The Impact of Hustle Plays in Winning NBA Games

By: Omar Chraibi

Agenda

- Introduction
- NBA Hustle statistics overview
- Data
- Statistical Methods (Regression models and Predictive Models)
- Takeaways
- Concluding Statements

Introduction

- What are we trying to answer?
- We are going to see if hustle plays lead to wins in NBA games.
- We will not be analyzing per team basis or who benefits the most, just a general overview of how impactful the hustle statistics can be when it comes to winning NBA games.
- We will be comparing the effects of the different hustle statistics.
- Eye Test vs. Numbers (intangibles vs tangibles)

Introduction

- What are Hustle plays? Diving into the Categories

TEAM	MIN	SCREEN ASSISTS	SCREEN ASSISTS PTS	DEFLECTIONS	OFF LOOSE BALLS RECOVERED	DEF LOOSE BALLS RECOVERED	LOOSE BALLS RECOVERED	% LOOSE BALLS RECOVERED OFF	% LOOSE BALLS RECOVERED DEF	CHARGES DRAWN	CONTESTED 2PT SHOTS	CONTESTED 3PT SHOTS	CONTESTED SHOTS
Atlanta Hawks	48.5	11.7	25.7	13.3	2.2	2.4	4.6	48.3	51.7	1.00	31.1	19.8	50.9
Boston Celtics	48.5	9.8	23.8	11.4	2.1	2.5	4.6	45.5	54.5	0.84	33.9	17.3	51.2
Brooklyn Nets	48.2	8.5	19.6	14.1	2.1	2.4	4.5	46.8	53.2	0.10	31.8	19.9	51.8
Charlotte Hornets	49.2	9.3	21.3	17.2	2.0	2.6	4.6	43.5	56.5	0.20	28.1	18.5	46.6
Chicago Bulls	48.3	9.7	21.6	16.8	2.3	2.7	5.0	45.8	54.2	0.63	29.1	20.7	49.7
Cleveland Cavaliers	49.3	9.2	22.0	11.6	2.6	2.3	4.9	52.7	47.3	0.84	30.9	18.5	49.4
Dallas Mavericks	48.6	6.1	14.3	14.3	2.1	2.7	4.8	43.7	56.3	0.33	32.7	16.7	49.4
Denver Nuggets	48.3	7.2	16.5	14.4	2.5	2.5	5.0	49.5	50.5	0.37	26.8	19.2	46.0
Detroit Pistons	48.0	9.4	21.0	13.4	2.7	2.3	5.0	54.3	45.7	0.19	30.6	20.1	50.7
Golden State Warriors	48.3	10.7	27.3	15.2	1.9	2.2	4.0	45.7	54.3	0.45	32.9	23.1	56.0
Houston Rockets	48.0	7.2	17.2	16.4	2.6	2.1	4.7	55.6	44.4	0.53	26.6	20.9	47.5
Indiana Pacers	48.0	8.3	19.2	14.7	3.1	3.2	6.3	49.1	50.9	0.72	31.7	18.6	50.3
LA Clippers	48.0	8.8	19.5	14.4	1.6	2.5	4.0	38.3	61.7	0.30	30.8	18.4	49.1
Los Angeles Lakers	48.3	5.4	12.3	15.7	2.1	1.7	3.8	54.4	45.6	0.72	33.4	20.7	54.1
Memphis Grizzlies	48.5	10.5	24.1	14.0	2.3	1.8	4.1	56.4	43.6	0.58	29.0	20.3	49.3

More on Hustle Plays

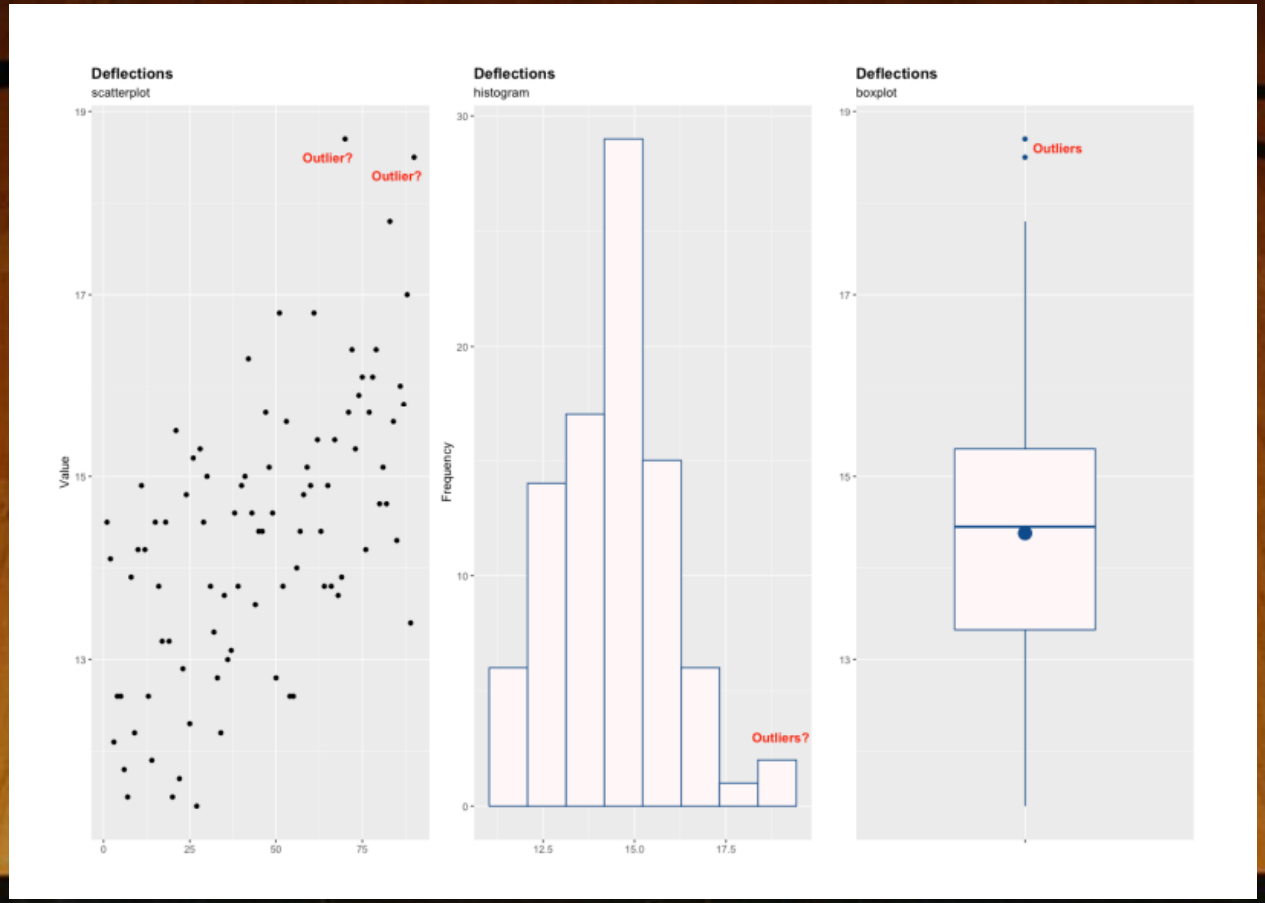
- The NBA started capturing and publishing hustle statistics since the 2016-2017 season.
- The underlying hypothesis is that hustle plays could have an influence on wins and losses.
- The purpose here is to test whether this hypothesis is true, this includes assessing which of these statistics are significant, and trying to quantify their effect.
- Data comes directly from the NBA website; we will be looking at hustle from a team perspective.
- We will be analyzing the relationship between these predictors and the target variable, which is the number of wins in a season.

Data

- Our data set spans from three NBA regular seasons 2016-2019, pre-COVID-19
- Given the dataset, we will have to clean some of the data including identifying and removing outliers.
- The following slide will show a visualization of outliers in our deflection statistic.
- We would adjust the values of spotted outliers by using winsorization, a statistical method that mitigates the effects of outliers by replacing them with less extreme values.
- Following this process, we will check the normality of each variable's frequency distribution (shape).

Identifying Outliers

- Here, we have an example of our deflection's statistic, here we can see some clear outliers using different types of visualizations.



Checking Normality of the Data

- We will use a Shapiro Wilk-Test to check that the data is normally distributed.

- We did find that all variables except for the variable charges has a non-normal distribution, based on the Shapiro-Wilk tests and drawing a line in the sand where the p-value is above or below the 0.05 threshold for significance.

```
> shapiro.test(hustle$charges)

        Shapiro-Wilk normality test

data:  hustle$charges
W = 0.95688, p-value = 0.004562
```

Correlation between Wins and Predictors

- Next, we will look to compute the correlation coefficients between the variable wins and the remaining variables and visualize the same with a correlation matrix.

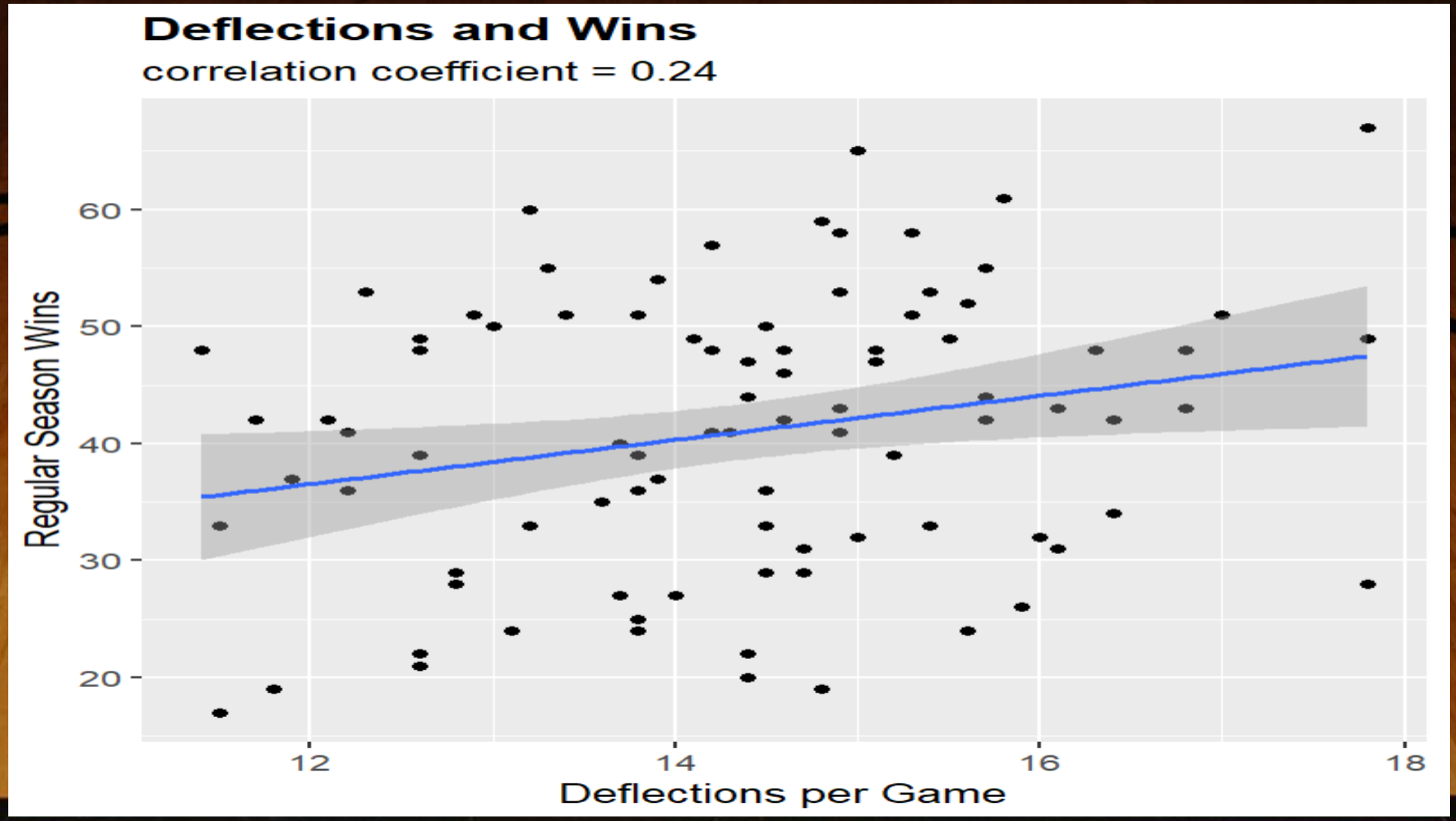
- Our purpose here is to identify which variables might be best fits, or not fits at all, as predictors in our linear regression model.

- This is mainly an exploratory step to visualize the linear relationship between wins and hustle statistics.

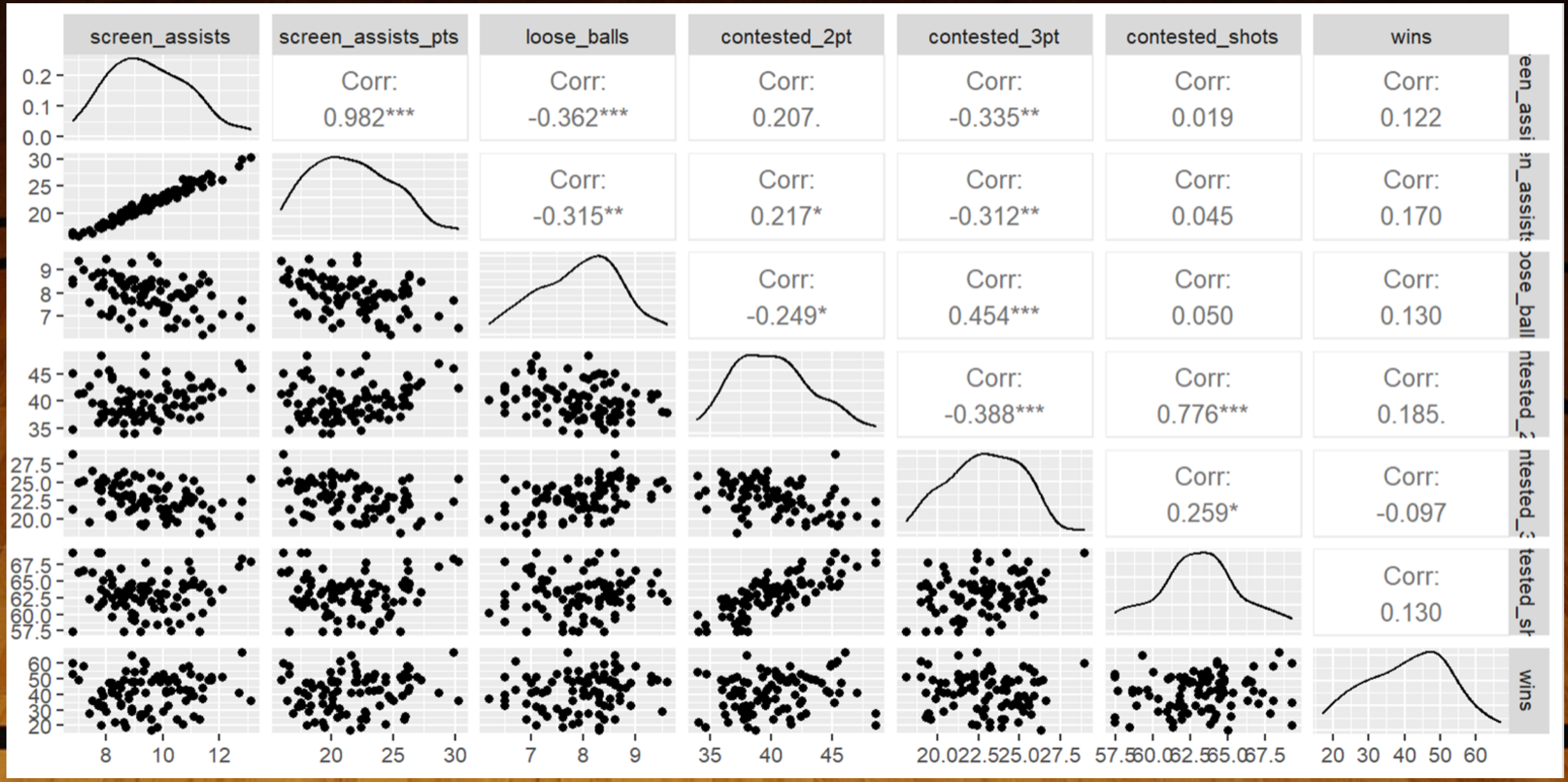
Correlation Results

- First, we looked at the correlation coefficient between wins and deflections. We found that the correlation coefficient between deflections and wins is 0.24, which is relatively weak.
- After, we calculated all correlation coefficients using a correlation coefficient matrix. We found that none of the remaining hustle variables has a strong correlation, one way or the other, with wins. (Deflections was the correlation coefficient in regard to wins)

Correlation Visualizations: Deflections and Wins



Correlations: Correlation Coefficient Matrix



Regression Analysis

- Now we will be using linear regression to test the linear relationship between wins and hustle statistics

- We will split the data into train and test data to reduce overfitting
- We created a linear regression model called fit1

```
fit1 <- lm(wins ~ screen_assists_pts + deflections + loose_balls + contested_2pt +
       contested_shots, data = train)
```

Regression Analysis fit1:

- Using the tidy function, we can get the coefficients estimates and p-values.
- Using the glance function, we can see that the adjusted r^2 is accounts for about 14% of variance in wins. (2 insignificant predictors)
- We can see the VIF function to check for multicollinearity (above 5), but none of our coefficients show sign of multicollinearity.

```
> tidy(fit1)
# A tibble: 6 × 5
  term                 estimate std.error statistic p.value
  <chr>                   <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)           -62.9       38.6      -1.63   0.108
2 screen_assists_pts      1.04       0.441     2.35   0.0219
3 deflections             2.23       0.882     2.53   0.0138
4 loose_balls             5.38       2.19      2.45   0.0170
5 contested_2pt           0.525      0.763     0.688  0.494
6 contested_shots        -0.241      0.790    -0.305  0.761
> glance(fit1)
# A tibble: 1 × 12
  r.squared adj.r.…¹ sigma stati…² p.value    df logLik   AIC   BIC
      <dbl>    <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.202    0.137  10.9    3.09  0.0150     5  -252.  518.  533.
> vif(fit1)
screen_assists_pts         deflections         loose_balls
          1.155995            1.062677            1.314189
     contested_2pt     contested_shots
          3.052934            2.637588
>
```

Regression Analysis continued: Fit2

- Since we could not establish a high significance of the hustle statistics, we will create a reduced model called fit2, removing the hustle statistics that have a lesser effect on winning

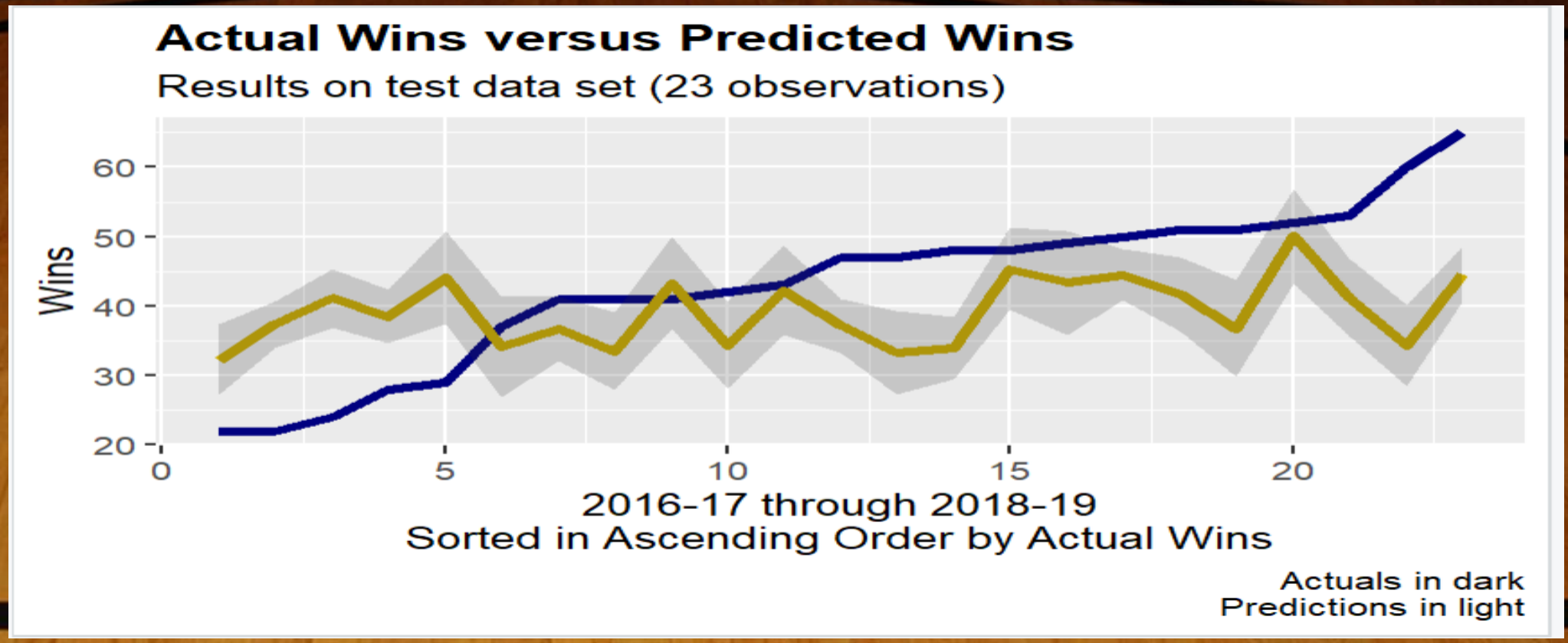
```
fit2 <- lm(wins ~ screen_assists_pts + deflections + loose_balls, data = train)
> tidy(fit2)
# A tibble: 4 × 5
  term                estimate std.error statistic p.value
  <chr>                  <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)            -56.1      25.5      -2.20 0.0317
2 screen_assists_pts       1.12      0.422     2.65 0.0101
3 deflections              2.35      0.859     2.74 0.00805
4 loose_balls              4.81      1.98      2.42 0.0182
> glance(fit2)
# A tibble: 1 × 12
  r.squared adj.r.…¹ sigma stati…² p.value    df logLik   AIC   BIC
      <dbl>    <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.194    0.156  10.8    5.06 0.00334     3  -252.  514.  525.
> vif(fit2)
screen_assists_pts        deflections        loose_balls
          1.085184           1.031128           1.098619
```

Comparing fit1 vs fit2

- We can see that fit2 is a better fit than fit1
- All fit2 predictors have a p-value lower than 0.05
- The Adjusted R^2 statistic for fit2 equals 0.16, versus 0.14 for fit1
- Fit2 has a lower AIC, meaning the model was better
- The average difference between actual and fitted wins is slightly greater in fit2 (8.43) versus fit1 (8.27);

Predictive modeling

- We used fit2 to develop the predictive model for regular season wins
- We plot Actual vs. Predicted Wins (Shaded area CI interval)
- fit2 does a much better job of predicting wins for teams that finished at or near .500 versus teams that had an extreme number of regular season wins. (low or high)



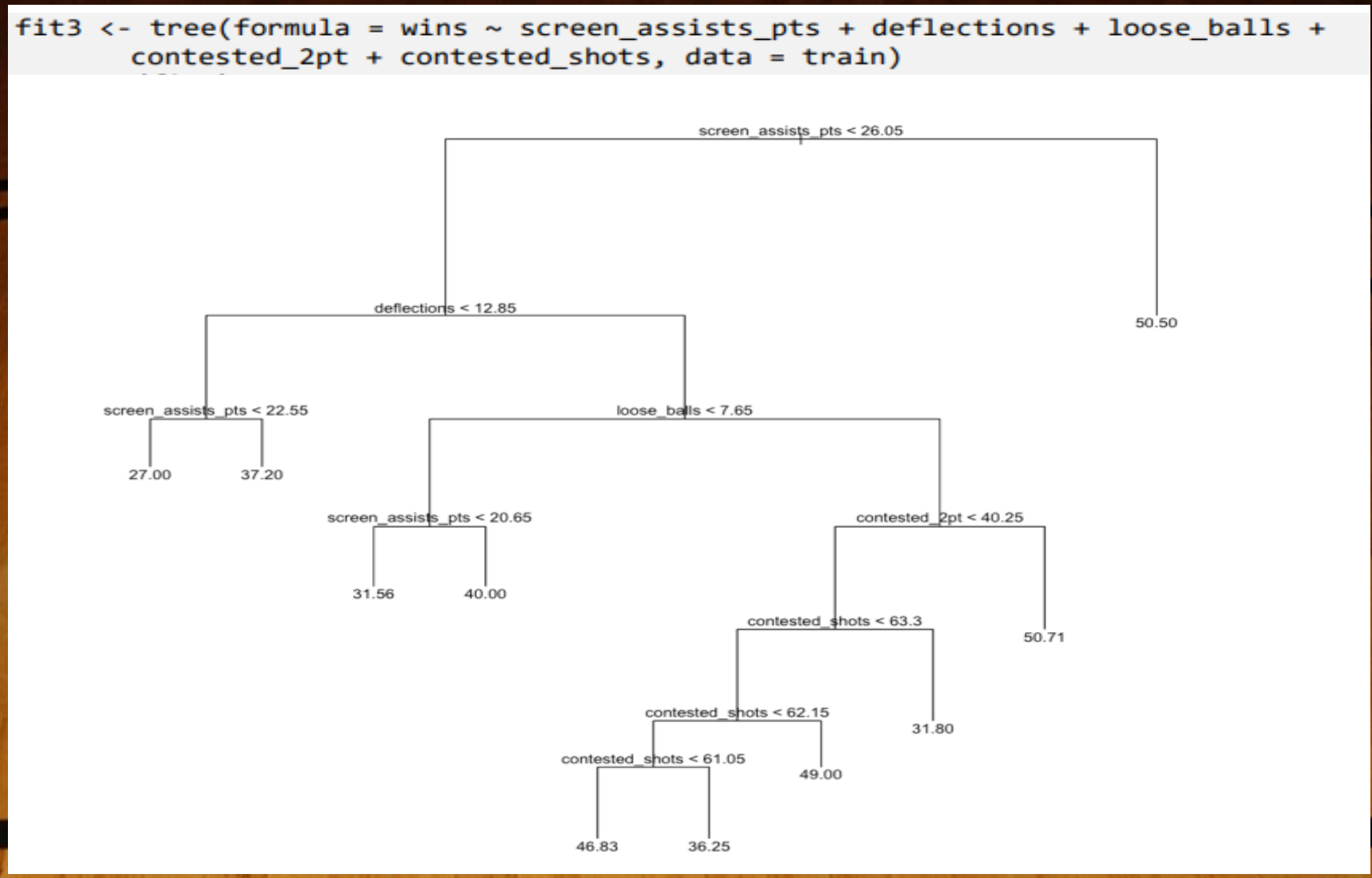
Takeaways after Predictive Modeling

- We wanted to identify if hustle statistics might have a statistically-significant impact on wins and to quantify that impact.
- We only able to account for about 16% variance in regular seasons wins using three hustle statistics: points off screens, deflections, and loose balls recovered.
- These findings seem to indicate that we need a wider data set to include additional statistics that contribute to winning or introduce a more accurate, interpretable model
- To interpret the results better, we will use a regression tree.

Regression Tree

- We will use a Regression tree, since they are relatively easy to construct and interpret. (41 is the average # of wins over the 3 seasons)

```
fit3 <- tree(formula = wins ~ screen_assists_pts + deflections + loose_balls +
       contested_2pt + contested_shots, data = train)
```



Regression Tree Results

- Teams that average more than 26.05 points off screens per game can be expected to win 50 or 51 regular season games.
- Alternatively, teams that average fewer than 26.05 points off screens per game can be expected to win anywhere between 27 and 51 regular season games, depending on other variables and other splits.
- Teams that average fewer than 26.05 points off screens and fewer than 12.85 deflections per game can be expected to win somewhere between 27 and 37 regular season games.
- Teams that average fewer than 26.05 points off screens but more than 12.85 deflections per game can be expected to win somewhere between 31 and 51 regular season games.

Summary: Part 1

- The purpose of this project was to determine which hustle statistics significantly impact regular season wins in the NBA

- First, we used a multiple linear regression model after a thorough analysis of the data. We mitigated and removed outliers, removed variables with a non-normal distribution, and checked for overfitting.

- Second, we only incorporated predictors with the strongest correlation of the dependent variable, and checking variables based on multicollinearity analysis.

- The two linear regression models did not explain or predict wins with much accuracy but identified 3 hustle statistics with the greatest effects that could have an impact on winning.

Summary: Part 2

- Our regression tree model confirmed the same variables identified by the linear regression models as being the most significant.
- To look for more accurate results, a random forest model was considered and applied, but the results did not significantly change the predictive model accuracy.
- This could be due to the lack of data provided (Only 3 seasons) and the small number of predictors (8 predictors). Generally, linear models need less data to achieve better results.

Conclusion

- We were able to determine that with predictor variables: points off screen assists, deflections, and loose balls, were the most impactful predictors. Statistically, positively led to wins and provided the greatest return on hustle effort.

- The recommendation for the next step is to use a wider data set that includes additional variables like shots made, shots attempted/made, turnovers margins etc. Variables that are better equipped for predicting wins.
- Linear Regression analysis might not accurately explain/predict wins but did provide useful statistics/insight. This exploratory analysis enabled us to identify which hustle statistics are the most significant, and which one should be discarded.

- We need more hustle data, as NBA only started recording hustle data in 2016, the current data set seems to be too small to infer significant insights.

Questions?