Nikhil Roy

Final Project

require(moments)

## Loading required package: moments

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

Reading In And Cleaning The Data:

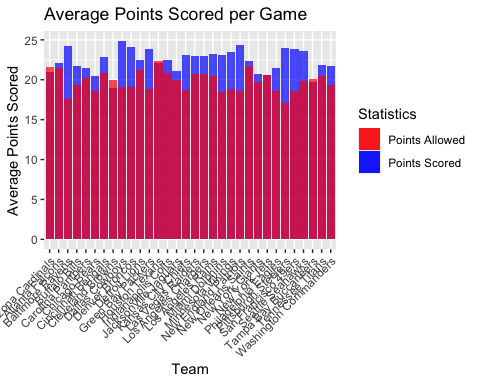
## Rows: 12,831  
## Columns: 12  
## $ season <int> 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970,…  
## $ month <int> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,…  
## $ day <int> 18, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 21, 27, 27,…  
## $ year <int> 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970,…  
## $ location <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…  
## $ game\_type <chr> "r", "r", "r", "r", "r", "r", "r", "r", "r", "r", "r", "r",…  
## $ home\_team <chr> "Los Angeles Rams", "New York Giants", "Los Angeles Charger…  
## $ away\_team <chr> "Arizona Cardinals", "Chicago Bears", "Indianapolis Colts",…  
## $ home\_score <int> 34, 16, 14, 31, 26, 7, 10, 0, 27, 3, 27, 7, 31, 20, 27, 34,…  
## $ away\_score <int> 13, 24, 16, 21, 17, 17, 25, 40, 10, 14, 14, 19, 21, 16, 27,…  
## $ tie <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…  
## $ score\_diff <int> 21, -8, -2, 10, 9, -10, -15, -40, 17, -11, 13, -12, 10, 4, …

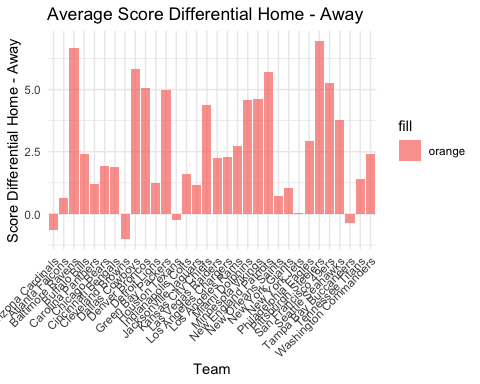
## Rows: 32  
## Columns: 5  
## $ team <chr> "Buffalo Bills", "Miami Dolphins", "New England Patri…  
## $ conference <chr> "afc", "afc", "afc", "afc", "afc", "afc", "afc", "afc…  
## $ division <chr> "afc\_east", "afc\_east", "afc\_east", "afc\_east", "afc\_…  
## $ rating <int> 1500, 1300, 1500, 1500, 1300, 1300, 1500, 1300, 1300,…  
## $ inaugural\_season <int> 1960, 1966, 1960, 1960, 1968, 1950, 1933, 1996, 2002,…

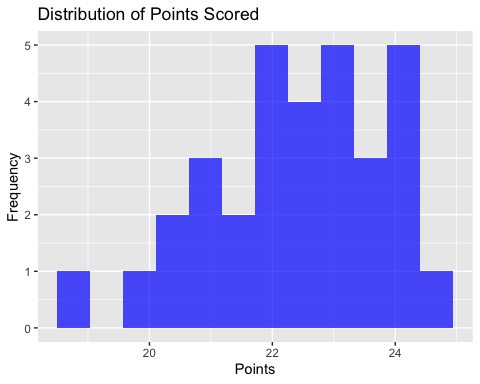
## Season Overview and Format:

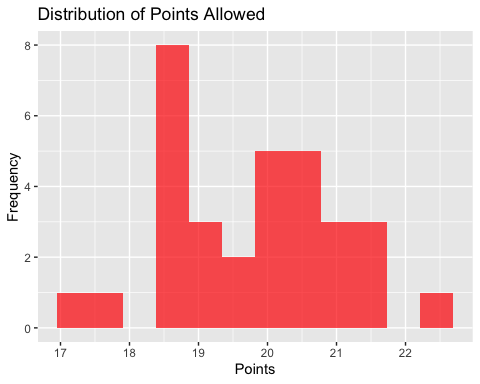
In the NFL, there are 32 teams, 8 divisions (AFC/NFC North, AFC/NFC East, AFC/NFC West, AFC/NFC South). Each of the 8 divisions contain 4 teams -> making 32 teams total. There is the regular season, playoffs, and superbowl. I decided to choose the 1970-2022 NFL season because it gives a broad and well enough data to organize the best teams and perform the many calculations. This also gets rid of any rules that influenced past game and removes the teams that are no longer in the leaugue currently.

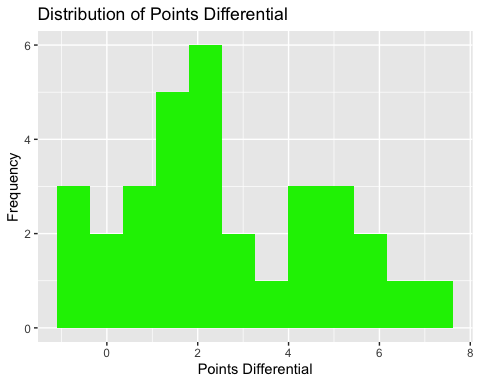
## Statistical Summary of Relevant Statistics:











## # A tibble: 1 × 6  
## mean\_points\_scored mean\_points\_allowed sd\_points\_scored sd\_points\_allowed  
## <dbl> <dbl> <dbl> <dbl>  
## 1 22.4 19.8 1.43 1.25  
## # ℹ 2 more variables: mean\_points\_diff <dbl>, sd\_points\_diff <dbl>

## Discussion of Relevant Statistics:

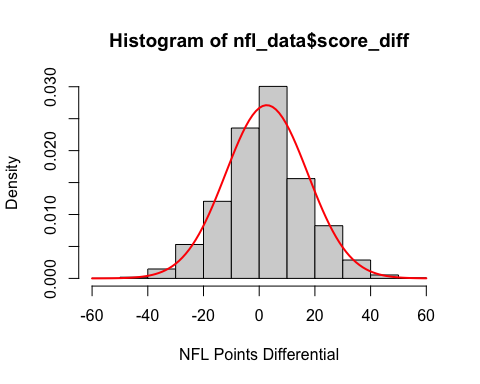
I decided to mutate and add some columns where I preformed some basic calculations in statsitics. For example, I created and summarized 3 columns: average points scored, average points allowed, and average point differential across each of the 32 teams. I decided to do this because it makes it cleaner to view each of the 32 teams and their statistics, rather than looking at 10 thousand + rows for each individual game. Thus, by creating summary caluclations like average (mean), it helps organize average points scored, allowed, and point differential for each team. Likewise, I created a new small table that incorperated the overall summary statistics for all 32 teams into 1 row. I have calculated the average and standard deviation of all 32 teams based on their points scored, points allowd, and point differentiable: both the average and standard deviation are shown and calclulated in the table above.

Looking at the distributions, the points scored looks slighly left skewed with an average of roughly 22 points. Looking at the distribution of points allowed, the plot contains values smaller than points scored indicating that teams score more points than they give up/allowed (as supported by the double stakced barchart called Average Points Per Game), and finally, the point differential is mostly positive - with more teams that are home tending to score more points since the point differential is positive (home score - away score).

The summary statsitics at the end shows a positive gap for points scored and points allowed across the averages in the 32 NFL teams, indicating that the points scored is greater on average than points allowed. This is heavily supported by the statistics showing a postive average point differential.

Likewise, the summary statistics in the table above indicate the average points scored is more than the average points allowed, making logical sense to the reason that point differential average is positive. Similarly, the standard deviation is greater for points scored than allowed, but pretty similar to each other.

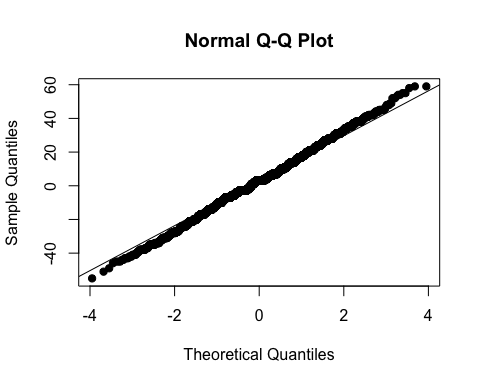
## Creation of Scoring Distribution:

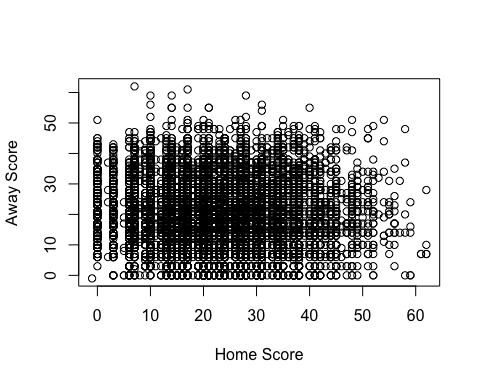


## [1] 2.695035

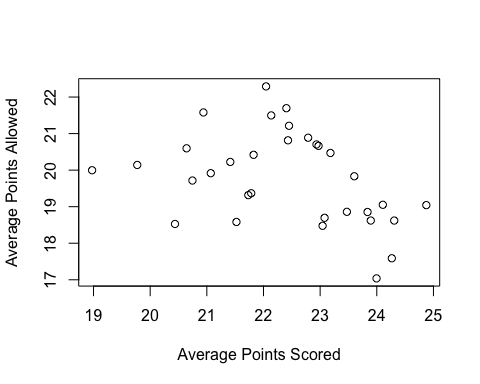
## [1] 14.71278

## [1] 0.03094876





## [1] -0.02492215



## [1] -0.3830689

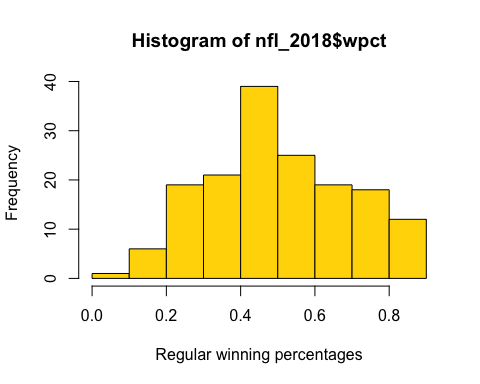
## Discussion of Scoring Distribution:

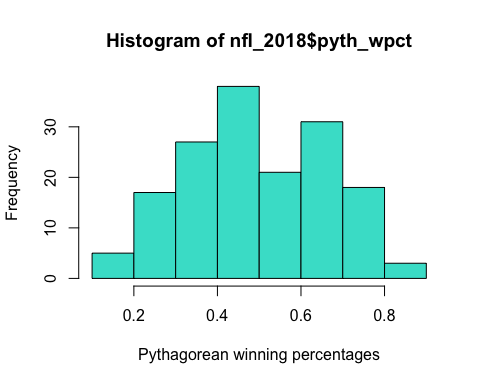
Looking at the section, the first plot displays the distribution of point differential across all games in the NFL From 1970 - 2022 season. As shown, it appears to be approximately a normal distribution, centered just over 1. I calculated the average and it was roughly 2.7. To help confirm and validate the normal shape of the data, I plotted a red curve over the bars and calculated the skewness, which was very close to 0 - so we know there is no skewness of any sort. Likewise, it makes sense for the point differentiable to be closer/occur most frequently as around 2-3 points than a large number because a lot of NFL games end in a close fashion. To conclude working with the normality of point differential, I plotted a normality plot on that variable. For the most part, the normality plot appears to be linear with the data points. It really starts to deviate from the diagonal line towards the beginning and end of the data points- and this makes sense since point differential is symmetric (ex. one team wins by 15 points so the other team loses by 15 points, making it + 15 for the winning teams and -15 for the losing team). Thus, we are safe to say we can assume normality of point differential across the NFL seasons 1970 - 2022.

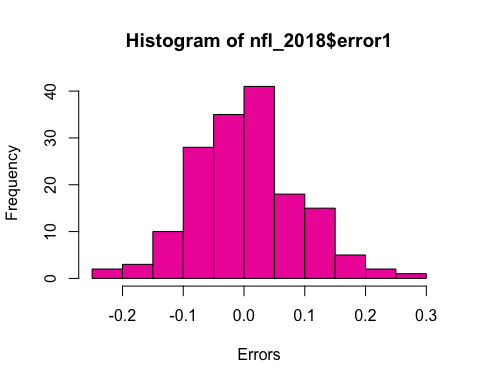
Furthermore, I plotted a scatterplot of home and away scores, and it visually shows no correlation. To validate this, I quickly calculated and displayed the correlation between these 2 variable below the scatterplot, and the value was very close to being 0 - indicating no correlation between home and away scores. However, one interesting thing I noticed when concluding this section was that there was a slightly moderate negative correlation (-0,383) between the averages of home and away scores, which is not what I expected, I believe this is because of the earlier summary sections that I calculated and visualized before, where teams typically scored more points than they allowed causing positive point differentials. Thus, this aligns with the negative moderate correlation where the higher points scored typically result in fewer points allowed for a particular game across 1970 - 2022.

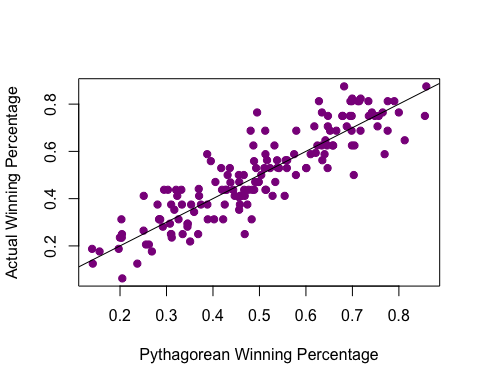
## Calculation of Pythagorean Win Percentages - Only using 2018 Season:

##   
## Call:  
## lm(formula = log(wins/losses) ~ 0 + log(scored/allowed), data = nfl\_2018)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.3520 -0.2314 0.0067 0.2666 1.1988   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## log(scored/allowed) 2.8554 0.1216 23.49 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4085 on 159 degrees of freedom  
## Multiple R-squared: 0.7763, Adjusted R-squared: 0.7749   
## F-statistic: 551.7 on 1 and 159 DF, p-value: < 2.2e-16









## Correlation between Pythagorean Win Percentage and Actual Win Percentage: 0.8929543

## Number of teams that overachieved: 82

## Number of teams that underachieved: 78

## 2022 Minnesota Vikings overachieved the most with a residual/error value of: 0.2697388

## 2020 Atlanta Falcons underachieved the most with a residual/error value of: -0.2183103

## Discussion of Pythagorean Win Percentages:

The histogram of the actual winning percentage is clearly normally distributed, centered at roughly 0.45 - 0.5. However, the Pythagorean winning percentage is a bit less normally distributed. If any skewness, it could possibly be a very slight right skew, centered at roughly 0.45 - 0.5 as well. Finally, The errors are right-skewed but centered at roughly 0, indicating very close residuals or differences between these 2 winning percentages. After plotting those 3 histograms, I created a scatterplot where I plotted the Pythagorean winning percentage with the actual winning percentage. The black line is linear with a constant slope of 1, so values above the line have a higher actual winning percentage than the Pythagorean winning percentage - and vice versa where values below the line have a higher Pythagorean winning percentage than the actual winning percentage. Just looking at the scatterplot visually, it appears that the data is along or very close to above and below the line. Finally, the correlation calculated was around 89.3% so there is a strong positive linear relationship between the 2 variables.

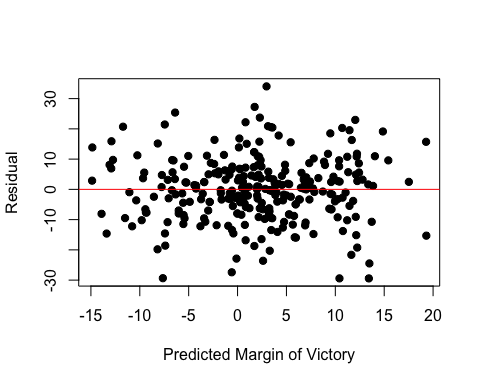
As for the teams that overachieved and underachieved, I looked at all the teams from the 2018 - 2022 season and calculated their Pythagorean win percentage, actual win percentage (generated in the dataset given), and their errors, which is the residual or difference between actual winning percentage and Pythagorean winning percentage. Thus, teams with a positive residual or error overachieved (actual > predicted Pythagorean), and teams with negative errors underachieved ((actual < predicted Pythagorean). Thus, the output displays that 82 teams overachieved and 78 teams underachieved. The team that overachieved the most was the 2022 Minnesota Vikings where they had 13 wins and only 4 losses making them have a 76.5% actual winning percentage, and the Pythagorean predicted they would only produce roughly 49.5% winning percentage. Thus, this team overachieved the most as they exceed the expectations the post, as calculated in the errors column. Out of their 13 wins, they won all 11 of their close games they were involved in that season, making them have more wins by winning close games which increased their actual winning percentage.

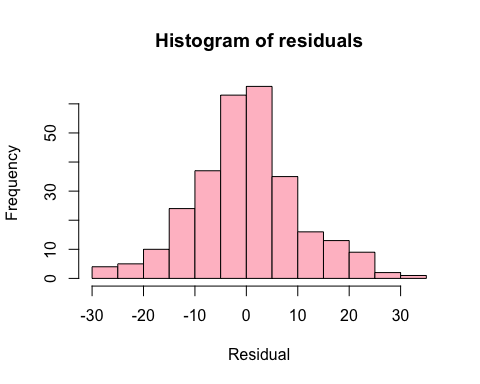
The 2020 Atlanta Falcons underachieved the most, with the largest negative error. The Pythagorean predicted a winning percentage of roughly 46.83%, but the 2020 Falcons actually achieved an actual winning percentage of 25%. Thus, this makes them the most underachieved team from the 2018 - 2022 seasons - because they have the most extreme negative error/residual amongst all teams in the subset of the NFL seasons.

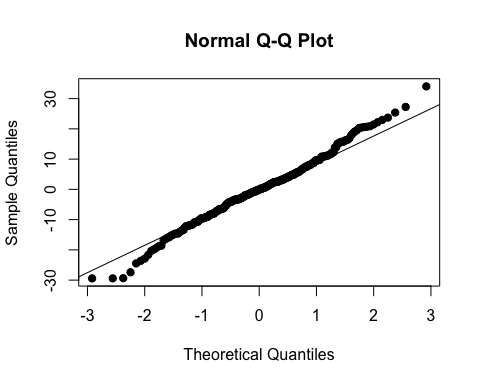
## Calculation of Bradley-Terry Ratings for Year 2022:

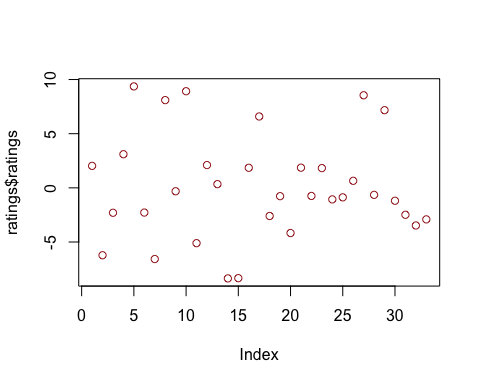
## ratings  
## Location 2.0272986  
## `Arizona Cardinals` -6.2221830  
## `Atlanta Falcons` -2.3008641  
## `Baltimore Ravens` 3.1035995  
## `Buffalo Bills` 9.3653047  
## `Carolina Panthers` -2.2812495  
## `Chicago Bears` -6.5729178  
## `Cincinnati Bengals` 8.0998082  
## `Cleveland Browns` -0.3152489  
## `Dallas Cowboys` 8.9203197  
## `Denver Broncos` -5.1066153  
## `Detroit Lions` 2.1045051  
## `Green Bay Packers` 0.3423099  
## `Houston Texans` -8.3598166  
## `Indianapolis Colts` -8.3359010  
## `Jacksonville Jaguars` 1.8522412  
## `Kansas City Chiefs` 6.5888328  
## `Las Vegas Raiders` -2.5979291  
## `Los Angeles Chargers` -0.7672892  
## `Los Angeles Rams` -4.1707106  
## `Miami Dolphins` 1.8614058  
## `Minnesota Vikings` -0.7474054  
## `New England Patriots` 1.8230081  
## `New Orleans Saints` -1.0650772  
## `New York Giants` -0.8761404  
## `New York Jets` 0.6498740  
## `Philadelphia Eagles` 8.5470571  
## `Pittsburgh Steelers` -0.6479671  
## `San Francisco 49ers` 7.1694510  
## `Seattle Seahawks` -1.1911873  
## `Tampa Bay Buccaneers` -2.4846853  
## `Tennessee Titans` -3.4738070  
## `Washington Commanders` -2.9107223

##   
## Call:  
## lm(formula = MoV ~ 0 + ., data = nfl\_data\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.432 -6.531 -0.024 5.650 34.037   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## Location 2.0273 0.6638 3.054 0.002500 \*\*   
## `Arizona Cardinals` -3.3115 3.9453 -0.839 0.402073   
## `Atlanta Falcons` 0.6099 3.8582 0.158 0.874530   
## `Baltimore Ravens` 6.0143 3.9109 1.538 0.125336   
## `Buffalo Bills` 12.2760 3.8440 3.194 0.001583 \*\*   
## `Carolina Panthers` 0.6295 3.9467 0.159 0.873406   
## `Chicago Bears` -3.6622 3.7188 -0.985 0.325676   
## `Cincinnati Bengals` 11.0105 3.8227 2.880 0.004313 \*\*   
## `Cleveland Browns` 2.5955 3.8600 0.672 0.501947   
## `Dallas Cowboys` 11.8310 3.4953 3.385 0.000825 \*\*\*  
## `Denver Broncos` -2.1959 3.9388 -0.557 0.577680   
## `Detroit Lions` 5.0152 3.7292 1.345 0.179874   
## `Green Bay Packers` 3.2530 3.7288 0.872 0.383809   
## `Houston Texans` -5.4491 3.7295 -1.461 0.145235   
## `Indianapolis Colts` -5.4252 3.7351 -1.452 0.147608   
## `Jacksonville Jaguars` 4.7630 3.6607 1.301 0.194404   
## `Kansas City Chiefs` 9.4996 3.7747 2.517 0.012466 \*   
## `Las Vegas Raiders` 0.3128 3.9399 0.079 0.936784   
## `Los Angeles Chargers` 2.1434 3.8760 0.553 0.580750   
## `Los Angeles Rams` -1.2600 3.9457 -0.319 0.749737   
## `Miami Dolphins` 4.7721 3.8821 1.229 0.220113   
## `Minnesota Vikings` 2.1633 3.6739 0.589 0.556499   
## `New England Patriots` 4.7337 3.9274 1.205 0.229208   
## `New Orleans Saints` 1.8456 3.9478 0.468 0.640536   
## `New York Giants` 2.0346 3.4921 0.583 0.560662   
## `New York Jets` 3.5606 3.9277 0.907 0.365512   
## `Philadelphia Eagles` 11.4578 3.4545 3.317 0.001044 \*\*   
## `Pittsburgh Steelers` 2.2628 3.9498 0.573 0.567237   
## `San Francisco 49ers` 10.0802 3.6982 2.726 0.006864 \*\*   
## `Seattle Seahawks` 1.7195 3.8978 0.441 0.659478   
## `Tampa Bay Buccaneers` 0.4260 3.8795 0.110 0.912642   
## `Tennessee Titans` -0.5631 3.7353 -0.151 0.880294   
## `Washington Commanders` NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.97 on 253 degrees of freedom  
## Multiple R-squared: 0.3214, Adjusted R-squared: 0.2356   
## F-statistic: 3.745 on 32 and 253 DF, p-value: 1.725e-09









## Discussion of Bradley-Terry:

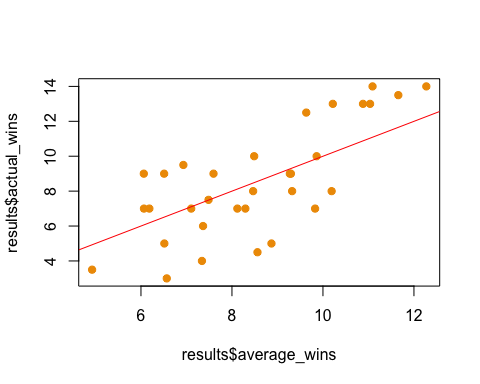
After running the Bradley Terry model and plotting the residuals and predicted margin of victory, there appears to be no correlation between the predicted margin of victory and residuals. This is supported by the histogram below the scatterplot where it is approximately normal and centered at roughly 0. The normality plot contains the points on the diagonal line pretty well in the center of the quintiles; however, it deviates towards the ends.

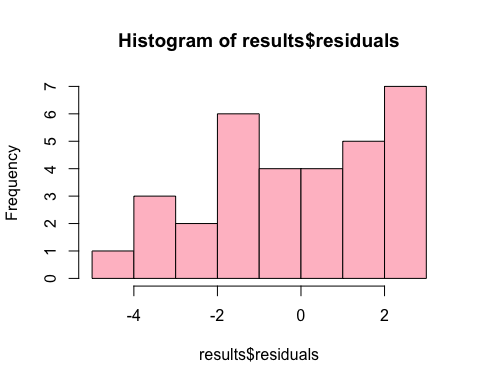
Now looking at the unadjusted Bradley Terry model (named model), Washington is labeled NA because it treated them as the baseline category. Thus, we can interpret the coefficients from the model as the number of points that team compared to Washington. For example, the best team, the Buffallo Bills were better than Washington the most because it has the highest positive coefficient value. However, the coefficients must be adjusted, which is what was done and stored in the “results” table. In this table, you can see that the best team was the Buffalo Bills, with the highest adjusted rating of 9.365. Similarly, the worst team with the largest adjusted value in the negative direction is the Houston Texans, having the lowest coefficient value of -8.36. There were a couple teams that were close to average (within + and - 1); these teams included: New York Jets, Green Bay Packers, Cleveland Browns, Pittsburgh Steelers, Minnesota Vikings, Los Angeles Chargers, and New York Giants.

## Simulated Regular Season Results:

## Simulate season

## [1] 1000  
## [1] 2000  
## [1] 3000  
## [1] 4000  
## [1] 5000  
## [1] 6000  
## [1] 7000  
## [1] 8000  
## [1] 9000  
## [1] 10000





Look Ahead of the table created that summarizes the actual win total, the simulated number of wins, and the simulated number of division titles

## Rows: 32  
## Columns: 7  
## $ team <chr> "Buffalo Bills", "Miami Dolphins", "New England Patrio…  
## $ starting\_elo <dbl> 1610.868, 1530.309, 1532.552, 1357.193, 1543.504, 1497…  
## $ ending\_elo <dbl> 1725.114, 1495.387, 1511.132, 1383.961, 1668.740, 1484…  
## $ actual\_wins <dbl> 13.5, 9.0, 8.0, 7.0, 12.5, 7.0, 9.0, 10.0, 3.5, 4.5, 9…  
## $ average\_wins <dbl> 11.6567, 9.2721, 8.4654, 6.0661, 9.6326, 8.1185, 7.594…  
## $ division\_titles <dbl> 0.7247, 0.1729, 0.0931, 0.0093, 0.3645, 0.1315, 0.0916…  
## $ residuals <dbl> 1.8433, -0.2721, -0.4654, 0.9339, 2.8674, -1.1185, 1.4…

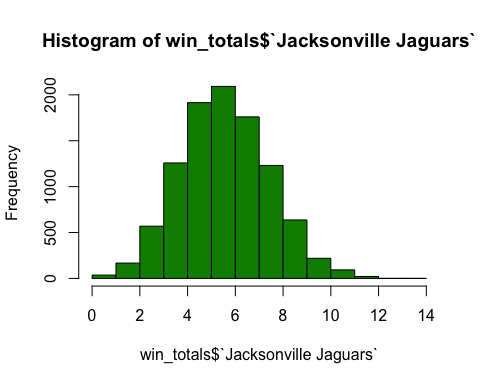
## Comparison of Simulation and Actual Regular Season:

For this section, I decided to simulate the 2022 NFL season amongst the 32 active teams. Plotting the simulated and actual wins, the scatterplot looks like a relatively positive linear correlation for the most part. There are a handful of points below the linear line; however, it mirrors for the handful of points above the line. The histogram of residuals is slightly left-skewed but close to approximately 0. Now, looking at the results of the Monte Carlo simulation, I stored 6 columns in the “results” table: 2022 season team, starting elo, ending elo, actual wins, average wins (simulated predicted wins), division titles, and residuals of actual wins subtracted by the simulated wins. Thus, the team with the greatest positive residual is the Jacksonville Jaguars, as they achieved more wins (9) than simulated/expected (6.06).

On the other side, the Indianapolis Colts had the smallest residual value - the greatest extreme value in the negative direction. This is very interesting because the Colts and Jaguars are in the same division - afc south. Thus, since the Jaguars had the greatest positive residual, they overachieved the most. However, their predicted or simulated wins was roughly 6. The Colts, on the other hand, had a higher predicted win total of around 8.56 wins. Thus, when we look at the divison\_winners table of the teams that won each of the 8 divisions, the model had the Colts winning the division because of their higher simulated average win total. However, they underperformed massively and the Jaguars overperformed, so the Jaguars truly won the division that year rather than the Colts since they had the most wins amongst the other 3 teams in the AFC South. Thus, I found this interesting how the team that underachieved (the Colts) was predicted to win the division, but in reality, the team that overachieved the most (The Jaguars) won that division in real life.

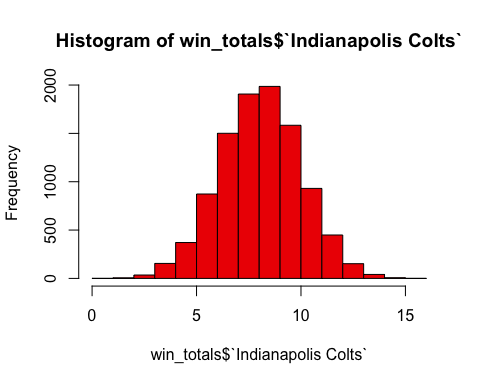
As for the division winners compared to who won the division, the model did pretty poorly in predicting who would win the division in the 2022 season. The teams that the model correctly predicted won their division were the Philadelphia Eagles, Kansas City Chiefs, and Minnesota Vikings. Thus, the Monte Carlo simulation correctly predicted 50% of the teams that truly won the division in the 2022 NFL season.

## Analysis of Overachieving Team - Jacksonville Jaguars:



As mentioned earlier, The Jacksonville Jaguars overachieved the most during the 2022 NFL season. Specifically, the simulation had them winning roughly 6.06 games (basically 6), but in reality, they won 9 games and won their respective division that year. When looking back at the Pythagorean winning percentage for the 2018 - 2022 seasons, the Jaguars in 2022 had a relatively small residual, where the Pythagorean winning percentage was slightly higher than their actual winning percentage. Likewise, when you sort the Pythagorean winning percentage from least to greatest, you can see the 2020 and 2021 Jaguars listed as the top 10 with the lowest simulated win percentage, so this definitely influenced their expected win percentage in the following season. Furthermore, the scatterplot displays the frequency of the win total the Monte Carlo simulation predicted the Jags based on 10,000 iterations. As illustrated, it is pretty normally distributed, centered at roughly 5 wins. 9 and more wins were no not as frequent as 3 - 7 wins were, making the Jaguars have such a high residual when comparing their simulated and actual winning percentage.

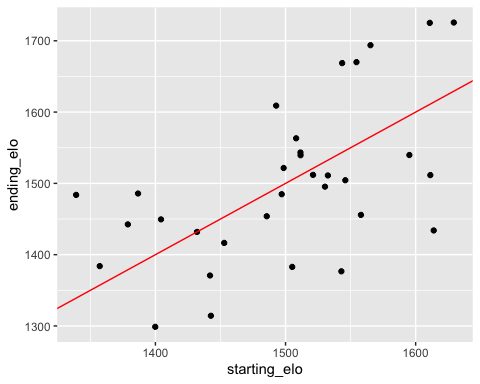
## Analysis of Underachieving Team - Indianapolis Colts:



The most underachieving team in the 2022 season based on the Monte Carlo simulation was the Indianapolis Colts, as mentioned earlier. Looking at the histogram plotted for the Colts, it is normally distributed, centered at roughly 7.5 - 8 wins, However, they actually got 4.5 wins that season, which very infrequent as compared to the other values. Thus, this resulted in a large negative residual as they deviated away from the simulated/expected wins. They were even expected to win the division that year, but failed to make the playoffs that season. Similar to the Jaguars, their residuals for the Pythagorean residuals is very minimal, meaning their Pythagorean win percentage was close to their actual winning percentage.

## Preseason and Postseason Elo Ratings For 2022 season:

## Warning in geom\_abline(a = 0, b = 1, col = "red"): Ignoring unknown parameters:  
## `a` and `b`



## $x  
## [1] "Preseason Elo Rating"  
##   
## $y  
## [1] "Postseason Elo Rating"  
##   
## $colour  
## [1] "Season"  
##   
## $title  
## [1] "Preseason vs. Postseason Elo Ratings"  
##   
## attr(,"class")  
## [1] "labels"

### Best 2 and Worst 2 Teams

## Best Team: Kansas City Chiefs

##   
## 2nd Best Team: Buffalo Bills

##   
## Worst Team: Houston Texans

##   
## 2nd Worst Team: Chicago Bears

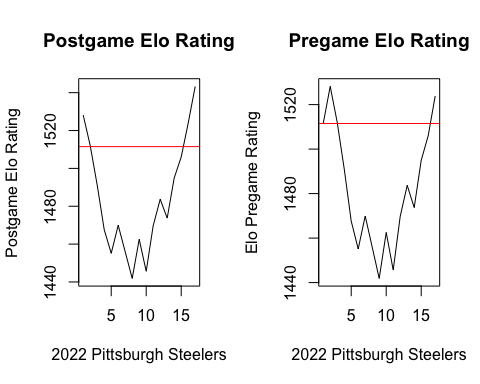
## team conference division rating wins rand  
## 2 Miami Dolphins afc afc\_east 1495.387 14 0.4116953  
## 7 Pittsburgh Steelers afc afc\_north 1543.314 12 0.5932269  
## 10 Indianapolis Colts afc afc\_south 1376.664 11 0.6285207  
## 14 Kansas City Chiefs afc afc\_west 1725.626 13 0.4108218  
## 19 Philadelphia Eagles nfc nfc\_east 1609.026 11 0.3140519  
## 24 Minnesota Vikings nfc nfc\_north 1563.244 11 0.2711165  
## 27 New Orleans Saints nfc nfc\_south 1504.401 10 0.6375820  
## 30 Los Angeles Rams nfc nfc\_west 1433.973 11 0.9352113

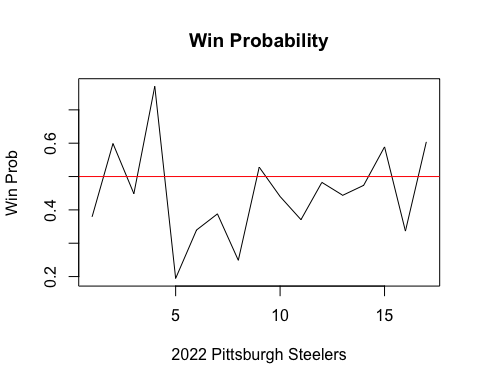
## Discussion of Elo Ratings:

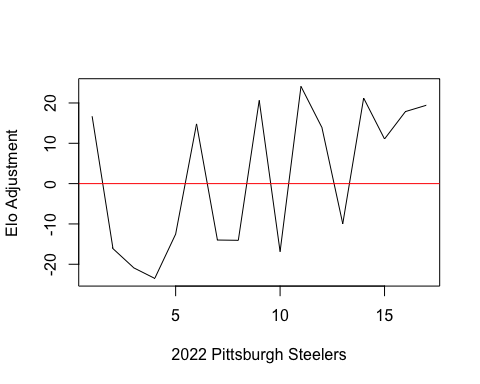
The scatterplot provided illustrates preseason elo against postseason elo during the 2022 NFL season across the 32 NFL teams. It looks pretty evenly distributed between the teams that finished with a lower elo than they started in the season, and vice versa. The two best teams during that season were the Kansas City Chiefs and the Buffalo Bills, as their ending elos were the highest. In contrast, the Houston Texans and the Chicago Bears were the 2 worst teams during the season, having the lowest elo rating at the end. This makes sense since both teams had the lowest wins during that season. The expected playoff teams were the Miami Dolphins, Pittsburgh Steelers, Indianapolis Colts, Kansas City Chiefs, Philadelphia Eagles, Minnesota Vikings, Los Angeles Rams, and New Orleans Saints.

The Jaguars and Eagles seemed to have improved and exceeded the expectations, which is shown in their larger residuals and improved ending elo, which they were predicted to have started with a sub 1500 elo. However, in reality, both teams won their respective divisions and made it to the playoffs, which was a big improvement throughout the season - as the Jaguars were predicted to lose a majority of their games and miss the playoffs, whereas the Eagles were not expected to be as dominant and achieve 14 wins during the 2022 season, which increased their elo ratings. On the other hand, the Colts and Rams had the biggest decline in their seasons because they were expected to win a majority of their games and win their respective divisions. However, they severely underperformed and neither team made the playoffs that year, which was a major decline as their ending elo ratings drastically declined compared to the start of the season. These two teams were meant to achieve a successful season but declined due to the many losses, dropping their elo rating and declining the chance of making the playoffs.

## Elo Ratings for One Team:







## Analysis of Elo Ratings for One Team:

I decided to analyze the 2022 Pittsburgh Steelers throughout the season. The side-by-side plots display the Steeler’s postgame and pre-game elo ratings throughout each of their 17 games during the season. Both of these graphs illustrate their starting elo entering the season (being roughly 1511.5 as denoted by the red horizontal line). Looking more closely at both of these graphs, it seemed for the majority of the season that the Pittsburgh Steelers fell below their starting elo rating as they had a rough beginning to start off their season. Specifically, in their first 5 games, the Steelers went 1-4, having 1 win and 4 losses. Thus, their pre-game and post-game elo rating dropped consistently to just under 1460. Then, it started to get inconsistent a bit during the middle of the season - as shown by the multiple spikes (ups and downs) during games 5 - 12. However, to close off their last 7 regular season games during the 2022 season, the Steelers won 6 games and only lost once. Thus, this consistently increased their elo ratings - to where their ending elo ratings for the season were surprisingly greater than their starting elo rating.

Finally, the 3rd plot illustrates the Steeler’s win probability for each of their 17 games. I thought this would be interesting to include because it sort of mimics what I talked earlier about their elo ratings during their inconsistent season. Specifically, they seemed to be the favorites during the beginning of the season, then they became big underdogs during the middle of their season, and their win probability started to increase and improve at the end of the season. Thus, the multiple visualizations were pretty interesting to see the changes in the elo ratings for the Pittsburgh Steelers throughout the 2022 NFL season because they started off poorly, creating their elo ratings to drop below their preseason elo rating; however, the last 7 games is where they improved dramatically, going 6 - 1, which brought up their final season elo rating passed the beginning - which looked nearly impossible to happen.

## Conclusions/Takeaways:

To wrap up my big data analysis project on the NFL seasons 1970 - 2022, there were a few interesting details worthwhile mentioning. To recap, I started by visually showing the average scoring distributions of each of the current 32 NFL teams - both points scored and points allowed throughout the span of 52 seasons/years. The visualization shows that teams generally, on average, score more points than they allow, while being able to inflate their offensive scoring and stabilizing a stronger defense. I supported these observations through several plots to help show the different aspects that relate to the scoring distribution. For instance, when we look at the scoring distributions of home and away teams, 28 out of the 32 NFL teams had a positive difference between home scores and away scores, suggesting the majority of teams score higher, on average, during their home games then they do on the road (at their opponents’ stadium). Likewise, a couple of histograms were plotted to show the distribution and trends of both points scored and allowed, respectively - where points scored were larger and had a bigger center than points allowed for a majority of teams. This makes logical sense since we later dived into elo ratings and simulations where it considered home field advantages when creating expected probabilities, elo ratings, and winning percentages.

After looking at the scoring distributions and summary statistics, I dived into the trend of this aspect. After plotting a histogram of the frequencies for score differentials in games (meaning the difference in the winning and losing team at the end of the game), I noticed that the trend could have been more systematic and more systematic. To further validate this, I created a normality plot where a good majority of points fit along the diagonal line, where it started to deviate a bit during both extremes. This can be solidified because huge blowouts do not occur that often - such as a 70 - 20 game that we had in the 2023 season where the Dolphins blew out the Broncos.

Moving on to the Pythagorean winning percentage, some things were to be expected such as the actual winning percentages, Pythagorean winning percentage, and errors between the two. The two winning percentages were roughly centered and had an average of close to 50%, which makes logical sense because a team either wins or loses, with ties being a rare occurrence. However, the errors were a bit skewed to the right, meaning errors generally had negative values close to -0.1 and -0.05. This suggests that the Pythagorean winning percentage is slightly overpredicting the actual winning percentage since the error is calculated by: actual wins - predicted wins. In other words, the Pythagorean winning percentage, on average, believes that teams have a higher winning percentage than they truly achieved. However, this error is very small and very close to 0, so I figured it was pretty accurate, abstractively. I decided to create a scatterplot against these two and validated the idea of both winning percentages being relatively close to each other. For the most part, the points follow along and are very close to the black line that has a slope of 1, meaning both actual and Pythagorean winning percentages were the same. Thus, this concluded that there was no worry about the very minimal negative error. Finally, when looking at the actual winning percentage against the predicted, the 2022 Minnesota Vikings overachieved the most - as they achieved a winning percentage of 76.5%, whereas the model only predicted them having a winning percentage of 49.5%. The reason behind this is that they had 11 close games, where they won all 11 games by 8 points or fewer. This is my hypothesis as to why they produced more wins than predicted by such a large residual than what the Pythagorean winning percentage predicted. In contrast, the 2020 Atlanta Falcons underachieved the most, where they did poorly during the season, only producing a winning percentage of 25%, compared to the Pytgaorean percentage predicting 46.83%.

Furthermore, I decided to only focus on the past season, 2022, to produce the Bradley Terry model, calculate the Elo ratings, and run the Monte Carlo simulation. Looking at the Bradley Terry model, specifically, the top ratings of the adjusted coefficient values make sense. For instance, the top values were the Buffalo Bills, Dallas Cowboys, Philadelphia Eagles, Kansas City Chiefs, Baltimore Ravens, San Fransisco 49ners, and Cincinnati Bengals, and all these teams made the playoffs that year. Comparing the results to the Elo ratings, the model predicted that the Kansas City Chiefs and Buffalo Bills were the top two teams that season, which is very much consistent with the Bradley Terry ratings. Similarly, if we look at the “results” data frame under the ending elo column, we can sort the ratings in descending order. Here, the ending elos line up very closely with the Bradley Terry model that identified the largest rating/coefficient. Thus, almost all the teams that had the largest values in the Bradley Terry model also have the largest ending elos with the exception of the Baltimore Ravens, where the Steelers and Packers ended with a higher elo rating. Looking closer, this should not be a big concern because the Packers barely had a higher ending elo rating by less than 0.4. Likewise, the Steelers upset many teams, such as Tom Brady and the Buccaneers, and ended the season going 6 - 1. Thus, this is my belief as to why they ended up with a slightly higher elo rating than the Ravens. However, the elo ratings aligned very well with the Bradley Terry ratings, having a majority of the teams going into the playoffs and winning their respective divisions.

Transitioning into the Monte Carlo simulation for the 2022 season, the top teams, in terms of division titles, were consistent with the Bradley Terry model and elo ratings calculated - such as the Chiefs and Bills being the top two teams that would win the most division titles. However, one very interesting thing I noticed was the Buccaneers being listed as number three for the most division titles, and there are two logical answers/hypotheses that I could come up with: the Buccaneers had Tom Brady and their division was relatively easier than the afc north, for instance. Looking at their residuals, which is the actual wins minus the predicted/simulated wins, the Buccaneers had a large negative residual (top 6), suggesting that the simulation predicted the team to have more wins than they truly had. To validate that, I checked these two values and noticed the model predicted, on average, of the Buccaneers having 10.19 wins, whereas they only had 8 wins that season. Thus, this makes sense to why the Buccaneers were in the top three for division titles. The two lowest teams the the end of season were the Texans and Bears. This is consistent with the Bradley Terry model as both teams were in the bottom 3, with negative rating values. Likewise, both teams had the least amount of wins during the season and the lowest elo ratings values. Overall, when looking at the 2022 NFL season, the top and bottom teams were all consistent across the Bradley Terry model, accessing the elo ratings, and running the Monte Carlo system.