Time Series Modeling

NFL Betting Model Creation and Analysis

Dev Misra

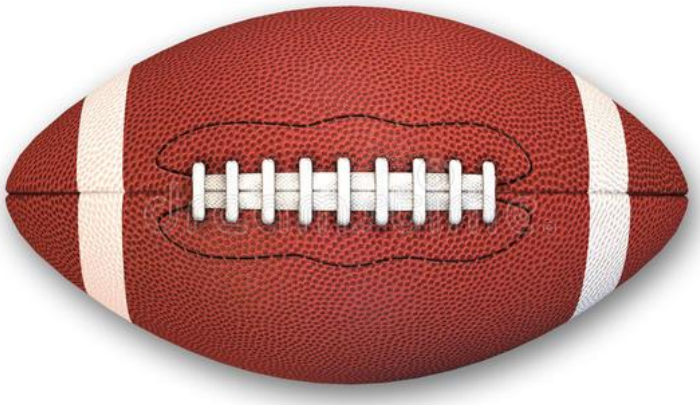


Table of Contents

Background

Sports betting is gaining popularity as it slowly becomes legal in more and more states within the US. One of the most popular sports to bet on is NFL football, which offers a multitude of different betting options, one of which is the Over/Under. Lines are posted by companies that own betting platforms and users can choose to bet on the Over or Under for Total Points scored in the game by both teams. This project analyzes historical NFL game data to create a model to predict the Total Points scored in Seattle Seahawks games next season, along with determining which variables are most instrumental towards this total. Due to the fact that more than half of all NFL teams play in open air stadiums, and that there are restrictions on when games can be played due to weather, I pinpointed a variety of weather conditions that impact the performance of teams and thus impact the total points scored. In addition, I incorporated offensive and defensive rankings for both teams in a game as variables to better predict team performance. My overall goal is to identify and leverage the most crucial factors affecting total points scored in NFL games to create a working model that predicts the total score of future games. Assuming accurate forecasts, an arbitrage opportunity is created when placing bets following the trend data.

NFL Betting (Over/Under)

Any given line created for the Over/Under represents the sum of both teams' final score and typically ranges from the 10s to the 60s. The line can also be set at half a point (for example, a line can be set at 41.5). If one were to take the Over on a line set at 41.5, the combined score would need to be 42 or above for the bet to pay out. Similarly, the Under would hit if the combined score was 41 or lower. Higher lines create greater risk for taking the over, and lower lines create more risk for taking the under, but doing so also offers greater odds and therefore better payouts. If an even semi-accurate model is created, so is a huge opportunity to profit, especially because many online betting platforms are fairly new and unregulated.

Model Creation

My analysis consisted of primarily using two models – a Holt-Winters model and an ARIMA model.

The Holt-Winters model was the first model that I attempted to use to predict the total score of future NFL games, and was chosen as a simple first attempt at using the time series data to create a forecasting model. It also helped me determine if there was any promise in attempting to create a model that would work properly with my data.

The ARIMA model was chosen mainly for its ability to consider extraneous variables that affect the response variable with *xreg* function. This is useful for incorporating the effects of environmental conditions and other variables into the final model. Because games are played in a variety of environments with differing conditions, seasonality doesn't really exist, and thus a SARIMA model would not be appropriate. This was also determined via the *auto.arima* function.

Data

Data Description

For this project, I used reliable, accurate datasets that included betting lines, game data, and weather data from 2006 to the current season. I split the main dataset into various different sections most ideal towards building a working model. Part of the dataset was used as training data, and the remaining was portioned as test data to back test the accuracy of my models. The specific steps taken in the data retrieval and cleansing process are outlined below.

Data Retrieval and Cleansing

In order to create a model to use for sports betting, I decided to use a combination of data sources available online that cover historical data for game results, betting data, and environmental factors for the National Football League (NFL). Some important variables that I needed the dataset to contain were: Date of Game, Home Team, Away Team, Home Score, Away Score, Total Score Open (the betting line), and offensive and defensive statistics for each team. I also needed to locate data for important environmental factors that could affect the outcome of the game, such as: Temperature, Humidity, Wind Speed, and Precipitation. I located data sources with the variables I was looking for online from three main sources -

1. http://www.aussportsbetting.com/data/historical-nfl-results-and-odds-data/
2. http://www.nflweather.com/en/archive
3. https://www.pro-football-reference.com/years/

These sources provide historical datasets that date back to the 2006 NFL season and contain the aforementioned variables.

The data set retrieved from the first website linked above combines data from all the NFL seasons from 2006 until current day. In the file are a multitude of variables: Date, Home Team, Away Team, Home Score, Away Score, Overtime?, Playoff Game?, Neutral Venue?, Home odds (broken into 12 separate variables), Away odds (broken into 12 separate variables), Home line, Over/Under Market, and Notes. Because of changes in the retrieval of bookkeeping data, many of the earlier seasons within the dataset do not contain as many variables representing odds data as the newer seasons do. To prime the dataset for analysis and ensure its consistency, these steps were performed:

I. Removal of irrelevant variables to keep only: Date, Home Team, Away Team, Home Score, Away Score, Overtime?, Playoff Game?, Neutral Venue?, and Total Score Open.

II. Since I am focusing on the Over/Under line, which is a combined team total, I combined Home Score and Away Score into Total Score and subtracted Total Score Open (betting line) to create my Differential value. If the differential value is -10, it means that the

Under hit by 10 points (i.e. Total Score Open was 32 points but the teams only scored a combined 22 points). Similarly, if the differential was 7, it means the Over hit by 7 points (i.e.Total Score Open was 15 points and the teams scored a combined 22 points).

The data set retrieved from the second website contains data for various weather conditions that existed for past NFL games. Through my research, I concluded that the most important weather variables that affect the outcome of football games are Wind Speed, Temperature, Humidity, and Precipitation. I used this data for a third step in my data retrieval process:

III. Incorporation of Wind Speed, Temperature, Humidity, and Precipitation columns.

After this process is complete, the initial dataset is now transformed into a collection of variables that are most important towards my data analysis process. This data logs all games starting from the 2006 season and ending in the 2020 season. I'm able to use this specific subset of data to further analyze how each individual variable specifically affects my chosen response variable (total score of the game).

Finally, in order to improve the accuracy of my models, I decided to include data representing the offensive and defensive capabilities of each team for each game. The rankings range from 1-32 and are reassessed for each year. If the 2nd ranked offense were to play against the 30th ranked defense, common sense dictates that the higher ranked offense will score a lot of points. On the other hand, the 31st ranked offense probably won't score many points against the 1st ranked defense. I located this data on the third website listed to complete the fourth step in my data retrieval and cleansing process:

IV. Incorporation of Seahawks Offensive/Defensive rating and Opponent Offensive/Defensive rating columns.

For the creation of my working model, I decided to randomly select a team using a random team selector online. This team was the Seattle Seahawks, and so a fifth and final step in the data cleansing process was completed:

V. Splitting and transferring of individual game and betting data of the Seattle Seahawks, into a new, separate file. This provides a working data set to leverage for creating my initial models.

Because my working model is based on data from a single team, the Seattle Seahawks, I split the data for the team into three main groups. I realize that players on the team – and therefore the performance of the team – varies drastically across years (simply due to the high-burn and constantly evolving nature of the sport); however, I incorporated offensive and defensive ranking data to minimize the possible effects of this. The three groups were as follows:

A. Data from beginning of 2006 to the end of 2012, which was used for the creation of my initial models.

B. Data from the end of 2012 to the beginning of 2017, which was used for checking the accuracy of the initial models.

C. Data from the beginning of 2017 to the current week in 2020, which was used for utilizing the models I created in order to predict scores for the remaining Seattle Seahawks games in the 2020 season.

The first few rows of **Dataset A** are depicted below:

	i..Date	Home.Team	Away.Team	Home.Score	Away.Score	Overtime.	Playoff.Game.	Neutral.Venue.	Total.Score.Open	Total.Score	Differential	Average.Wind	Average.Temperature
1	9/17/2006	Seattle Seahawks	Arizona Cardinals	21	10	0	0	0	47.0	31	-16.0	5.500000	64.00
2	9/24/2006	Seattle Seahawks	New York Giants	42	30	0	0	0	42.5	72	29.5	8.500000	70.75
3	10/1/2006	Chicago Bears	Seattle Seahawks	37	6	0	0	0	36.5	43	6.5	8.000000	65.50
4	10/22/2006	Seattle Seahawks	Minnesota Vikings	13	31	0	0	0	40.5	44	3.5	4.500000	63.00
5	10/29/2006	Kansas City Chiefs	Seattle Seahawks	35	28	0	0	0	37.0	63	26.0	11.250000	75.25
6	11/6/2006	Seattle Seahawks	Oakland Raiders	16	0	0	0	0	34.0	16	-18.0	12.250000	59.00
7	11/12/2006	Seattle Seahawks	St. Louis Rams	24	22	0	0	0	43.5	46	2.5	9.750000	48.50
8	11/19/2006	San Francisco 49ers	Seattle Seahawks	20	14	0	0	0	43.5	34	-9.5	4.500000	59.75
9	11/27/2006	Seattle Seahawks	Green Bay Packers	34	24	0	0	0	42.0	58	16.0	6.750000	30.50
10	12/3/2006	Denver Broncos	Seattle Seahawks	20	23	0	0	0	40.5	43	2.5	4.000000	23.33
11	12/14/2006	Seattle Seahawks	San Francisco 49ers	14	24	0	0	0	37.5	38	0.5	21.000000	53.25
12	12/24/2006	Seattle Seahawks	San Diego Chargers	17	20	0	0	0	45.5	37	-8.5	8.250000	44.50
13	12/31/2006	Tampa Bay Buccaneers	Seattle Seahawks	7	23	0	0	0	37.0	30	-7.0	7.500000	82.25
14	1/6/2007	Seattle Seahawks	Dallas Cowboys	21	20	0	1	0	48.5	41	-7.5	10.250000	43.50
15	1/14/2007	Chicago Bears	Seattle Seahawks	27	24	1	1	0	37.0	51	14.0	13.500000	36.00
16	9/9/2007	Seattle Seahawks	Tampa Bay Buccaneers	20	6	0	0	0	41.0	26	-15.0	9.000000	73.50
17	9/23/2007	Seattle Seahawks	Cincinnati Bengals	24	21	0	0	0	50.0	45	-5.0	2.250000	59.50
18	9/30/2007	San Francisco 49ers	Seattle Seahawks	3	23	0	0	0	41.0	26	-15.0	16.500000	67.50
19	10/7/2007	Pittsburgh Steelers	Seattle Seahawks	21	0	0	0	0	41.5	21	-20.5	0.750000	82.75
20	10/14/2007	Seattle Seahawks	New Orleans Saints	17	28	0	0	0	43.0	45	2.0	4.000000	62.25

Average.Humidity	Sky	Precipitation	Seahawks.DF.Ranking	Seahawks.OF.Ranking	Opponent.DF.Ranking	Opponent.OF.Ranking
52.50000	overcast/mostly cloudy	None	19	14	29	19
55.50000	clear	None	19	14	24	11
62.75000	clear/cloudy mix	None	19	14	3	2
50.00000	clear	None	19	14	14	26
26.75000	clear	None	19	14	12	15
85.25000	overcast/mostly cloudy	light rain	19	14	18	32
81.50000	overcast/mostly cloudy	light rain	19	14	28	10
84.50000	overcast/mostly cloudy	light rain	19	14	32	24
86.75000	overcast/mostly cloudy	light snow	19	14	25	22
56.66667	clear/cloudy mix	None	19	14	9	17
76.25000	overcast/mostly cloudy	light rain	19	14	32	24
82.50000	overcast/mostly cloudy	light rain	19	14	7	1
57.00000	overcast/mostly cloudy	None	19	14	21	31
71.25000	overcast/mostly cloudy	None	19	14	20	4
72.00000	overcast/mostly cloudy	None	19	14	3	2
39.25000	clear	None	6	10	3	19
58.00000	overcast/mostly cloudy	None	6	10	24	11
48.25000	clear/cloudy mix	None	6	10	20	32
56.25000	clear	None	6	10	2	9
60.00000	overcast/mostly cloudy	None	6	10	25	13

Preliminary Analysis of Data

In this section, I begin by examining online research discussing the effects of various weather conditions on NFL games. This is followed by a heatmap breakdown and analysis of some of these weather conditions, specifically wind and temperature, to determine exactly how instrumental they are towards determining the total number of points scored in a game. Finally, I conduct a decomposition on the data to identify any potential trends or patterns to account for within the model

Environmental Conditions

There is scientific evidence of weather affecting sports games played outdoors. According to an article from weather.com, cold weather has a notable effect specifically in NFL games. The author of the article mentions that passing completions drop by about 2% during games played in extreme cold. This could partially be due to the fact that the weather affects not only the players, but also the football and the field conditions. One of the most notable changes for the football itself created by bad weather is the fact that the ball pressure reduces by about 20% in the extreme cold. Extremely cold weather, as one can imagine, also has a pronounced effect on players. Cold temperatures cause the body to send less blood to its externalities in order to keep a stable core temperature at the location of the vital organs. This can lead to a potential loss of grip strength, affecting both the quarterback and players holding the ball. Cold air also makes it harder for players to intake a sufficient amount of oxygen, as their respiratory systems battle increased irritation. Less oxygen can lead to more injuries – because the muscles don't receive a sufficient amount of oxygen, they aren't able to stretch, and stiffen to conserve heat and energy. This is evidenced further in the article – "NFL games average a 1.79% fumble rate. During games played in extreme cold, that rate increases to 2.42%, a 35% increase." (Rae, 2016)1

Another environmental factor that affects total points scored in a game is wind speed. One source located online discusses how wind speed affects NFL quarterbacks. After a statistical analysis, the article concludes that when the wind speed is recorded at less than 10 mph, NFL quarterbacks complete 60.31% of their passes on average. This figure drops to 54.65% when wind speeds are at least 20 mph. Under these same categories, the average touchdown percentage drops from 4.29% to 3.58%, while interception percentage increases from 2.99% to 3.11%. Wind also impacts the kicking of field goals for extra points – the NFL's average successful field goal percentage is 83.8% when wind speeds are slower than 10 mph. When wind gusts exceed 20 mph, the average successful field goal percentage drops to 76.9%. (Cheema, 2020)2

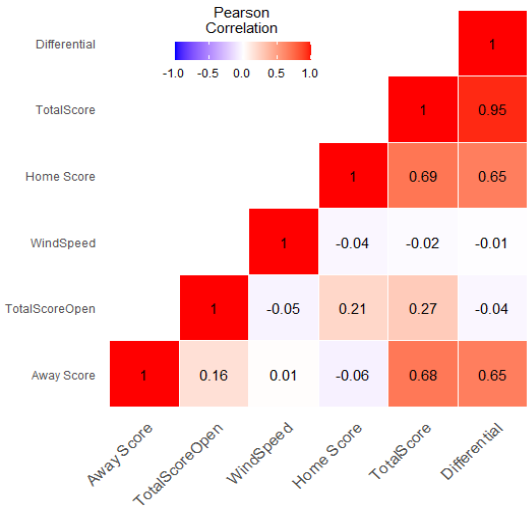
[1] https://weather.com/wunderground/news/news/how-cold-temps-affect-nfl-games

[2] https://www.thespax.com/nfl/analyzing-the-effect-of-weather-in-the-nfl/

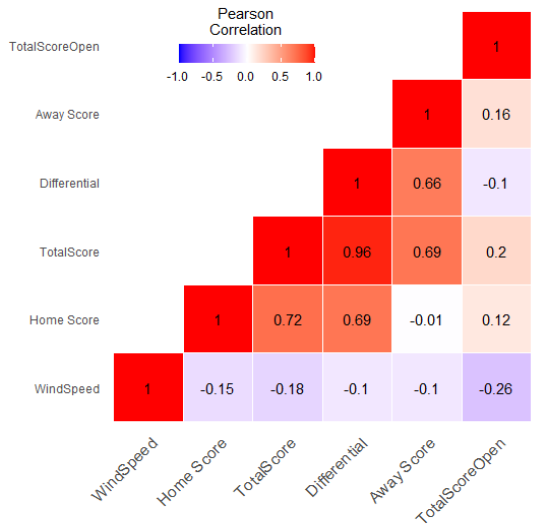
Heatmap Analysis

Wind Heatmaps

Low Wind Heatmap:



High Wind Heatmap:



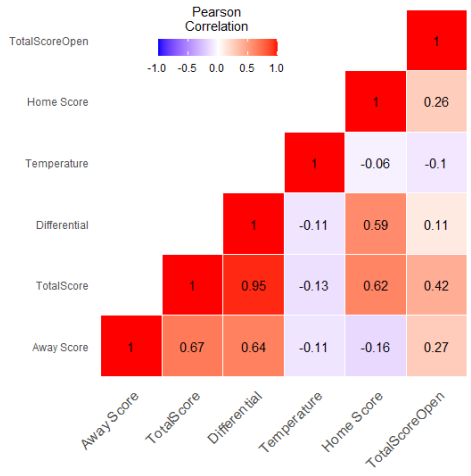
Above are two heatmaps created from the dataset – one for Low Wind and one for High Wind – highlighting the correlation between wind speed and important response variables.

As can be seen in the first heatmap, low wind speeds have very little to no effect on the overall score of the game. Low wind, for reference, is defined as when the average wind speed for the game was 6 miles per hour or under (taken from a sample of 443 games). When following the rows/columns labeled Wind Speed, one can observe a fairly small and insignificant negative correlation with the HomeScore (-.04), TotalScore (-.02), and TotalScoreOpen (-.06) values.

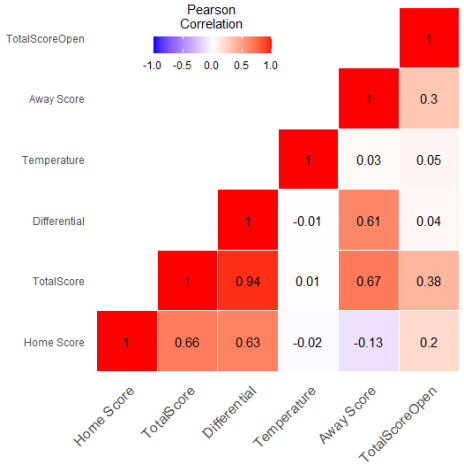
On the contrary, when analyzing the second High Wind heatmap (high wind is defined as when the average wind speed for the duration of the game 13 miles per hour or more and was taken from a sample of 268 games), a more pronounced negative correlation can be observed. In this matrix, the correlation metrics for HomeScore (-0.15), TotalScore (-0.18), and TotalScoreOpen (-0.26) are much larger (in absolute terms) than they are within the Low Wind heatmap. This suggests that high wind speeds lead to a greater decrease in these scores. From the figures above, the total score had a net decrease in correlation of (0.16) and the total score open had a net decrease of (0.20) when the wind speed increased from “low wind” to “high wind”. The response variable seemingly least impacted by wind speed was Differential. Its net decrease in correlation of only (0.09) between low and high wind speeds suggests that the difference in final scores between teams doesn’t change as much due to wind speed. One can potentially profit on this by taking the Over on the base spread (Differential) offered in high wind games.

Temperature Heatmaps

Very Hot Weather Heatmap:



Very Cold Weather Heatmap:



The above heatmaps highlight the correlation between temperature and various response variables. It can be observed that games played on very hot days (>80°F, n = 153) seem to exhibit a somewhat weak negative correlation between Temperature and Total Score (-0.13) and on very cold days (<30°F, n = 89) an almost nonexistent positive correlation (0.01).

Similar observations can be made with Temperature and Total Score Open. On very hot days, temperature and the total score open are weakly negatively correlated (-0.10), but on very cold days, they are very weakly positively correlated (0.05).

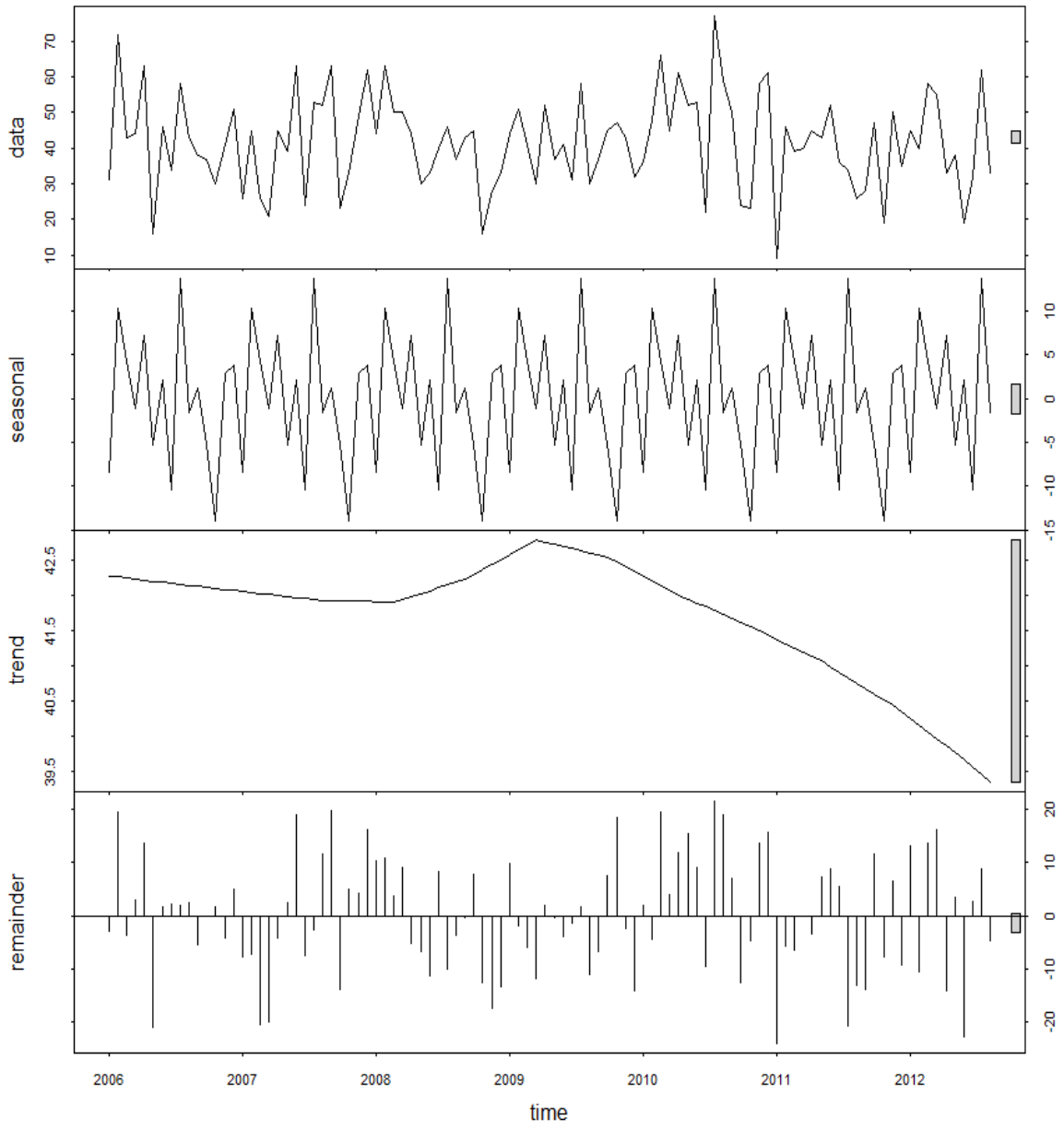
On very hot days, Temperature and Differential are weakly negatively correlated (-0.11), and on very cold days, temperature and differential are very weakly negatively correlated (-0.01).

Although the correlations observed in these heatmaps are not very strong, there is an observable difference between these correlations on very hot days versus very cold ones. One can see that for games played at >80°F, the TotalScore, TotalScoreOpen, and Differential response variables tend to decrease as temperatures increase. For games played at <30°F, the TotalScore and TotalScoreOpen values either rise or do not change as temperatures increase. Essentially, the differential falls when it gets hotter on very hot days, but it may not fall as it gets colder on days that are already very cold.

While the correlations in this matrix does not allow me to forecast or predict the total score on their own, they do provide some insight towards the relationships between variables as the time series models are leveraged to further analyze trends in the data.

Total Score Decomposition

The first diagnostic test that I performed was a decomposition of the total score by using the *stl* function. This function decomposes time series data into seasonal, trend, and irregular components. As can be seen in the decomposition, a seasonal component seems to exist in the time series. This is counterintuitive, as there is no real seasonality in the total scores of NFL games, and can likely be explained by the fact that other important variables such as weather conditions (which will span a wide range) have not yet been factored into the data. There is no clear trend in the time series, and there is no clear trend in the decomposed remainder.



Initial Models

Utilizing the Seattle Seahawks data, I created three main working models: a Multiple Regression model, a Holt-Winters model, and an ARIMA model.

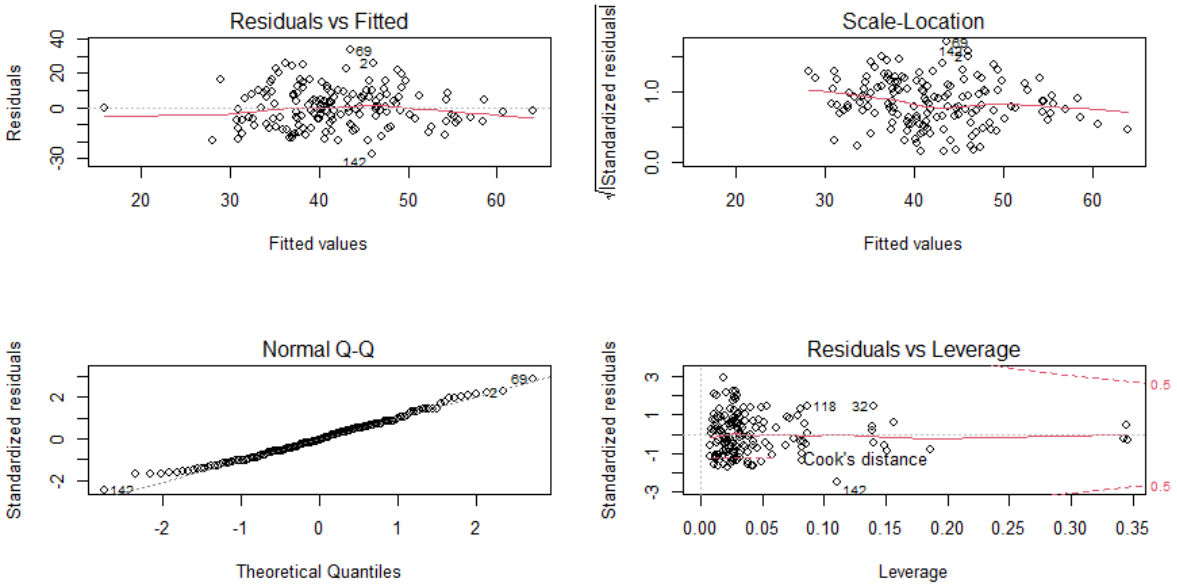
Multiple Regression Model

After performing a multiple regression analysis on all the variables in the dataset and a backwards stepwise function on the regression, I received this output:

```
Step:  AIC=782.67
total.score ~ overtime + avg.temp + precip + opp.def + opp.off

           Df Sum of Sq   RSS    AIC
<none>                  20231 782.67
- overtime  1     560.5 20791 784.99
- avg.temp  1     690.9 20922 785.98
- precip    3    1907.1 22138 790.91
- opp.def   1    1688.8 21920 793.34
- opp.off   1    4499.7 24731 812.40

Call:
lm(formula = total.score ~ overtime + avg.temp + precip + opp.def +
    opp.off)

Coefficients:
     (Intercept)          overtime          avg.temp  preciplight snow        precipNone
         52.2901            8.7095           -0.1427           16.3536            1.1833
      precipsnow          opp.def          opp.off
         -33.6282            0.3913           -0.6109
```

When using all the variables, the AIC of the regression was 793.11, but after removing insignificant variables, the model improved, and the AIC was reduced to 782.67. After performing a new regression on these variables, I produce the following output:

```
>   summary(fit2)

Call:
lm(formula = total.score ~ overtime + avg.temp + precip + opp.def +
    opp.off)

Residuals:
    Min      1Q  Median      3Q     Max
-27.013  -8.209  -0.807   7.368  33.391

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       52.29010    4.60206  11.362  < 2e-16 ***
overtime           8.70949    4.27252   2.038 0.043257 *
avg.temp          -0.14265    0.06303  -2.263 0.025052 *
preciplight snow  16.35362    7.46947   2.189 0.030114 *
precipNone         1.18325    3.08565   0.383 0.701915
precipsnow       -33.62822   12.08114  -2.784 0.006070 **
opp.def            0.39135    0.11059   3.539 0.000536 ***
opp.off           -0.61088    0.10576  -5.776 4.26e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.61 on 150 degrees of freedom
Multiple R-squared:  0.2858,    Adjusted R-squared:  0.2525
F-statistic: 8.575 on 7 and 150 DF,  p-value: 8.006e-09
```

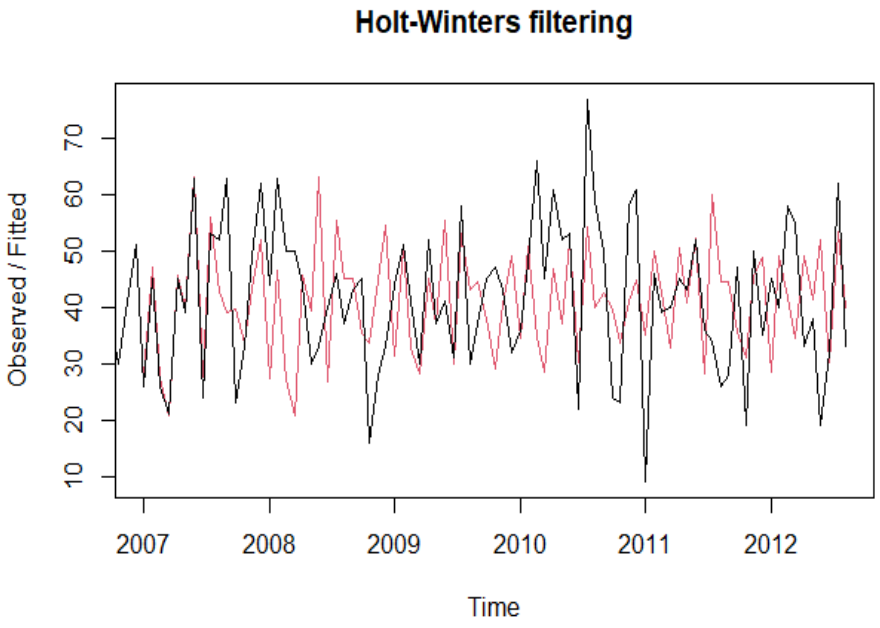


Based on the p-value of 8.006e-9, it can be concluded that the model is statistically significant, with the most significant variables being the opposing team's offense and defense scores. Each increase in the numerical rank of the opposing team's defense (higher = worse) amounts to a .391 increase in the total score, while each increase in the numerical rank of the opposing team's offense amounts to a .611 decrease in the total score. Based on the adjusted R^2, however, it can be observed that this regression model only accounts for about 25% of the variation in the response variable.

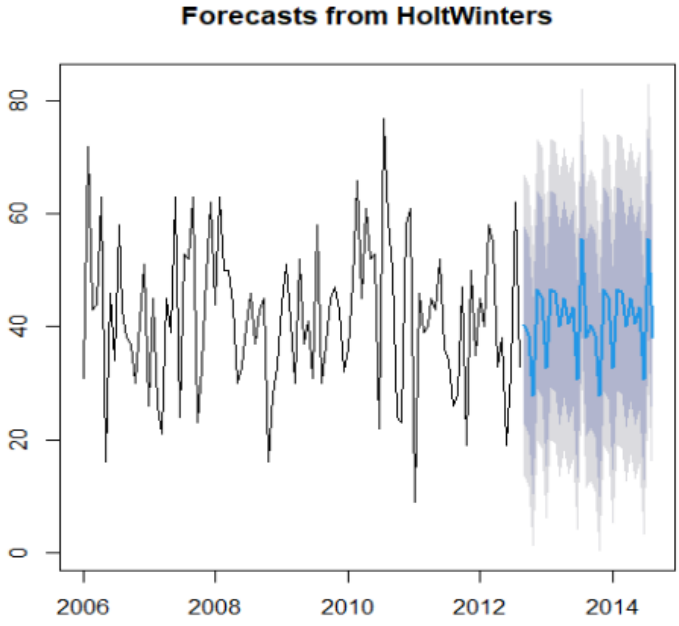
Holt-Winters Model

```
> scorecast
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = ts.total)

Smoothing parameters:
 alpha: 0
 beta : 0
 gamma: 0.2640427
```

Based on the decomposition function, a pattern is highlighted in the seasonality component. Running the Holt-Winters function indicates that there actually is seasonality as shown by the gamma. In my filtering and forecasting model, I set the gamma to .26 and perform a Holt-Winters exponential smoothing.



From the smoothing, it can be observed that the forecasts do not exactly agree well with the data, but nonetheless, I attempted to forecast onto a test set to see its accuracy. To create the Holt-Winters model, I split the first 100 games of the Seahawks dataset into the training set and used the next 30 games as a test set. The accuracy of this forecast is discussed in a later section.



ARIMA Model

To create the ARIMA model, I split half of the Seahawks games dataset (158 games starting in 2006 and ending in early 2017) into a training set and the remaining half into a test set. The total score for each game was turned into a time series model. The plot of this time series can be observed in the bottom-left figure below. After this, I used the *auto.arima* function to determine the best-fitting ARIMA model parameters. The following code was run:

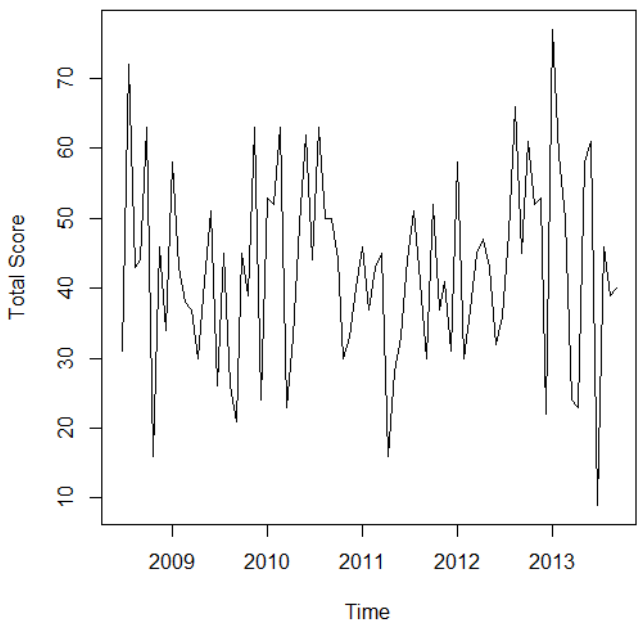
```
fitm2 <- auto.arima(m1, xreg = t2)
```

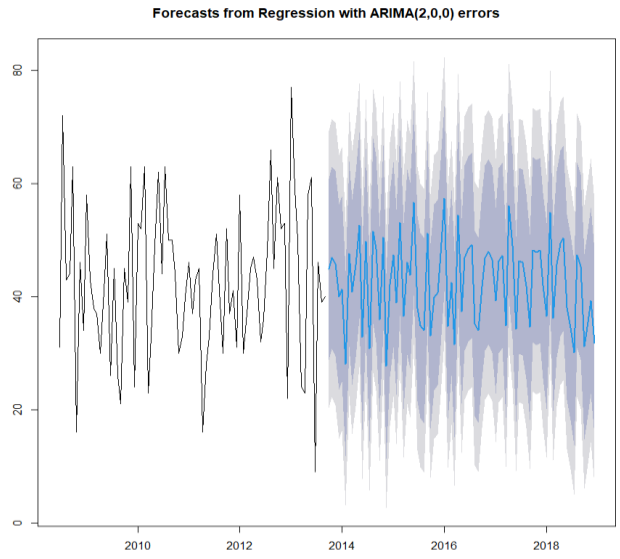
The most notable parameter used in this function – and one critical towards the creation of a more accurate model – is the *xreg* parameter. This parameter allows for the inclusion of extraneous variables that affect the response variable (in this case total score) to create a better fitting model. The *t2* variable represents an array with same length as the time series of the extraneous variables that affect each data point; specifically, these extraneous variables are: Wind Speed, Temperature, Humidity, Seahawks Defense Rank, Seahawks Offense Rank, Opponent Defense Rank, and Opponent Offense Rank. The *auto.arima* function chose an ARIMA(2,0,0) model as the most appropriate for the time series data.

After the creation of the ARIMA model, I created an ARIMA forecast, again utilizing the *xreg* function. The code used is as follows:

```
f1 <- forecast(fitm2, xreg = t2)
```

The data points of this generated forecast are what was used to determine the accuracy of this model (via comparison with actual game data from the test set), which will be expanded on in a later section. I compared the plot of the total score in the left figure against the plot of the forecast in the right figure below.





Model Diagnosis

To assess the accuracy of the initial models that were created, I ran a variety of diagnostic tests that determine the level of fit.

Holt-Winters Model

After running the Holt-Winters model, I decided to use a Ljung-Box test to determine whether the autocorrelations are different from zero (if the model fits properly or not). The Box test outputs the following result:

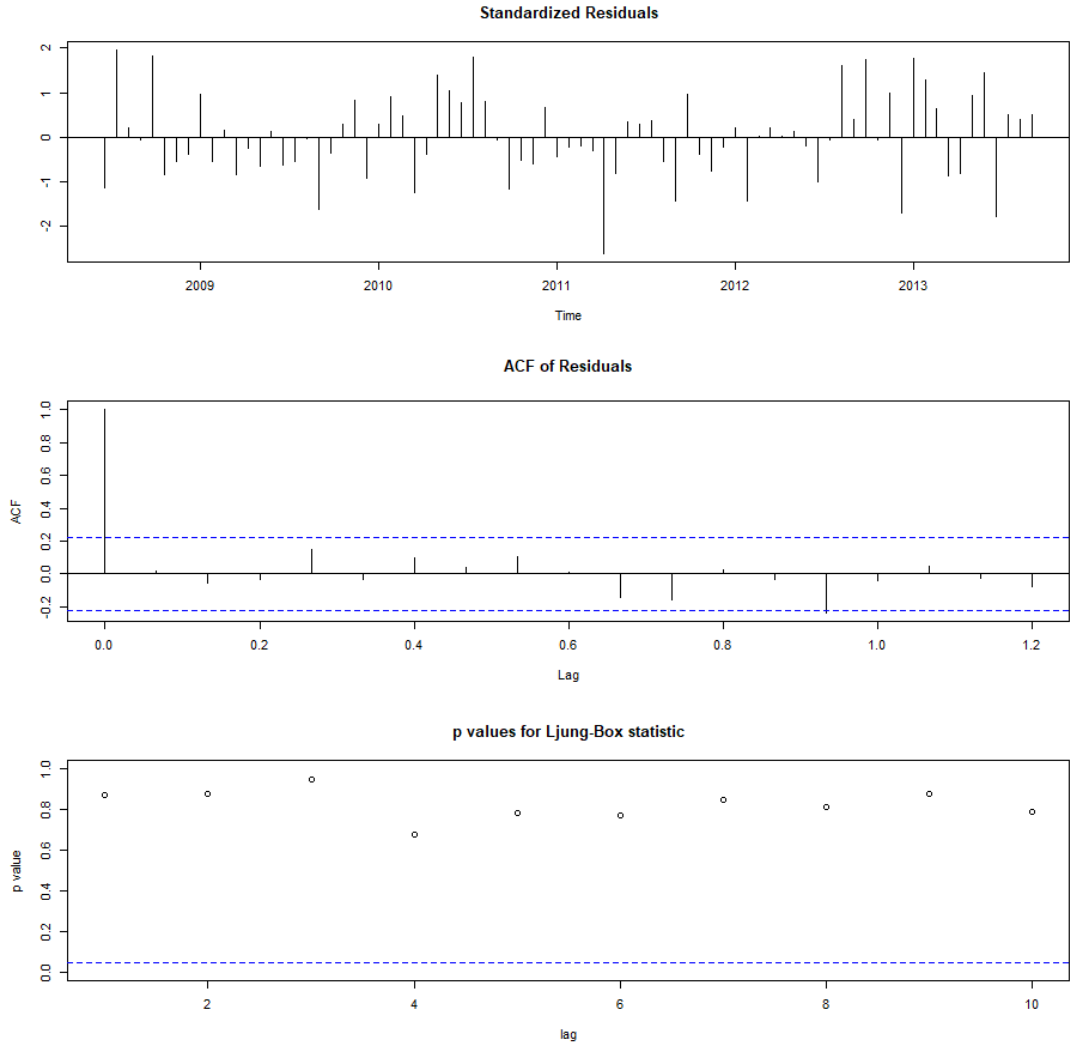
```
        Box-Ljung test

data:  scorecast2$residuals
X-squared = 31.495, df = 20, p-value = 0.04899
```

The p-value of 0.04899 is lower than the alpha value of 0.05, a standard measurement used to determine the outcomes of a hypothesis test. This means that there is sufficient evidence to reject the null hypothesis (that the model is a good fit for the data) in favor of the alternate hypothesis, meaning that the Holt-Winters model cannot be said to be a good fit for the time series. This is another reason as to why I incorporated an ARIMA model for predicting the total score.

ARIMA Model

For the ARIMA model, I began by again running an Ljung-Box test. This time, unlike with the Holt-Winters model, the p-value is calculated as 0.7251. This means that the null hypothesis cannot be rejected in favor of the alternate hypothesis, and thus the ARIMA model can be described as a good fit for the time series. As mentioned previously, the Ljung-Box test is primarily testing for if the autocorrelations of the residuals are different from zero. As can be seen in the ACF plot in the *tsdiag* plots in the figure below, the ACF of the residuals does not significantly differ from 0. The plot of p-values from the Ljung-Box tests are also observed to remain consistently well above the alpha value of 0.05.

```
> Box.test(fitm2$residuals, lag = 20, type = "Ljung-Box")

        Box-Ljung test

data:  fitm2$residuals
X-squared = 15.863, df = 20, p-value = 0.7251
```



Along with the Ljung-Box test and the *tsdiag* plots, I also calculated the AIC and the σ^2 values for the ARIMA model.

```
> fitm2$aic
[1] 634.1381

> fitm2$sigma2
[1] 155.2395
```

As can be seen above, the AIC is calculated as 634.1381 and the σ^2 (MSE) is calculated as 155.2395, which are both lower values than what was initially achieved with the Multiple Regression model.

Model Accuracy

In order to test the accuracy of my models, the following code was used:

```
projectedScores <- f1$mean

realLine <- data[n, Total Score Open]

projectedDifferential = projectedScores - realLine

realDifferential <- data[n, Differential]

productDiff <- projectedDifferential * realDifferential

posProductDiff <- productDiff[which(productDiff > 0)]

length(posProductDiff)/length(projectedScores)
```

**n* = `length(projectedScores)` more rows from where data ends and forecast begins

The code begins by storing the total scores projected by the forecast into the variable projectedScores. A projected Differential is calculated from this value and the actual total score line data from the corresponding value in the test set. This projected Differential value is multiplied by the actual differential value, leaving either a positive or negative number. If the number is positive, that means that the predicted total score and the actual total score were either both above or both below the actual Total Line Open for the game, and is counted as a successful prediction. The total number of correct predictions is divided by the total number of projected scores to receive the accuracy of the model.

Holt-Winters Model

In the Holt-Winters Model, the value represented by *n* was 101:130. Out of 30 total predictions, 12 were correct predictions, giving the model an accuracy rate of **40%**.

ARIMA Model

In the ARIMA model, the value represented by *n* was 80:158. Out of 79 total predictions, 44 were correct predictions, giving the model an accuracy rate of **55.70%**.

Current Season Upcoming Games Forecast

Because the ARIMA model gave the highest accuracy rate, I determined that it would be the best model to use to predict the total scores of the remaining Seahawks games within the 2020 season. The training set for this data is all Seahawks games from 2017 until current day, or about 3.94 total seasons. By using the *predict* function, I am able to input three important parameters: the model to predict with, the number of predictions, and the new set of *xreg* variables with the same number of rows as the number of predictions. The code is as follows:

```
data2 <- read.csv("seahawkspredict.csv")

seasonforecast <- data2[,4:10]
seasonforecast <- data.matrix(seasonforecast)

a1 <- predict(f1, n.ahead = 4, newxreg = seasonforecast, se.fit = TRUE)
a1$mean[1:4]
```

Because there are four remaining Seahawks games left in the season, the *n.ahead* parameter is set to 4. The *seasonforecast* variable input for the *newxreg* parameter is an array with the weather conditions and other variables expected at the time of these games. It is important to note that, due to the nature of weather, the values in the array will not be 100% accurate; however, they can be updated very easily to reflect changes in weather conditions. The values stored within *seasonforecast* are shown as the highlighted columns in the table below as a reference. Columns that are highlighted in blue are subject to change by the date of the game (weather conditions) while columns that are highlighted in green are constant.

Date	Opponent	Wind (mph)	Temp (°F)	Humidity (%)	Seahawks Defence Rank	Seahawks Offence Rank	Opponent Defence Rank	Opponent Offense Rank
12/13/20	Jets	9	53	61	26	4	29	32
12/20/20	Washington	7	54	70	26	4	7	24
12/27/20	Rams	8	52	77	26	4	4	20
1/3/21	49ers	9	57	80	26	4	11	21

The output of this prediction is displayed on the following page.

ARIMA Prediction Output

```
> a1$mean[1:4]
[1] 40.95890 51.98391 47.11397 45.42473
```

Date	Home Team	Away Team	Total Points Prediction
12/13/2020	Seattle Seahawks	New York Jets	41 (41.95890)
12/20/2020	Washington	Seattle Seahawks	52 (51.98391)
12/27/2020	Seattle Seahawks	Los Angeles Rams	47 (47.11397)
1/3/2021	San Francisco 49ers	Seattle Seahawks	45 (45.42473)

Above is a table outlining the predicted total points scored for the final four Seattle Seahawks games in the regular season. Based on the betting lines created for these games, a betting person can use this information to help in placing bets. For example, if the Total Line Open for the Seahawks vs Jets game is released as 45 points, my model suggests that the "Under" bet should be taken, as it predicts the score will be 41 points. Likewise, if the Total Line Open for the game is set at 40 points, my model suggests that taking the "Over" bet would be lucrative.

Conclusions

With the accuracy of my ARIMA model being 55.7%, if I were to place 100 bets on Over/Under lines with equal units per bet, I would be able to turn a profit of 5 units ($1000 per bet would result in a $5000 profit). Sports betting is never a guarantee, so anything over a 50% win rate on bets with binary outcomes can be considered successful - especially if it is sustainable. Additionally, this model can be continuously updated with Offensive and Defensive rankings. Weather data can be added as close to the game time as possible to create the most accurate prediction.