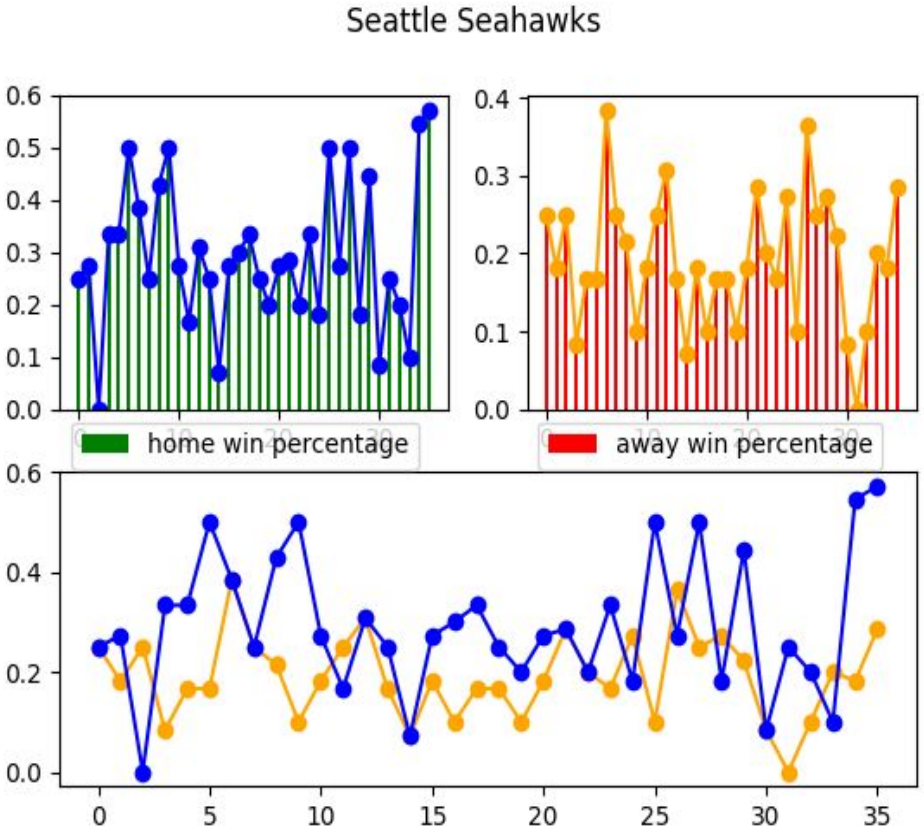
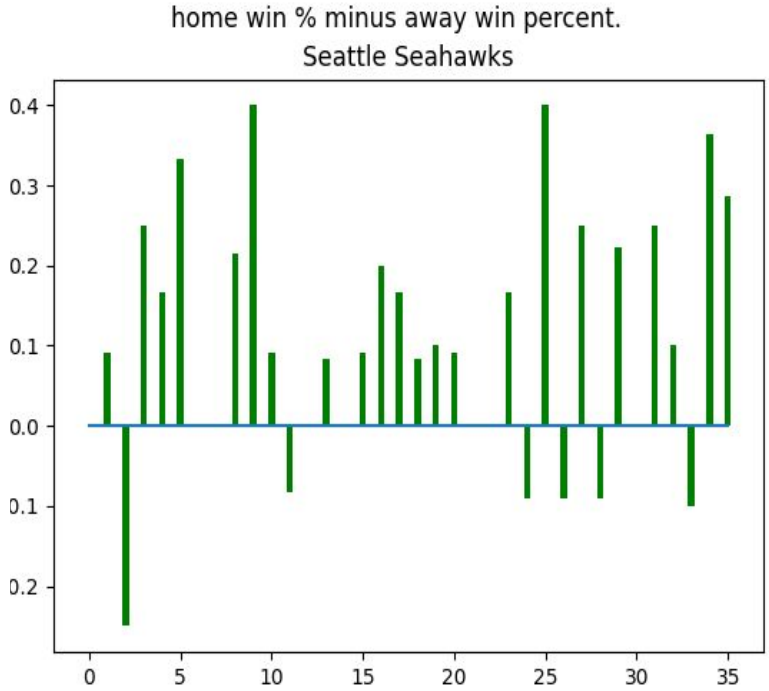
Football And The Home Advantage

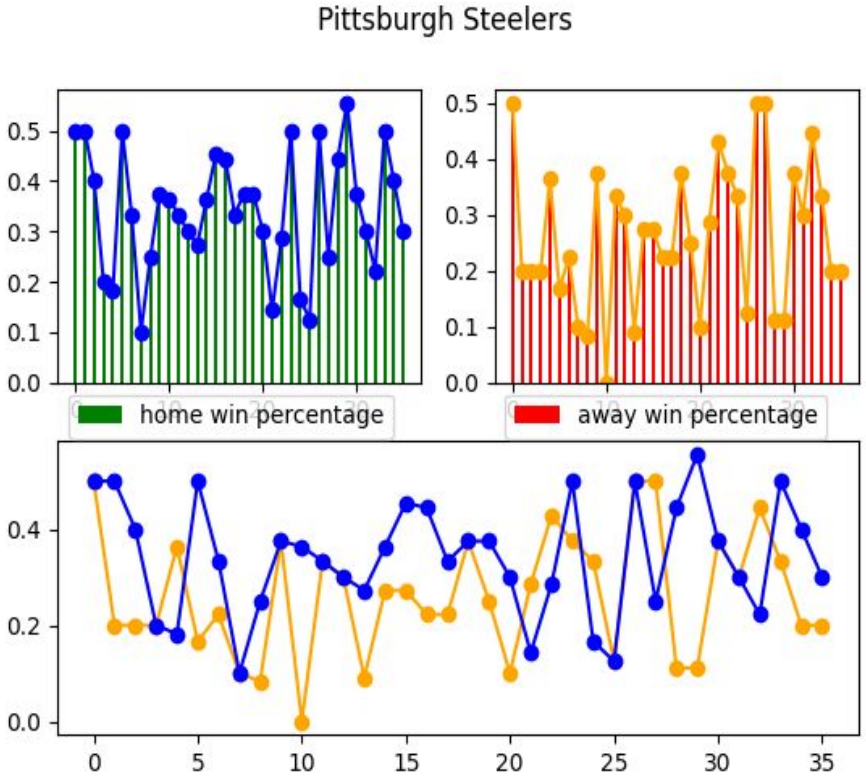
Noah Frahm, Amin Zamani, Priscilla Maryanski, Antonio Pano Flores

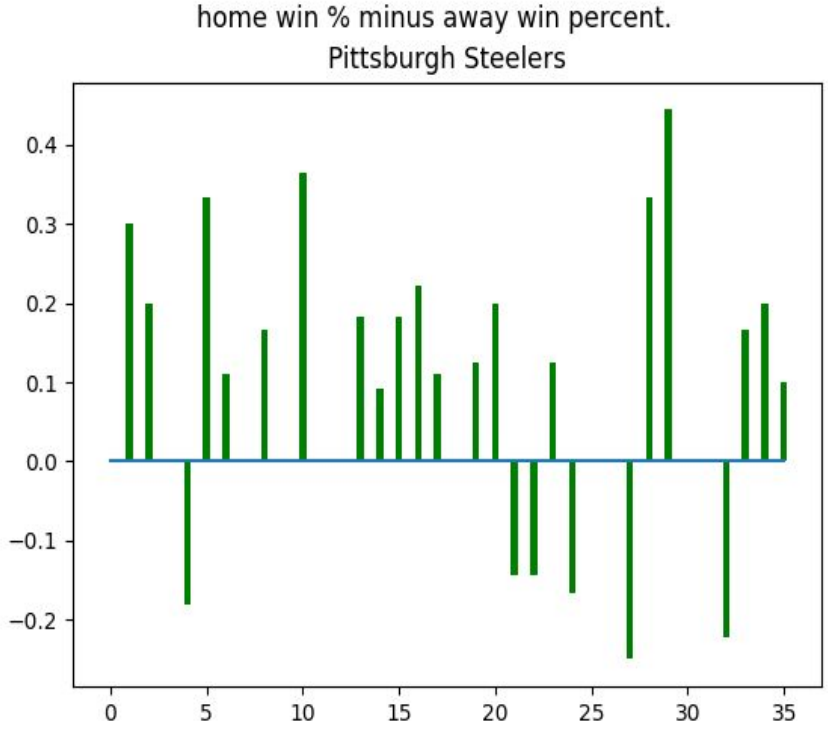
Hypothesis

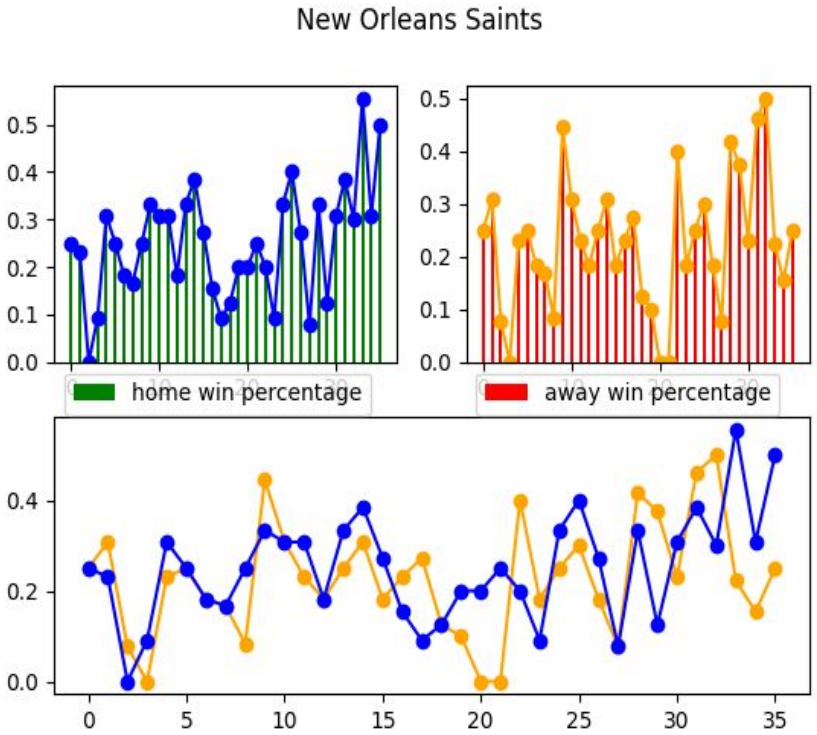
Playing on Home turf increases the football teams chance of winning.

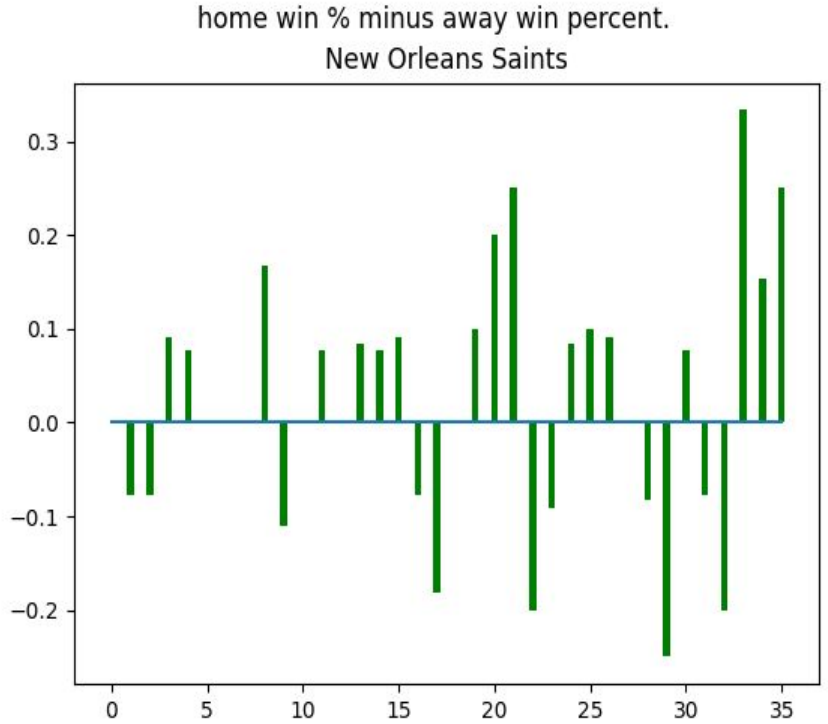

Seattle Seahawks
home win percentage
away win percentage


home win % minus away win percent.
Seattle Seahawks


Pittsburgh Steelers
home win percentage
away win percentage


home win % minus away win percent.
Pittsburgh Steelers


New Orleans Saints
home win percentage
away win percentage


home win % minus away win percent.
New Orleans Saints

What do the graphs mean

These graphs show that there is clearly a home advantage, for each team more than half of the data over the 35 year period shows that the team had a greater win percentage in home games than away games.

We thought this was interesting and decided to try and make several machine learning models to see if we could accurately predict the outcome of a game using football stats and game location data.

It is to be noted that the data shows a big range of values. It is odd to see a low home win percentage one season and then one that almost doubles it the next. This can maybe be explained by looking into other variables beyond just game location such as player age, fan attendance, and the teamself.

Hypothesis Test

SeahawksMean < - 40.23077 : The mean of all the Seahawks games' total score.

Seahawks <- 14.34198 : The standard deviation of all the Seahawks games' total score.

sqrt(nrow(Seahawks)) <- 3.605551 : The square root of the number of Seahawk games.

2014 Super Bowl Total Score = 43 + 8 : Total score of the 2014 Super Bowl.

T <- (51 - 40.23077) / (14.34198/ 3.605551)

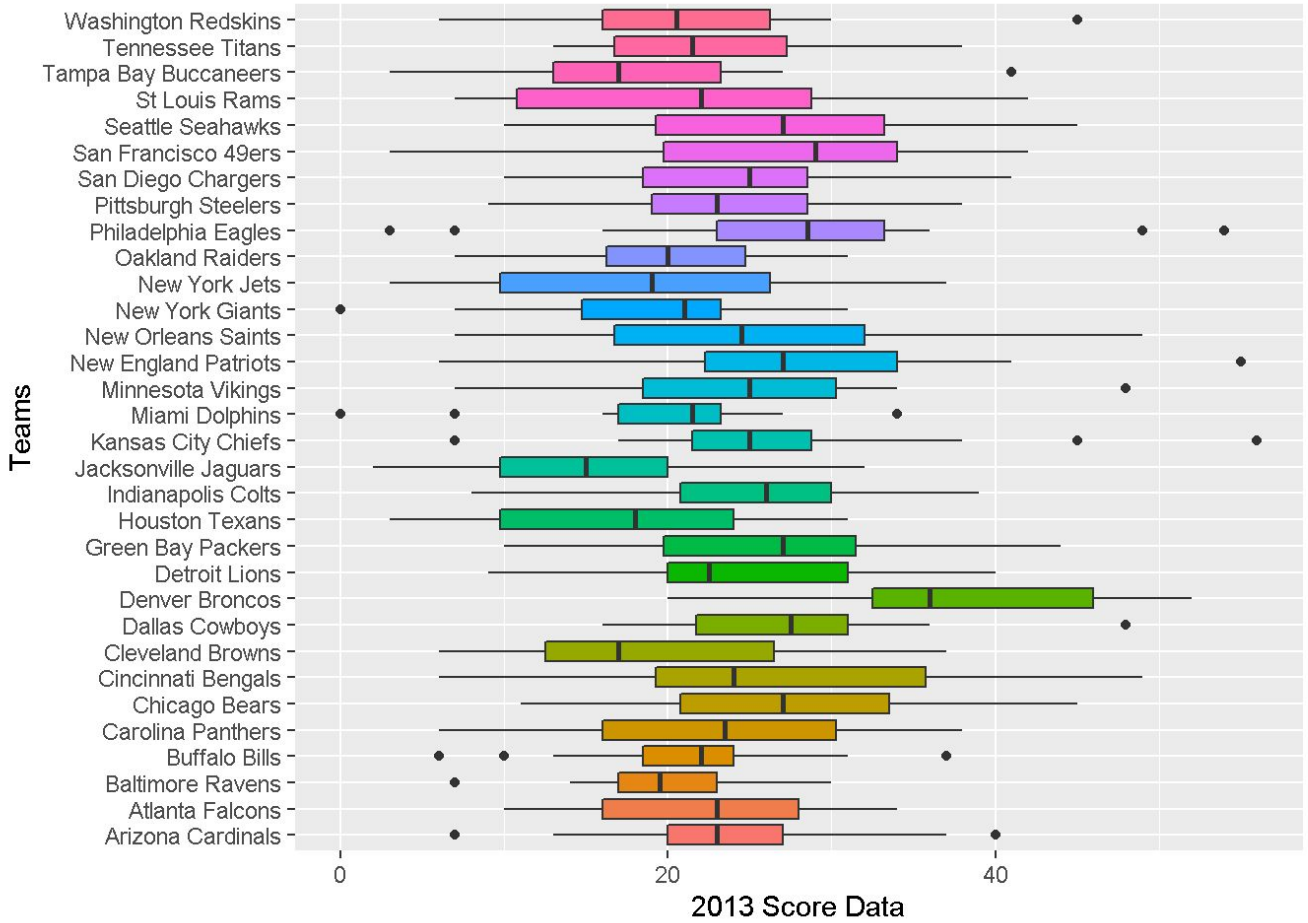
Hypothesis Test continued...

After acquiring T: 2.707368

P Value <- pt(2.707368, nrow(Seahawks) - 1)

P Value : 0.990476

0.99 > 0.05 and so we fail to reject the null hypothesis. Thus, concluding that the 2014 Super Bowl for the Seahawks was “just another game”.



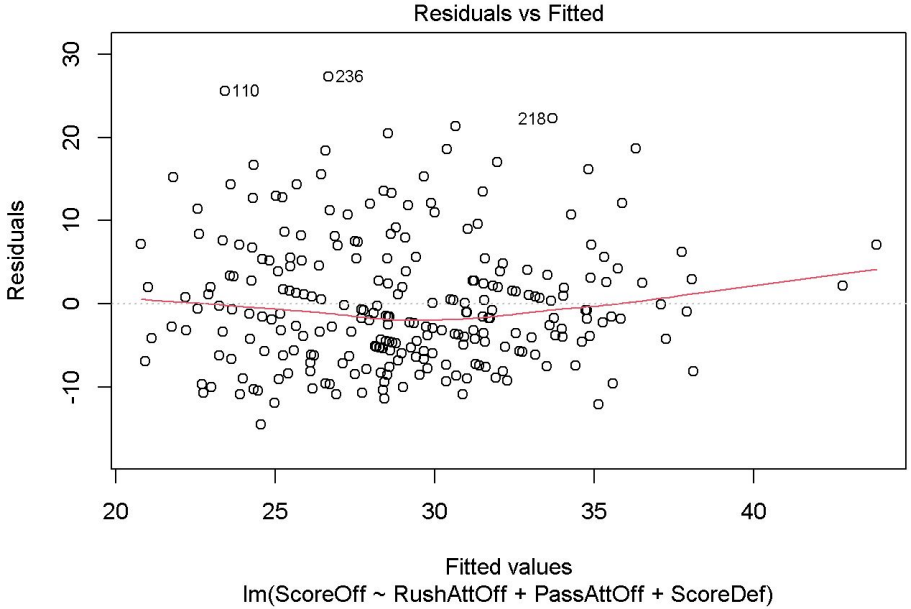
Outliers don't need to be removed.

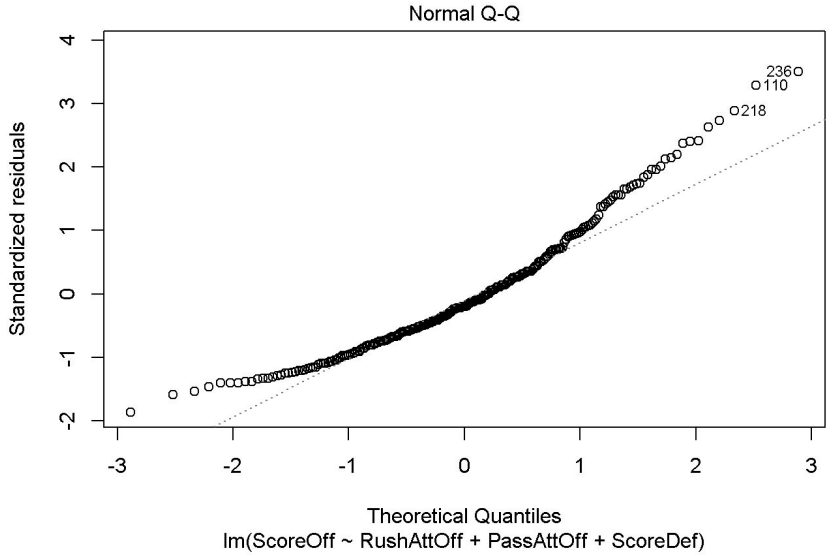
Broncos played above the rest during the regular season.

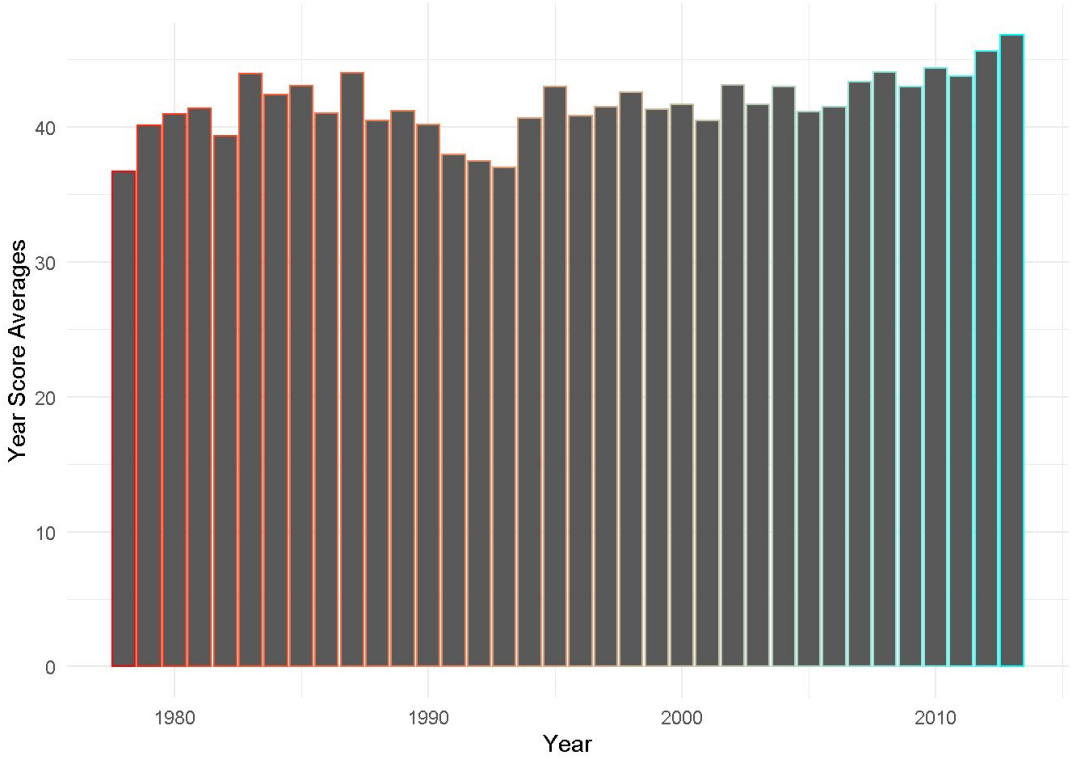
Visually a mean score of around 20 for all teams.

```
##
## Call:
## lm(formula = ScoreOff ~ RushAttOff + PassAttOff + ScoreDef, data = NFL2013.F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.557  -5.611  -1.540   3.978  27.323
##
## Coefficients:
##             Estimate Std. Error t value           Pr(>|t|)
## (Intercept) 12.85764    3.80217   3.382           0.000835 ***
## RushAttOff   0.21975    0.07439   2.954           0.003434 **
## PassAttOff   0.02367    0.06652   0.356           0.722277
## ScoreDef     0.48329    0.06145   7.865 0.000000000000108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.815 on 252 degrees of freedom
## Multiple R-squared:  0.217,  Adjusted R-squared:  0.2076
## F-statistic: 23.27 on 3 and 252 DF,  p-value: 0.0000000000002484
```

lm = 12.85 + .22(x1) + .02(x2) + .48(x3)


Residuals vs Fitted
Residuals
Fitted values
lm(ScoreOff ~ RushAttOff + PassAttOff + ScoreDef)
110
236
218
30
20
10
0
-10
20
25
30
35
40


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(ScoreOff ~ RushAttOff + PassAttOff + ScoreDef)
236
110
218
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3



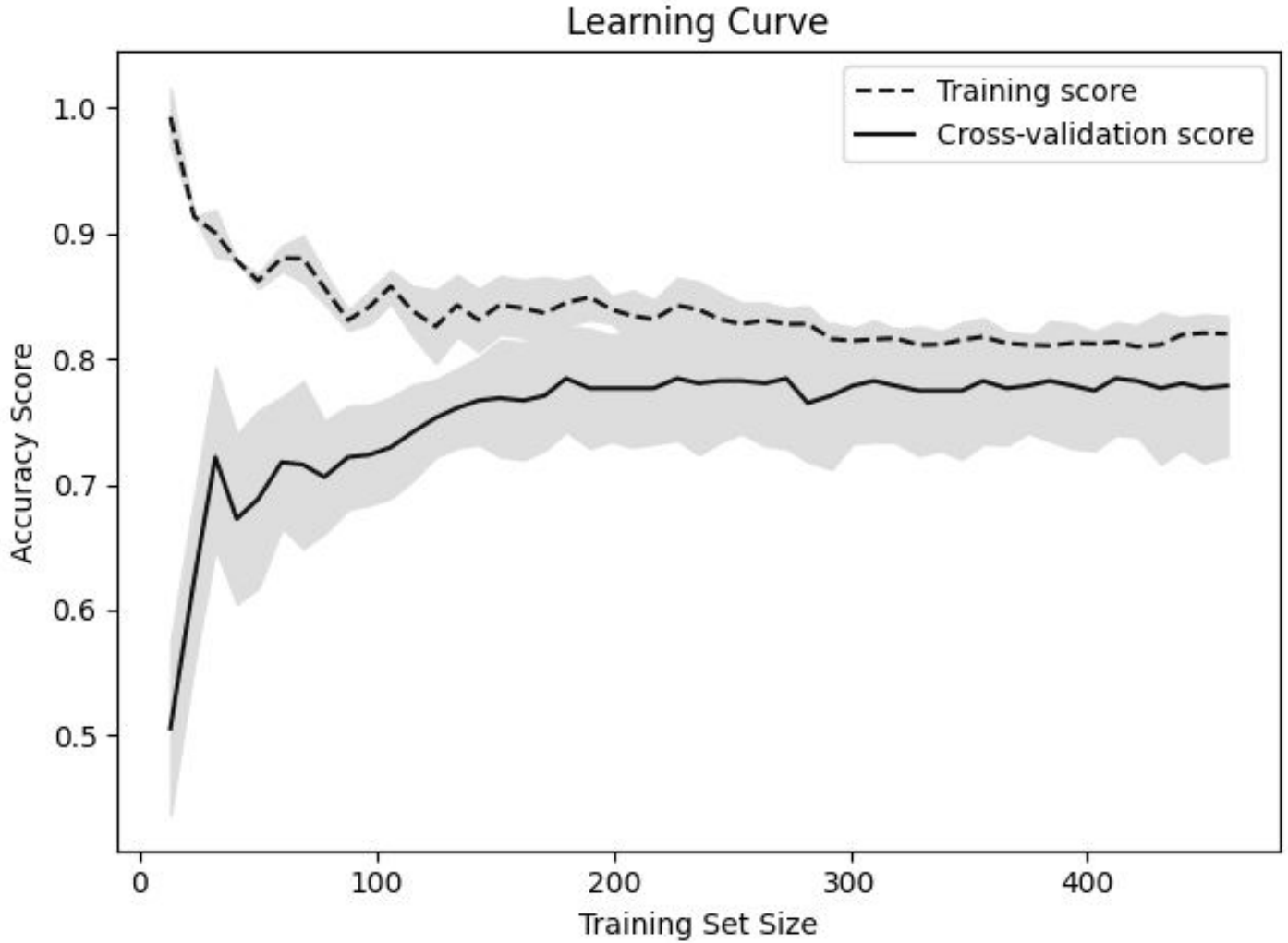
Years span from 1978 - 2013.

Amin's model: the beginning

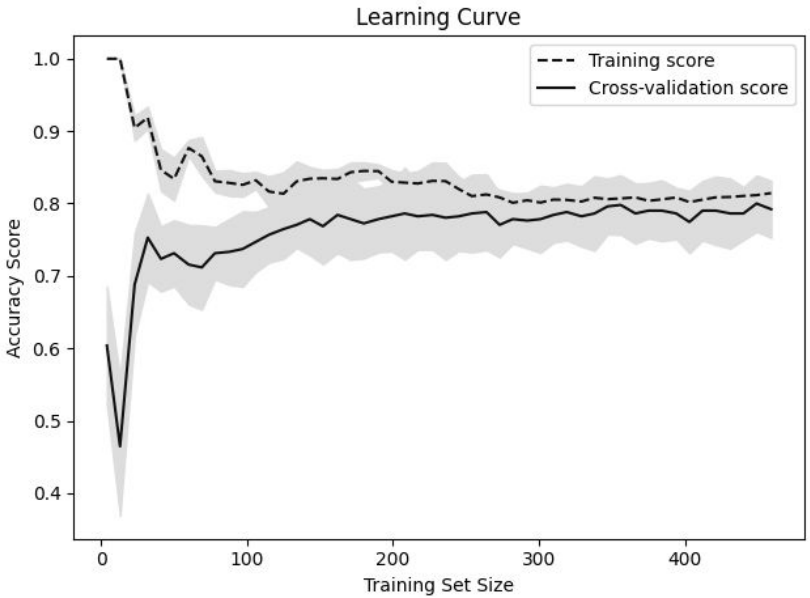
- Using the features:
  - RushAtOff
  - PassAtOff
  - ScoreDef
  - Home or Away
- To predict: Outcome (Win/Loss)
- This was because Noah's graphs showed there was a correlation between how often teams won depending on whether or not they were they home or visiting team.

Amin's model

- First got rid of the rows of data with nan in them for their win/loss outcome.
- Transformations
  - Use LabelEncoder to convert "Win/Loss" and "H/V" to numerical values
  - Linear dimensionality reduction: Incremental Principal Component Analysis
  - Standard Scaler: To standardize some of the features so that no one feature is having major influence on the output
- Estimator: C-Support Vector Classification


Learning Curve
Training score
Cross-validation score
Accuracy Score
1.0
0.9
0.8
0.7
0.6
0.5
0
100
200
300
400
Training Set Size

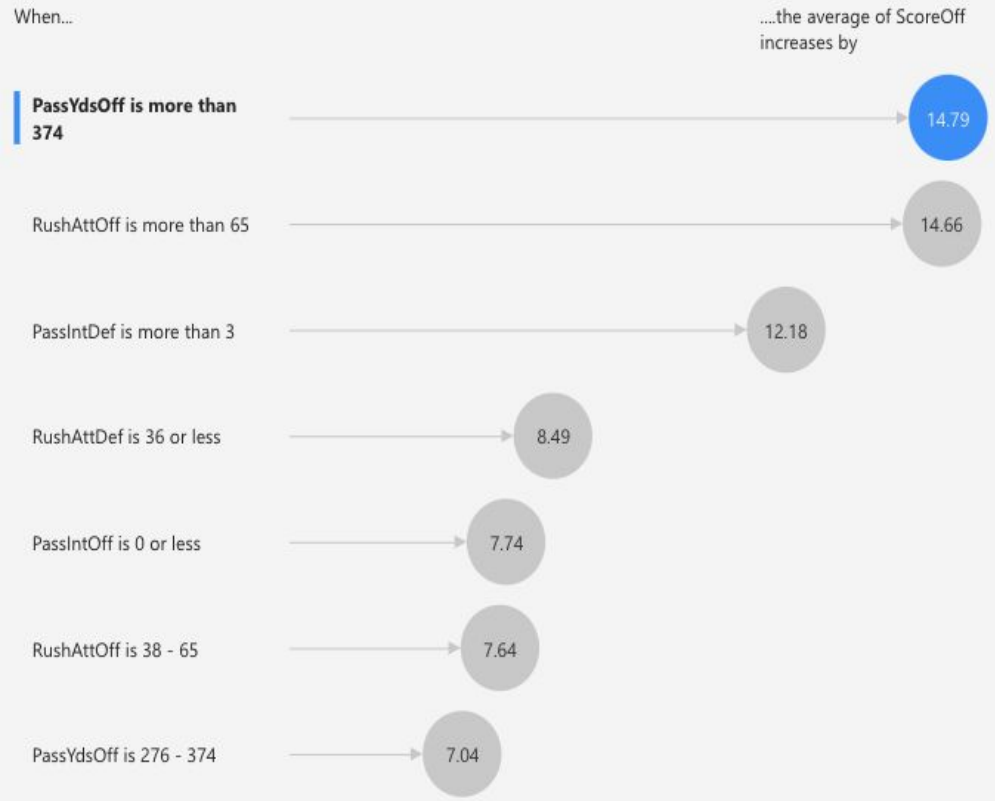
K-nearest neighbors

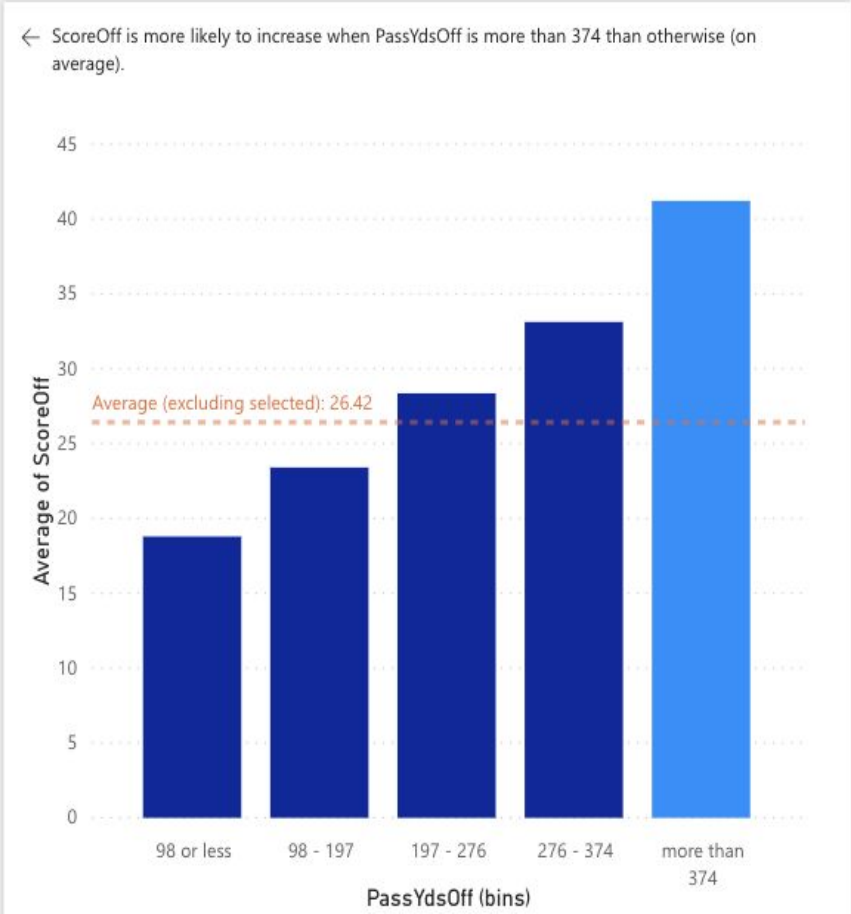


Priscilla's Key Influencers Model

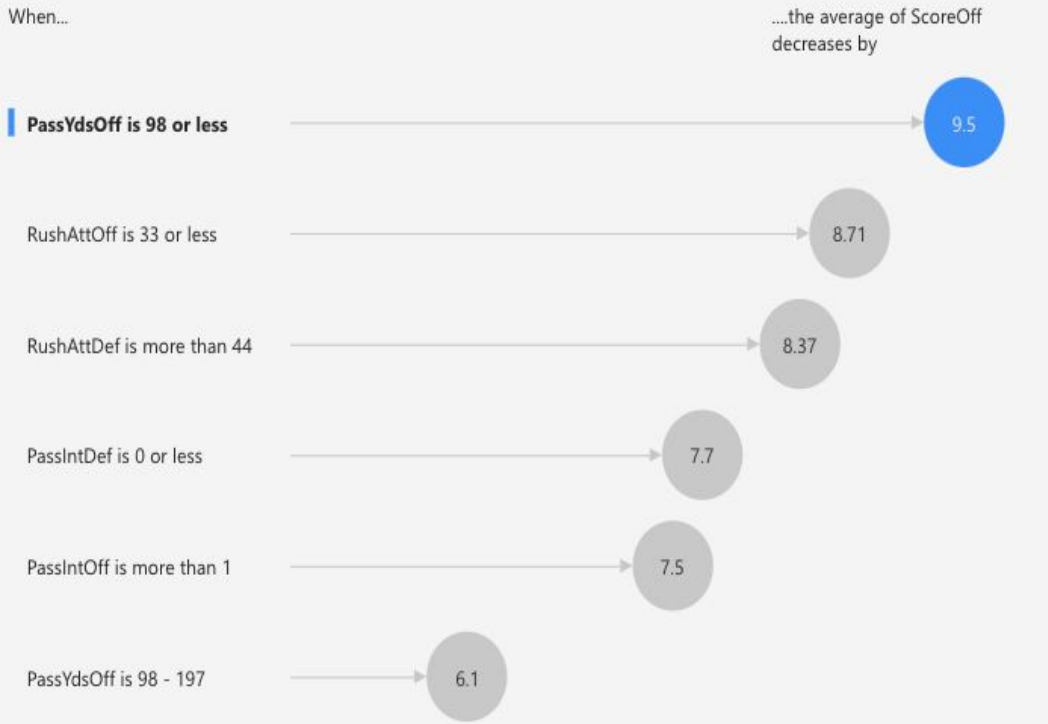
College Football Edition

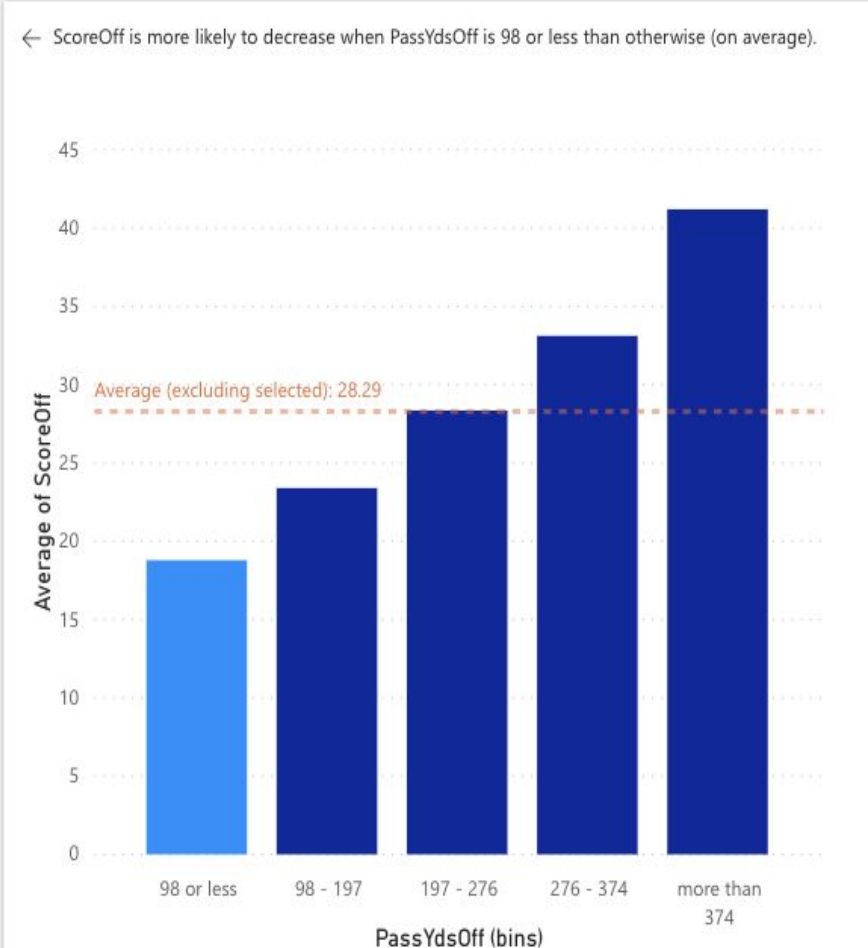
What influences the **home** team's score to **increase**?





What influences the **home** team's score to **decrease**?





The **home** team's score is **considered high** in one instance of 36.31 points.

Under what conditions will the **home** team's score have the **best chances** of being **high**?

Segment 1

PassIntDef is greater than 0

PassIntOff is less than or equal to 1

PassYdsOff is less than or equal to 98 or is greater than 197

RushAttDef is less than or equal to 36

In segment 1, the average ScoreOff is 36.31. This is 8.95 units higher than the overall average, 27.36.

Segment 1 contains 232 data points (16.3% of the data).

The **home** team's score is considered **low** at three instances of 11.61, 18.82, and 18.85 points (average low score of 16.43).

Under what conditions will the **home** team's score have the **best chances** of being **low**?

Segment 1

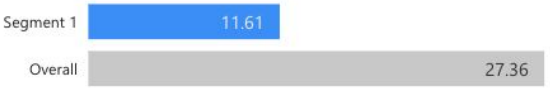
PassIntOff is greater than 0

PassYdsOff is greater than 98 and is less than or equal to 197

RushAttDef is greater than 36

RushAttOff is less than or equal to 33

In segment 1, the average ScoreOff is 11.61. This is 15.75 units lower than the overall average, 27.36.



Segment 1 contains 90 data points (6.3% of the data).

Segment 2

PassIntOff is greater than 0

PassYdsOff is less than or equal to 276 or is greater than 374

PassYdsOff is less than or equal to 98 or is greater than 197

RushAttDef is greater than 36

RushAttOff is less than or equal to 33

In segment 2, the average ScoreOff is 18.82. This is 8.54 units lower than the overall average, 27.36.

Segment 2 contains 130 data points (9.1% of the data).

Segment 3

PassIntOff is greater than 0

PassYdsOff is less than or equal to 276 or is greater than 374

RushAttDef is greater than 44

RushAttOff is greater than 33

In segment 3, the average ScoreOff is 18.85. This is 8.51 units lower than the overall average, 27.36.

Segment 3 contains 95 data points (6.7% of the data).