates:
mean of x mean of y 
0.5700557 0.4926658 
```

Timespan Figure:

Figure 1:


TrendsinKeyPredictorVariablesOverTime
avg_E
avg_ERA
avg_R
avg_RA
AverageValue
Year