Regression Analysis

(You are expected to give a descriptive title)

Contents

1 Introduction

The National Football League (NFL) is a syndicate of 32 teams which over the last several decades has become not only America's favorite past time but also one of the most profitable leagues globally. In 2023, the league generated over 20 billion dollars in revenue and held 93 spots in the top 100 most watched broadcasts. A large part of the success the league enjoys is due to its scarcity. The regular season is only 18 weeks long, where each team plays 17 games along with one bye week for rest. This is in sharp contrast to other sports leagues where teams might play hundreds of games.

Although the NFL season is short, there is an abundance of data generated each game and over the span of a season. Finding ways to leverage this data is important to the success and health of a team and its

players. More recently, fantasy leagues and the growing popularity of sports gambling have amplified the significance of this data for fans and Wall Street investors seeking profitable opportunities.

For this project, we selected a relatively simple data set encompassing box score statistics for each NFL team during the 2023 regular season gathered from pro-football-reference (https://www.pro-football-reference.com/). Our data consists of 544 observations and 25 features (23 predictors and 2 response) which are listed below. Our goal is simple:

1. Can we use box score statistics to predict the points scored by a team via linear regression?
2. Can we use box score statistics to predict the result of the game via logistic regression?

`Team` : Name of the team
`Week` : Week of the season
`Day` : Day the game was played
`Date` : Date the game was played
`Time` : Time the game was played
`Result` : W if the won or L if they lost
`OT` : If the game went into overtime
`Rec` : Win-loss record
`isHome` : Weather the team played at home or away
`Opp` : Opponent the team played against
`Tm_score` : Points scored
`Opp_score` : Points scored by opponent
`1stD_Off` : First downs gained by offense
`TotYd_Off` : Total yardage gained by offense
`PassY_Off` : Passing yardage gained by offense
`RushY_Off` : Rushing yardage gained by offense
`TO_Off` : Turnovers by the offense
`1stD_Def` : First downs given up by defense
`TotY_Def` : Total yardage given up by defense
`PassY_Def` : Passing yardage given up by defense
`RushY_Def` : Rushing yardage given up by defense
`TO_Def` : Turnovers caused by defense
`OffenseExp` : Expected points by offense
`DefenseExp` : Expected points by defense
`SpTms_Exp` : Expected points by special teams

Note: The expected points features are calculated via play by play data. According to pro-football-reference, expected points represent the estimated point value at the start of a given play, based on down, distance, and field position.

2 EDA and Feature Engineering

2.1 Cleaning Data and Feature Engineering

A quick glance at our data, shows us that almost all of our columns have missing values. We handle this in two ways:

1. The 32 missing values in many of our columns occur because this data includes the bye where no data is generated. We drop these rows.

2. The other missing values are due to the way our source inputs the data, leaving things blank when an event does not occur. For example, in the OT column, if the game goes into overtime, OT is listed, otherwise, the cell is left blank. Similarly, if a team does not generate a turnover, the entry is blank. We fill these missing values in as zero.

```
    Team                Week          Day                Date
Length:576         Min.   : 1.0   Length:576         Length:576
Class :character   1st Qu.: 5.0   Class :character   Class :character
Mode  :character   Median : 9.5   Mode  :character   Mode  :character
                   Mean   : 9.5
                   3rd Qu.:14.0
                   Max.   :18.0

    Time              boxscore            Result              OT
Length:576         Length:576         Length:576         Length:576
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character

    Rec               atHome              Opp              Tm_score
Length:576         Length:576         Length:576         Min.   : 0.00
Class :character   Class :character   Class :character   1st Qu.:16.00
Mode  :character   Mode  :character   Mode  :character   Median :21.00
                                                         Mean   :21.77
                                                         3rd Qu.:28.00
                                                         Max.   :70.00
                                                         NA's   :32
  Opp_score       X1stD_Off       TotYd_Off       PassY_Off
Min.   : 0.00   Min.   : 6.00   Min.   : 58.0   Min.   : -9.0
1st Qu.:16.00   1st Qu.:16.00   1st Qu.:273.0   1st Qu.:167.8
Median :21.00   Median :19.00   Median :335.0   Median :215.0
Mean   :21.77   Mean   :19.26   Mean   :331.6   Mean   :218.9
3rd Qu.:28.00   3rd Qu.:23.00   3rd Qu.:389.0   3rd Qu.:268.0
Max.   :70.00   Max.   :33.00   Max.   :726.0   Max.   :472.0
NA's   :32      NA's   :32      NA's   :32      NA's   :32
  RushY_Off         TO_Off        X1stD_Def        TotY_Def
Min.   : 17.0   Min.   :1.000   Min.   : 6.00   Min.   : 58.0
1st Qu.: 77.0   1st Qu.:1.000   1st Qu.:16.00   1st Qu.:273.0
Median :107.0   Median :2.000   Median :19.00   Median :335.0
Mean   :112.7   Mean   :1.877   Mean   :19.26   Mean   :331.6
3rd Qu.:141.0   3rd Qu.:2.000   3rd Qu.:23.00   3rd Qu.:389.0
Max.   :350.0   Max.   :6.000   Max.   :33.00   Max.   :726.0
NA's   :32      NA's   :185     NA's   :32      NA's   :32
  PassY_Def       RushY_Def         TO_Def        OffenseExp
Min.   : -9.0   Min.   : 17.0   Min.   :1.000   Min.   :-35.860
1st Qu.:167.8   1st Qu.: 77.0   1st Qu.:1.000   1st Qu.: -6.798
Median :215.0   Median :107.0   Median :2.000   Median :  1.990
Mean   :218.9   Mean   :112.7   Mean   :1.877   Mean   :  1.628
3rd Qu.:268.0   3rd Qu.:141.0   3rd Qu.:2.000   3rd Qu.:  9.895
Max.   :472.0   Max.   :350.0   Max.   :6.000   Max.   : 48.650
NA's   :32      NA's   :32      NA's   :185     NA's   :32
  DefenseExp        SpTms_Exp
```

```
Min.   :-48.650   Min.   :-12.380
1st Qu.: -9.895   1st Qu.: -2.845
Median : -1.990   Median :  0.000
Mean   : -1.628   Mean   :  0.000
3rd Qu.:  6.798   3rd Qu.:  2.845
Max.   : 35.860   Max.   : 12.380
NA's   :32        NA's   :32
```

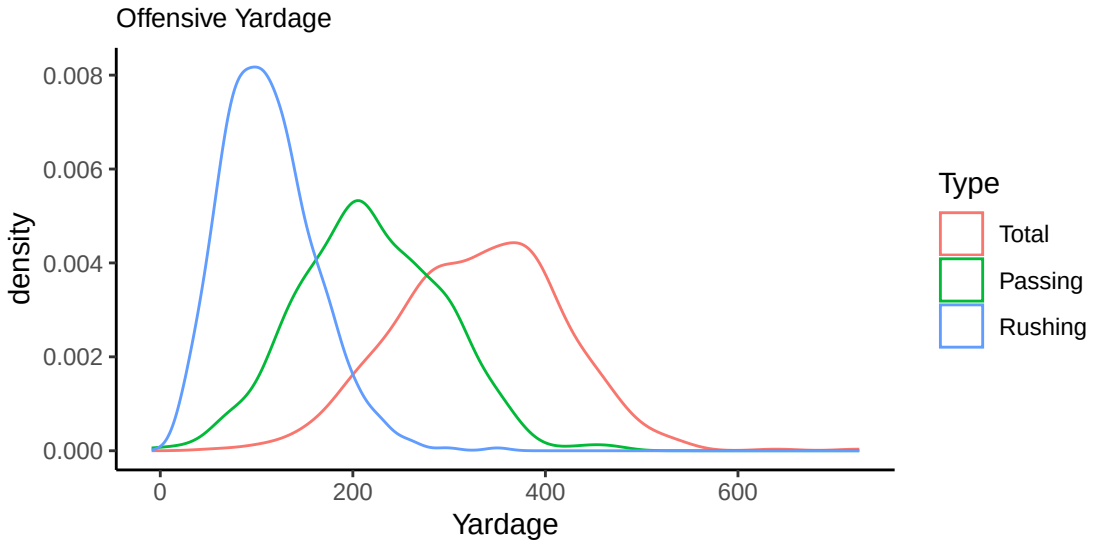
To clean our data, we set our binary variables to be either 1 or 0. We also create, two new binary variables.

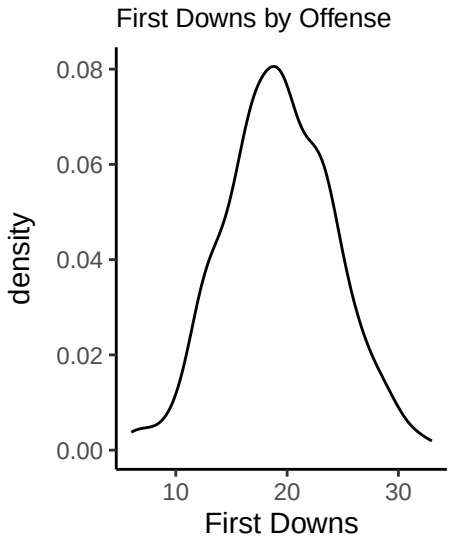
1. We use the `Time` column to create a new variable called `isPrimeTime` which lets us know whether or not the game was played at night on national television.
2. A `isThursday` feature is created from the `Day` column which lets us know if the game was played on Thursday. Thursday night games are played on a short week which can lead to sloppy play and increased risk of injury.

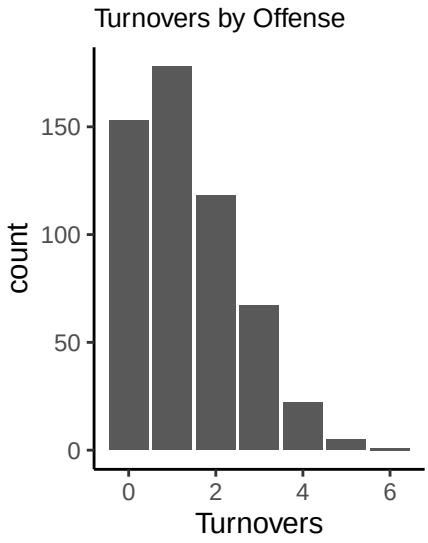
After dropping unnecessary columns and renaming other columns, we are left with the following features: `OT`,`atHome`, `Tm_score` , `Opp_score`, `1stD_Off`, `TotYd_Off`, `PassY_Off`, `RushY_Off`, `TO_Off`, `1stD_Def`, `TotY_Def`, `PassY_Def`, `RushY_Def`, `TO_Def`, `OffenseExp`, `DefenseExp`, `SpTms_Exp`, `isPrimeTime`, and `isThursday`.

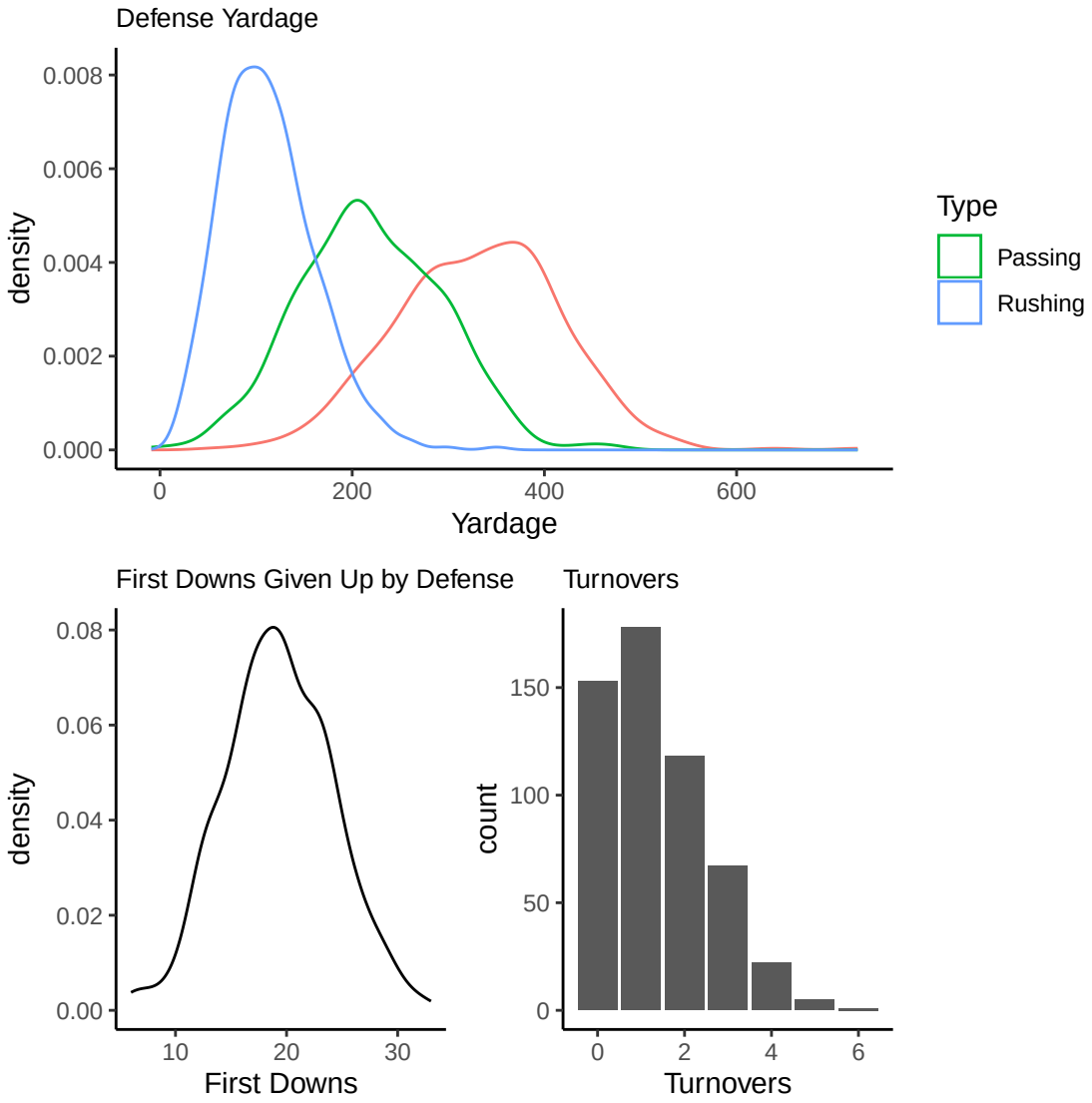
2.2 Visualising Distributions

The distribution for our continuous numerical variables for offense and defense are shown below. The first down rate, total yardage, and passing yards look to be normally distributed. The rushing yards have a slight skew but nothing major catches the eye. Turnovers on the other hand have a noticeable right skew which is to be expected.

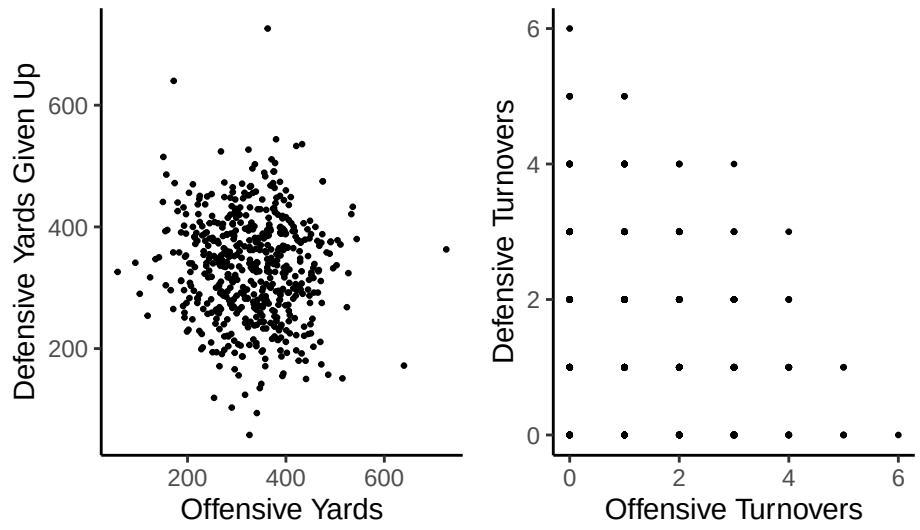

Offensive Yardage
density
0.008
0.006
0.004
0.002
0.000
0
200
400
600
Yardage
Type
Total
Passing
Rushing


First Downs by Offense
density
0.08
0.06
0.04
0.02
0.00
10
20
30
First Downs

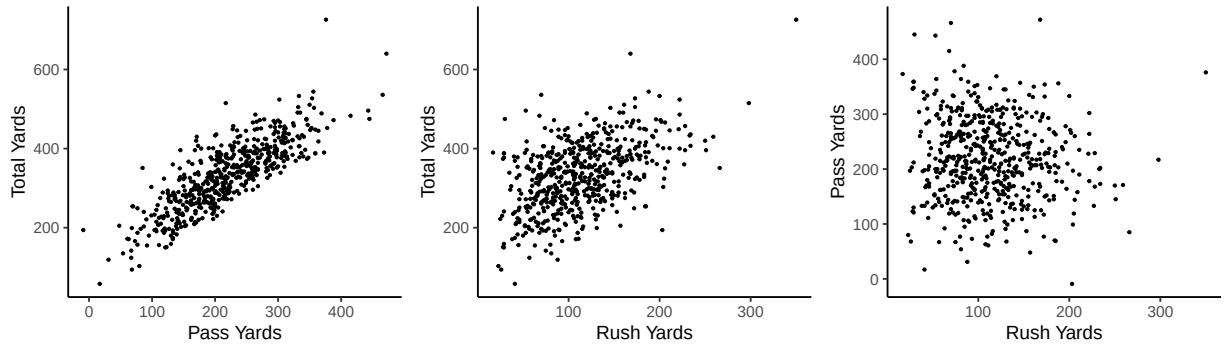

Turnovers by Offense
count
150
100
50
0
0
2
4
6
Turnovers



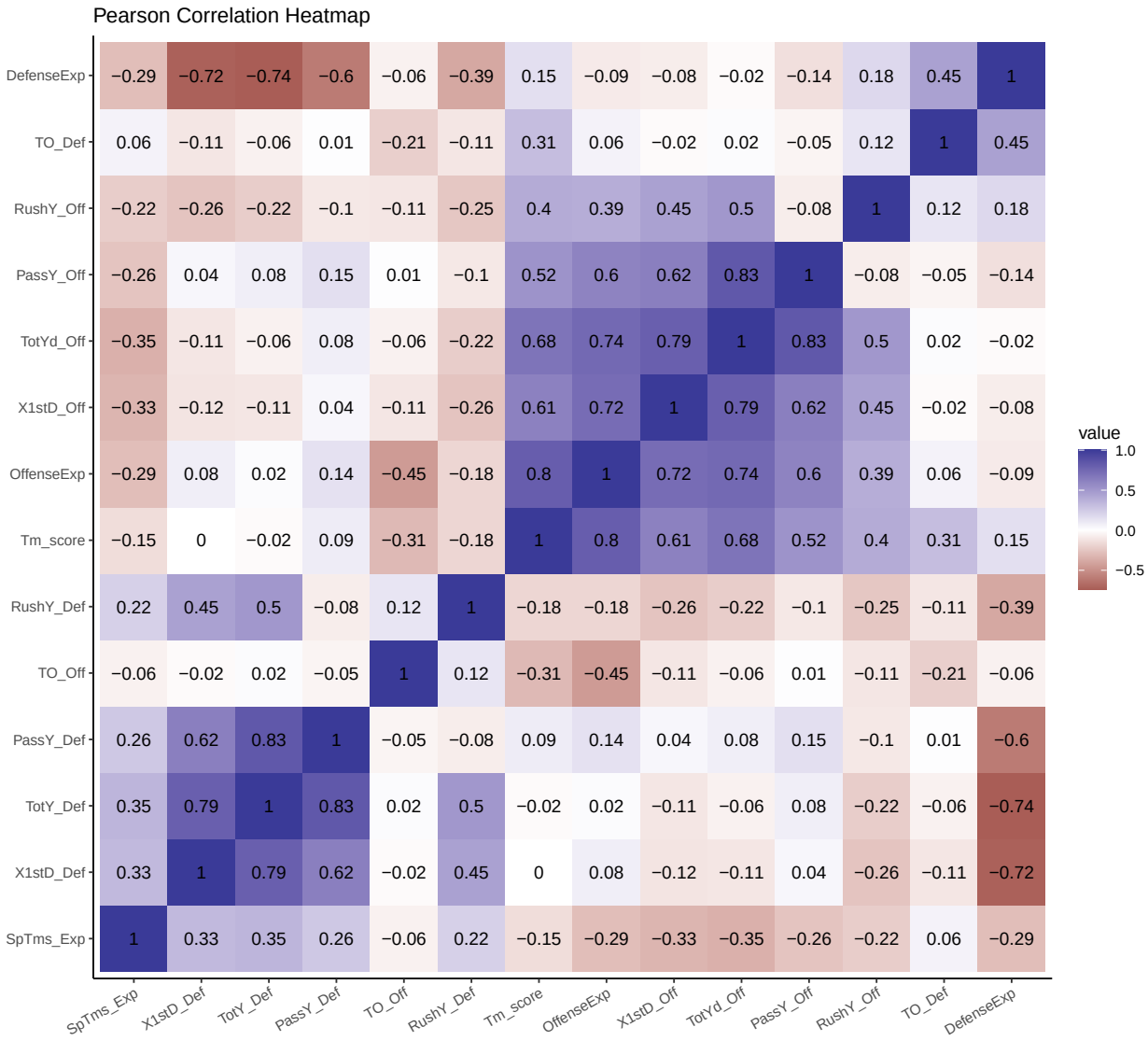
Comparing the defensive plots vs the offensive plots shows that the distributions are exactly the same but this is to be expected. Since two teams play in a game, we have observations for both teams from the same game. For example, suppose the Eagles play the Giants and gain 400 total yards. This will show up as 400 under `TotYd_Off` but will show up again in another observation of the Giants as `TotY_Def`. We see in the two plots below that the offensive stats do not correlate with the defensive stats for each observation.



It should be expected that passing and rushing yards are correlated with total yards which is indeed the case. Surprisingly, there is little correlation between rush and pass yards.

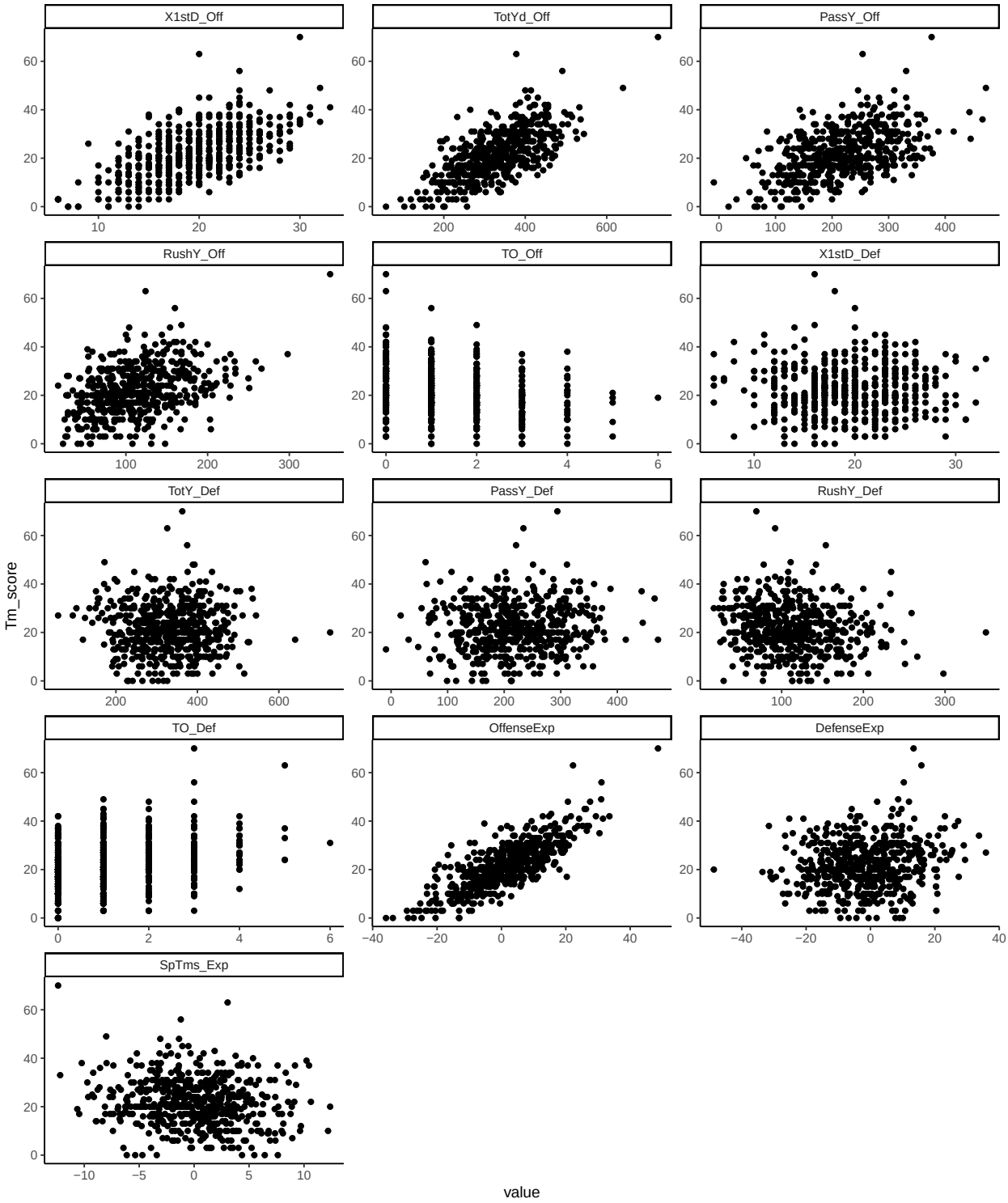


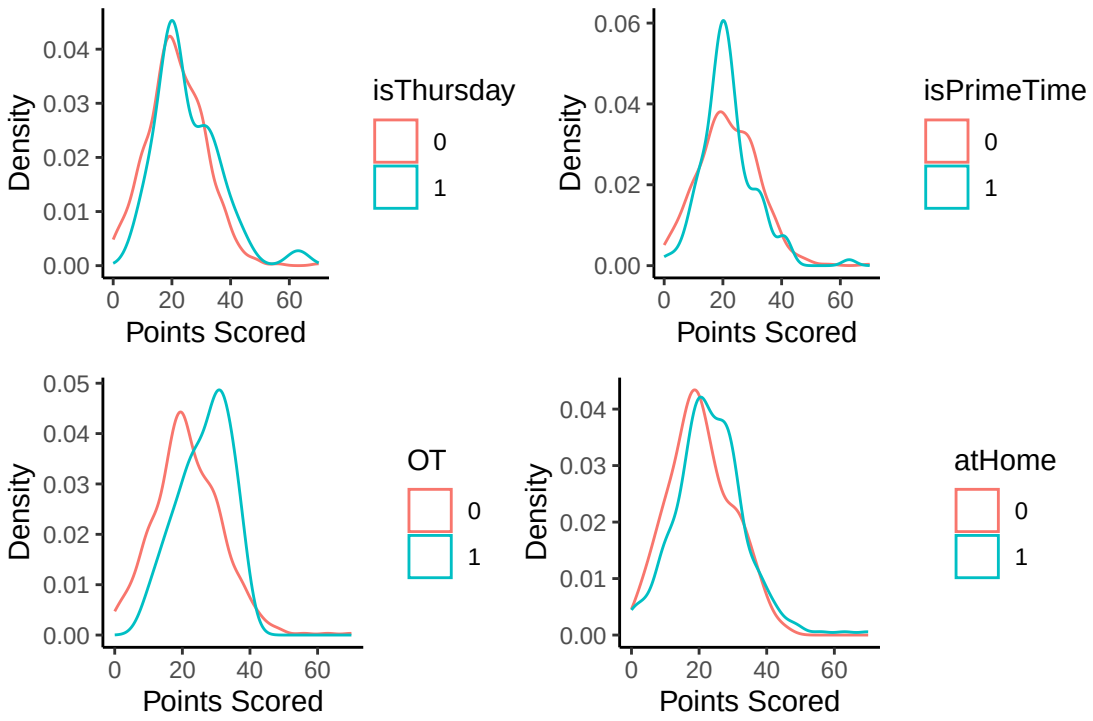
To further capture relationships between our numerical variables we create a Pearson correlation plot. We see that offensive expected points feature is highly correlated with yardage. The same applies for defensive expected points and yardage given up. And of course the total yardage is dependent on rushing and passing yardage.



2.3 Relationship to response variables

We would now like to take a closer look at the relationship between our features and the response variables.To capture this relationship, we create a scatter plot for each numerical variable against our response variable. For our categorical variables, we create density plots.

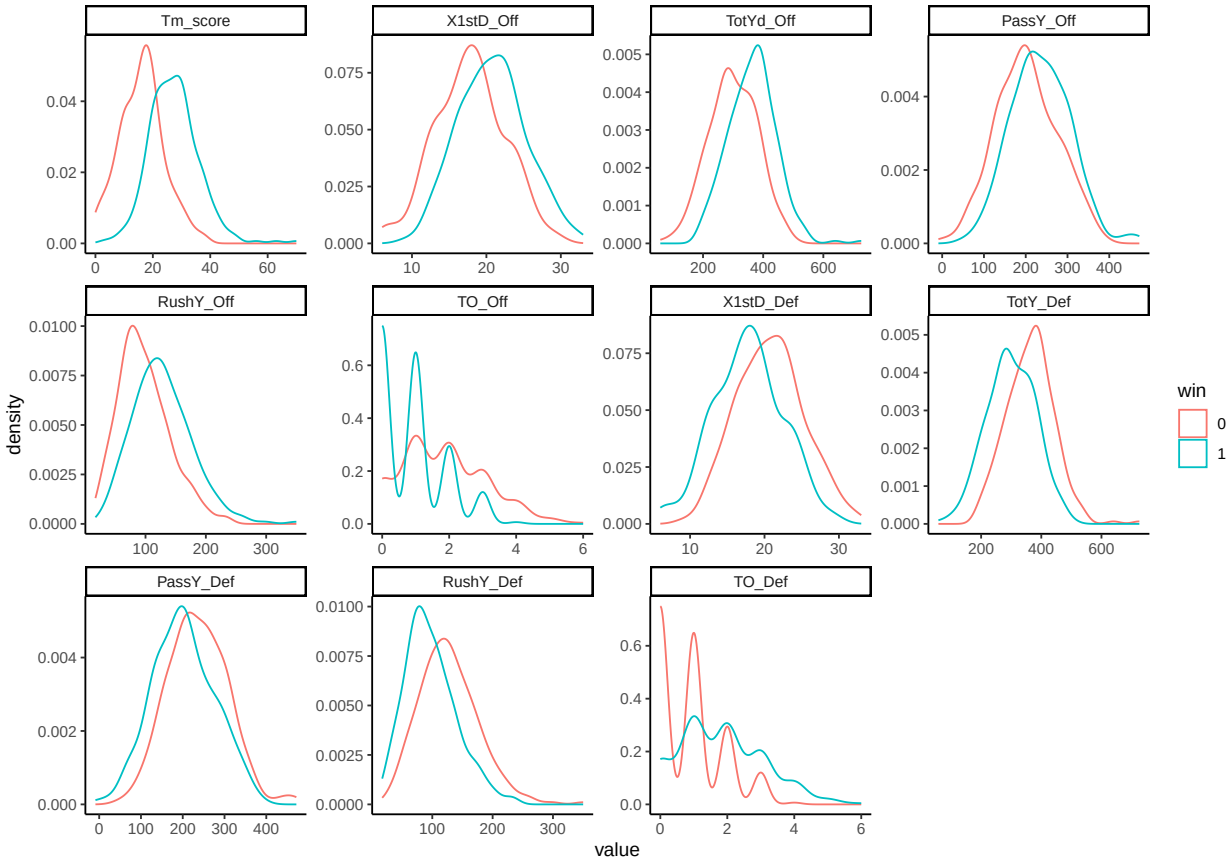

X1stD_Off
TotYd_Off
PassY_Off
RushY_Off
TO_Off
X1stD_Def
TotY_Def
PassY_Def
RushY_Def
TO_Def
OffenseExp
DefenseExp
SpTms_Exp
Tm_score
value



Observations:

1. Defensive statistics have no bearing on the points scored which is not too surprising since defenses rarely score points.
2. Special teams expected points has no correlation with points scored. This is a bit surprising since, field goals and extra points are a significant chunk of points scored during a normal game.
3. Turnovers may have a slight correlation.
4. A difference in the distribution between games that went into overtime vs those that did not. Similarly, home vs away games have a slight difference in distributions.

Plotting our features against our second response variable, we see that almost all features have a difference in there distribtions for a win vs a loss. The two turnover features look a bit odd and they may pose problems downstream.



3 Linear Regression

Points are primarily scored by the offense and a small amount by the special teams via extra points, field goals, punt and kick returns. In rare cases, points can be scored by the defense. For this to happen, a turnover is required. This fact helps us choose |.3| as the cutoff for selecting features from the correlation plot. This decision is also supported by our scatter plots.

3.1 Model One

The first model we build is a naive model to set a baseline. From our visualization, we know that offensive expected points was the highest correlated feature to our response variable so we choose this instead of other offensives stats. Let O =`OT`,H=`atHome`, P=`Tm_score` , F_O=`1stD_Off`, Y_O = `TotYd_Off`, P_O = `PassY_Off`, R_O =`RushY_Off`, TO_O =`TO_Off`, F_D =`1stD_Def`, T_D =`TotY_Def`, P_D =`PassY_Def`, R_D = `RushY_Def`, TO_D =`TO_Def`, E_O =`OffenseExp`, E_D = `DefenseExp`, E_S`SpTms_Exp`, N = `isPrimeTime`, and T = `isThursday`.

Our first model:

$$M1 : \text{Points} = \alpha_0 + \alpha_1 E_D + \alpha_2 E_O + \alpha_3 E_S + \alpha_4 F_O + \alpha_5 H + \alpha_6 TO_O + \alpha_7 TO_D + \alpha_8 O + \alpha_9 T + \alpha_{10} N + \epsilon$$

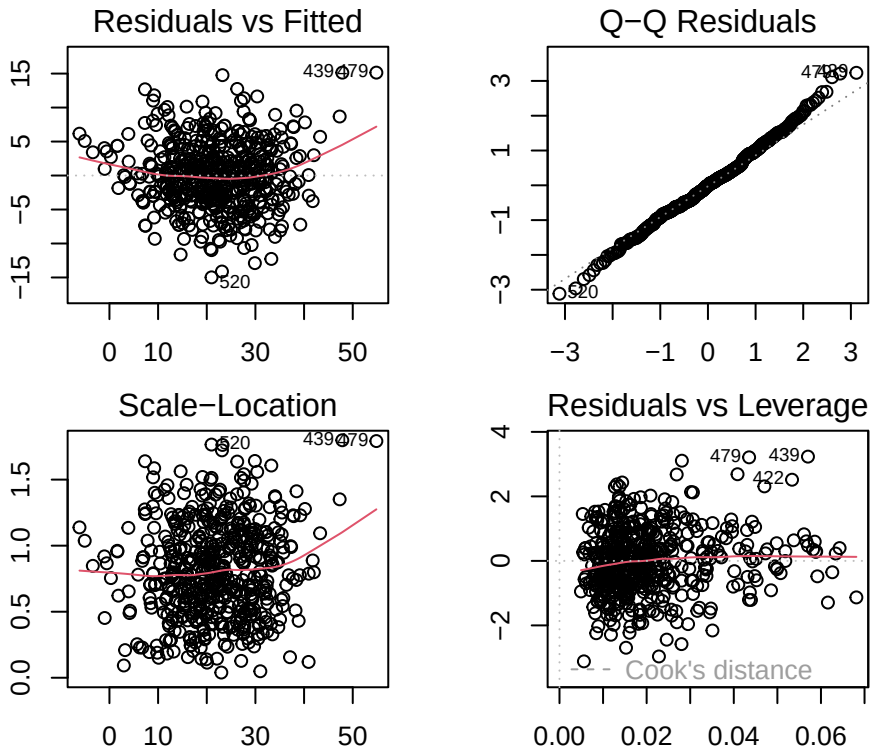


Table 1: Summarized statistics of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.3393576	1.2557828	10.622345	0.0000000
SpTms_Exp	0.4601147	0.0584145	7.876726	0.0000000
OffenseExp	0.7110176	0.0300953	23.625539	0.0000000
DefenseExp	0.1865435	0.0215198	8.668459	0.0000000
X1stD_Off	0.1696519	0.0662606	2.560374	0.0107304
TO_Off	1.3859644	0.2171200	6.383402	0.0000000
TO_Def	1.6217397	0.2063290	7.859970	0.0000000
atHome1	0.1725996	0.4206915	0.410276	0.6817683
OT1	4.1180071	0.9824969	4.191369	0.0000325
isThursday1	4.2122865	0.9246353	4.555619	0.0000065
isPrimeTime1	-1.5391315	0.5751517	-2.676044	0.0076789

From the output of our first model, we see that playing at home had a small effect on the outcome. Furthermore, we know from our correlation plots, that the defensive and special teams had little relationship to our response variable. Our second model is focused on offensive stats and we remove the `atHOME` variable.

3.2 Model 2

Our second model:

$$M2 : \text{Points} = \alpha_0 + \alpha_1 E_O + \alpha_2 F_O + \alpha_3 TO_D + \alpha_4 TO_O + \alpha_5 O + \alpha_6 T + \alpha_7 N + \epsilon$$

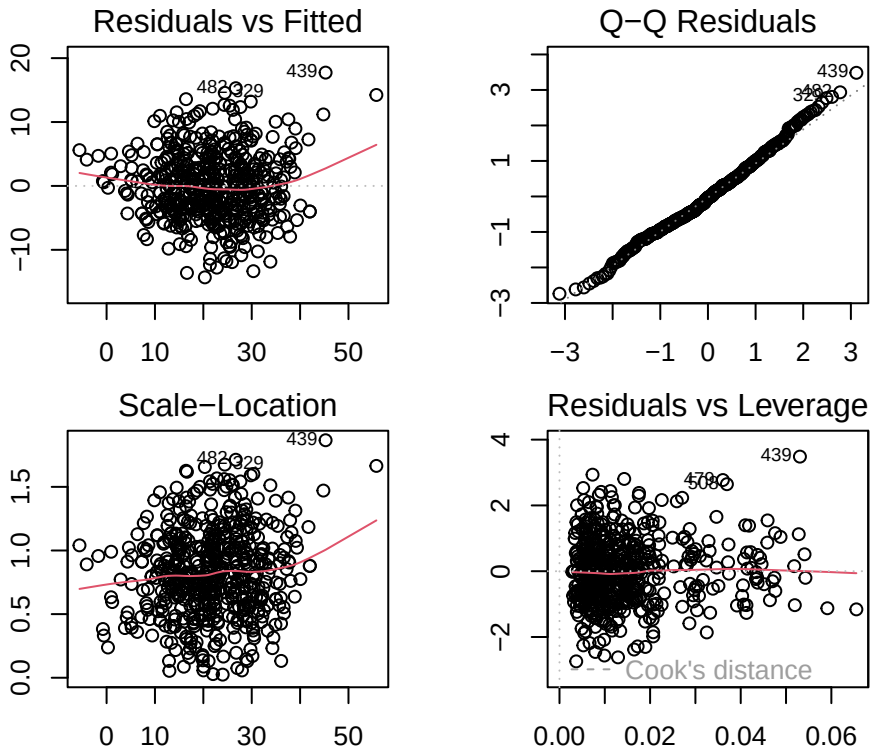


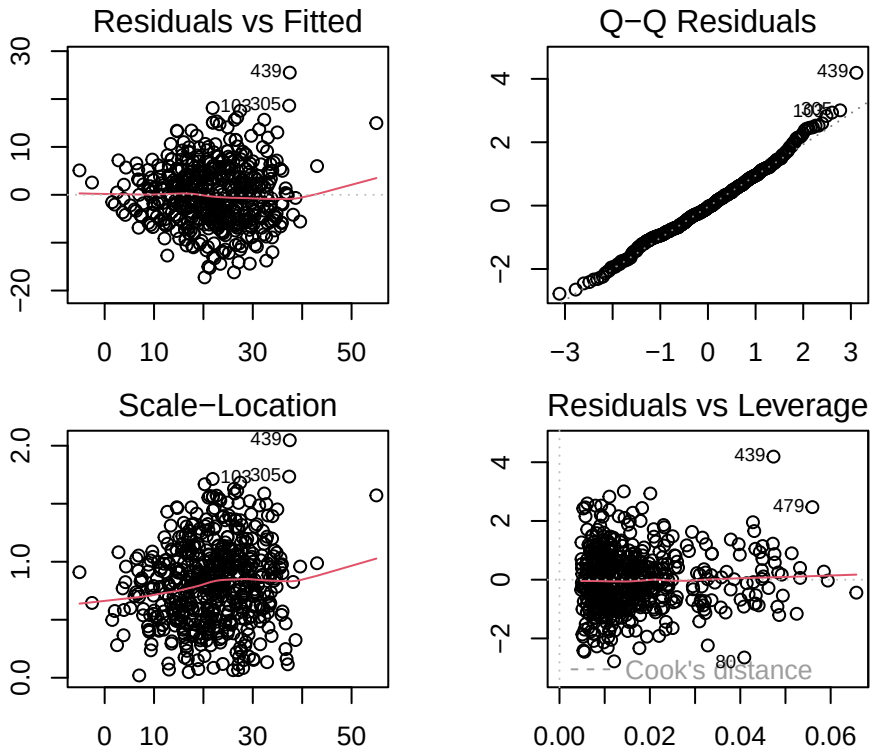
Table 2: Summarized statistics of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.2627618	1.3295888	9.975085	0.0000000
OffenseExp	0.6353635	0.0314334	20.212975	0.0000000
X1stD_Off	0.1325279	0.0713863	1.856490	0.0639322
TO_Off	0.9979134	0.2306389	4.326735	0.0000181
TO_Def	2.5389342	0.1947000	13.040237	0.0000000
isThursday1	4.0789361	1.0026368	4.068209	0.0000545
OT1	4.0403746	1.0652762	3.792795	0.0001659
isPrimeTime1	-1.5557811	0.6239310	-2.493515	0.0129489

3.3 Model Three

The next two models we build, we replace offensive expected points by the three yardage statistics. The third model uses `TotYd_Off` and fourth model will use `RushY_Off` + `PassY_Off`.

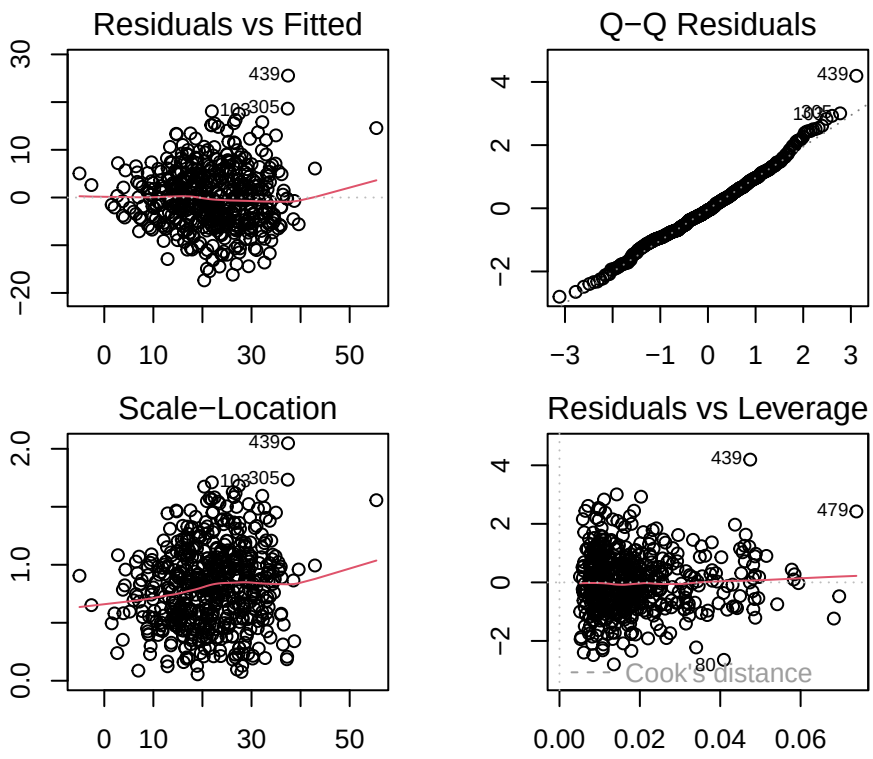
$$M3 : \text{Points} = \alpha_0 + \alpha_1 Y_O + \alpha_2 F_O + \alpha_3 TO_D + \alpha_4 TO_O + \alpha_5 O + \alpha_6 T + \alpha_7 N + \alpha_8 H + \epsilon$$



3.4 Model Four

In our fourth model, we use both rush and passing yards instead of total yards.

$$M4 : \text{Points} = \alpha_0 + \alpha_1 R_O + \alpha_1 P_O + \alpha_3 F_O + \alpha_4 TO_D + \alpha_5 TO_O + \alpha_6 O + \alpha_7 T + \alpha_8 N + \alpha_9 H + \epsilon$$



The respective R^2 values for our four models are:

1. naiveModel = 0.772688,
2. offModel = 0.7309864,
3. yardModel = 0.6181533,
4. typeYardage_model = 0.6183322

Our models based on expected value features perform much better than the ones using yardage. This may be because the expected value features are based on other information and thus are more informative.

4 Logistic regression

For our logistic models, we will omit the three expected points features because they capture the margin of points between two teams. The first model we build includes all features aside from expected points since our exploration showed a difference in distribution for each feature.

4.1 Model One

In our fourth model, we use both rush and passing yards instead of total yards.

$$M1 : \text{win} = \alpha_0 + \alpha_1 P + \alpha_3 F_O + \alpha_4 F_D + \alpha_5 TO_D + \alpha_6 TO_O + \alpha_7 Y_O + \alpha_8 Y_D + \alpha_9 O + \alpha_{10} T + \alpha_{11} N + \alpha_{12} H + \epsilon$$

Table 3: Summary of the significant tests of the logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.0183789	1.1132013	2.7114402	0.0066992
Tm_score	0.2588536	0.0337682	7.6656022	0.0000000
X1stD_Off	-0.1537508	0.0556692	-2.7618654	0.0057472
TotYd_Off	0.0094882	0.0033489	2.8332499	0.0046077
TO_Off	-0.9399173	0.1513878	-6.2086741	0.0000000
X1stD_Def	-0.0715834	0.0518345	-1.3809988	0.1672793
TotY_Def	-0.0207389	0.0033062	-6.2727850	0.0000000
TO_Def	0.7652793	0.1590230	4.8123814	0.0000015
atHome1	-0.2030294	0.2935336	-0.6916734	0.4891425
isThursday1	-0.3598061	0.6806176	-0.5286464	0.5970507
isPrimeTime1	0.0162286	0.4045573	0.0401145	0.9680019
OT1	-0.7039025	0.5495380	-1.2808988	0.2002292

From the above output, we can see that our binary variables and first downs given up by the defense have little impact and are not significant. So in our second model, we remove them.

4.2 Model Two

$$M1 : \text{win} = \alpha_0 + \alpha_1 P + \alpha_2 F_O + \alpha_3 TO_D + \alpha_4 TO_O + \alpha_5 Y_O + \alpha_6 Y_D + \epsilon$$

Table 4: Summary of the significant tests of the logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.3184603	0.9886073	2.345178	0.0190180
Tm_score	0.2397413	0.0311415	7.698446	0.0000000
X1stD_Off	-0.1501170	0.0552602	-2.716547	0.0065967
TotYd_Off	0.0101172	0.0032726	3.091486	0.0019916
TO_Off	-0.8967764	0.1461624	-6.135480	0.0000000
TotY_Def	-0.0233840	0.0026075	-8.968079	0.0000000
TO_Def	0.8302725	0.1545586	5.371894	0.0000001

5 Predictive Modeling

5.1 Linear Regression

We pick our first and third linear regression models to perform predictive modeling. First we split our data 80 : 20 between training and test sets. From there, we perform 5-fold cross validation on our training set and then use our test set to get final results for our chosen model.

Table 5: Model 1 - Naive

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
TRUE	5.026822	0.7612303	3.913466	0.2671396	0.0541707	0.1868635

Table 6: Model 2 - Total Yards

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
TRUE	6.386885	0.5989067	5.002793	0.5652721	0.0733344	0.4164388

We see that model 1 performs better across all metrics and therefore chose it to be our final model. The final results are:

```
     RMSE        R2      MAE
1 4.44901 0.7876264 3.652442
```

5.2 Logistic Regression Cross Validation

We use our training data to perform 5-fold cross validation and then use our test set to get final results for our chosen model.

Table 7: Model 1 - All Features

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.8508098	0.7015463	0.0448226	0.0897408

Table 8: Model 2 - Some Features

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.8738767	0.7476918	0.0196936	0.0393681

Cohen's kappa and accuracy are high in both models. We see little drop off in the second, simpler model and thus choose it as our final model. We now use our entire training data to fit the model and test it on our test data. A cufusion matrix is provided for the final model, along with various statistics to measure model performance, with accuracy at 90%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 48  8
         1  4 48

               Accuracy : 0.8889
                 95% CI : (0.814, 0.9413)
    No Information Rate : 0.5185
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7781

 Mcnemar's Test P-Value : 0.3865

            Sensitivity : 0.9231
            Specificity : 0.8571
```

```
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.9231
             Prevalence : 0.4815
         Detection Rate : 0.4444
   Detection Prevalence : 0.5185
      Balanced Accuracy : 0.8901

       'Positive' Class : 0
```

6 Conclusion

Our models worked fairly well on a limited data set but could use improvement. More data never hurts. Perhaps our models, particularly our linear regression model, would perform better if we had data across multiple years or if we had more features like time of possession, loss of downs, 3rd and 4th down conversions, etc.

One thing not mentioned in the above sections is the Box-Cox transformation. This was attempted for the linear regression but had little effect on performance and occasionally hurt the performance. Therefore, it was left out.

Furthermore, the expected points features were the best performers for linear regression which is less than ideal since they are opaque variables. It is not clear how they are calculated from other statistics by pro-football-reference and it would be better to have the actual statistics instead.