Aditya Kumar & Vivek Vasireddy

2022-05-11

How much does a QB Impact His Team's Success in the NFL?

Background

Football is America's most popular sport and is a huge part of this country's culture. The game starts with a coin toss, and the team that wins the coin toss chooses if they want to defer (where they kick the ball off to the opposing team and receive the ball when the second half starts) or to receive the ball. Once the kickoff receiver is tackled by the kicking team or calls a fair catch, the game begins wherever the ball is "downed" (this is the spot where the player was tackled). The offense then has 4 chances to advance ten yards, before they get another four chances. The offense can advance the ball with forward passes or running the ball. This continues until the offense reaches the end zone. At any given time, each team can only have up to 11 players on the field, so the defense can only use 11 players to stop the opposing team's offense.

Project Introduction

In this project, we wanted to examine how important a QB is to a team's success in the regular season. We first do some exploratory data analysis on a few QB metrics, so that we can get a basic understanding of the relationship between certain variables. Then, we use linear regression and stepwise regression to see how strong those specific relationships are between our predictor variables and the predicted variable. Furthermore, we will discover which predictors are the most significant and have the most impact on a team's success. With this information, people can possibly use this data to determine what kinds of QBs to sign or draft, or whether they need to have an elite QB at all since teams can still win by having good defense, rushing, etc.

Glossary

- TD = Touchdown (When the offensive team gets in to the endzone, this is worth 6 points)
- INT = Interception (When the offensive teams throws it to the defensive teams and the possession is switched)
- Sk = Sack (When the defensive team is able to make it past the offensive linemen and tackle the quarterback behind the line)
- Att = Attempts (When the offensive team attempts to throw the ball)
- Comp = Completion (When the offensive teams throws the ball and the offensive player catches is)
- QBR = QB Rating (A rating given to every team's Quarterback based on factors such as TD's, INT's, completions, attempts and yards thrown)
- 4QC = 4th Quarter Comebacks (When a team is trailing in the 4th quarter and is able to take the lead before the game ends)
- EXP = Expected Points (The Expected Points added by a teams offense)

```
overallNFL.df <- read.csv("2002-2021.csv")
overallAvgNFL.df <- read.csv("2002-2021Avg.csv")
```

```
summary(overallNFL.df)
```

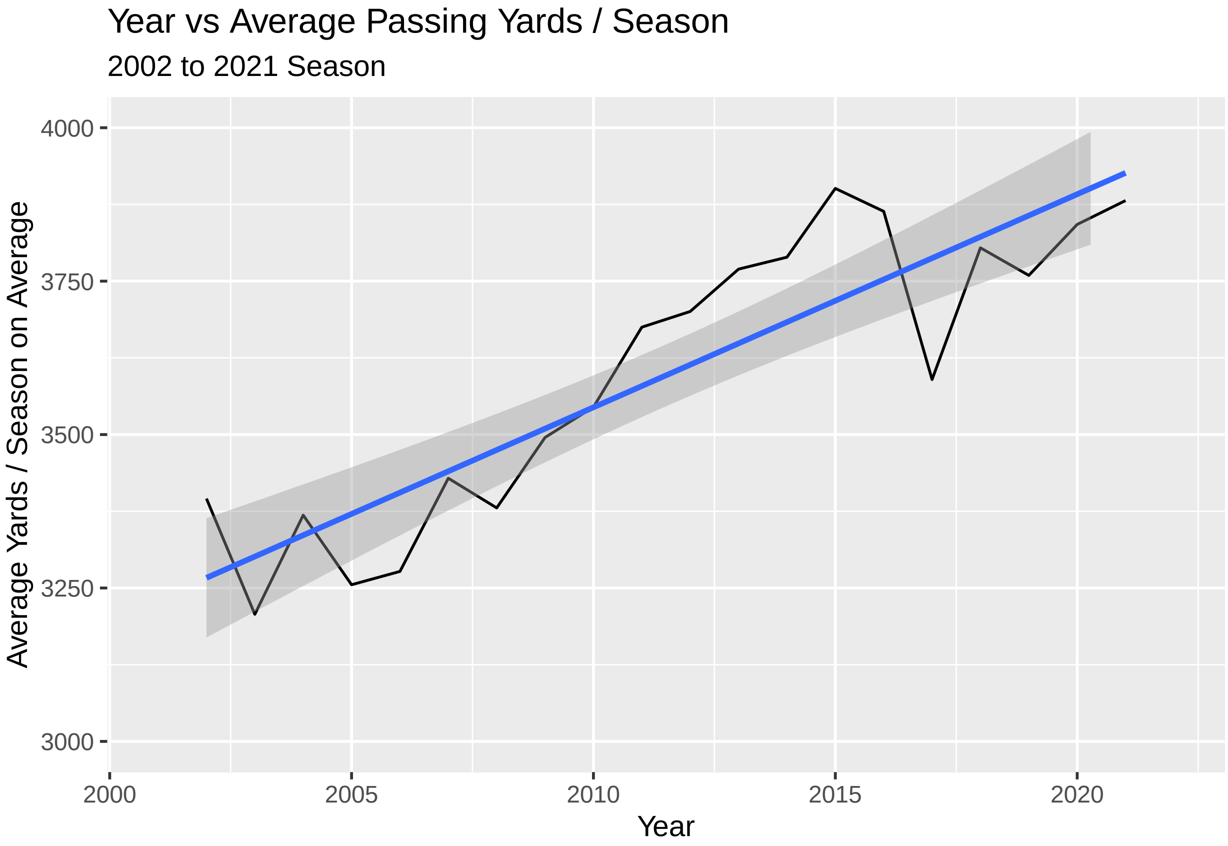
```
##       Year           Tm                  G              Cmp
##  Min.   :2002   Length:640         Min.   :16.00   Min.   :204.0
##  1st Qu.:2007   Class :character   1st Qu.:16.00   1st Qu.:299.0
##  Median :2012   Mode  :character   Median :16.00   Median :334.0
##  Mean   :2012                      Mean   :16.05   Mean   :335.8
##  3rd Qu.:2016                      3rd Qu.:16.00   3rd Qu.:371.0
##  Max.   :2021                      Max.   :17.00   Max.   :492.0
##       Att             Cmp.            Yds             TD             TD.
##  Min.   :358.0   Min.   :48.00   Min.   :1898   Min.   : 7.00   Min.   :1.400
##  1st Qu.:504.0   1st Qu.:58.50   1st Qu.:3140   1st Qu.:18.00   1st Qu.:3.400
##  Median :544.0   Median :61.60   Median :3576   Median :23.00   Median :4.100
##  Mean   :544.4   Mean   :61.54   Mean   :3596   Mean   :23.49   Mean   :4.306
##  3rd Qu.:588.0   3rd Qu.:64.90   3rd Qu.:4036   3rd Qu.:28.00   3rd Qu.:5.100
##  Max.   :740.0   Max.   :73.40   Max.   :5444   Max.   :55.00   Max.   :9.700
##       Int             Int.            Lng             Y.A
##  Min.   : 2.00   Min.   :0.400   Min.   :40.00   Min.   :5.10
##  1st Qu.:12.00   1st Qu.:2.100   1st Qu.:63.00   1st Qu.:6.50
##  Median :15.00   Median :2.700   Median :73.00   Median :7.00
##  Mean   :14.88   Mean   :2.751   Mean   :71.38   Mean   :7.04
##  3rd Qu.:18.00   3rd Qu.:3.400   3rd Qu.:80.00   3rd Qu.:7.60
##  Max.   :32.00   Max.   :5.500   Max.   :99.00   Max.   :9.30
##       AY.A             Y.C             Y.G             Rate
##  Min.   : 3.600   Min.   : 8.80   Min.   :118.6   Min.   : 53.60
##  1st Qu.: 5.800   1st Qu.:10.80   1st Qu.:195.0   1st Qu.: 76.58
##  Median : 6.600   Median :11.40   Median :222.9   Median : 85.30
##  Mean   : 6.663   Mean   :11.44   Mean   :224.1   Mean   : 85.58
##  3rd Qu.: 7.500   3rd Qu.:12.00   3rd Qu.:251.9   3rd Qu.: 93.80
##  Max.   :10.500   Max.   :14.20   Max.   :340.3   Max.   :122.60
##        Sk            Yds.1            Sk.              NY.A
##  Min.   :11.00   Min.   : 64.0   Min.   : 1.800   Min.   :4.100
##  1st Qu.:29.00   1st Qu.:184.0   1st Qu.: 5.000   1st Qu.:5.600
##  Median :36.00   Median :233.0   Median : 6.300   Median :6.100
##  Mean   :36.68   Mean   :238.8   Mean   : 6.372   Mean   :6.182
##  3rd Qu.:44.00   3rd Qu.:289.0   3rd Qu.: 7.700   3rd Qu.:6.800
##  Max.   :76.00   Max.   :484.0   Max.   :14.500   Max.   :8.500
##      ANY.A            X4QC            GWD             EXP               WIN
##  Min.   :2.500   Min.   :0.000   Min.   :0.000   Min.   :-237.06   Min.   : 0
##  1st Qu.:5.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: -54.87   1st Qu.: 6
##  Median :5.800   Median :2.000   Median :2.000   Median :  12.68   Median : 8
##  Mean   :5.831   Mean   :1.925   Mean   :2.536   Mean   :  14.81   Mean   : 8
##  3rd Qu.:6.700   3rd Qu.:3.000   3rd Qu.:3.250   3rd Qu.:  77.69   3rd Qu.:10
##  Max.   :9.600   Max.   :8.000   Max.   :8.000   Max.   : 287.73   Max.   :16
##      EXP.G              TD.INT
##  Min.   :-14.8163   Min.   : 0.2917
##  1st Qu.: -3.4292   1st Qu.: 1.0817
##  Median :  0.7875   Median : 1.5385
##  Mean   :  0.9118   Mean   : 1.8720
##  3rd Qu.:  4.8556   3rd Qu.: 2.2798
##  Max.   : 17.9831   Max.   :16.0000
```

```
summary(overallAvgNFL.df)
```

```
##       Year        AvgYdsSea       AvgWins        AvgYdsGame         AvgAtt
##  Min.   :2002   Min.   :3207   Min.   :7.906   Min.   :200.4   Min.   :511.1
##  1st Qu.:2007   1st Qu.:3392   1st Qu.:7.969   1st Qu.:212.0   1st Qu.:528.3
##  Median :2012   Median :3632   Median :7.969   Median :226.3   Median :545.3
##  Mean   :2012   Mean   :3596   Mean   :8.000   Mean   :224.1   Mean   :544.4
##  3rd Qu.:2016   3rd Qu.:3793   3rd Qu.:8.000   3rd Qu.:235.9   3rd Qu.:559.8
##  Max.   :2021   Max.   :3901   Max.   :8.469   Max.   :243.8   Max.   :584.8
##     AvgRate
##  Min.   :78.30
##  1st Qu.:82.75
##  Median :84.95
##  Mean   :85.69
##  3rd Qu.:89.53
##  Max.   :93.60
```

Part 1 - Exploratory Data Analysis

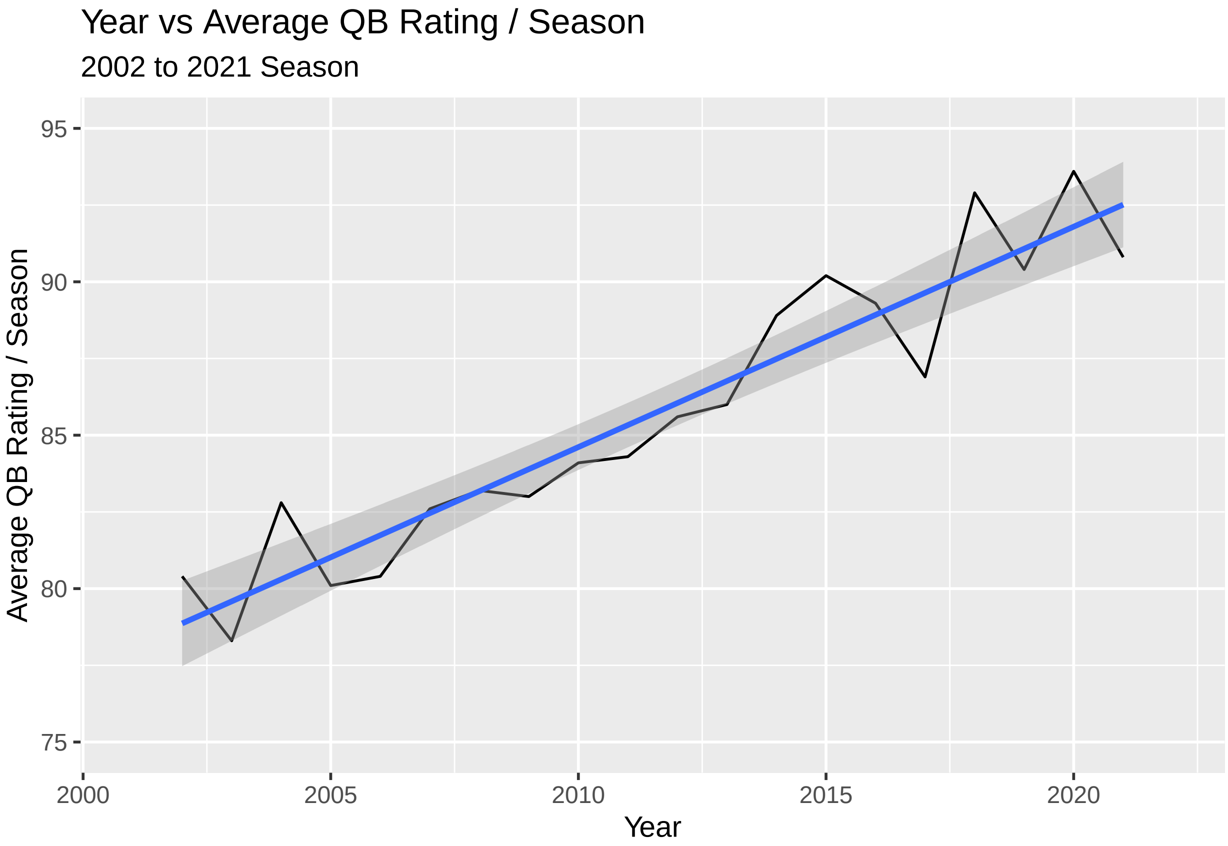
The data we are exploring covers the past 20 years and over that time the NFL has massively changed going from a run first game to becoming a passing first game. In doing so, we created plots that would describe just how the NFL game has changed. The metrics we used below were the Average Passing Yards Per Season, the change in Average QB Rating (A metic that is based off how the QB plays in a game) and the average attempts per season.



```
##
## Call:
```

```
## lm(formula = Year ~ AvgYdsSea, data = overallAvgNFL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9073 -2.0008 -0.2519  1.7612  5.6548
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.929e+03  9.882e+00 195.224  < 2e-16 ***
## AvgYdsSea   2.288e-02  2.742e-03   8.342 1.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.755 on 18 degrees of freedom
## Multiple R-squared:  0.7945, Adjusted R-squared:  0.7831
## F-statistic: 69.58 on 1 and 18 DF,  p-value: 1.344e-07
```

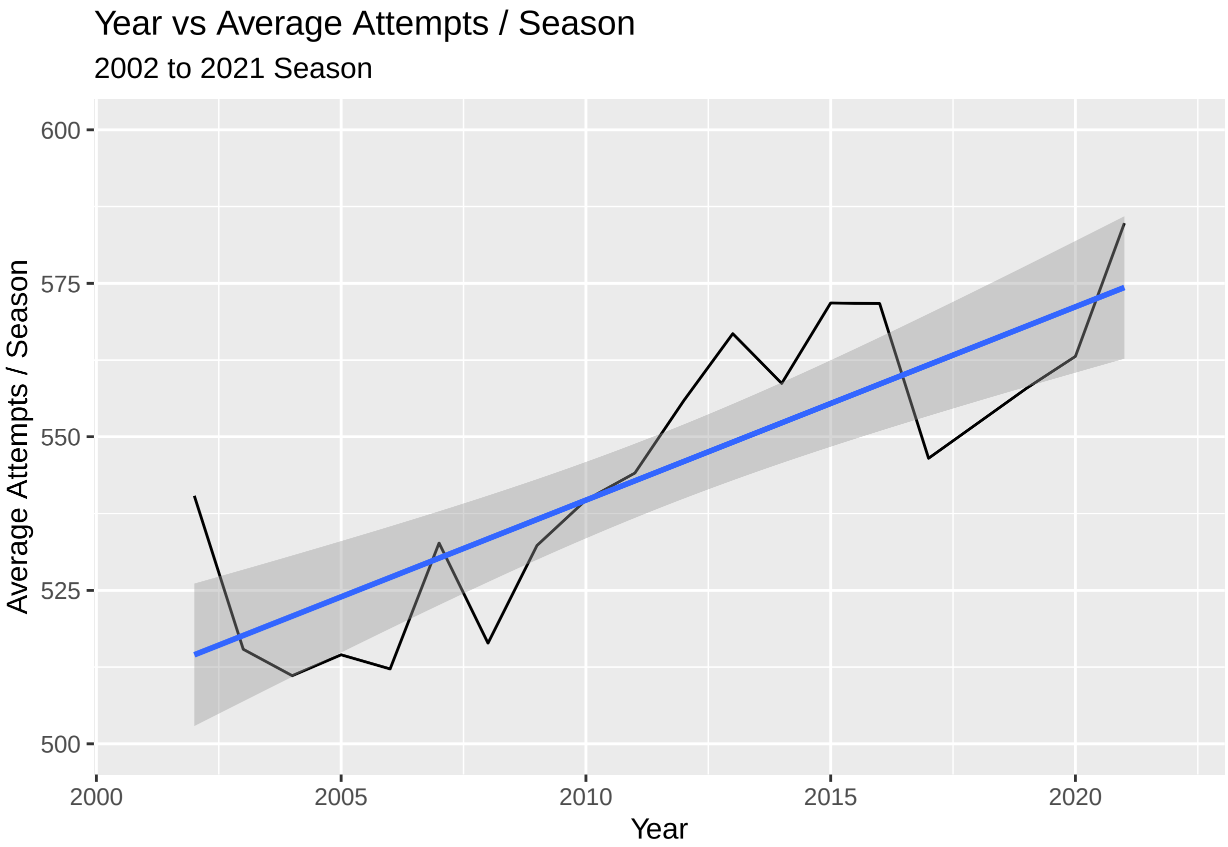
- In the graph that compares the Average Passing Yards per season and the year we can see that is has a positive trendline and a reasonably high adjusted r value. This is indicative that this is a good fit and there exists a positive correlation. Using the graph we can use this to show how NFL teams are looking to pass more and we don't expect this to change any time soon.



```
##
## Call:
## lm(formula = Year ~ AvgRate, data = overallAvgNFL.df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -3.9208 -1.3411  0.4461  1.0676  4.0015 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1905.3755     8.8212  216.00  < 2e-16 ***
## AvgRate        1.2385     0.1028   12.05 4.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.019 on 18 degrees of freedom
## Multiple R-squared:  0.8896, Adjusted R-squared:  0.8835 
## F-statistic: 145.1 on 1 and 18 DF,  p-value: 4.742e-10
```

- Here we have the graph between the QB Rating and the year. With this, we can see that a positive correlation exists between these variables given the graph is going upwards. From the linear model we can see that there is a high R squared value and the p value is less than 0.05 indicating that this is significant.


Year vs Average Attempts / Season

2002 to 2021 Season

```
## 
## Call:
## lm(formula = Year ~ AvgAtt, data = overallAvgNFL.df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.6208 -1.9542 -0.3143  1.8319  5.0418 
## 
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.892e+03  1.886e+01 100.342  < 2e-16 ***
## AvgAtt      2.193e-01  3.461e-02   6.335 5.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.382 on 18 degrees of freedom
## Multiple R-squared:  0.6904, Adjusted R-squared:  0.6732
## F-statistic: 40.13 on 1 and 18 DF,  p-value: 5.715e-06
```

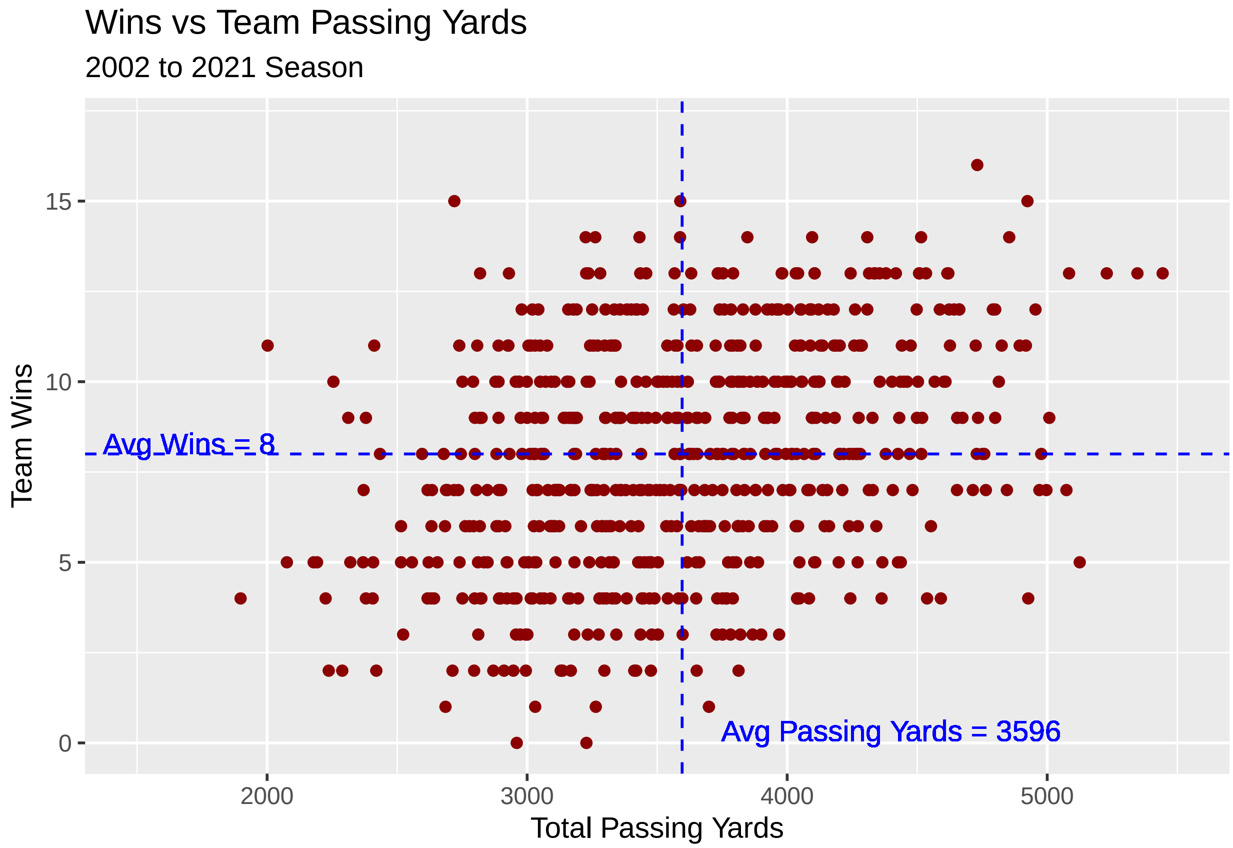
- Within the graph for the Average Attempts and the Year there exists a Positive correlation. With this linear model we can also see that there is a reasonably high adjusted R squared value. So like the other two graphs before this we can see that there is a high positive correlation within these.

Section Conclusion

The graphs above appear to show that the league is attempting to throw the ball more and completing more passes, all whilst the average QB play has appeared to have improved. This demonstrates that teams are putting more emphasis on QB's, and are trying to make them the focal point of their teams more now than in seasons past. In other words, teams are asking more of their QB's. So if the teams are becoming more QB reliant, what areas do the QBs have to excel at to help their teams win more?

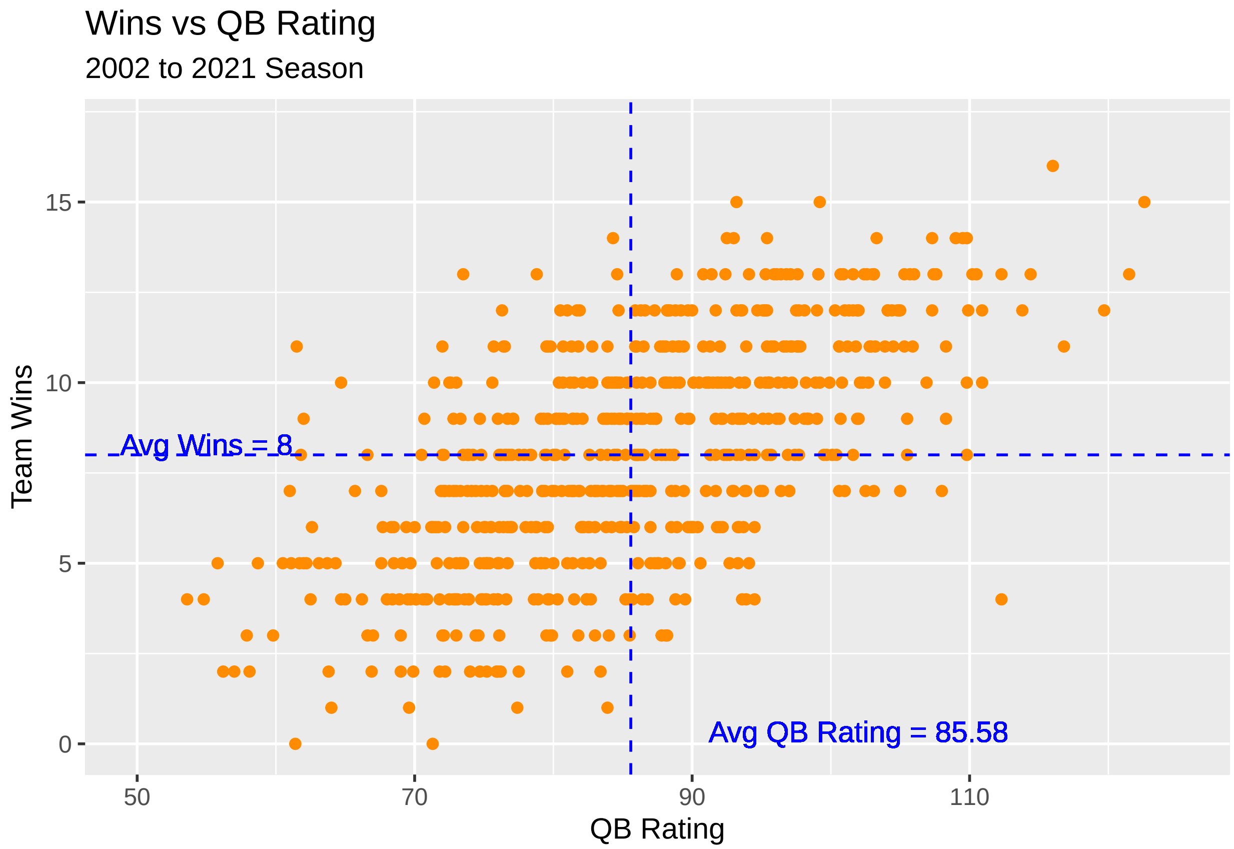
Part 2 - Predictive Modeling

Part 2a - Linear Regression Models



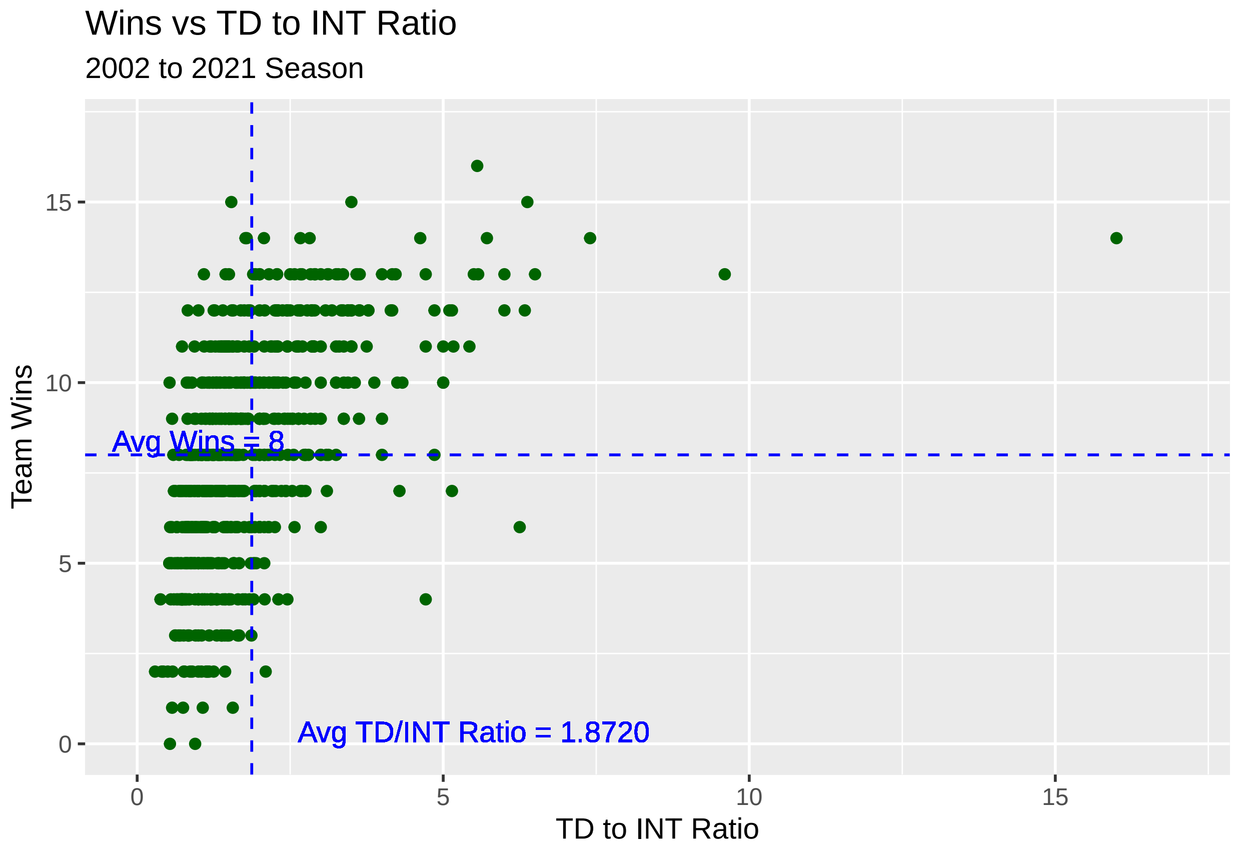
```
##
## Call:
## lm(formula = Yds ~ WIN, data = overallNFL.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1801.09  -423.96    -7.34   394.26  1735.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3045.438     63.956  47.618   <2e-16 ***
## WIN           68.878      7.458   9.235   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 582.5 on 638 degrees of freedom
## Multiple R-squared:  0.1179, Adjusted R-squared:  0.1165
## F-statistic: 85.28 on 1 and 638 DF,  p-value: < 2.2e-16
```

- The first linear model we plotted was between Passing Yards over the Season against the Wins a team got. We did this in order to possible see if our thought that a team who is more successful is better at passing. In this case it is given that the top right is heavily populated with these points. With the model above having a p-value of less than 0.05, there does seem to be a significance when it comes to passing yards and wins. So, teams with an above average passing yards tend to have more wins.



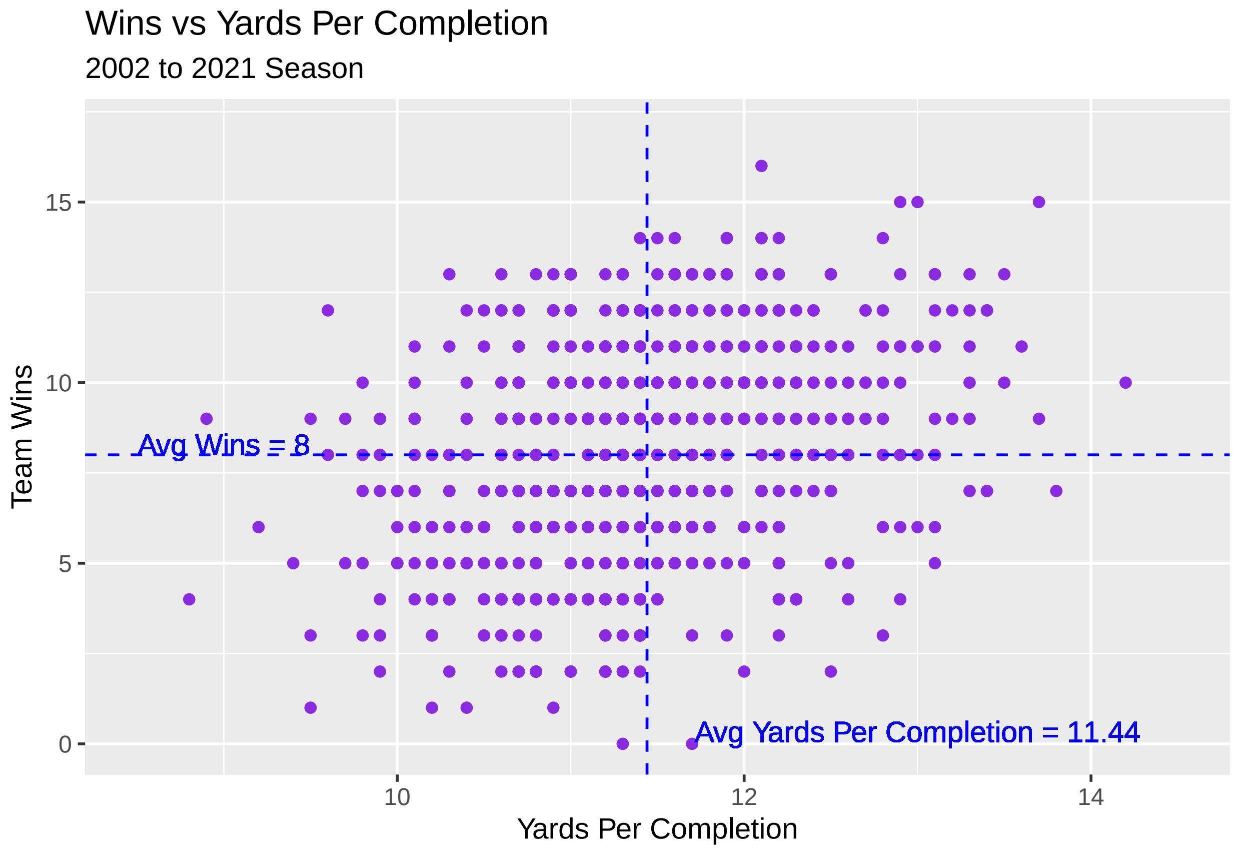
```
##
## Call:
## lm(formula = Rate ~ WIN, data = overallNFL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.338  -6.355  -0.184   6.418  36.394
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.2304     1.0295   64.33   <2e-16 ***
## WIN           2.4189     0.1201   20.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.377 on 638 degrees of freedom
## Multiple R-squared:  0.3888, Adjusted R-squared:  0.3879
## F-statistic: 405.9 on 1 and 638 DF,  p-value: < 2.2e-16
```

- This plot shows the QB Rating against the Wins. Here we can see that the graph exhibits an overall positive linear growth with the teams that have less wins also have a significantly lower passer rating than the teams with more wins. This is about what we expected and with the p value being very small this is significant. With this the conclusion can be drawn thatteams with an above average QB Rating would tend to have more wins.



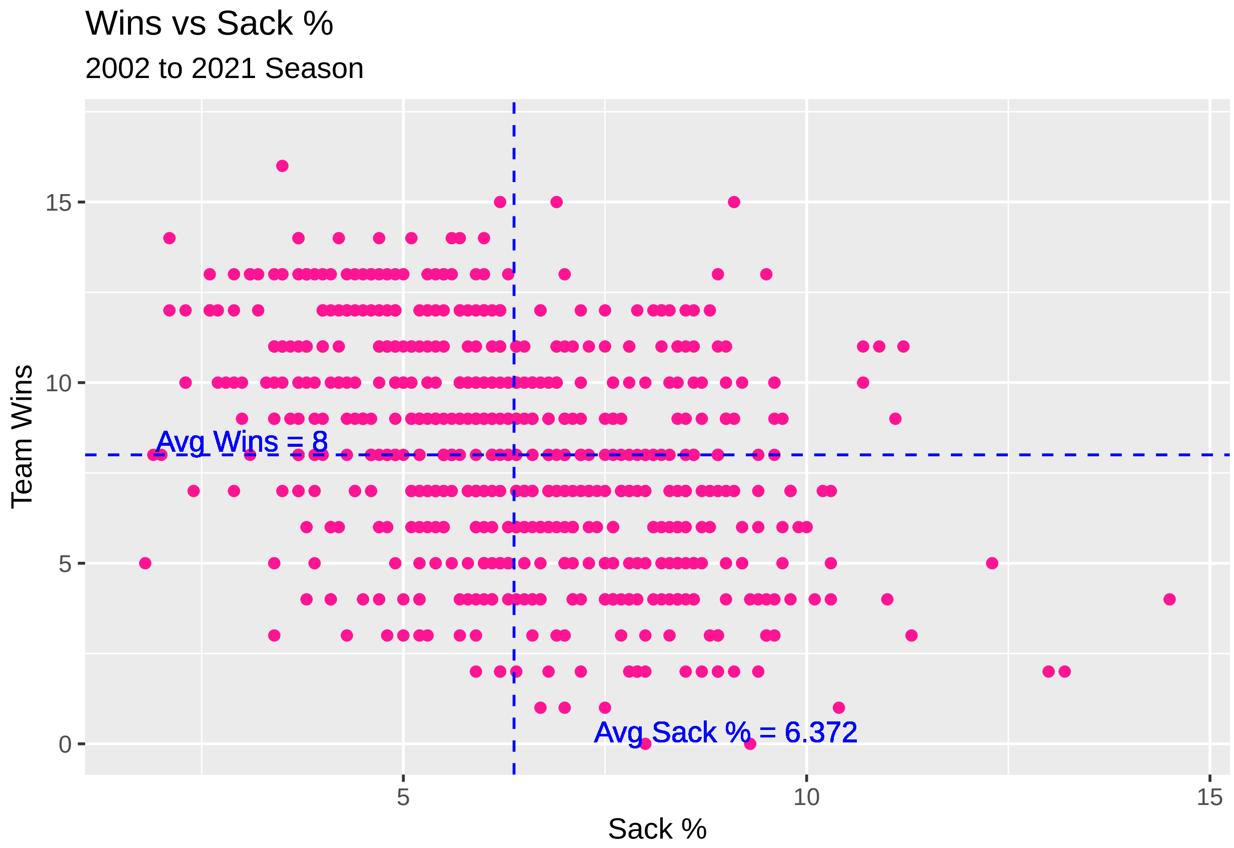
```
##
## Call:
## lm(formula = TD.INT ~ WIN, data = overallNFL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9324 -0.6027 -0.1948  0.3888 12.7983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09910    0.11790   0.841    0.401
## WIN          0.22162    0.01375  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 638 degrees of freedom
## Multiple R-squared:  0.2894, Adjusted R-squared:  0.2883
## F-statistic: 259.8 on 1 and 638 DF,  p-value: < 2.2e-16
```

- This graph above compares the TD to INT ratio against the Wins. The TD to Int ratio measures how many TDs a team throws compared to the amount of interceptions. Hence, the formula is just TD/INT. With this we can see that teams which throw a lot of TDs and few Ints tend to win more which is logical. Keeping the ball would mean more less chances for the opposition to score and also allows for the offensive team to have the ball more. However, if a team throws a similar amount of TDs and Ints, they won’t win more even if they score a lot.



```
##
## Call:
## lm(formula = Y.C ~ WIN, data = overallNFL.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63259 -0.51964 -0.03906  0.45446  2.57389
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.69087    0.08839 120.954   <2e-16 ***
## WIN          0.09352    0.01031   9.073   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.805 on 638 degrees of freedom
## Multiple R-squared:  0.1143, Adjusted R-squared:  0.1129
## F-statistic: 82.33 on 1 and 638 DF,  p-value: < 2.2e-16
```

- This graph compares Yards Per completion against the Wins. Yards per completion exists to show how many yards a team gained on average from the completions thrown. Teams with higher yards per completion usually have more solid and efficient QB play. In this case, with the p-value being so low we can see that this is in fact a significant factor.

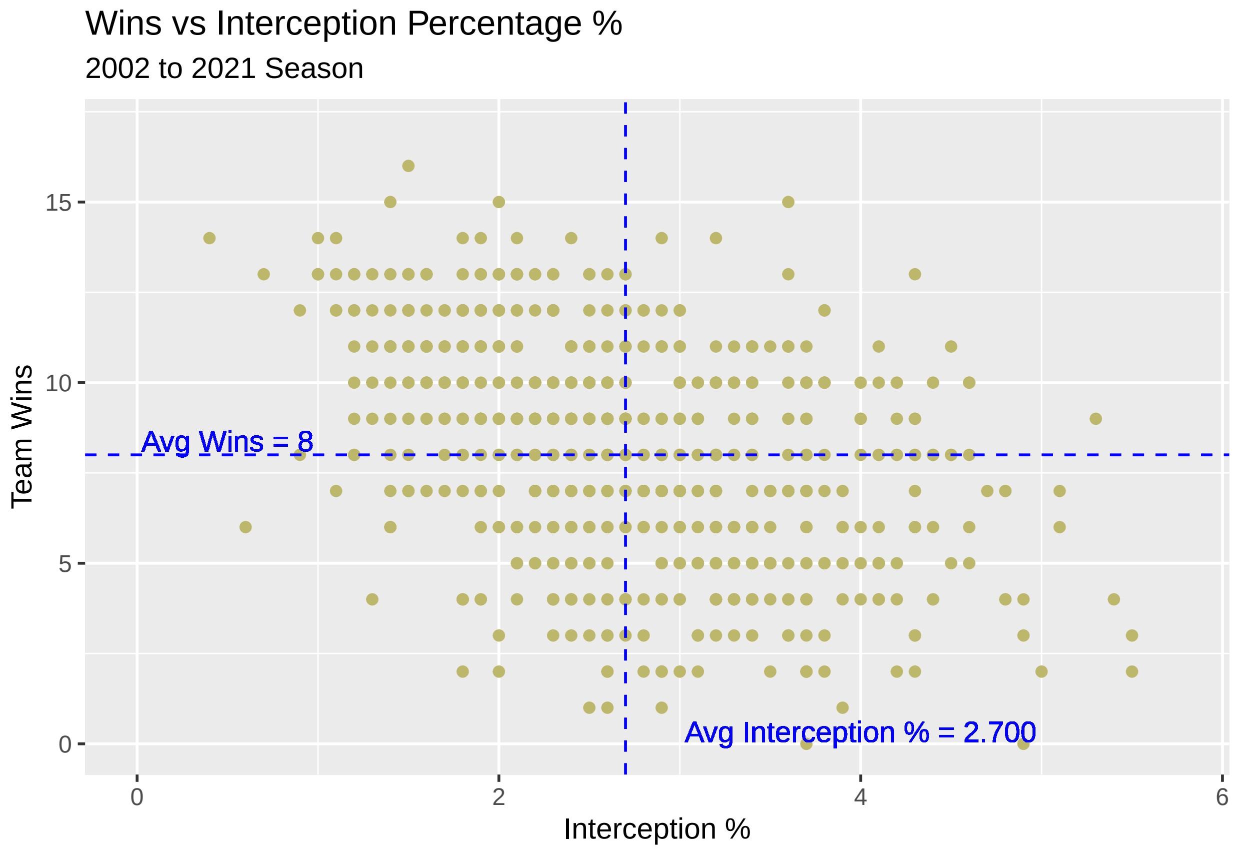


Wins vs Sack %

2002 to 2021 Season

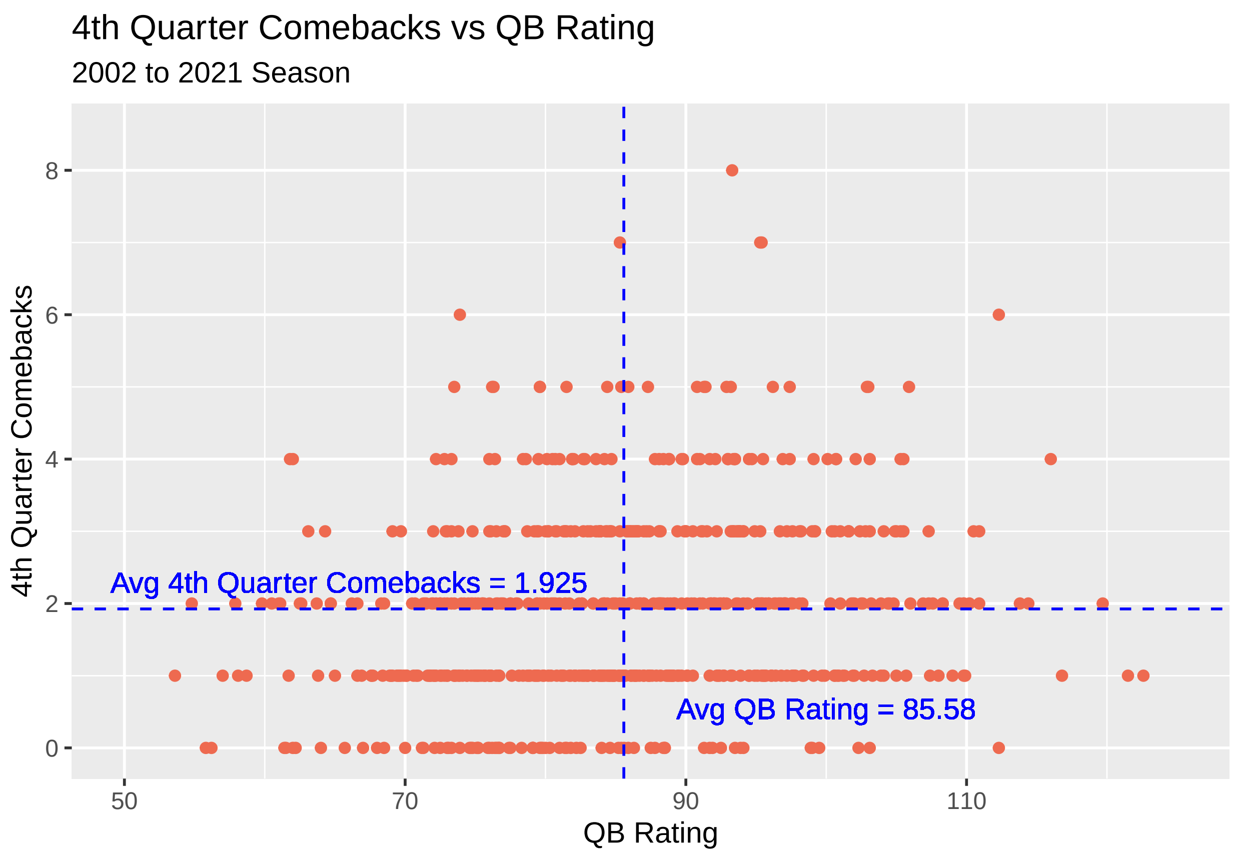
```
##
## Call:
## lm(formula = Sk. ~ WIN, data = overallNFL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3043 -1.2074 -0.1043  1.0839  7.1516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.32482    0.19607   42.46   <2e-16 ***
## WIN         -0.24410    0.02287  -10.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.786 on 638 degrees of freedom
## Multiple R-squared:  0.1516, Adjusted R-squared:  0.1502
## F-statistic:   114 on 1 and 638 DF,  p-value: < 2.2e-16
```

- Here we graphed the Sack % against the Team Wins. In this case we see that the teams in the top left quadrant are the most successful. Which makes sense as teams that protect the QB better or teams that have QBs that can avoid the sack end up winning more games. You would expect a better team to have a lower sack % and allow the team to not be knocked back and progress more which is the case in the graph.



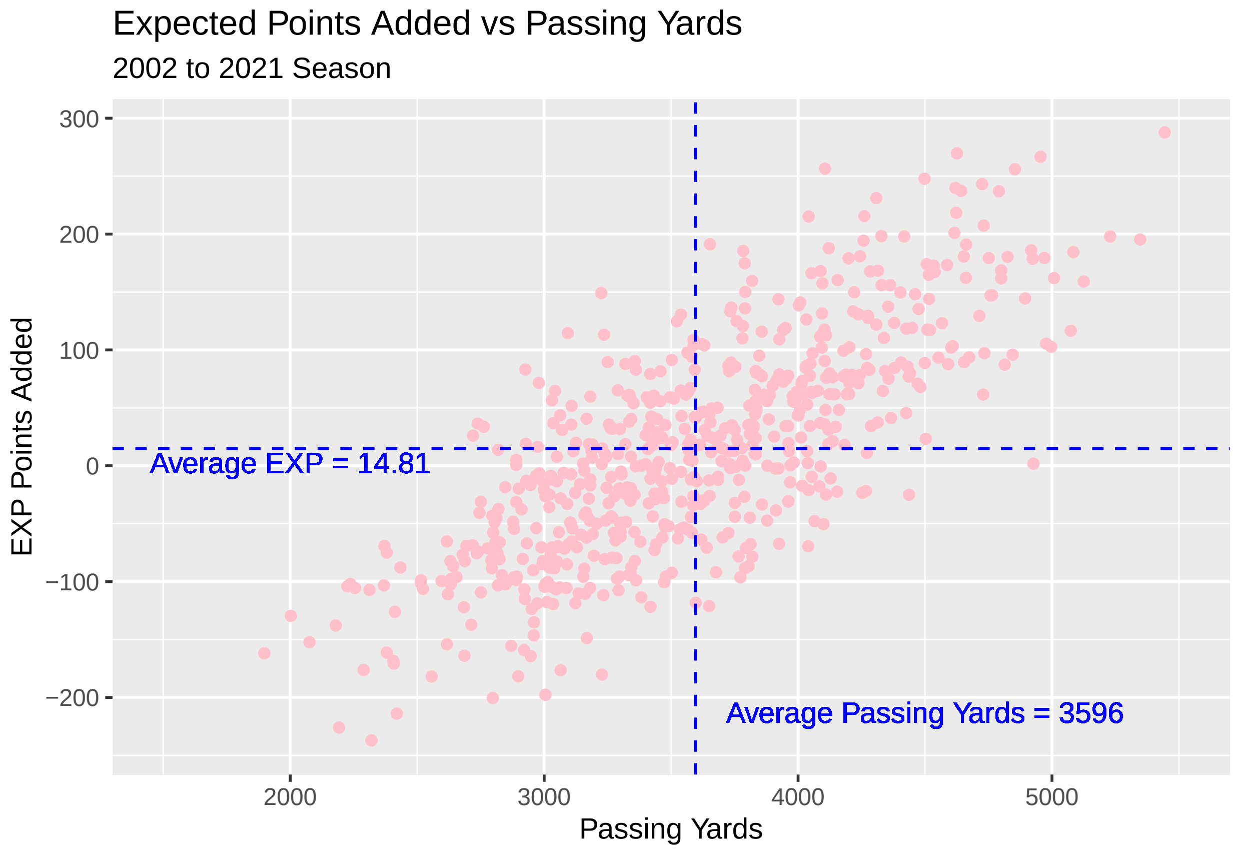
```
##
## Call:
## lm(formula = Int. ~ WIN, data = overallNFL.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41177 -0.59994 -0.04204  0.45105  2.67901
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.79335    0.08928   42.49   <2e-16 ***
## WIN         -0.13026    0.01041  -12.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8132 on 638 degrees of freedom
## Multiple R-squared:  0.197,  Adjusted R-squared:  0.1958
## F-statistic: 156.5 on 1 and 638 DF,  p-value: < 2.2e-16
```

- Similar to the graph of Sack Percentage, the teams that are the most successful are the ones that are able to keep the ball without turning it over often. As such, the graph resembles this as the data shows how the teams that have a lower INT % are often more successful.



```
##
## Call:
## lm(formula = Rate ~ X4QC, data = overallNFL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.918  -8.238  -0.498   7.772  38.702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  82.0783     0.8117 101.120  < 2e-16 ***
## X4QC          1.8200     0.3459   5.261 1.95e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.74 on 638 degrees of freedom
## Multiple R-squared:  0.04158,    Adjusted R-squared:  0.04008
## F-statistic: 27.68 on 1 and 638 DF,  p-value: 1.955e-07
```

- Here, we wanted to examine if having a higher rated QB makes a difference in pressure situations (can they help their team come from behind). Although there's a significance, the r-squared value is low which means that the data is extremely varied. Hence, it's hard to determine a conclusive relationship between the two variables because of the variance. If r value is low and significance is present, it goes to show QB is important but not the end all be all.



```
##
## Call:
## lm(formula = Yds ~ EXP, data = overallNFL.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1058.94  -289.42     7.31   262.70  1396.93
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3520.5999    15.8796  221.71   <2e-16 ***
## EXP            5.1210     0.1687   30.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 396.7 on 638 degrees of freedom
## Multiple R-squared:  0.5909, Adjusted R-squared:  0.5902
## F-statistic: 921.4 on 1 and 638 DF,  p-value: < 2.2e-16
```

- This compares the EXP points added against the yards thrown. We can see that this has a positive linear graph and with the p-value being less than 0.05 we can see that this is significant. Adjusted r squared value indicates that model is a decent fit. This is indicative that having more passing yards can help teams score more points which goes with the basic logic that more offense will lead to more points.

- Next, we wanted to implement a stepwise regression analysis. Since we have many different predictor variables, we felt a stepwise regression would be the best to compile all the data and allow us to present it as such.

Part 2b - Stepwise Regression

```
fit.all <- lm(WIN ~ ., data = overallNFL.df)
summary(fit.all)
```

```
##
## Call:
## lm(formula = WIN ~ ., data = overallNFL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9784 -1.3950  0.1043  1.2768  5.7364
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.5860042 19.5382433  -0.286   0.7751
## Cmp          0.0085833  0.0445488   0.193   0.8473
## Att         -0.0090683  0.0270150  -0.336   0.7372
## Cmp.         0.0963588  0.6277151   0.154   0.8780
## Yds         -0.0006982  0.0039355  -0.177   0.8592
## TD           0.1789831  0.1362484   1.314   0.1895
## TD.         -0.8489270  1.7686984  -0.480   0.6314
## Int         -0.1602529  0.1702484  -0.941   0.3469
## Int.         0.6447395  2.0087756   0.321   0.7483
## Y.A         -0.6169447  2.3125203  -0.267   0.7897
## AY.A         1.3980191  2.5320218   0.552   0.5811
## Y.C          1.4679046  1.0891454   1.348   0.1782
## Y.G         -0.0174439  0.0386245  -0.452   0.6517
## Rate         0.1737649  0.5781034   0.301   0.7638
## Sk           0.0048739  0.0733280   0.066   0.9470
## Yds.1       -0.0122005  0.0051301  -2.378   0.0177 *
## Sk.         -0.4581693  0.5122018  -0.895   0.3714
## NY.A        -0.0371177  2.3806677  -0.016   0.9876
## ANY.A       -2.8158694  2.1401753  -1.316   0.1888
## X4QC         0.1391770  0.1044552   1.332   0.1832
## GWD          0.4341745  0.0961035   4.518 7.49e-06 ***
## EXP          0.0522148  0.0824188   0.634   0.5266
## EXP.G       -0.7437962  1.3189346  -0.564   0.5730
## TD.INT       0.2122600  0.1199400   1.770   0.0773 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.936 on 616 degrees of freedom
## Multiple R-squared:  0.6215, Adjusted R-squared:  0.6073
## F-statistic: 43.97 on 23 and 616 DF,  p-value: < 2.2e-16
```

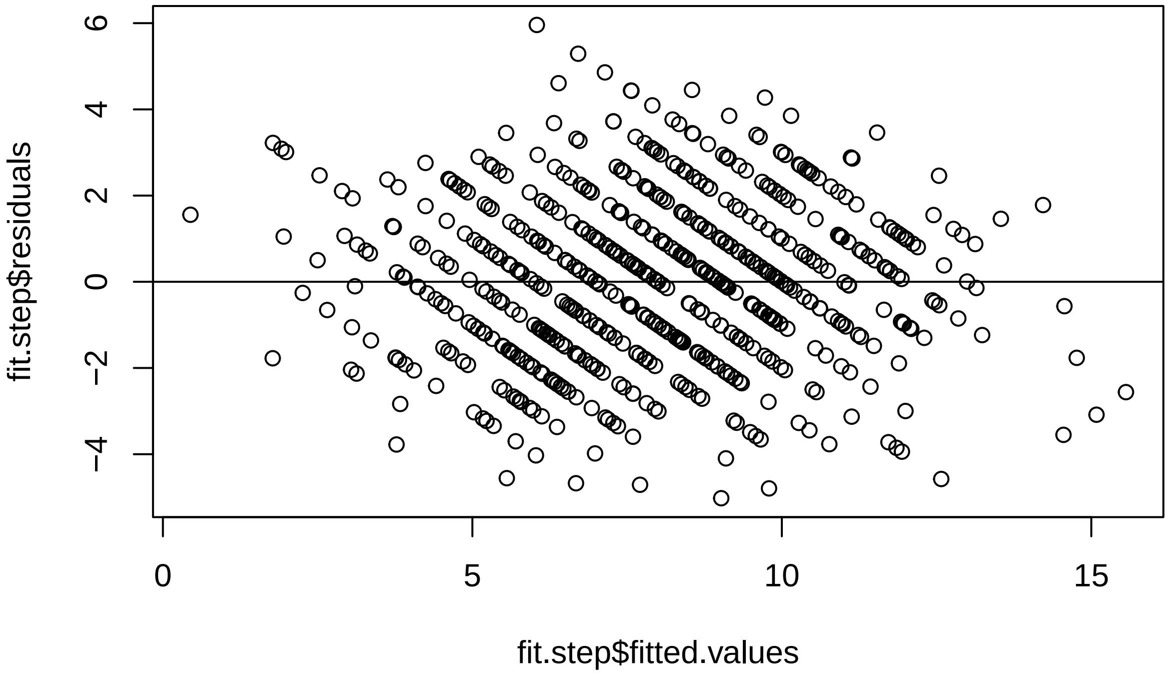
- The null hypothesis is that the data is normally distributed which we want. From this, we can accept the null hypothesis as the p-value is greater than 0.05 which is what we want to see.

```
fit.step <- step(fit.all)
```

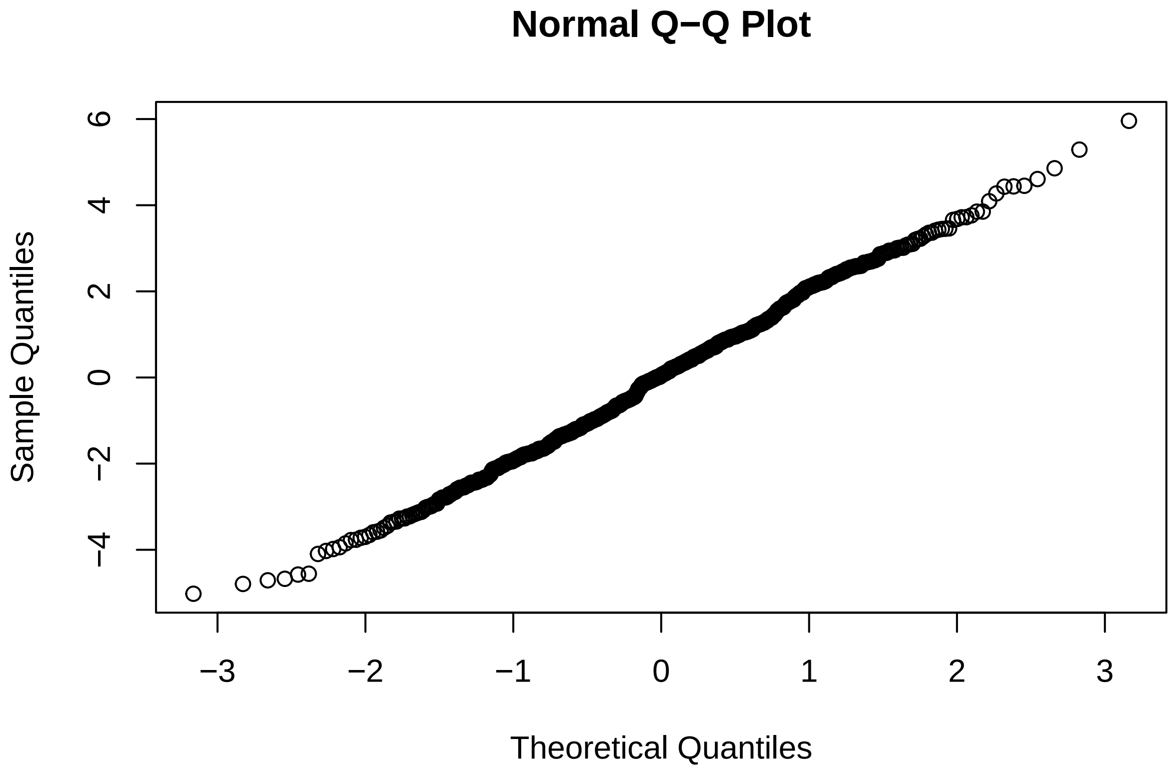
```
## Start:  AIC=869.21
## WIN ~ Cmp + Att + Cmp. + Yds + TD + TD. + Int + Int. + Y.A +
##     AY.A + Y.C + Y.G + Rate + Sk + Yds.1 + Sk. + NY.A + ANY.A +
##     X4QC + GWD + EXP + EXP.G + TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - NY.A    1     0.001 2309.1 867.21
## - Sk      1     0.017 2309.1 867.21
## - Cmp.    1     0.088 2309.2 867.23
## - Yds     1     0.118 2309.2 867.24
## - Cmp     1     0.139 2309.2 867.25
## - Y.A     1     0.267 2309.4 867.28
## - Rate    1     0.339 2309.4 867.30
## - Int.    1     0.386 2309.5 867.32
## - Att     1     0.422 2309.5 867.33
## - Y.G     1     0.765 2309.8 867.42
## - TD.     1     0.864 2309.9 867.45
## - AY.A    1     1.143 2310.2 867.53
## - EXP.G   1     1.192 2310.3 867.54
## - EXP     1     1.505 2310.6 867.63
## - Sk.     1     2.999 2312.1 868.04
## - Int     1     3.321 2312.4 868.13
## - TD      1     6.469 2315.6 869.00
## - ANY.A   1     6.489 2315.6 869.01
## - X4QC    1     6.655 2315.7 869.05
## - Y.C     1     6.809 2315.9 869.09
## <none>                2309.1 869.21
## - TD.INT  1    11.740 2320.8 870.46
## - Yds.1   1    21.201 2330.3 873.06
## - GWD     1    76.509 2385.6 888.07
##
## Step:  AIC=867.21
## WIN ~ Cmp + Att + Cmp. + Yds + TD + TD. + Int + Int. + Y.A +
##     AY.A + Y.C + Y.G + Rate + Sk + Yds.1 + Sk. + ANY.A + X4QC +
##     GWD + EXP + EXP.G + TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - Sk      1     0.016 2309.1 865.21
## - Cmp.    1     0.087 2309.2 865.23
## - Yds     1     0.126 2309.2 865.24
## - Cmp     1     0.142 2309.2 865.25
## - Y.A     1     0.274 2309.4 865.29
## - Rate    1     0.338 2309.4 865.30
## - Int.    1     0.398 2309.5 865.32
## - Att     1     0.423 2309.5 865.33
## - Y.G     1     0.764 2309.8 865.42
## - TD.     1     0.869 2310.0 865.45
## - AY.A    1     1.157 2310.2 865.53
## - EXP.G   1     1.191 2310.3 865.54
## - EXP     1     1.504 2310.6 865.63
## - Int     1     3.391 2312.5 866.15
## - Sk.     1     3.543 2312.6 866.19
## - TD      1     6.548 2315.6 867.02
## - X4QC    1     6.657 2315.8 867.05
## - Y.C     1     7.138 2316.2 867.19
## <none>                2309.1 867.21
## - ANY.A   1     7.537 2316.6 867.30
## - TD.INT  1    11.765 2320.8 868.46
## - Yds.1   1    25.396 2334.5 872.21
## - GWD     1    76.606 2385.7 886.10
##
## Step:  AIC=865.21
## WIN ~ Cmp + Att + Cmp. + Yds + TD + TD. + Int + Int. + Y.A +
##     AY.A + Y.C + Y.G + Rate + Yds.1 + Sk. + ANY.A + X4QC + GWD +
##     EXP + EXP.G + TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - Cmp.    1     0.097 2309.2 863.24
## - Yds     1     0.120 2309.2 863.25
## - Cmp     1     0.128 2309.2 863.25
## - Y.A     1     0.283 2309.4 863.29
## - Rate    1     0.332 2309.4 863.31
## - Int.    1     0.390 2309.5 863.32
## - Att     1     0.417 2309.5 863.33
## - Y.G     1     0.777 2309.9 863.43
## - TD.     1     0.863 2310.0 863.45
## - AY.A    1     1.144 2310.2 863.53
## - EXP.G   1     1.179 2310.3 863.54
## - EXP     1     1.490 2310.6 863.63
## - Int     1     3.391 2312.5 864.15
## - TD      1     6.532 2315.6 865.02
## - X4QC    1     6.648 2315.8 865.05
## - Y.C     1     7.123 2316.2 865.19
## <none>                2309.1 865.21
## - ANY.A   1     7.782 2316.9 865.37
## - TD.INT  1    11.793 2320.9 866.47
## - Sk.     1    22.607 2331.7 869.45
## - Yds.1   1    28.511 2337.6 871.07
## - GWD     1    76.751 2385.9 884.14
##
## Step:  AIC=863.24
## WIN ~ Cmp + Att + Yds + TD + TD. + Int + Int. + Y.A + AY.A +
##     Y.C + Y.G + Rate + Yds.1 + Sk. + ANY.A + X4QC + GWD + EXP +
##     EXP.G + TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - Yds     1     0.142 2309.3 861.28
## - Y.A     1     0.285 2309.5 861.32
## - Cmp     1     0.324 2309.5 861.33
## - Att     1     0.713 2309.9 861.44
## - Y.G     1     0.773 2310.0 861.46
## - AY.A    1     1.048 2310.2 861.53
## - EXP.G   1     1.201 2310.4 861.57
## - Int.    1     1.307 2310.5 861.60
## - EXP     1     1.516 2310.7 861.66
## - Rate    1     2.839 2312.0 862.03
## - TD.     1     3.277 2312.5 862.15
## - Int     1     3.356 2312.6 862.17
## - TD      1     6.443 2315.7 863.02
## - X4QC    1     6.617 2315.8 863.07
## <none>                2309.2 863.24
## - ANY.A   1     8.173 2317.4 863.50
## - Y.C     1     8.263 2317.5 863.53
## - TD.INT  1    11.705 2320.9 864.48
## - Sk.     1    23.195 2332.4 867.64
## - Yds.1   1    30.155 2339.4 869.54
## - GWD     1    77.088 2386.3 882.26
##
## Step:  AIC=861.28
## WIN ~ Cmp + Att + TD + TD. + Int + Int. + Y.A + AY.A + Y.C +
##     Y.G + Rate + Yds.1 + Sk. + ANY.A + X4QC + GWD + EXP + EXP.G +
##     TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - Cmp     1     0.217 2309.6 859.34
## - Y.A     1     0.387 2309.7 859.39
## - Att     1     0.875 2310.2 859.52
## - AY.A    1     0.953 2310.3 859.54
## - EXP.G   1     1.087 2310.4 859.58
## - Int.    1     1.270 2310.6 859.63
## - EXP     1     1.434 2310.8 859.68
## - Y.G     1     2.504 2311.8 859.97
## - TD.     1     3.139 2312.5 860.15
## - Rate    1     3.288 2312.6 860.19
## - Int     1     3.413 2312.8 860.23
## - TD      1     6.583 2315.9 861.10
## - X4QC    1     6.778 2316.1 861.16
## <none>                2309.3 861.28
## - Y.C     1     8.124 2317.5 861.53
## - ANY.A   1     9.177 2318.5 861.82
## - TD.INT  1    11.744 2321.1 862.53
## - Sk.     1    24.136 2333.5 865.93
## - Yds.1   1    31.231 2340.6 867.88
## - GWD     1    76.973 2386.3 880.26
##
## Step:  AIC=859.34
## WIN ~ Att + TD + TD. + Int + Int. + Y.A + AY.A + Y.C + Y.G +
##     Rate + Yds.1 + Sk. + ANY.A + X4QC + GWD + EXP + EXP.G + TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - Y.A     1     0.420 2310.0 857.46
## - AY.A    1     0.804 2310.4 857.56
## - Att     1     1.052 2310.6 857.63
## - EXP.G   1     1.388 2310.9 857.73
## - EXP     1     1.797 2311.4 857.84
## - Y.G     1     2.293 2311.8 857.98
## - Int.    1     2.882 2312.4 858.14
## - Int     1     4.631 2314.2 858.62
## - TD.     1     6.557 2316.1 859.16
## - X4QC    1     6.827 2316.4 859.23
## <none>                2309.6 859.34
## - TD      1     7.519 2317.1 859.42
## - Y.C     1     7.909 2317.5 859.53
## - Rate    1     8.788 2318.3 859.77
## - ANY.A   1     9.565 2319.1 859.99
## - TD.INT  1    11.563 2321.1 860.54
## - Sk.     1    24.350 2333.9 864.05
## - Yds.1   1    31.264 2340.8 865.95
## - GWD     1    76.920 2386.5 878.31
##
## Step:  AIC=857.46
## WIN ~ Att + TD + TD. + Int + Int. + AY.A + Y.C + Y.G + Rate +
##     Yds.1 + Sk. + ANY.A + X4QC + GWD + EXP + EXP.G + TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - AY.A    1     0.533 2310.5 855.60
## - Att     1     0.827 2310.8 855.69
## - EXP.G   1     1.166 2311.2 855.78
## - EXP     1     1.548 2311.5 855.89
## - Int.    1     2.693 2312.7 856.20
## - Y.G     1     2.786 2312.8 856.23
## - Int     1     5.273 2315.2 856.92
## - TD.     1     6.442 2316.4 857.24
## - X4QC    1     6.956 2316.9 857.38
## <none>                 2310.0 857.46
## - Y.C     1     7.732 2317.7 857.60
## - TD      1     8.211 2318.2 857.73
## - Rate    1     8.699 2318.7 857.86
## - ANY.A   1    11.250 2321.2 858.57
## - TD.INT  1    11.743 2321.7 858.70
## - Sk.     1    26.795 2336.8 862.84
## - Yds.1   1    34.903 2344.9 865.06
## - GWD     1    76.574 2386.6 876.33
##
## Step:  AIC=855.6
## WIN ~ Att + TD + TD. + Int + Int. + Y.C + Y.G + Rate + Yds.1 +
##     Sk. + ANY.A + X4QC + GWD + EXP + EXP.G + TD.INT
##
##          Df Sum of Sq    RSS    AIC
## - Att     1     1.094 2311.6 853.91
## - EXP.G   1     1.394 2311.9 853.99
## - EXP     1     1.814 2312.3 854.11
## - Int.    1     2.274 2312.8 854.23
## - Y.G     1     2.441 2313.0 854.28
## - Int     1     5.122 2315.6 855.02
## - TD.     1     6.473 2317.0 855.40
## - X4QC    1     6.897 2317.4 855.51
## <none>                 2310.5 855.60
## - TD      1     7.786 2318.3 855.76
## - ANY.A   1    10.951 2321.5 856.63
## - TD.INT  1    11.795 2322.3 856.86
## - Y.C     1    18.900 2329.4 858.82
## - Rate    1    20.946 2331.5 859.38
## - Sk.     1    27.129 2337.6 861.08
## - Yds.1   1    35.257 2345.8 863.30
## - GWD     1    76.788 2387.3 874.53
##
```

```
## Step:  AIC=853.91
## WIN ~ TD + TD. + Int + Int. + Y.C + Y.G + Rate + Yds.1 + Sk. + 
##     ANY.A + X4QC + GWD + EXP + EXP.G + TD.INT
## 
##          Df Sum of Sq    RSS    AIC
## - EXP.G   1     0.636 2312.2 852.08
## - EXP     1     0.963 2312.6 852.17
## - X4QC    1     6.541 2318.2 853.72
## <none>                2311.6 853.91
## - TD      1     7.444 2319.1 853.97
## - TD.     1     7.447 2319.1 853.97
## - ANY.A   1    10.063 2321.7 854.69
## - Int     1    10.859 2322.5 854.91
## - TD.INT  1    11.570 2323.2 855.10
## - Int.    1    12.207 2323.8 855.28
## - Y.G     1    13.460 2325.1 855.62
## - Y.C     1    22.987 2334.6 858.24
## - Rate    1    24.663 2336.3 858.70
## - Sk.     1    26.199 2337.8 859.12
## - Yds.1   1    45.432 2357.0 864.36
## - GWD     1    77.331 2388.9 872.97
## 
## Step:  AIC=852.08
## WIN ~ TD + TD. + Int + Int. + Y.C + Y.G + Rate + Yds.1 + Sk. + 
##     ANY.A + X4QC + GWD + EXP + TD.INT
## 
##          Df Sum of Sq    RSS    AIC
## - X4QC    1     6.400 2318.7 851.85
## <none>                2312.2 852.08
## - ANY.A   1     9.810 2322.1 852.79
## - Int     1    10.239 2322.5 852.91
## - TD.INT  1    11.365 2323.6 853.22
## - TD.     1    11.763 2324.0 853.33
## - Int.    1    11.881 2324.1 853.36
## - TD      1    12.120 2324.4 853.43
## - Y.G     1    20.263 2332.5 855.67
## - Y.C     1    23.898 2336.1 856.66
## - Rate    1    25.650 2337.9 857.14
## - Sk.     1    26.147 2338.4 857.28
## - EXP     1    29.141 2341.4 858.10
## - Yds.1   1    45.758 2358.0 862.63
## - GWD     1    78.265 2390.5 871.39
## 
## Step:  AIC=851.85
## WIN ~ TD + TD. + Int + Int. + Y.C + Y.G + Rate + Yds.1 + Sk. + 
##     ANY.A + GWD + EXP + TD.INT
## 
##          Df Sum of Sq    RSS    AIC
## <none>                2318.7 851.85
## - ANY.A   1      9.43 2328.1 852.45
## - Int     1     10.35 2329.0 852.70
## - TD.     1     11.24 2329.9 852.95
## - TD.INT  1     11.31 2329.9 852.97
## - TD      1     11.65 2330.3 853.06
## - Int.     1      11.74 2330.4 853.09
## - Y.G      1      19.67 2338.3 855.26
## - Y.C      1      23.27 2341.9 856.25
## - Rate     1      24.63 2343.3 856.62
## - Sk.      1      26.28 2344.9 857.07
## - EXP      1      31.21 2349.9 858.41
## - Yds.1    1      43.68 2362.3 861.80
## - GWD      1     368.91 2687.6 944.35
```

```
plot(fit.step$residuals~fit.step$fitted.values)
abline(h=0)
```



```
qqnorm(fit.step$residuals)
```



```
shapiro.test(fit.step$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.step$residuals
## W = 0.9968, p-value = 0.2353
```

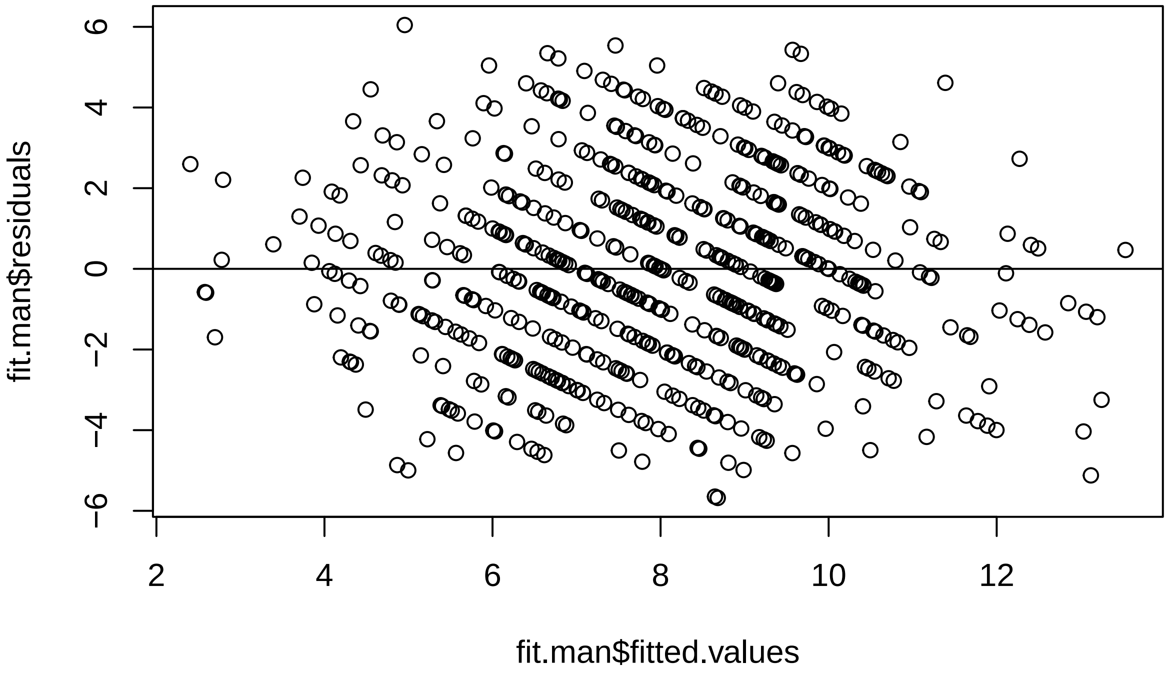
- Noted that the p-value is 0.2353 which indicates that the residuals are normally distributed which we want to see.

```
fit.man <- lm(WIN ~ GWD + Yds.1 + Y.C, data = overallNFL.df)
summary(fit.man)
```

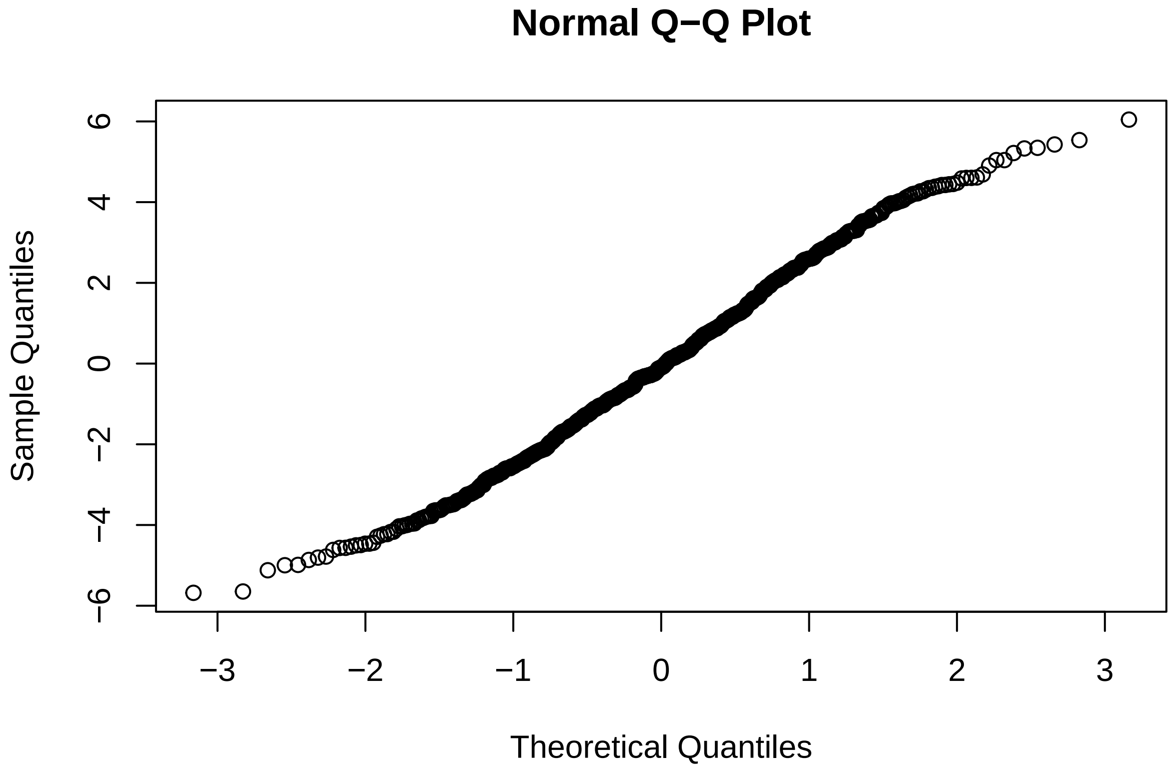
```
##
## Call:
## lm(formula = WIN ~ GWD + Yds.1 + Y.C, data = overallNFL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6803 -1.7018 -0.0935  1.7452  6.0448
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.034491   1.350764  -0.026     0.98
## GWD          0.687058   0.065568  10.479  < 2e-16 ***
## Yds.1       -0.016204   0.001294 -12.520  < 2e-16 ***
## Y.C          0.888350   0.112342   7.908 1.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.391 on 636 degrees of freedom
## Multiple R-squared:  0.4038, Adjusted R-squared:  0.4009
## F-statistic: 143.6 on 3 and 636 DF,  p-value: < 2.2e-16
```

```
plot(fit.man$residuals~fit.man$fitted.values)
abline(h=0)
```



```
qqnorm(fit.man$residuals)
```



```
shapiro.test(fit.man$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.man$residuals
## W = 0.99127, p-value = 0.0007895
```

- Noted that the p-value is very low at 0.0007895 which indicates that the residuals for the fit.man are not normally distributed.

Part 2c - ARIMA

Yards Test

```
yds.vec <- overallAvgNFL.df$AvgYdsSea
summary(yds.vec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3207    3392    3632    3596    3793    3901
```

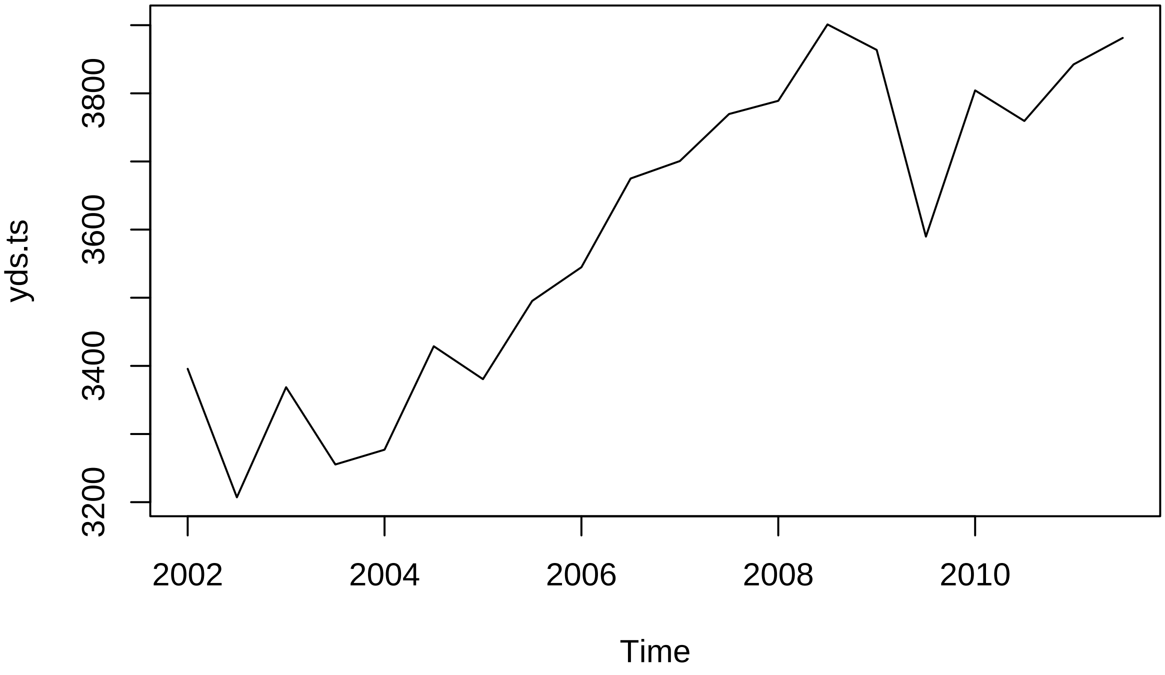
```
na.approx(yds.vec)
```

```
##  [1] 3395.7 3207.1 3368.7 3255.3 3277.0 3428.8 3380.5 3495.3 3544.8 3675.0
## [11] 3700.6 3769.6 3789.0 3901.1 3863.7 3589.7 3804.3 3759.4 3842.4 3881.3
```

```
summary(yds.vec)
```

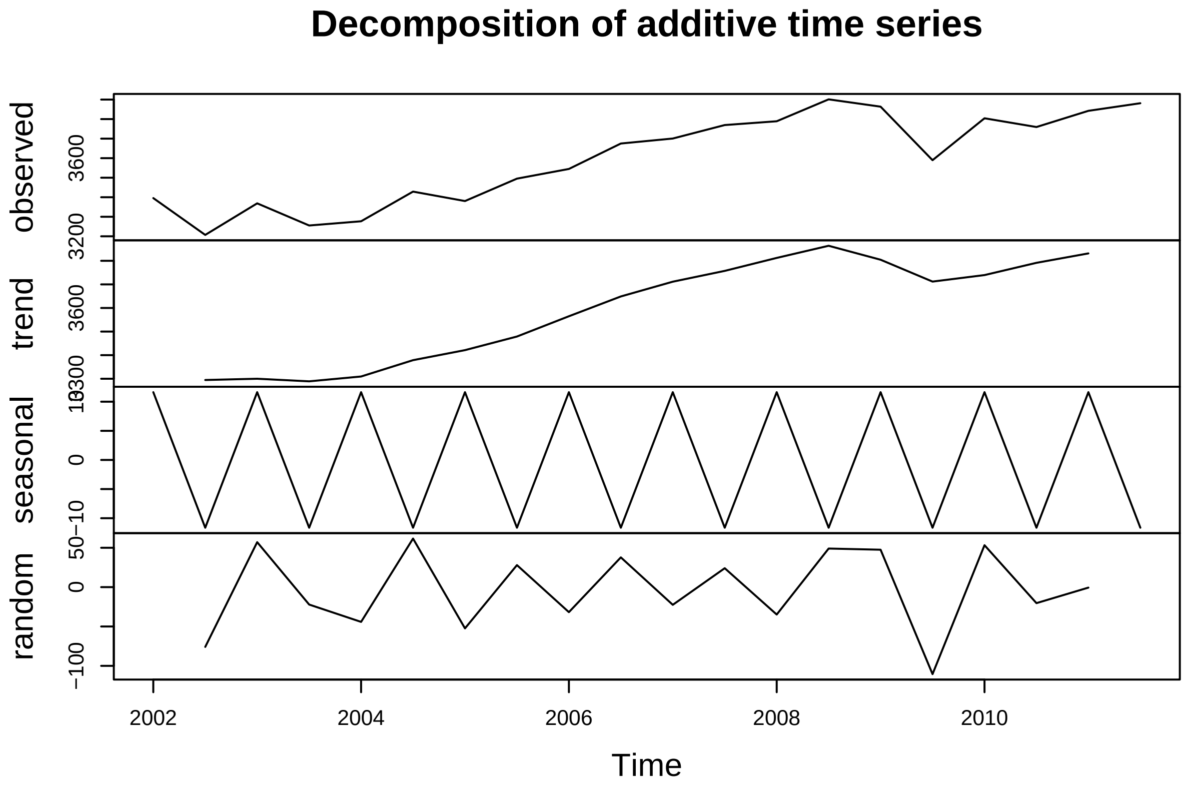
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3207    3392    3632    3596    3793    3901
```

```
yds.ts <- ts(yds.vec, frequency = 2, start = c(2002))
plot(yds.ts)
```

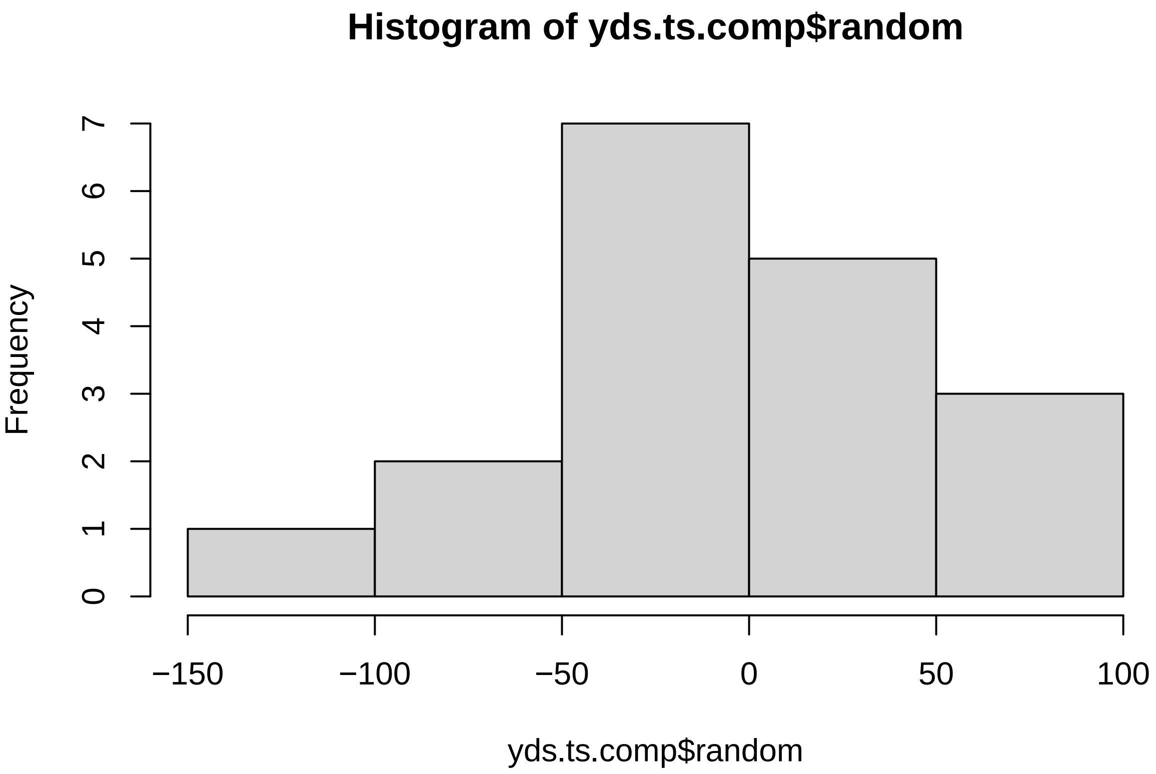


- Shows the time series graph for Year against the average passing yards per season in a time series format.

```
yds.ts.comp <- decompose(yds.ts)
plot(yds.ts.comp)
```

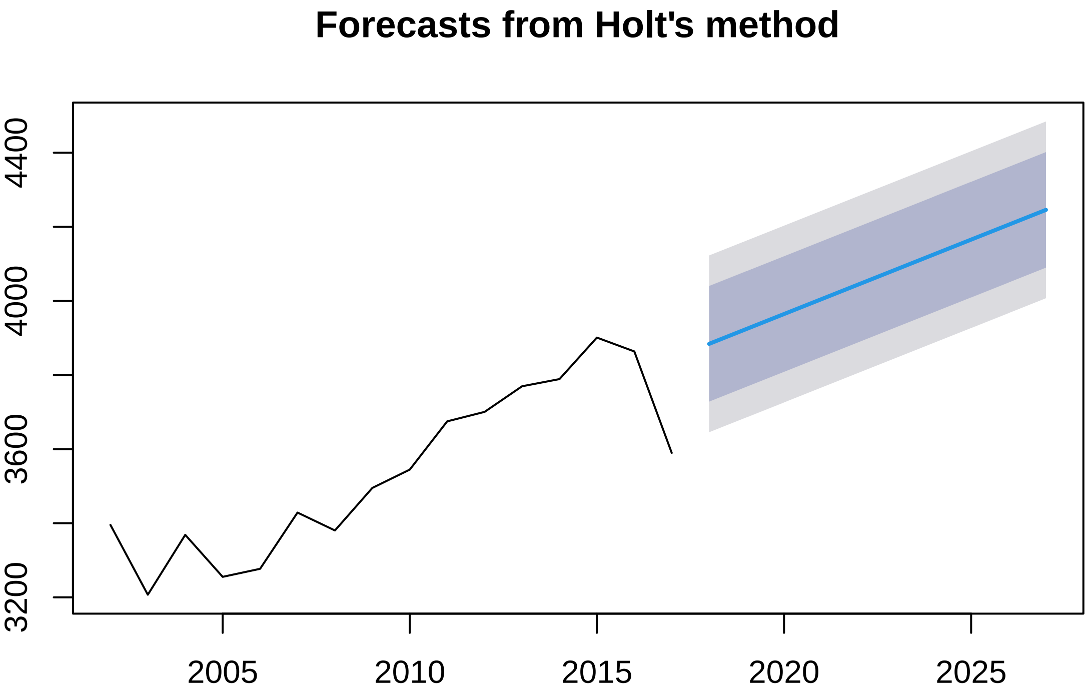


```
hist(yds.ts.comp$random)
```



- Histogram shows that the results are about normally distributed

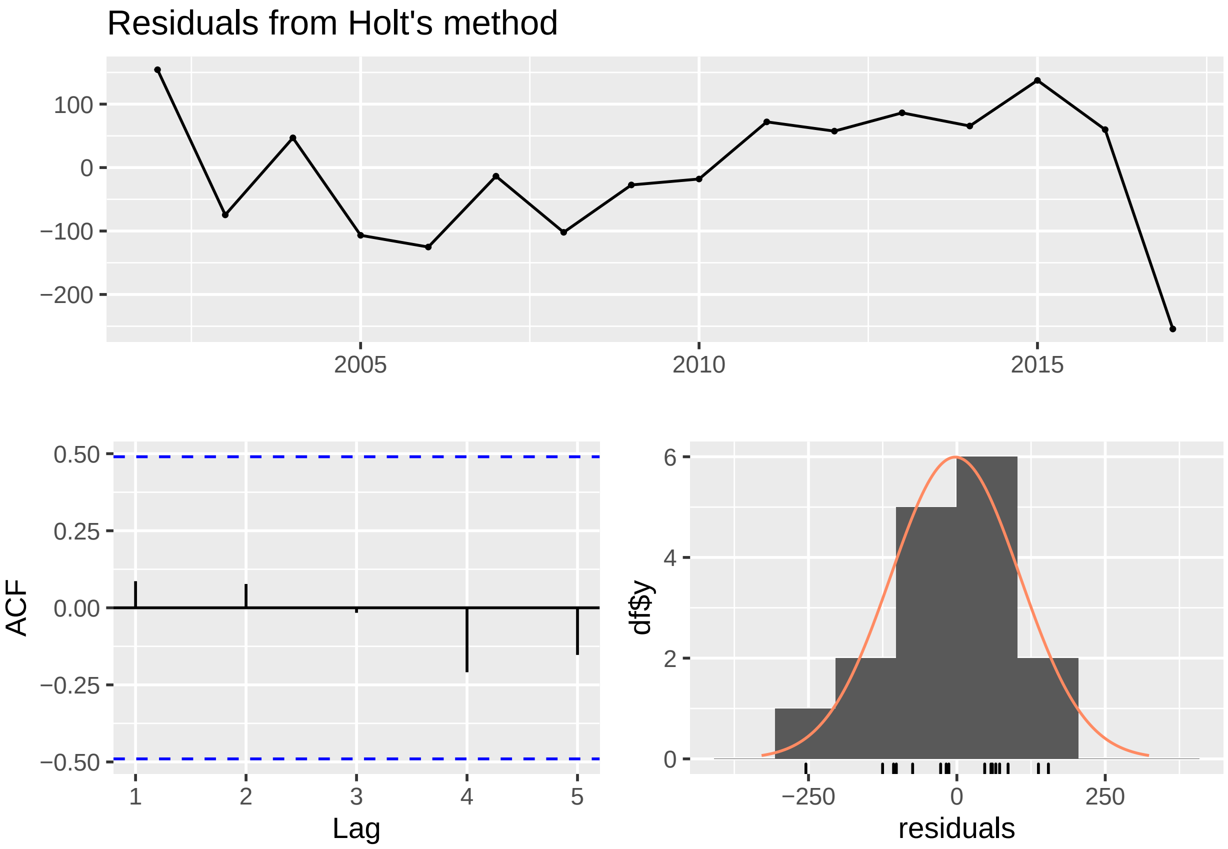
```
modelYds.hw <- holt(y = trainYds.data, h = 10)
plot(modelYds.hw)
```



- In this case we took our graph of yards thrown per season on average and used the holt() method from R in order to predict how the passing offense NFL would change over the next 10 years. We have a linear line that helps to show how the passing offense will grow over the next 10 years.

- We originally planned to use the holt-winters method however with our data being annual, it didn't allow for us to use holt-winters as we were not able to convert the data to be quarterly, monthly, weekly or daily without messing up the data given how the NFL season only runs for a certain time during the year.

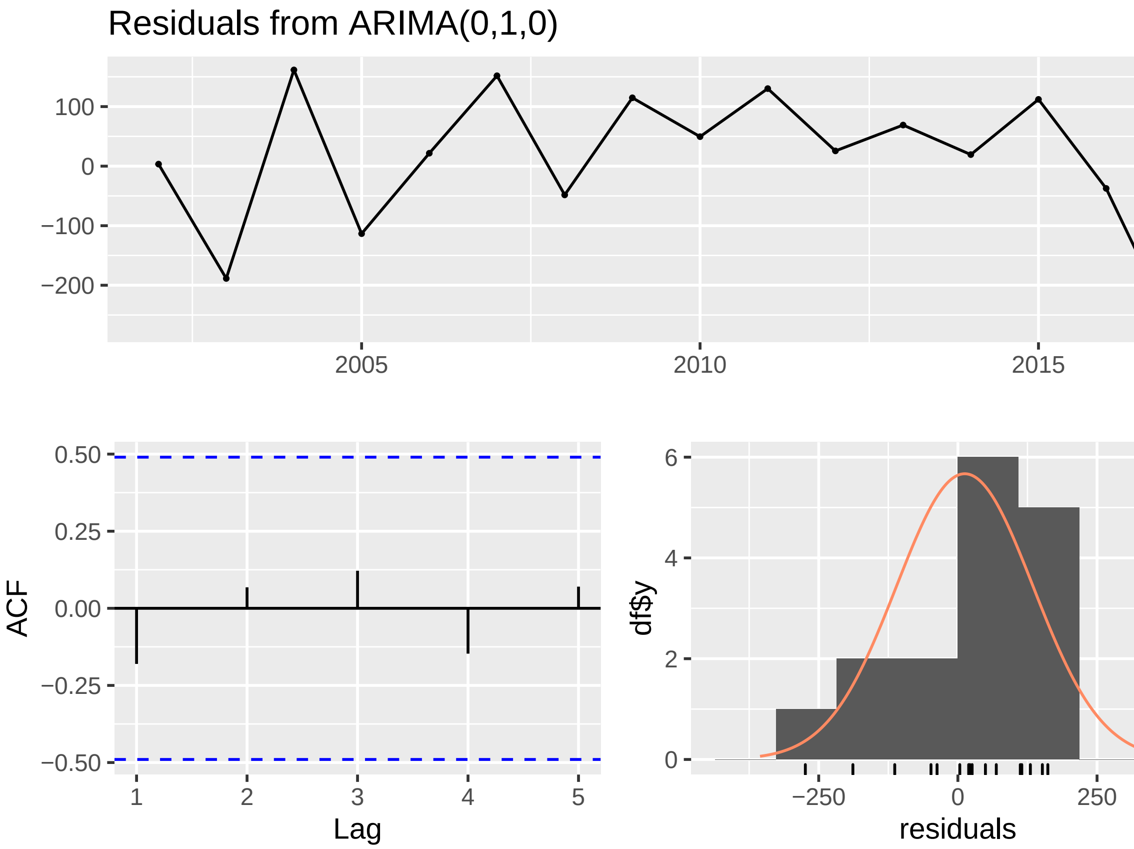
```
checkresiduals(modelYds.hw)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from Holt's method
## Q* = 6.1237, df = 3, p-value = 0.1057
##
## Model df: 4.   Total lags used: 7
```

- Running a test to check our residuals we can see that the data has a zero mean and has a roughly normal distribution. Our p-value was greater than 0.05 which is consistent with the data we had in the stepwise regression and the linear regression modeling for the yards.

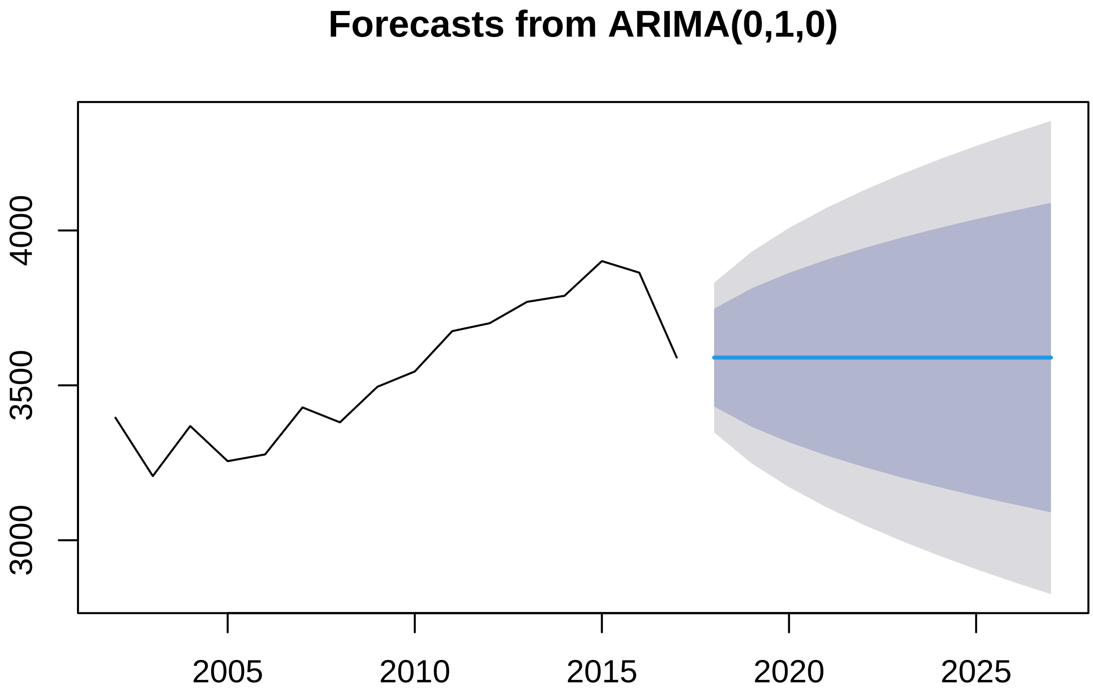
```

arimaYds.fit <- auto.arima(trainYds.data)
checkresiduals(arimaYds.fit)
```



Arima Testing - Yards

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,0)
## Q* = 1.0477, df = 3, p-value = 0.7897
##
## Model df: 0.   Total lags used: 3
```

```
arimaYds.pred <- forecast(arimaYds.fit, h = 10)
plot(arimaYds.pred)
```



Attempts Test

Here we created an ARIMA model that tested the Attempts and possibly saw how that would change in the future.

```
att.vec <- overallAvgNFL.df$AvgAtt
summary(att.vec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   511.1   528.3   545.3   544.4   559.8   584.8
```

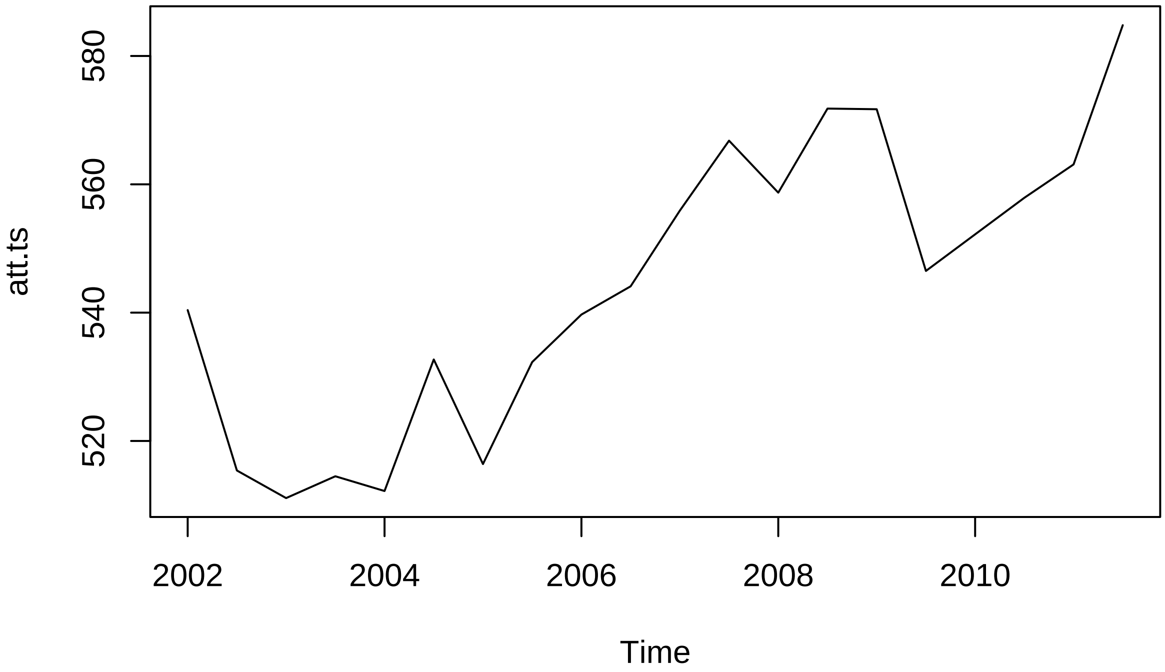
```
na.approx(att.vec)
```

```
##  [1] 540.4 515.4 511.1 514.5 512.2 532.7 516.4 532.3 539.7 544.1 555.9 566.8
## [13] 558.7 571.8 571.7 546.5 552.2 557.9 563.1 584.8
```

```
summary(att.vec)
```

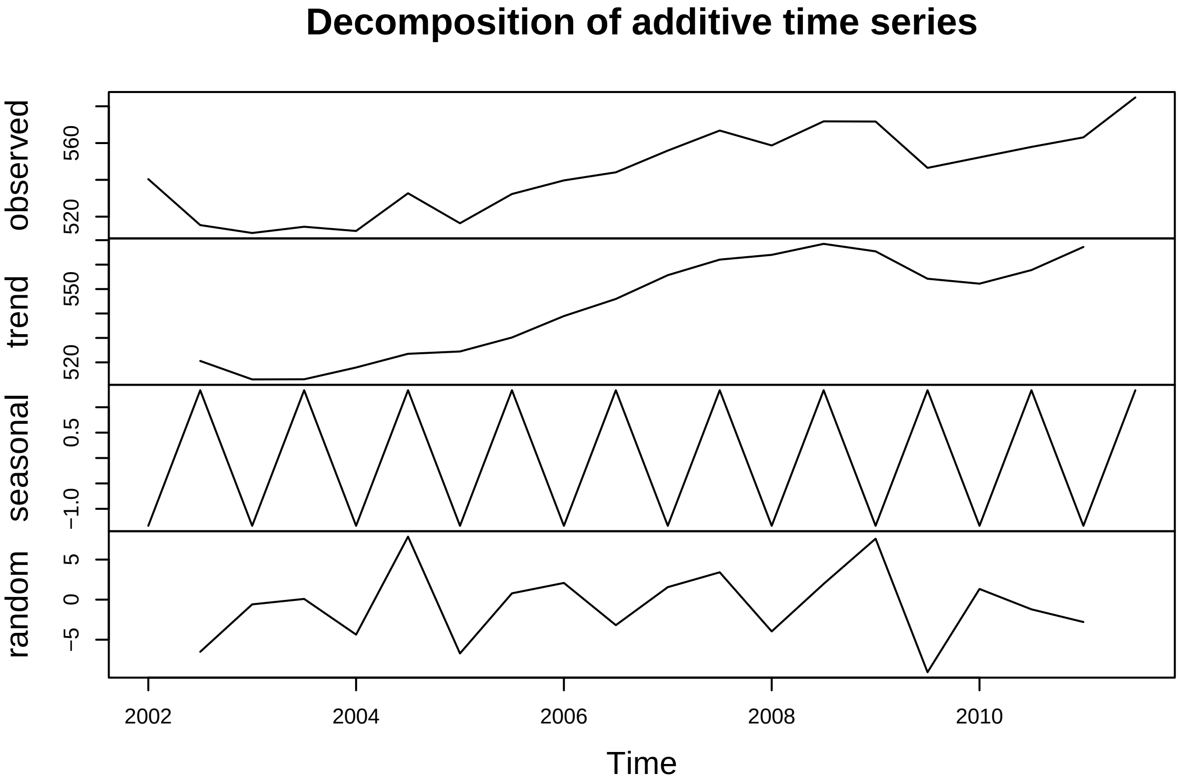
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   511.1   528.3   545.3   544.4   559.8   584.8
```

```
att.ts <- ts(att.vec, frequency = 2, start = c(2002))
plot(att.ts)
```

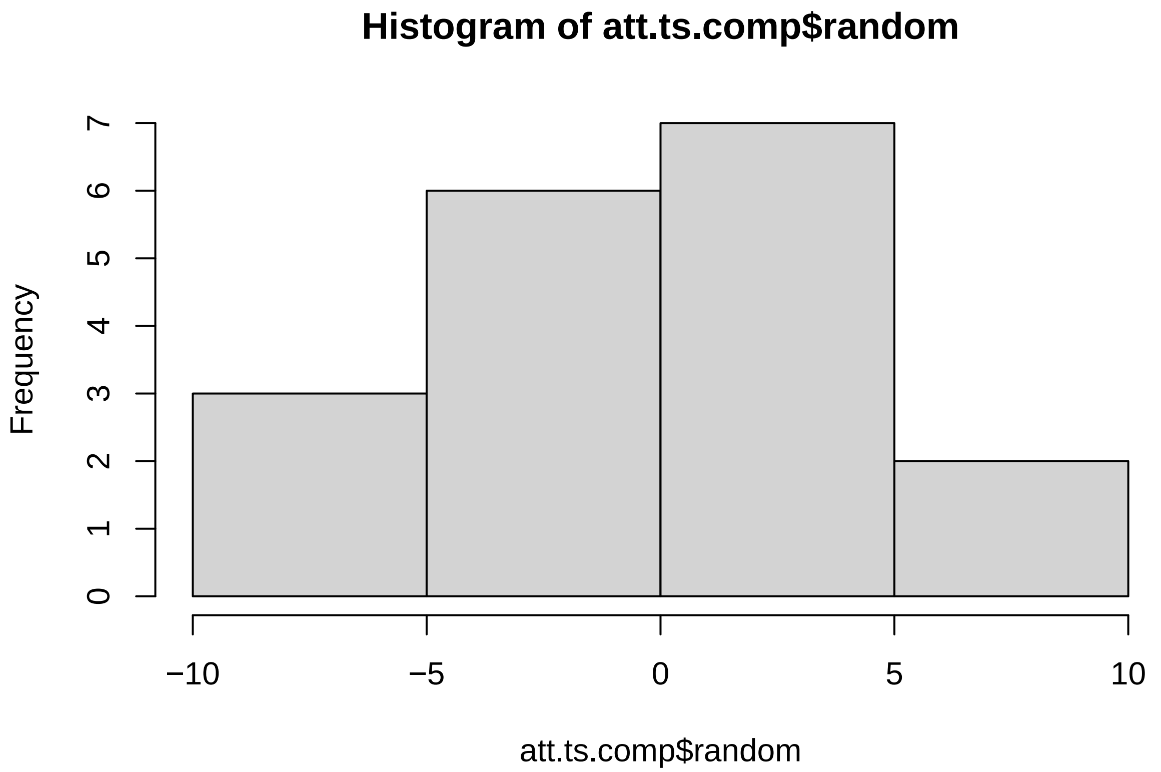


- This graph is just the attempts per year against the the year so that we can get a sense of how the attempts changed over the past 20 years.

```
att.ts.comp <- decompose(att.ts)
plot(att.ts.comp)
```



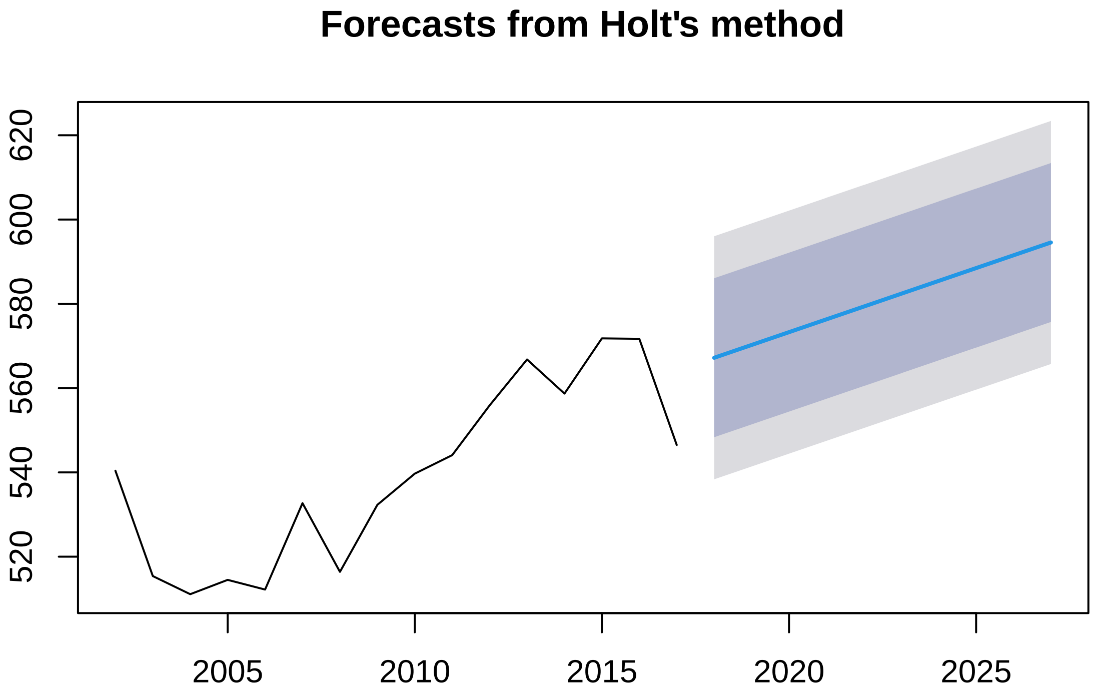
```r
hist(att.ts.comp$random)
```



- The histogram of the decomposition shows that the results about normally distributed

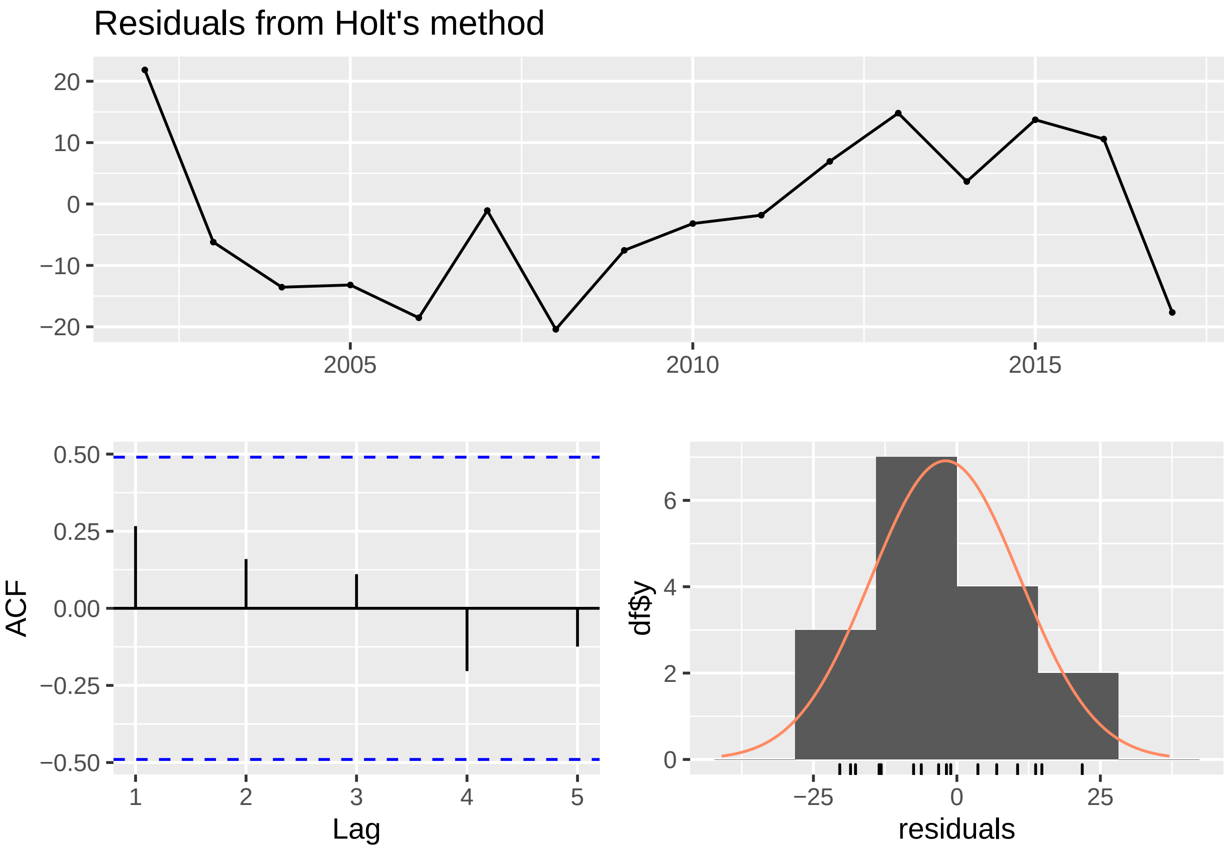
```r
trainAtt.data <- ts(att.ts, frequency = 1, start = c(2002), end = c(2017))
testAtt.data <- ts(att.ts, frequency = 1, start = c(2018), end = c(2021))
```

```r
modelAtt.hw <- holt(y = trainAtt.data, h = 10)
plot(modelAtt.hw)
```



- This graph compares the attempts thrown over the years and uses that data to further predict the attempts and how they'll be in the next 10 years. In doing so we have a linear line that helps us predict how the NFL will change in terms of passing yards in the next 10 years.
- We originally planned to use the holt-winters method however with our data being annual, it didn't allow for us to use holt-winters as we were not able to convert the data to be quarterly, monthly, weekly or daily without messing up the data given how the NFL season only runs for a certain time during the year.

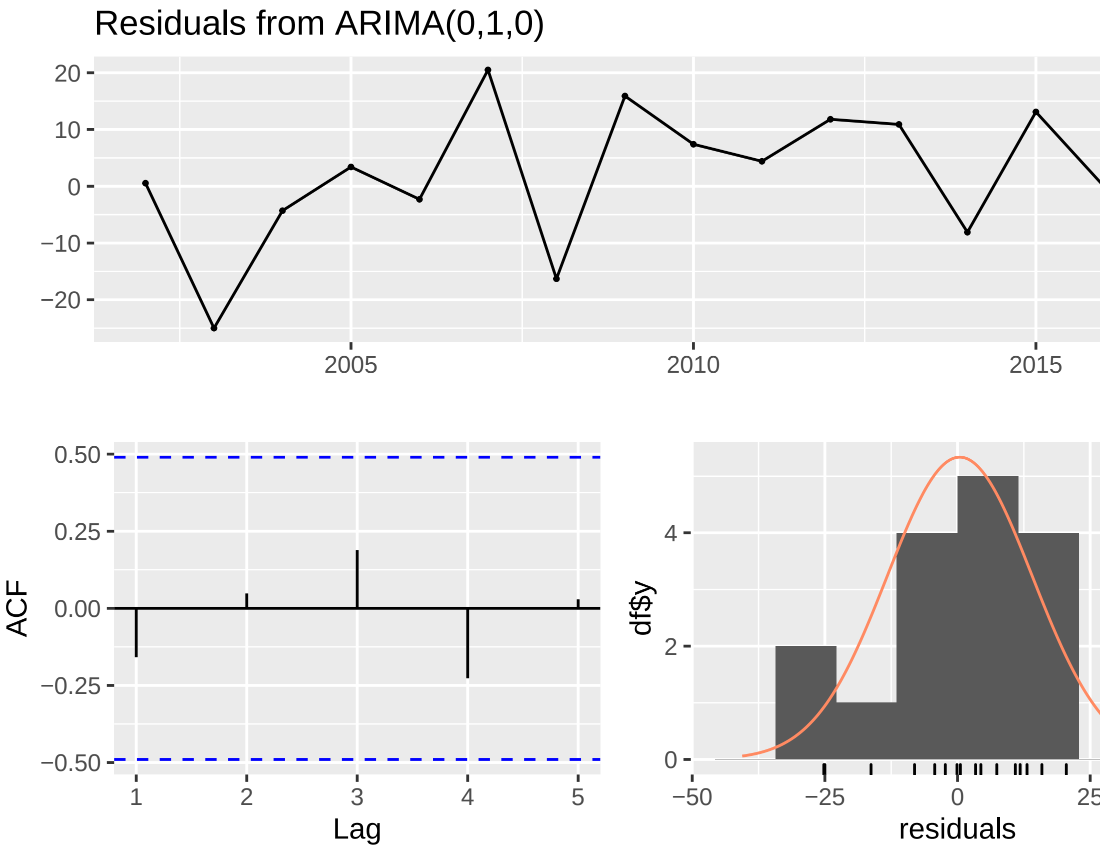
```
checkresiduals(modelAtt.hw)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from Holt's method
## Q* = 9.5746, df = 3, p-value = 0.02255
##
## Model df: 4.   Total lags used: 7
```

- Running a test to check our residuals we can see that the data has a zero mean and has a roughly normal distribution. The p-value we had here was less than 0.05 which indicates that it's fitting our threshold and that the data is significant

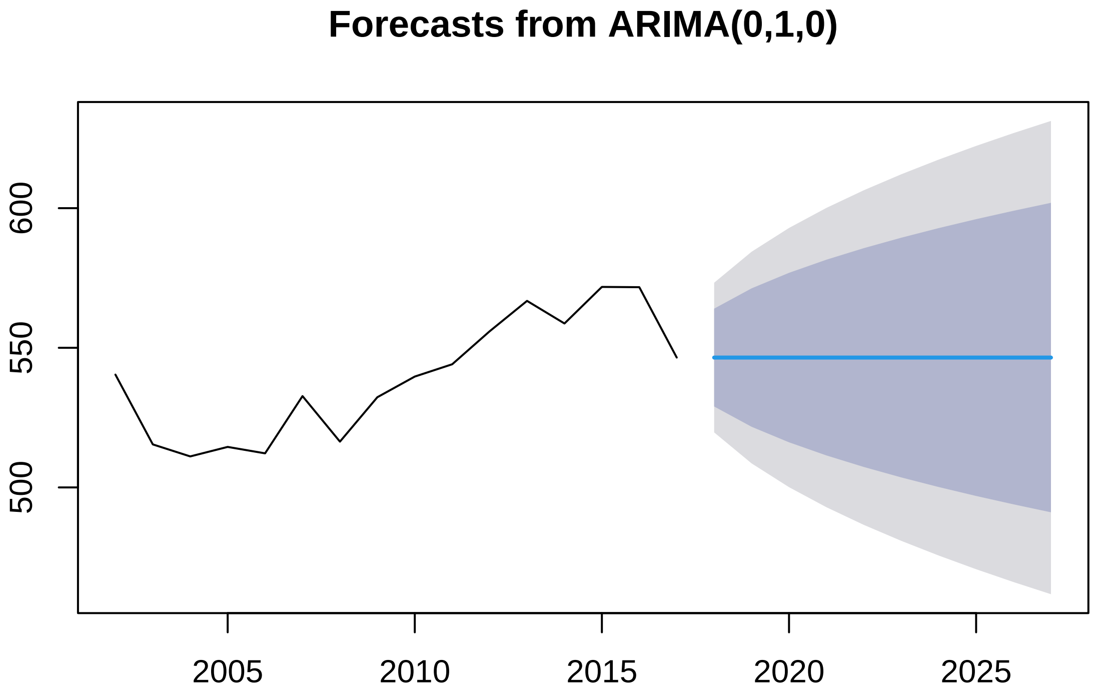
```
arimaAtt.fit <- auto.arima(trainAtt.data)
checkresiduals(arimaAtt.fit)
```



arima testing - Attempts

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,0)
## Q* = 1.3216, df = 3, p-value = 0.724
##
## Model df: 0.   Total lags used: 3
```

```
arimaAtt.pred <- forecast(arimaAtt.fit, h = 10)
plot(arimaAtt.pred)
```



```
accuracy(modelYds.hw, testYds.data)
```

Accuracy Test

```
##                         ME     RMSE       MAE         MPE      MAPE      MASE
## Training set   -2.628491 105.3826  87.59581  -0.1672567  2.484481 0.8659135
## Test set     -637.777645 646.1166 637.77764 -19.3656031 19.365603 6.3046426
##                     ACF1 Theil's U
## Training set  0.08640178        NA
## Test set     -0.46423092  4.363167
```

```
accuracy(arimaYds.pred, testYds.data)
```

```
##                        ME     RMSE       MAE        MPE     MAPE     MASE
## Training set   12.33723 119.3257  95.04973  0.2952487 2.719495 0.939598
## Test set     -283.00000 293.5486 283.00000 -8.6189404 8.618940 2.797548
##                    ACF1 Theil's U
## Training set -0.1802169        NA
## Test set     -0.7492214  2.009951
```

```
accuracy(modelAtt.hw, testAtt.data)
```

```
##                        ME     RMSE      MAE        MPE     MAPE      MASE
## Training set  -1.974793 12.74195 10.92019 -0.4344939 2.035785 0.9709716
## Test set     -51.410233 53.41783 51.41023 -9.9453045 9.945304 4.5711529
##                    ACF1 Theil's U
## Training set 0.26646528        NA
## Test set     0.07680067  4.185328
```

```
accuracy(arimaAtt.pred, testAtt.data)
```

```
##                        ME     RMSE      MAE         MPE     MAPE      MASE
## Training set   0.415025 13.23778 10.57752  0.04605554 1.968041 0.9405031
## Test set     -26.150000 28.64254 26.15000 -5.07720249 5.077202 2.3251334
##                      ACF1 Theil's U
## Training set -0.158694049        NA
## Test set      0.001194421  2.302481
```