NHL Statistical Analysis

Armaan Atwal, Vikram Brar, Rupin Sehgal, William Spencer

Haskayne School of Business, University of Calgary

BTMA 431: Gathering, Wrangling, And Analyzing Data in R

Professor Jun Ma

April 14, 2023

Tables of Contents

Introduction

The motivation behind this project was to gather, wrangle, and analyze datasets in R to develop insights regarding the chosen data. In doing so, possible suggestions, inferences, and conclusions can be made that otherwise would have been unnoticed. For our specific project, we decided to analyze statistics from the National Hockey League (NHL), a major professional sports league spanning across Canada and the United States. Our project is focused on the breakdown and analysis of multiple datasets containing NHL statistics. These datasets contained numerous different performance metrics for both the teams in the NHL as well as the individual players. Examples of these performance metrics include goals, assists, shots on goal, points, as well as other key metrics. We undertook this project with the intention of gaining insight into the teams and players of the NHL in order to develop an understanding of possible trends and patterns within the data while also uncovering statistical insights that potentially may not have been obvious beforehand. To demonstrate our developed understanding of the data, our project focused on a few key areas of analysis. Examples of which include examining the relationship between different key metrics along with the impact these metrics had on team or individual player success. To create a compelling story with the data, our team used statistical techniques to first interpret the underlying patterns in the data. After that, our team used data visualization tools to create detailed charts and graphs along with other unique visuals. These visuals made interpreting the statistical findings much simpler and most audiences will be able to quickly and easily understand our insights. Overall, the story we created from data provided very valuable insights regarding the performance of teams and players in the NHL. From the gathered data, and the subsequent analysis, our team will provide recommendations and conclusions. These recommendations and conclusions can be used by coaches, staff, players, team owners, or even sports bettors to make more strategic decisions to not only increase their success but also their profitability.

Methodology

In our approach to better understand the NHL, we first had to find data relevant to our analysis. We wanted to explore the NHL from both the perspective of teams as well as players so we initially settled on using NHL.com, thinking that the official website would be the best place to get high quality data for our research. The problem we came across is that we struggled to scrape data off of NHL.com since the website did not store its data tables in a format we were familiar with such as HTML format, but instead, was stored and accessed the data using a JSON format. Unfettered by this temporary set back we quickly found hockey-reference.com which had most of the data we needed stored in HTML format so we began scraping the data. Once we had settled on using hockey-reference.com for both our team and player data we had to find the tag of the tables where the data was stored for the player table. This was under "#stats" and for the team table it was under "#active_franchises". From here, the scraping was quite simple and straightforward. You would first read the link, then anchor to the correct css node ("stats" or "active_franchises" in our case), and finally pull data from the html table(s) within that node. With this approach we only had to change the url, and the css node in our code to perform both webpage scrapes. This was great, but particularly the player data was very messy and had multiple headers within the dataframe we scraped to, and both had several repeated entries, and missing data entries we would need to clean. With this completed, our data collection was nearly finished, but we still needed salary information to answer some of our questions, so we sought out supplementary data from kaggle. For the sake of simplicity, we filtered for only salary and player names before exporting as csv. After this we combined the first and last name columns in Excel before importing it into R. This data importantly contained both player names, and salary so we were able to combine this without other player data to satisfy all our data needs for our analysis.

For our analysis we used a variety of different methods to generate insights into the data. Starting with question one we first started by grouping players into one of two groups, either Veteran (over 25 years old) or Young (under 25 years old), and recorded which group each player belonged to (See Appendix A). This information was used in a linear regression to see if we could predict their mean points, and mean wins by their age group. With the understanding we gained from the first parts of this question we sought to see if we could better understand what decided a player's potential earnings (See Appendix B). To compute this we had to mutate a new column for our data to represent points per second of each player. This is where we utilized the supplementary salary data. This allowed us to plot the average salary by age, as well as plotting a normalized chart which allowed us to see how salary play time, and points per second changed by age (See Appendix C & D). With the new understanding from these visualizations we set out to use a linear regression to predict player salary usings all applicable data that we had collected (See Appendix E).

We then moved onto our second question where we explored the effects different game statistics have on the number of points players score. For this we also used a linear regression where we attempted to predict points, by using goals, assists, games played, penalty minutes, game winning goals, shots on goal, blocks, hits, and age. The output summary of the linear model indicated what metrics can be determined as significant predictors of points scored and what metrics are not (See Appendix F). Regression visuals were constructed for each of the predictors illustrating their significance in a visual manner. The sub question for question two seeks to better understand the connections between average goals scored and different age groups. For this all players were broken up by unique ages (all 28 year olds were averaged as one age group, all 29 year olds in another), and plotted using a heat map to understand where the most goals were occurring (See Appendix G). This was an interactive heat map made in tableau where you could use a slider to adjust the average points from the age ranges you were viewing, which is useful for excluding very new, and very old players who were not contributing much besides goals scored (See Appendix H).

For the final question we sought insight into the potential to predict long term health, and performance of a team, by trying to predict the number of Stanley cups they had won over their lifetime. In performing this we first sought to check the correlation between the potential predictors and total number of Stanley cups won per team (See Appendix I). We would use this insight to then explore if any of the weakly correlated (under 0.7 correlation was considered weak) predictors and if any presented a non-linear relation that could be exploited. To undertake this, we plotted all of the weakly correlated predictors and visually assessed them for potential other relations (See Appendix J). Once all potential relations had been established, we started with a linear regression using only the strongly correlated predictors. The issue with this is many predictors were not statistically significant to the model so in subsequent regressions these non-significant predictors were dropped from the model (See Appendix K). Once only statistically significant predictors remained (See Appendix L), we tested this model for any higher power relationships from the poorly correlated data we investigated earlier (See Appendix M). The subsequent reduction in adjusted R squared between my first and final model was significant so we looked to machine learning to see if a better model could be generated (See

Appendix N). We settled on using a stepwise function to ensure only relevant predictors were added to the linear regression, and used a bi-directional search to ensure that only relevant predictors were added and that they were removed if they lost relevance to other predictors. We also used AIC scoring in judging a predictors relevance since the previous model only had a single statistically significant predictor, because of this we wanted the model to try and be more lenient in what predictors it selected, compared to the more restrictive BIC scoring. The resulting model still had some statistically non-significant predictors (See Appendix O), but once these were removed the model mostly maintained its exceptional adjusted R squared value unlike the previous linear regressions (See Appendix P & Q & R).

Results

During our analysis of NHL data we uncovered several interesting results. During our exploration of the impact of veteran vs young players in the NHL on a teams performance we discovered that despite generally older players populating most of the teams, teams which balanced the number of new versus veteran players saw a more consistent mean win and points spread (See Appendix A). This was contrasted by the teams which only focussed on veteran players and saw large variances and that no single conclusion could be made about them, unlike the balanced teams (See Appendix A). We then sought to further explore this insight by trying to use the data to predict mean wins and points using the age groups we created. This regression told us that both age groups were statistically significant predictors of mean wins and points, and that specifically being a veteran player resulted in you being more likely to wins/points score (as seen by the large positive coefficient of 42.74), and that being a young player is likely to have a lesser positive impact of 29.12 your average wins/points scored (as seen by it's coefficient

-13.62, 42.74 - 13.62 = 29.12) (See Appendix B). Furthermore, this regression explained about 29% of the variation in means wins and points, which surprised us that only this predictor could manage to explain a third of variation (See Appendix B). The final explorations in the first part of our question was an attempt to deepen our understanding of what determines a player's potential earnings. Our sub question, focused on finding the important variables that influence a player's wage. The following variables were significant to predicting compensation from our second linear regression model:

- Age: Age of a player
- +/-: statistic used to measure a player's overall impact on the game. It's calculated by subtracting the number of goals the player was on the ice for when their team allowed a goal (minus) from the number of goals they were on the ice for when their team scored a goal (plus).
- Point shares: statistic that attempts to assign a value to a player's contributions to their team's success. It's calculated by taking the number of points the team earns while the player is on the ice and dividing it by the total number of points the team earns in a season. This gives a percentage value, which is multiplied by the total number of points available in the league (normally 1230 for an 82-game season) to get the player's point share.
- Faceoff wins and losses: statistic measures a player's success rate in winning faceoffs during games. A faceoff occurs when the puck is dropped in the center of the rink and two opposing players try to gain possession of it. The player who wins the faceoff gets possession of the puck for their team. Faceoff wins and losses are tracked separately.

Over 55 percent of variation (r squared value) in our dataset indicates that performance does have a correlation with salary, but players get paid dramatically more with age; peaking from 28-35 years (See Appendix C). This bar graph leads the audience to the conclusion that players 28-35 are superior in skill as they are paid the most, but is that correct (See Appendix C)? With this information, we became curious if this was also the age range where players were the most efficient with their ice time and therefore potentially worth more to the team. To answer this we made a new graphic with normalized values for play time points per second and salary over the ages of different players (See Appendix D). To our surprise we found that players under 27 had some of the highest ice times, as well as the highest points per second they could expect during their career, despite them being the lowest paid they would ever be (See Appendix D). This led us to believe that players were the most efficient when they were under 27, their ice time and points per second would level out when the players were 29-34, and any player remaining after 34 would begin to see improvements in line with what is seen from younger players (See Appendix D). It's clear that young players provide the most results on the ice for the team and as they age their time playing and efficiency declines. Inversely, their salary increases dramatically with age and does not fluctuate like time on ice or player efficiency does with age, even if performance cannot justify their high pay. Finally we created a regression trying to predict player salary and discovered that a players position seems to be irrelevant to salary, but age, +/- (+1 for goals scored for the team while this player was on the ice -1 for goals scored on the team while this player was on the ice), average goals, point shares, face off wins, and face off losses were all key predictors of a player's earnings (See Appendix E). As an example the model predicts that for every year the player continues playing they can expect an increase of about $533,000 in annual salary (See Appendix E). This model managed to explain 56% of the variation in player salaries suggesting that these predictors contain the majority of the predictive data to accurately predict someone's salary (See Appendix E).

The results of the regression output for the question "What game characteristics have a significant impact on a player's points?" will now be discussed. From the output summary, it can be determined that Goals, Assists, Hits, and Penalty minutes are all significant predictors of a player's points scored. Initially, our team had assumed that Penalty minutes along with Hits would not be significant. Penalty minutes, and Hits had a very low p-values of 0.07668, and 0.00362 respectively (See Appendix F). Upon further investigation however, these significant p-values began to make sense in the context of points scored. In an attempt to maximize their ability to score goals, players must spend as much time as they possibly can on the ice playing. An explanation for penalty minutes being a significant predictor to goals scored is that penalty minutes are spent in the penalty box, which is off the ice, and dramatically limits a player's total time on the ice playing. Likewise, illegal Hits are a major contributor to in-game fouls. These illegal hits are actions that often require the player to subsequently spend time in the penalty box again taking away from their time on ice. Goals and assists were expected to be significant, and their extremely small p-values of less than 2e-^16 indicate that they are very crucial predictors of points scored (See Appendix F). However, upon further analysis we concluded that the way our question was structured may be influencing these results. Points scored is a measured metric. A player's points only consist of two statistics, their goals scored, and their assists made. By including both goals and assists as predictors, the overall significance of the regression output is lowered because it creates an artificially low Residual Standard Error of 3.405e^-14 (See Appendix F). What we can take away from this analysis is when predicting a metric that is directly determined by other metrics, including these determining metrics in a regression model will give you artificially high confidence in your regression models accuracy.

After seeing that goals were one of the most significant variables found in our regression model from the first part of question two, we decided to see how the average goals scored per player changes across all age groups in the current season. We decided to group up the ages so we can see which age stands out against the rest. Goals are one of the most important variables when it comes to hockey as it's important to have goal scorers amongst your team. Without scoring enough goals, a team won't win and will likely hinder their performance in the season. We modeled our question via a for loop in R, where it looped through each unique age of each player in our data frame by using the unique function and calculated the average goals scored by each of them and then grouped them in their respective age group within the loop. The results we found were mostly surprising. Ages 37 and 38 were found to have the most average goals respective to all other age groups (See Appendix G). The age group with the lowest number of average goals was 18, with only 1.33. A clearer explanation of this is that there aren't many 18-year old's ready to enter the league as they are still improving their skills in developmental leagues such as the WHL, AHL and European leagues. It is unusual to see older players have many more goals than younger players. An underlying reason for this to the current 2022-2023 season is that the 2nd all time goal scorer is actively playing. Alex Ovechkin, 37 years old and has 42 goals this season (NHL, 2023). Age group 26 was the third highest group with 11.75 goals on average this season. A lot of these age groups are close to each other regarding average goals, concluding there isn't much differentiation between nearby age groups seen together. After seeing the unusual number of goals scored by the older age groups, we concluded that there were some outliers in the model. This second visualization has an interactive slider function applied to each age group, forcing to show age groups with average points between 15 and 35 only (See Appendix H). This way, it takes out some outliers where the goals weren’t seen as significant for that age group or was by luck. We now see that age group 26 leads all age groups filtered by this measure. This can be explained since most players tend to peak around their mid 20s. A reason why there are some age groups seen as outliers, is that there aren't as many players in their age group compared to others, so the results are skewed, depending on the number of players available from the data set.

In our final exploration we sought to gain insights into a teams long term performance by attempting to predict all time Stanley cup wins for each team in the NHL. In the set up for this regression we learned that years a team played, games played, wins, ties, year in the playoffs, and years at the top of there division were highly correlated with the number of Stanley cup wins for a team (See Appendix I). We also discovered that none of the weakly correlated predators had any non-linear higher power relationships. Yet, trying to get these highly correlated predictors into a model of only statistically significant predictors was a challenge as the removal of predictors quickly led to a large deterioration in the adjusted R squared value of the model (See Appendix N). To combat this issue we turned to machine learning using a stepwise function set to a bi-directional search to uncover which predictors we really needed. This model when finally narrowed down to only statistically significant predictors indicated that years a team played, and years a team finished at the top of there conference were the most important predictors (as measured by the size of the coefficients on each 0.85, and 0.79 respectively), and that games played had a slight, but significant negative influence on Stanley cup wins (a coefficient of -0.01) (See Appendix Q). The model was quite good at predicting actual Stanley cup wins, given its adjusted R squared of 0.7639, and it's normally distributed residuals (See Appendix F). The model was most accurate for middling teams, having a harder time predicting very small or very large numbers of Stanley cup wins (See Appendix T).

Discussion

Data collection was one of the main takeaways we took from this project. Not having enough data to use gave us a disadvantage regarding getting the best results in an accurate manner. If we had oversampled our data to include more player data from different seasons, we'd have less outliers regarding the different age groups and see a more accurate result in our questions. Using a larger dataset that contains more players and seasons might produce more precise and reliable results. A larger dataset would also allow us to create a training/test split of data from testing our models to determine their accuracy, sensitivity, AUC, and other performance measures via the confusion matrix. A limitation towards the data collection process also affected how much data we could collect and expand onto with our questions. With a low availability of scrape friendly websites that have the NHL statistics needed for our models, it was much harder to gather data in support of our questions. It would be much easier to use csv files which would not only help speed up the process but allow more information to be used for our models. Furthermore, text analytics and scraping online discussion forums and websites such as reddit and twitter may allow us to observe qualitative trends in player and team popularity trades and news in real time. This would allow us to further concentrate our understanding about the NHL and its participants. Model diversity is the last takeaway, where if we had more models, it would allow for better comparisons and higher quality analysis overall. Our models may be enhanced by integrating more relevant predictors. Our models also allow us to create a basic understanding of NHL insights, but it is important to verify these results with other sources as there are many qualitative factors that may affect results.Future work on the topic might focus on increasing the model's precision and using more sophisticated methods, including machine learning algorithms, to analyze the data. These models could include but not be limited to KNN (supervised and unsupervised models), Support Vector Machines, Principal Component Analysis, various other models, and preprocessing efforts. Two of the challenges faced throughout the analysis were the limited availability of data and the need to modify the data to match the research. The dataset lacked certain information that might have aided the study, such as player performance in postseason or international games. One of the project's shortcomings is the investigation's limited focus on only two concerns. To provide a more accurate picture of team performance and player earnings, future study may expand the research to include additional elements such as team expenses or player injuries. Furthermore, other sources of income such as marketing sponsors, concession stand earnings, etc. could help further our understanding and models' predictive ability. Not only that, but increased statistics about play-by-play information could help our understanding of mean wins and points scored by players. This would allow us to gather more records and categories of data that could potentially be significant in prediction.

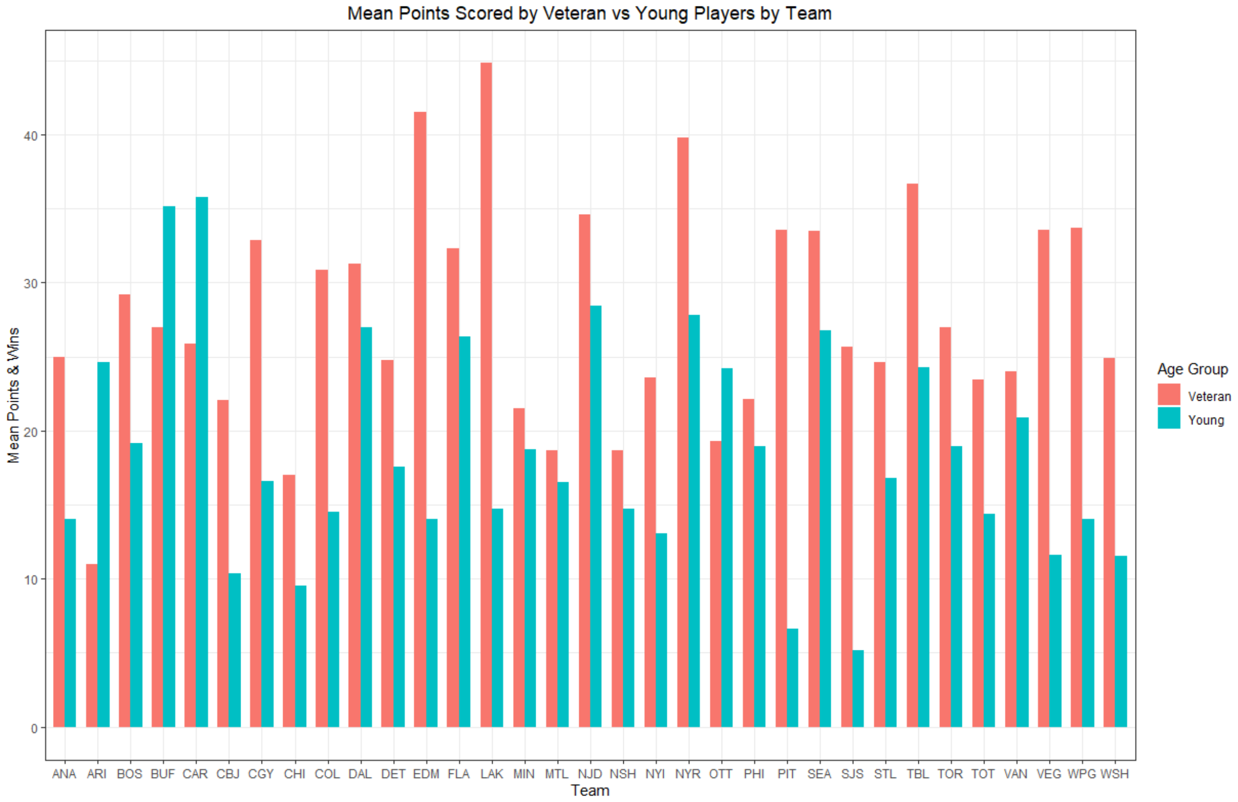
Conclusion

In summation, the findings we obtained through the analysis conducted can be used by NHL staff and team management to improve their teams on-ice performance. Potential recommendations can be developed based on the individual questions we have formulated. Based on the interpretation of question one's results, our recommendation to optimize team performance would be to ensure there is a balance between young and veteran players. For the first sub-question it was a regression analysis determined that age groups were a significant predictor of mean wins and points. The conclusion that can be drawn from this analysis is that veteran players will likely have a more significant impact on a teams mean wins/points than a younger player. For the final sub-question to question one, our goal was to determine what metrics determine a player's salary as well. The metrics we discovered to have a significant impact on a player's salaries include their age, their point shares, their faceoff wins and losses, and finally the (+/-) metric. Our second question sought to answer what game metrics would be a significant predictor of a player's Points in a season. The summary for the regression we built indicated that Goals, Assists, Hits, and Penalty minutes were all significant. Building off this, our sub-question for question two aimed to answer what then average goals scored across all ages would be. The highest goals scored per age was surprisingly age 37 with an average of 15.71. Lastly, Our third question aimed to predict Stanley cup wins for each team. To answer this question, we built another regression model. After numerous attempts, we concluded that the two most significant predictors for a Stanley cup are a teams total years played and the number of years they have finished in top of their confrences, with a slight moderating effect from total games each team played.

List of Appendices

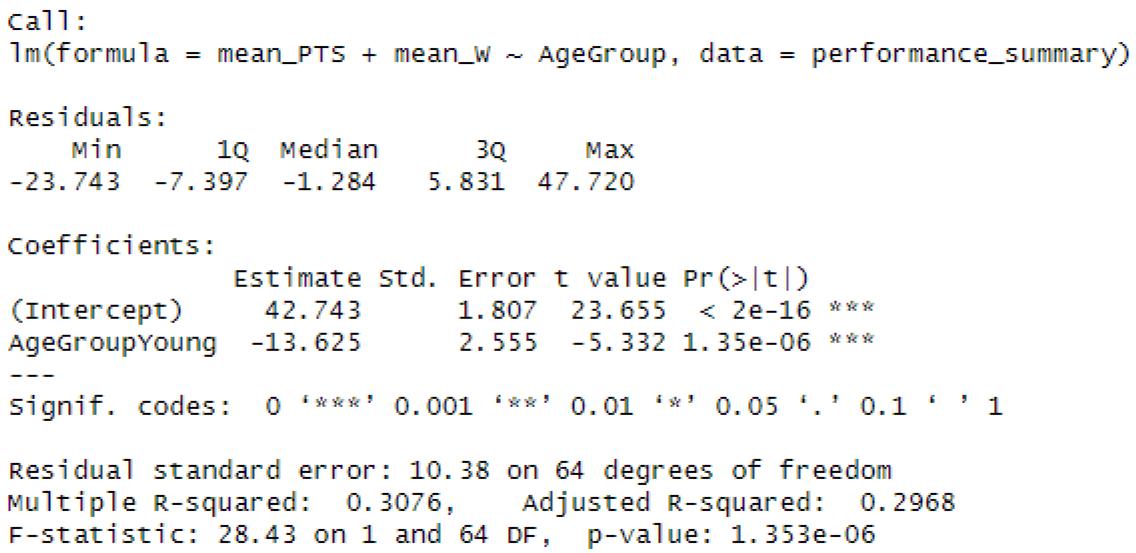
Appendix A

Visualization of mean points & wins broken down by veteran or young players, and team



Appendix B

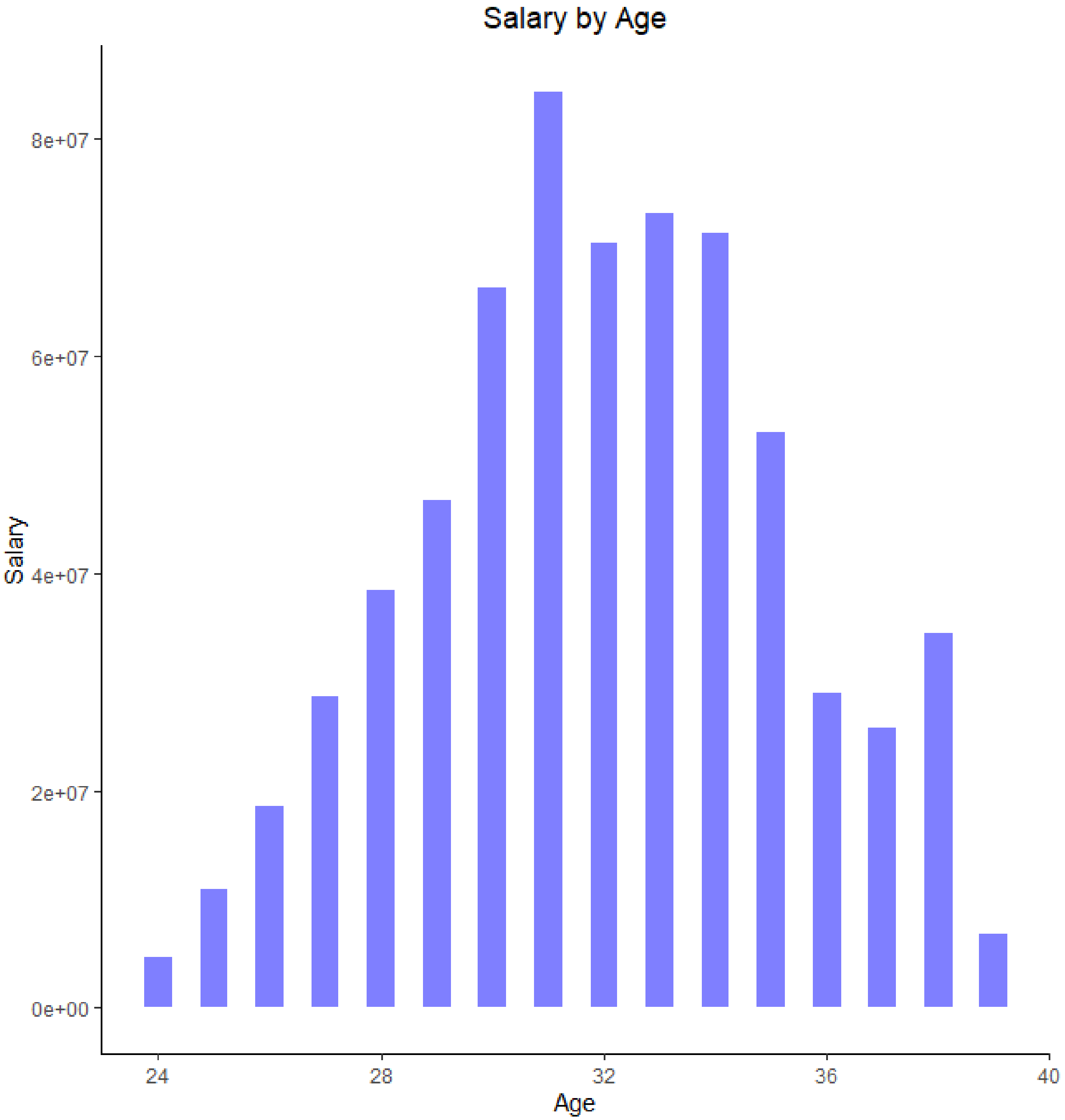
Summary of model



```
Call:
lm(formula = mean_PTS + mean_W ~ AgeGroup, data = performance_summary)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.743  -7.397  -1.284   5.831  47.720 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     42.743      1.807  23.655  < 2e-16 ***
AgeGroupYoung  -13.625      2.555  -5.332 1.35e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.38 on 64 degrees of freedom
Multiple R-squared:  0.3076,    Adjusted R-squared:  0.2968 
F-statistic: 28.43 on 1 and 64 DF,  p-value: 1.353e-06
```

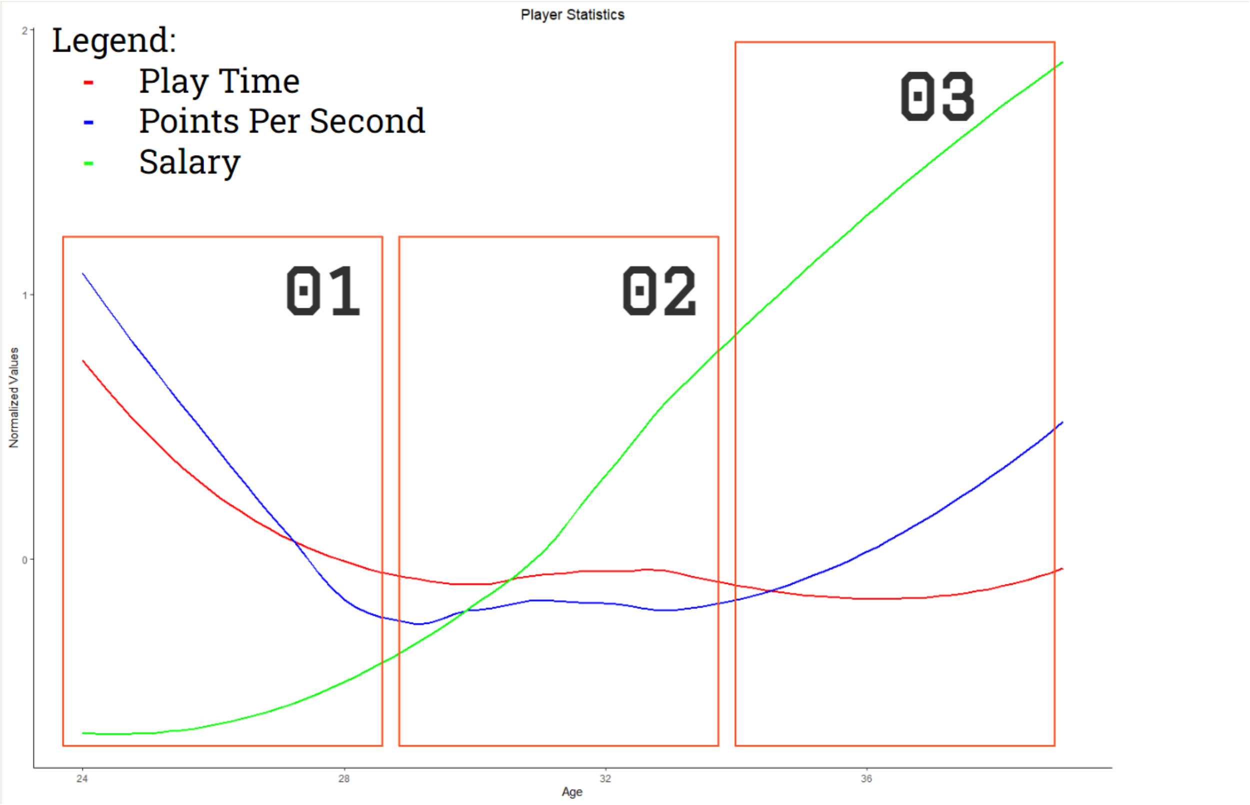
Appendix C

Visual demonstration of salary by age



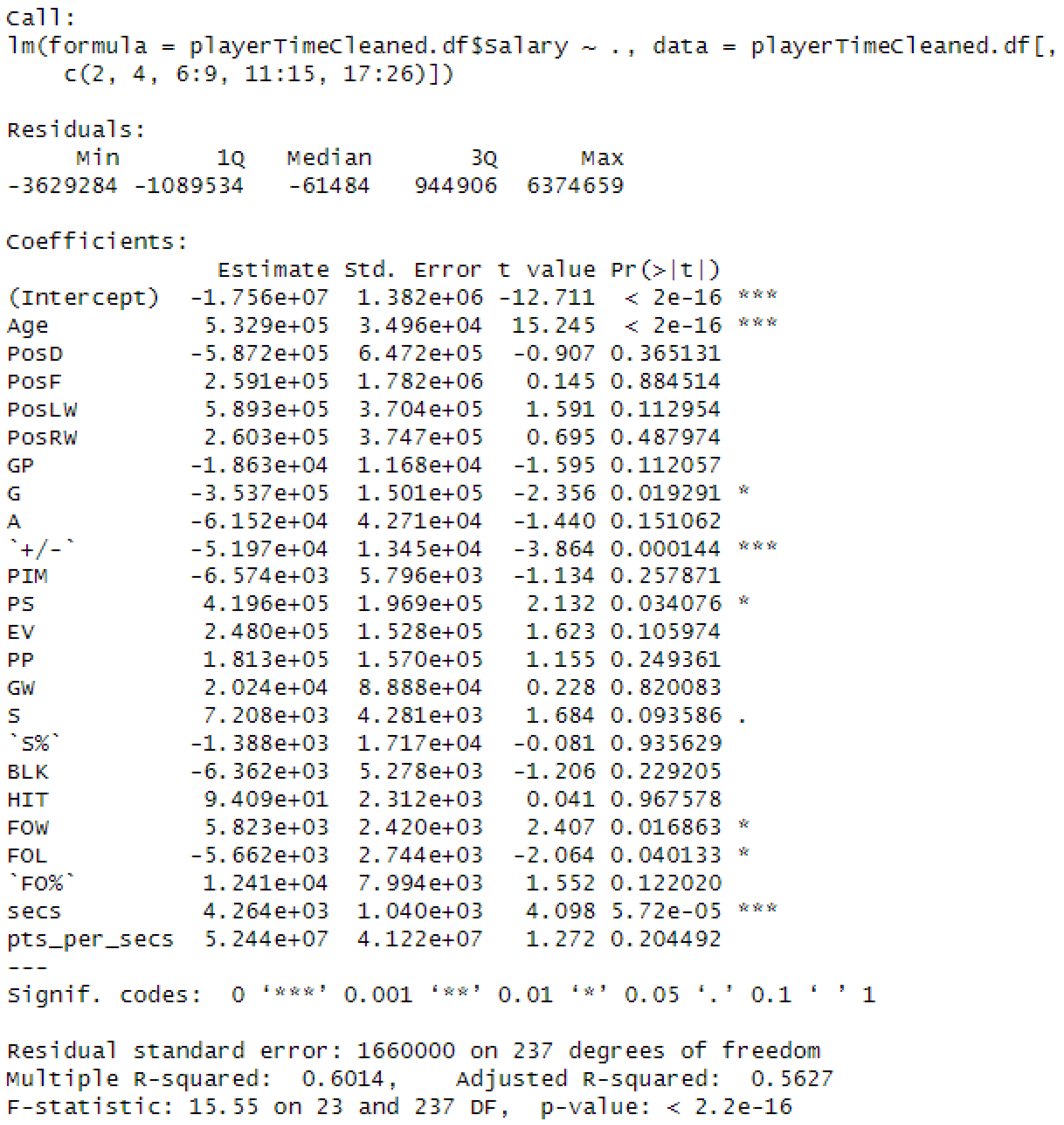
Appendix D

Breakdown of player efficiency measured by play time, points per second, and salary over player ages. The graphic is broken down into 3 main phases.



Appendix E

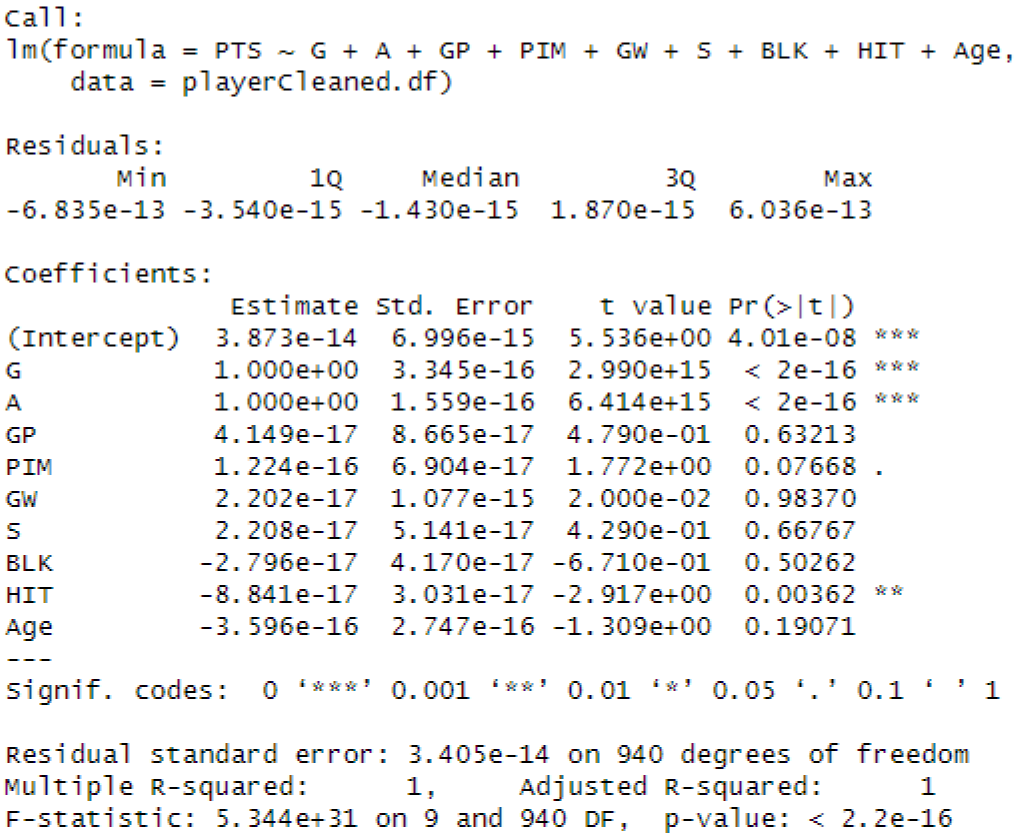
Summary of salarypred Model



```
Call:
lm(formula = playerTimeCleaned.df$Salary ~ ., data = playerTimeCleaned.df[,
    c(2, 4, 6:9, 11:15, 17:26)])

Residuals:
     Min       1Q   Median       3Q      Max
-3629284 -1089534   -61484   944906  6374659

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.756e+07  1.382e+06 -12.711  < 2e-16 ***
Age           5.329e+05  3.496e+04  15.245  < 2e-16 ***
PosD         -5.872e+05  6.472e+05  -0.907 0.365131
PosF          2.591e+05  1.782e+06   0.145 0.884514
PosLW         5.893e+05  3.704e+05   1.591 0.112954
PosRW         2.603e+05  3.747e+05   0.695 0.487974
GP           -1.863e+04  1.168e+04  -1.595 0.112057
G            -3.537e+05  1.501e+05  -2.356 0.019291 *
A            -6.152e+04  4.271e+04  -1.440 0.151062
`+/-`        -5.197e+04  1.345e+04  -3.864 0.000144 ***
PIM          -6.574e+03  5.796e+03  -1.134 0.257871
PS            4.196e+05  1.969e+05   2.132 0.034076 *
EV            2.480e+05  1.528e+05   1.623 0.105974
PP            1.813e+05  1.570e+05   1.155 0.249361
GW            2.024e+04  8.888e+04   0.228 0.820083
S             7.208e+03  4.281e+03   1.684 0.093586 .
`S%`         -1.388e+03  1.717e+04  -0.081 0.935629
BLK          -6.362e+03  5.278e+03  -1.206 0.229205
HIT           9.409e+01  2.312e+03   0.041 0.967578
FOW           5.823e+03  2.420e+03   2.407 0.016863 *
FOL          -5.662e+03  2.744e+03  -2.064 0.040133 *
`FO%`         1.241e+04  7.994e+03   1.552 0.122020
secs          4.264e+03  1.040e+03   4.098 5.72e-05 ***
pts_per_secs  5.244e+07  4.122e+07   1.272 0.204492
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1660000 on 237 degrees of freedom
Multiple R-squared:  0.6014,	Adjusted R-squared:  0.5627
F-statistic: 15.55 on 23 and 237 DF,  p-value: < 2.2e-16
```

Appendix F

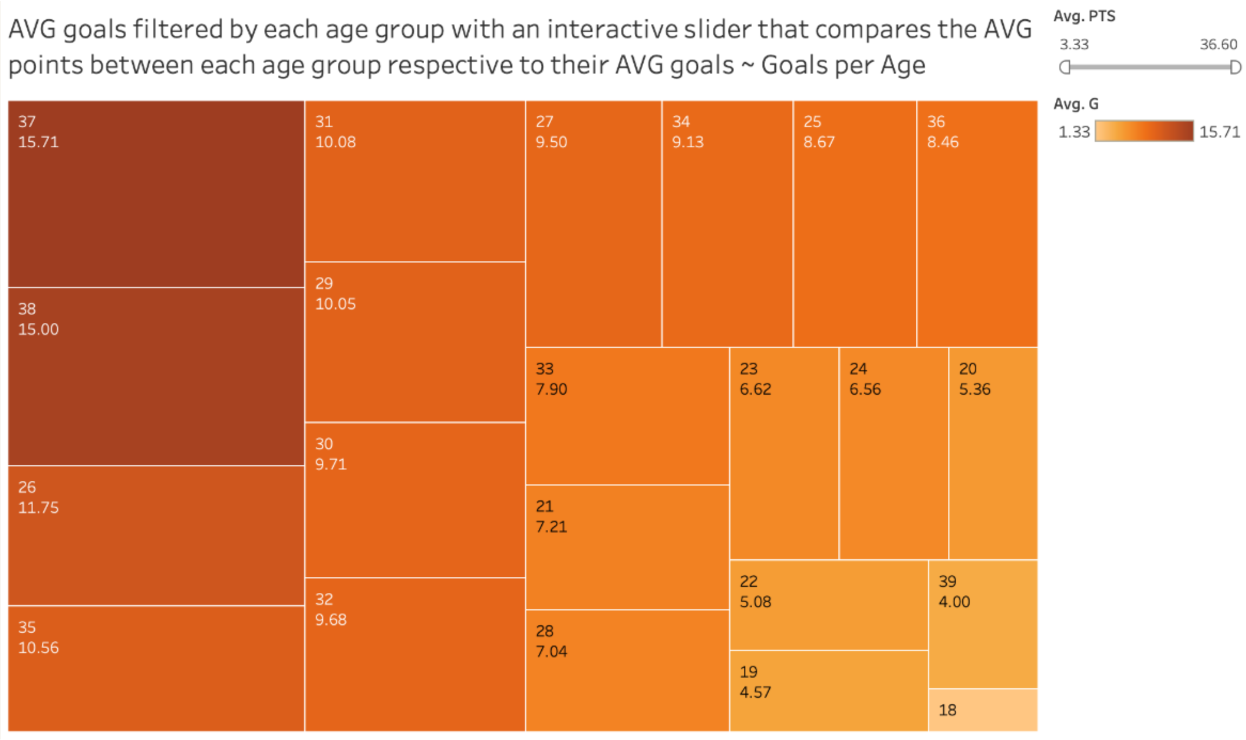
Summary of Q2corFit.1 model



```
Call:
lm(formula = PTS ~ G + A + GP + PIM + GW + S + BLK + HIT + Age, 
    data = playerCleaned.df)

Residuals:
       Min         1Q     Median         3Q        Max 
-6.835e-13 -3.540e-15 -1.430e-15  1.870e-15  6.036e-13 

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)    
(Intercept)  3.873e-14  6.996e-15  5.536e+00 4.01e-08 ***
G            1.000e+00  3.345e-16  2.990e+15  < 2e-16 ***
A            1.000e+00  1.559e-16  6.414e+15  < 2e-16 ***
GP           4.149e-17  8.665e-17  4.790e-01  0.63213    
PIM          1.224e-16  6.904e-17  1.772e+00  0.07668 .  
GW           2.202e-17  1.077e-15  2.000e-02  0.98370    
S            2.208e-17  5.141e-17  4.290e-01  0.66767    
BLK         -2.796e-17  4.170e-17 -6.710e-01  0.50262    
HIT         -8.841e-17  3.031e-17 -2.917e+00  0.00362 ** 
Age         -3.596e-16  2.747e-16 -1.309e+00  0.19071    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.405e-14 on 940 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 5.344e+31 on 9 and 940 DF,  p-value: < 2.2e-16
```

Appendix G

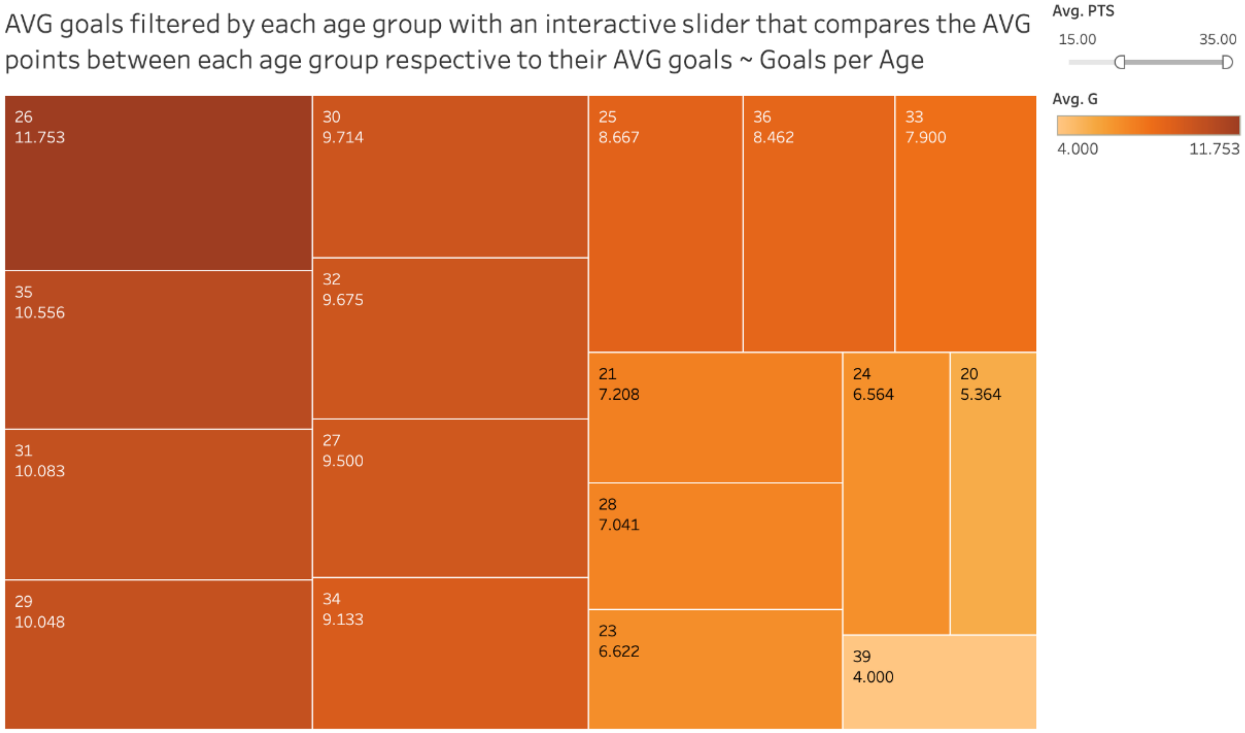
Heat map showing average goals by each age group



AVG goals filtered by each age group with an interactive slider that compares the AVG points between each age group respective to their AVG goals ~ Goals per Age

Appendix H

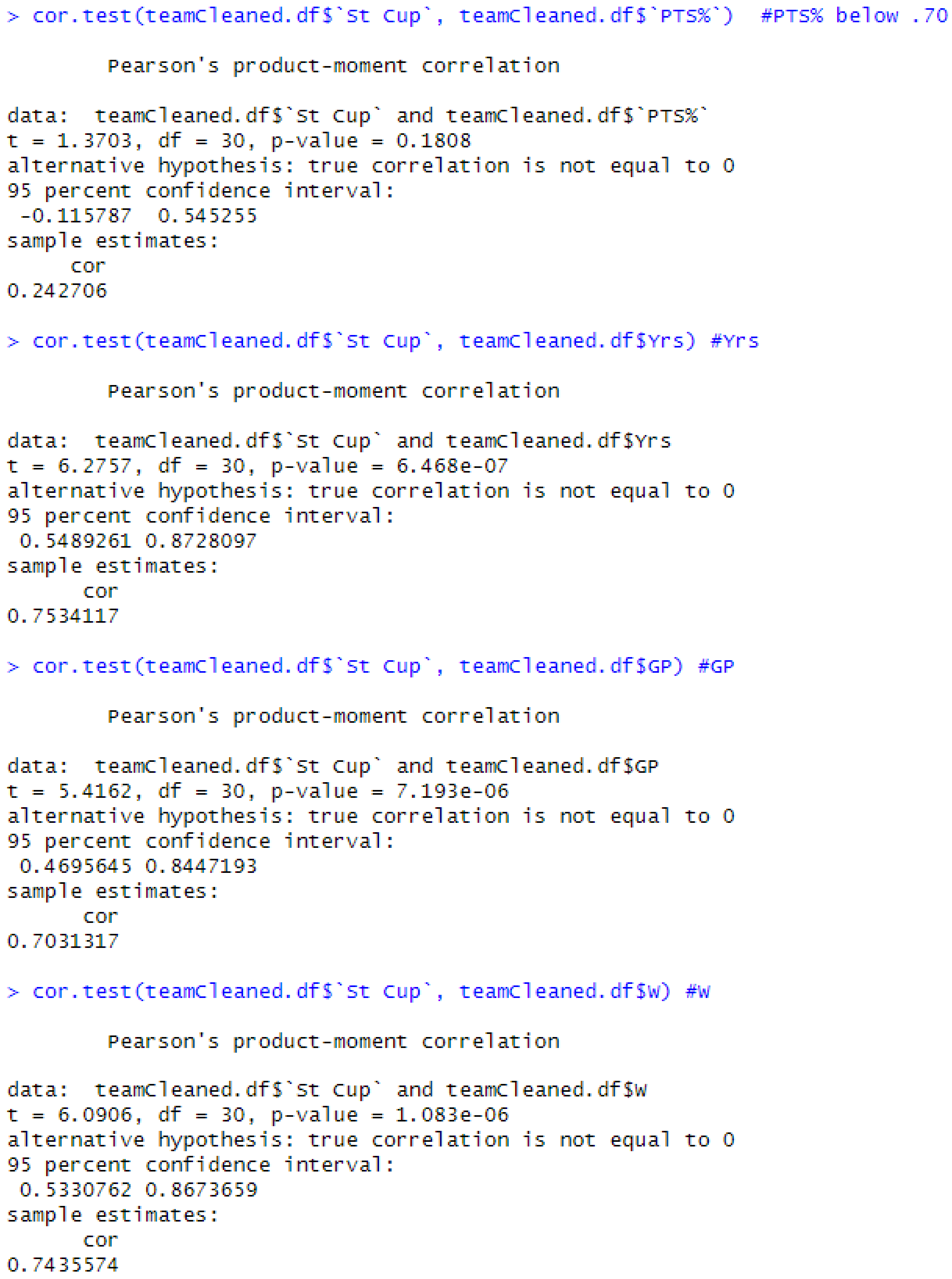
Heat map with slider set to eliminate outliers



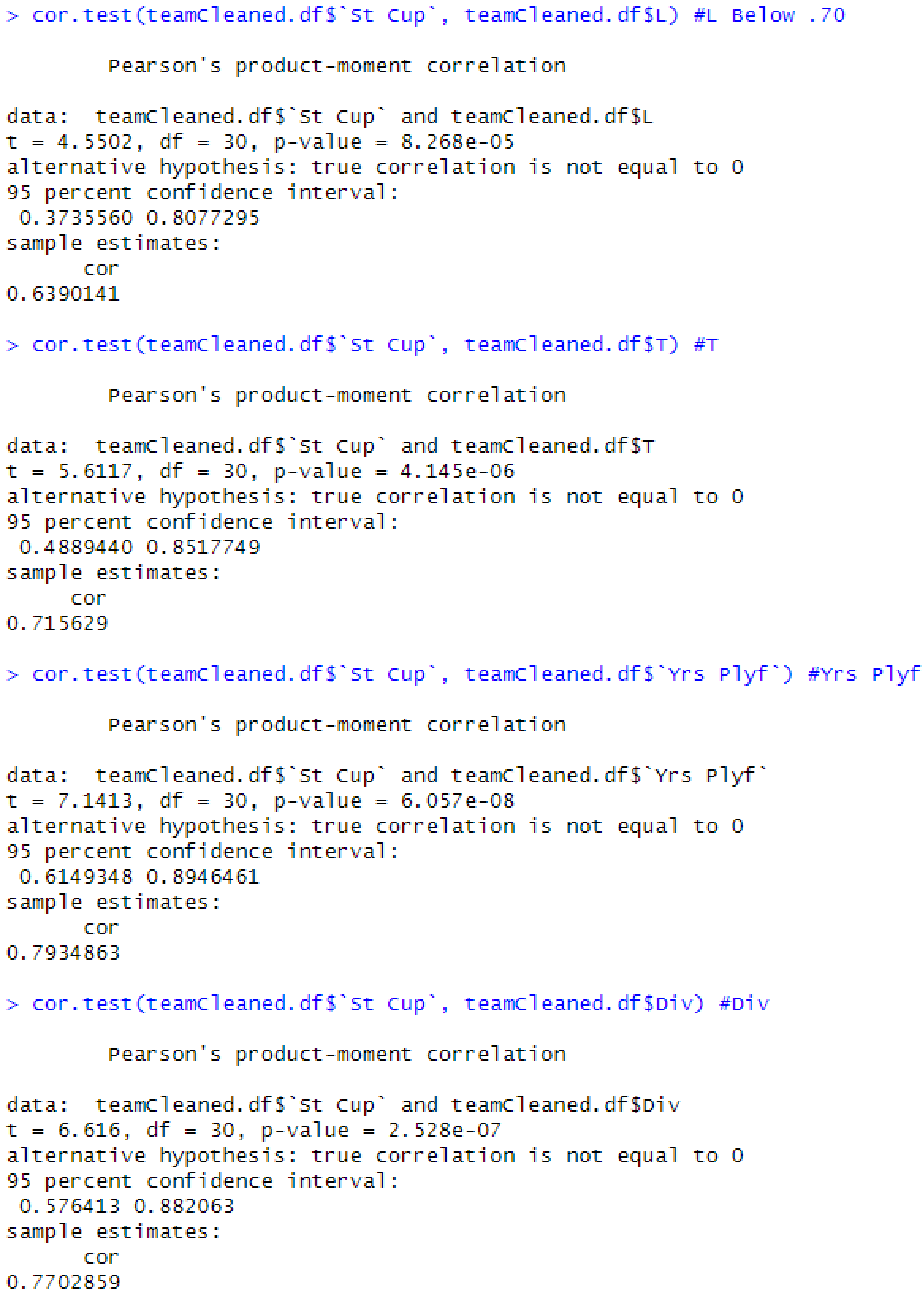
AVG goals filtered by each age group with an interactive slider that compares the AVG points between each age group respective to their AVG goals ~ Goals per Age

Appendix I

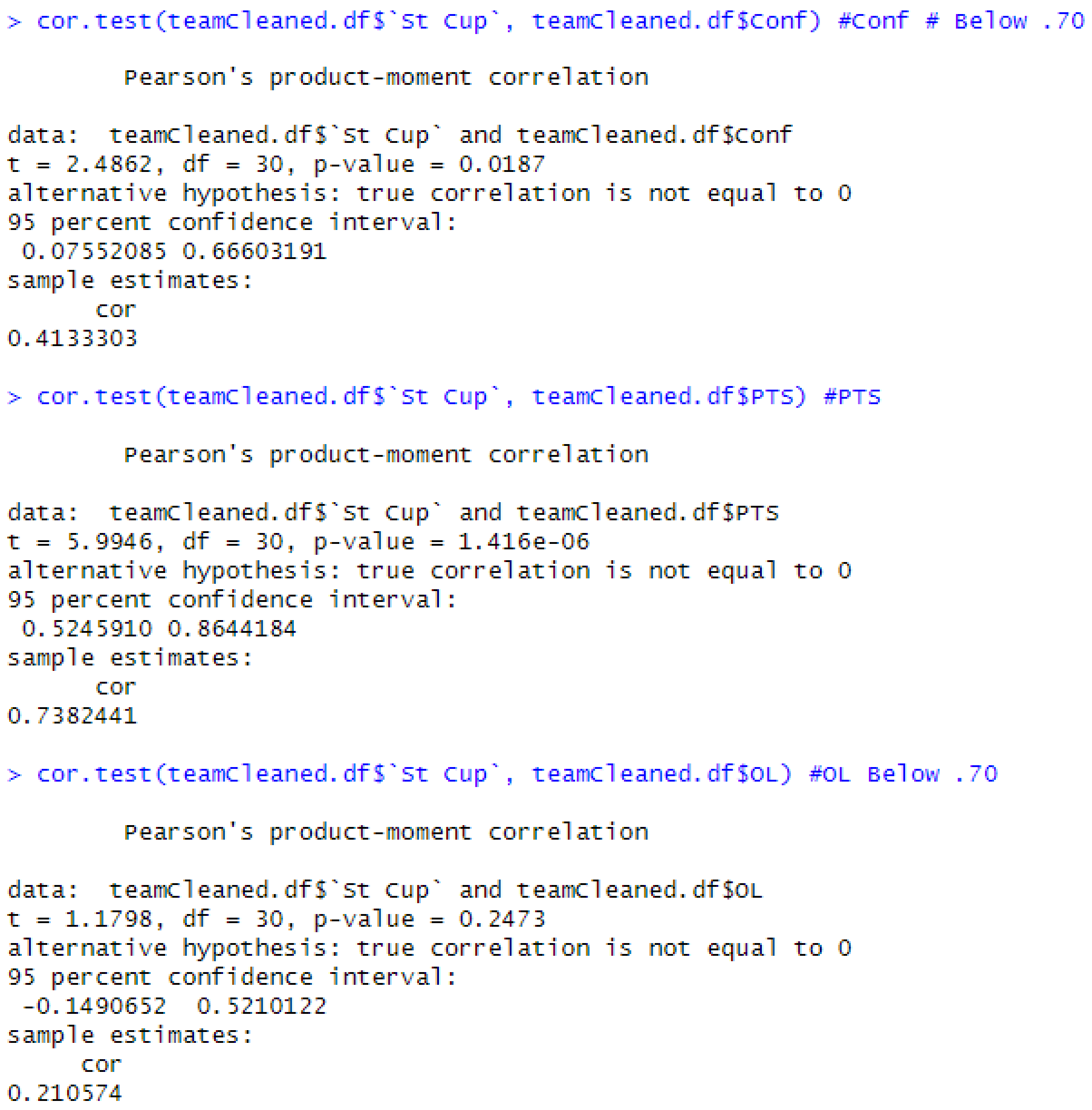
Results from Correlation Tests in Q3



```
> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$`PTS%`)  #PTS% below .70

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$`PTS%`
t = 1.3703, df = 30, p-value = 0.1808
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.115787  0.545255
sample estimates:
     cor 
0.242706 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$Yrs) #Yrs

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$Yrs
t = 6.2757, df = 30, p-value = 6.468e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5489261 0.8728097
sample estimates:
      cor 
0.7534117 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$GP) #GP

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$GP
t = 5.4162, df = 30, p-value = 7.193e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4695645 0.8447193
sample estimates:
      cor 
0.7031317 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$W) #W

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$W
t = 6.0906, df = 30, p-value = 1.083e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5330762 0.8673659
sample estimates:
      cor 
0.7435574 
```



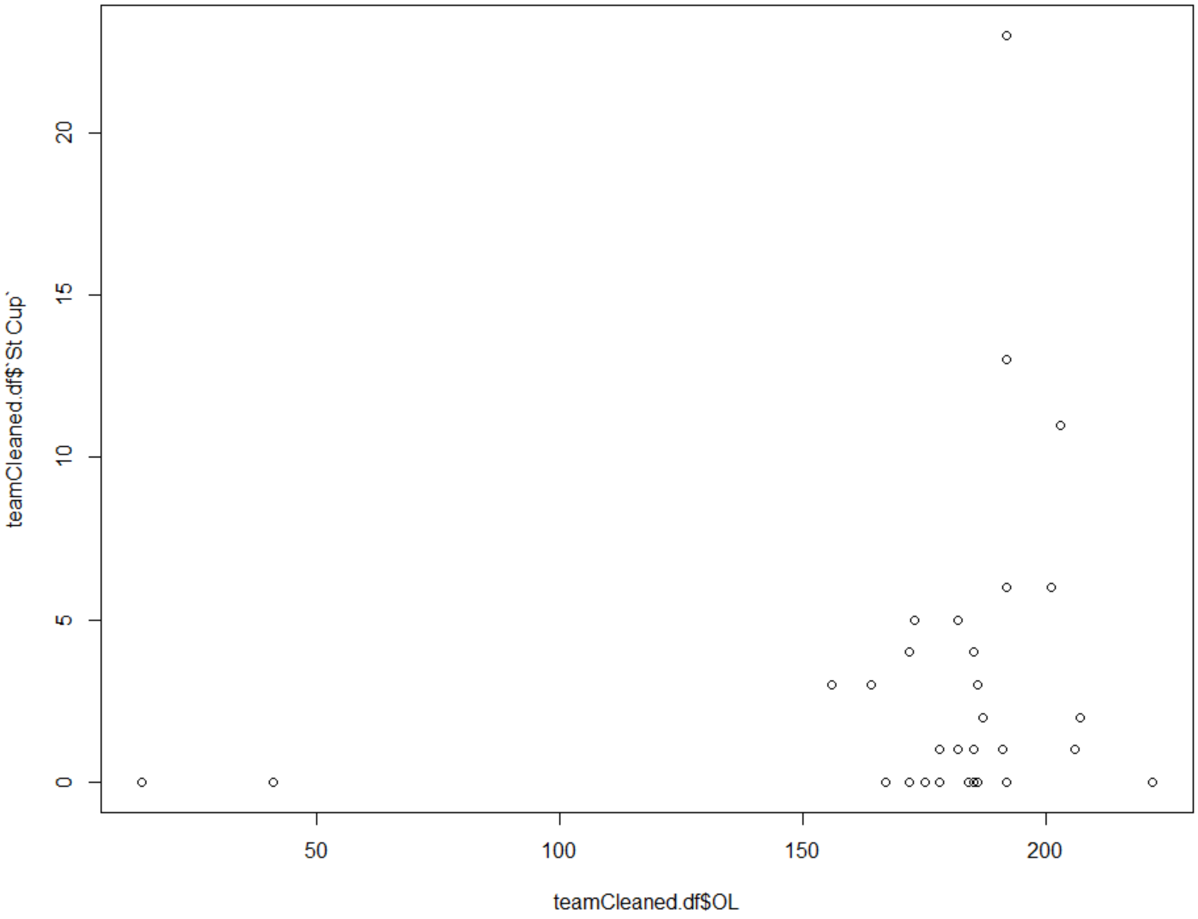
```
> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$L) #L Below .70

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$L
t = 4.5502, df = 30, p-value = 8.268e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3735560 0.8077295
sample estimates:
      cor 
0.6390141 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$T) #T

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$T
t = 5.6117, df = 30, p-value = 4.145e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4889440 0.8517749
sample estimates:
     cor 
0.715629 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$`Yrs Plyf`) #Yrs Plyf

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$`Yrs Plyf`
t = 7.1413, df = 30, p-value = 6.057e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6149348 0.8946461
sample estimates:
      cor 
0.7934863 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$Div) #Div

	Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$Div
t = 6.616, df = 30, p-value = 2.528e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.576413 0.882063
sample estimates:
      cor 
0.7702859 
```



```
> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$Conf) #Conf # Below .70

        Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$Conf
t = 2.4862, df = 30, p-value = 0.0187
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07552085 0.66603191
sample estimates:
      cor 
0.4133303 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$PTS) #PTS

        Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$PTS
t = 5.9946, df = 30, p-value = 1.416e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5245910 0.8644184
sample estimates:
      cor 
0.7382441 

> cor.test(teamCleaned.df$`St Cup`, teamCleaned.df$OL) #OL Below .70

        Pearson's product-moment correlation

data:  teamCleaned.df$`St Cup` and teamCleaned.df$OL
t = 1.1798, df = 30, p-value = 0.2473
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1490652  0.5210122
sample estimates:
     cor 
0.210574 
```

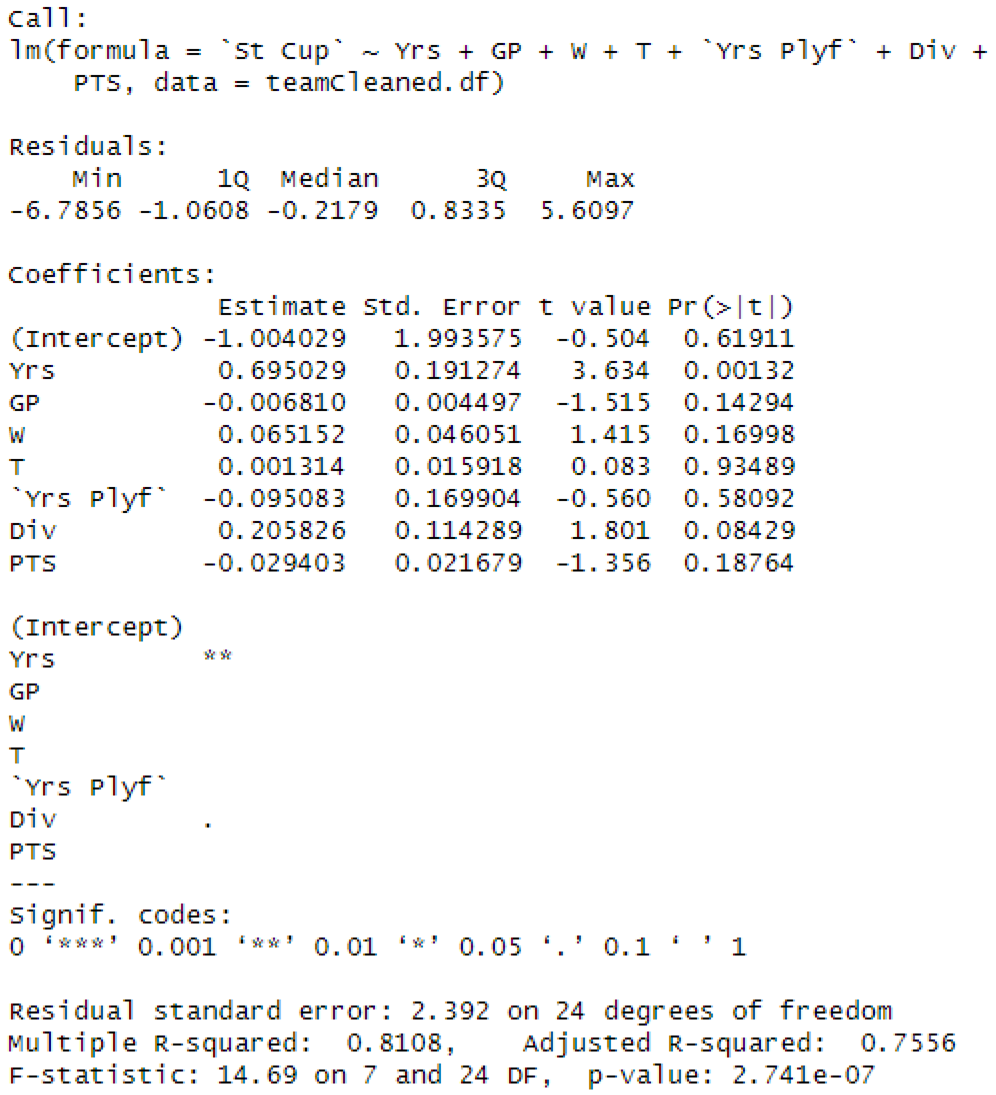
Appendix J

Plot of weakly correlated Overtime/shootout losses by Stanley Cup Wins



Appendix K

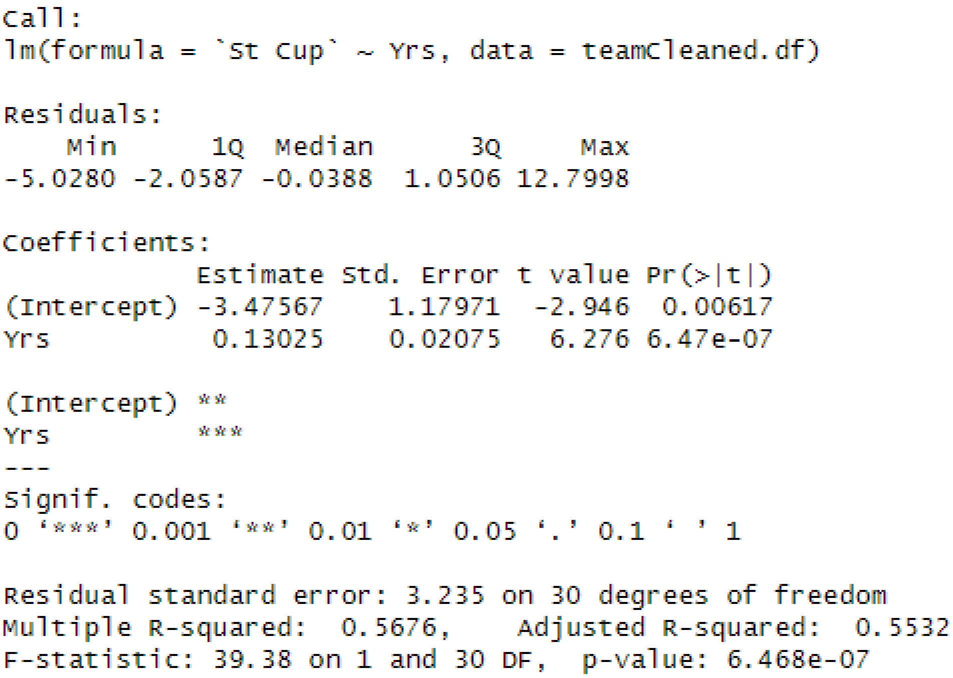
Summary of corFit model



```
Call:
lm(formula = `St Cup` ~ Yrs + GP + W + T + `Yrs Plyf` + Div + 
    PTS, data = teamCleaned.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7856 -1.0608 -0.2179  0.8335  5.6097 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.004029   1.993575  -0.504  0.61911
Yrs          0.695029   0.191274   3.634  0.00132
GP          -0.006810   0.004497  -1.515  0.14294
W            0.065152   0.046051   1.415  0.16998
T            0.001314   0.015918   0.083  0.93489
`Yrs Plyf`  -0.095083   0.169904  -0.560  0.58092
Div          0.205826   0.114289   1.801  0.08429
PTS         -0.029403   0.021679  -1.356  0.18764

(Intercept)   
Yrs         **
GP            
W             
T             
`Yrs Plyf`    
Div         . 
PTS           
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.392 on 24 degrees of freedom
Multiple R-squared:  0.8108,	Adjusted R-squared:  0.7556 
F-statistic: 14.69 on 7 and 24 DF,  p-value: 2.741e-07
```

Appendix L

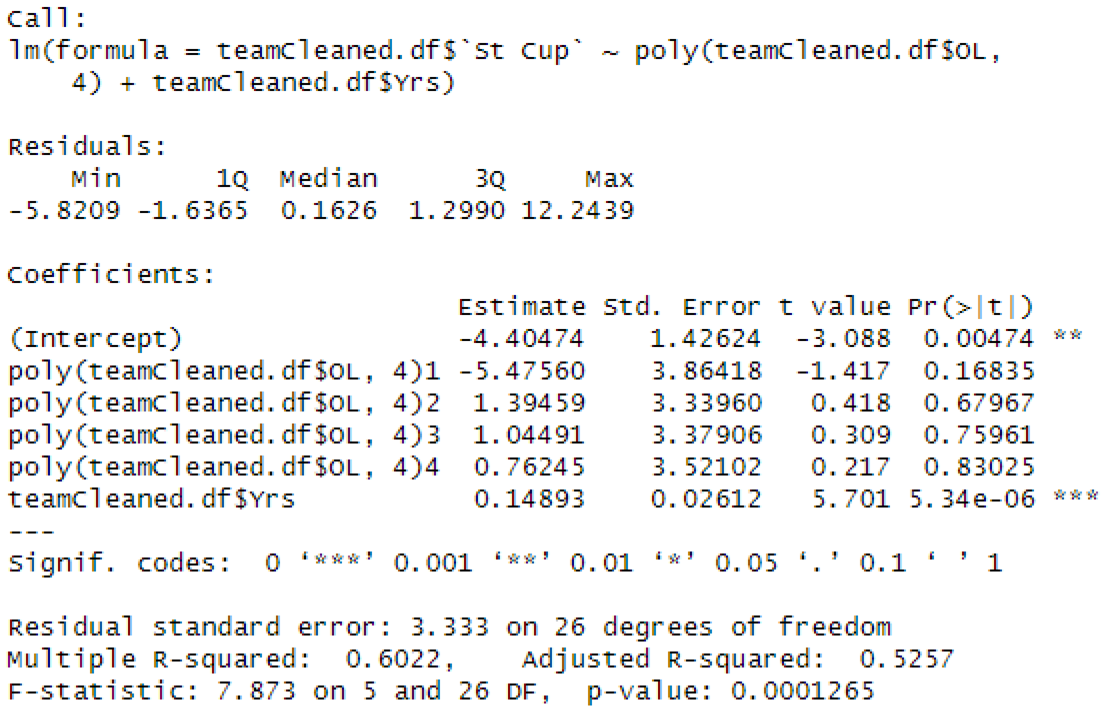
Summary of yrsFit Model



```
Call:
lm(formula = `St Cup` ~ Yrs, data = teamCleaned.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0280 -2.0587 -0.0388  1.0506 12.7998 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.47567    1.17971  -2.946  0.00617
Yrs          0.13025    0.02075   6.276 6.47e-07

(Intercept) ** 
Yrs         ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.235 on 30 degrees of freedom
Multiple R-squared:  0.5676,	Adjusted R-squared:  0.5532 
F-statistic: 39.38 on 1 and 30 DF,  p-value: 6.468e-07
```

Appendix M

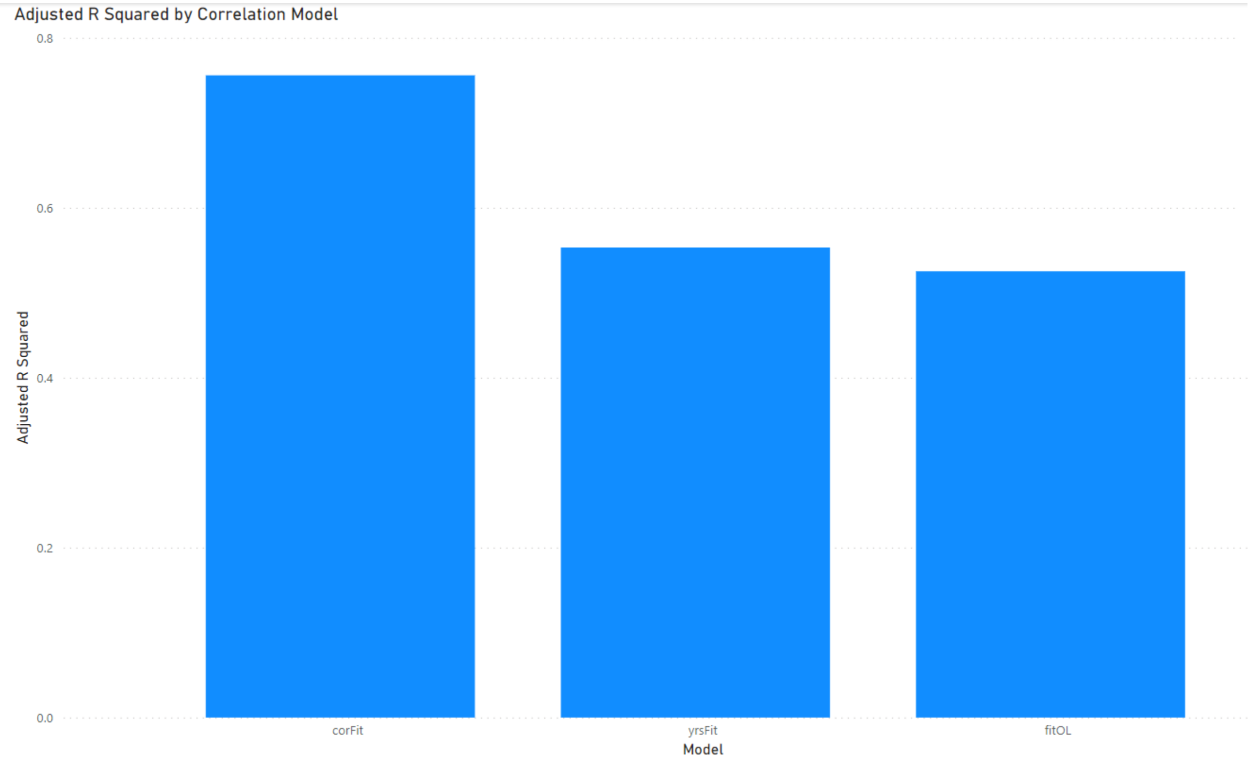
Summary of fitOLPower Model



```
Call:
lm(formula = teamCleaned.df$`St Cup` ~ poly(teamCleaned.df$OL,
    4) + teamCleaned.df$Yrs)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8209 -1.6365  0.1626  1.2990 12.2439

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -4.40474    1.42624  -3.088  0.00474 **
poly(teamCleaned.df$OL, 4)1 -5.47560    3.86418  -1.417  0.16835
poly(teamCleaned.df$OL, 4)2  1.39459    3.33960   0.418  0.67967
poly(teamCleaned.df$OL, 4)3  1.04491    3.37906   0.309  0.75961
poly(teamCleaned.df$OL, 4)4  0.76245    3.52102   0.217  0.83025
teamCleaned.df$Yrs           0.14893    0.02612   5.701 5.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.333 on 26 degrees of freedom
Multiple R-squared:  0.6022,    Adjusted R-squared:  0.5257
F-statistic: 7.873 on 5 and 26 DF,  p-value: 0.0001265
```

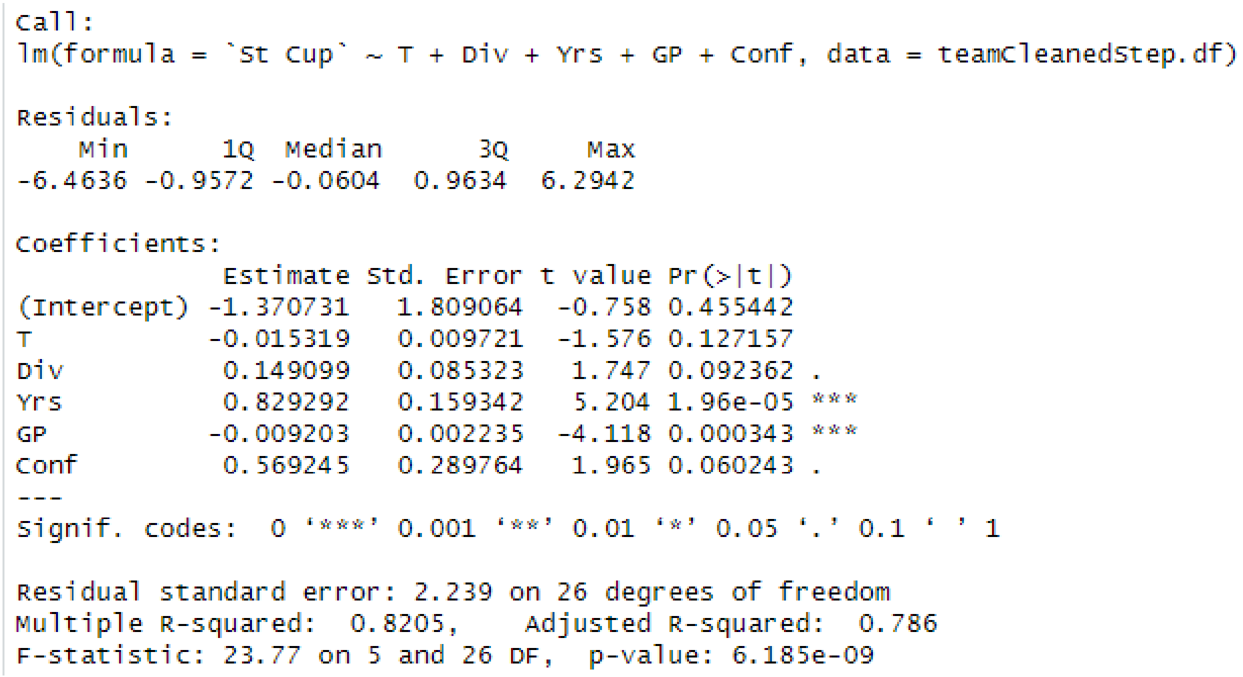
Appendix N

Visual of Adjusted R Squared of Correlation Based Models



Appendix O

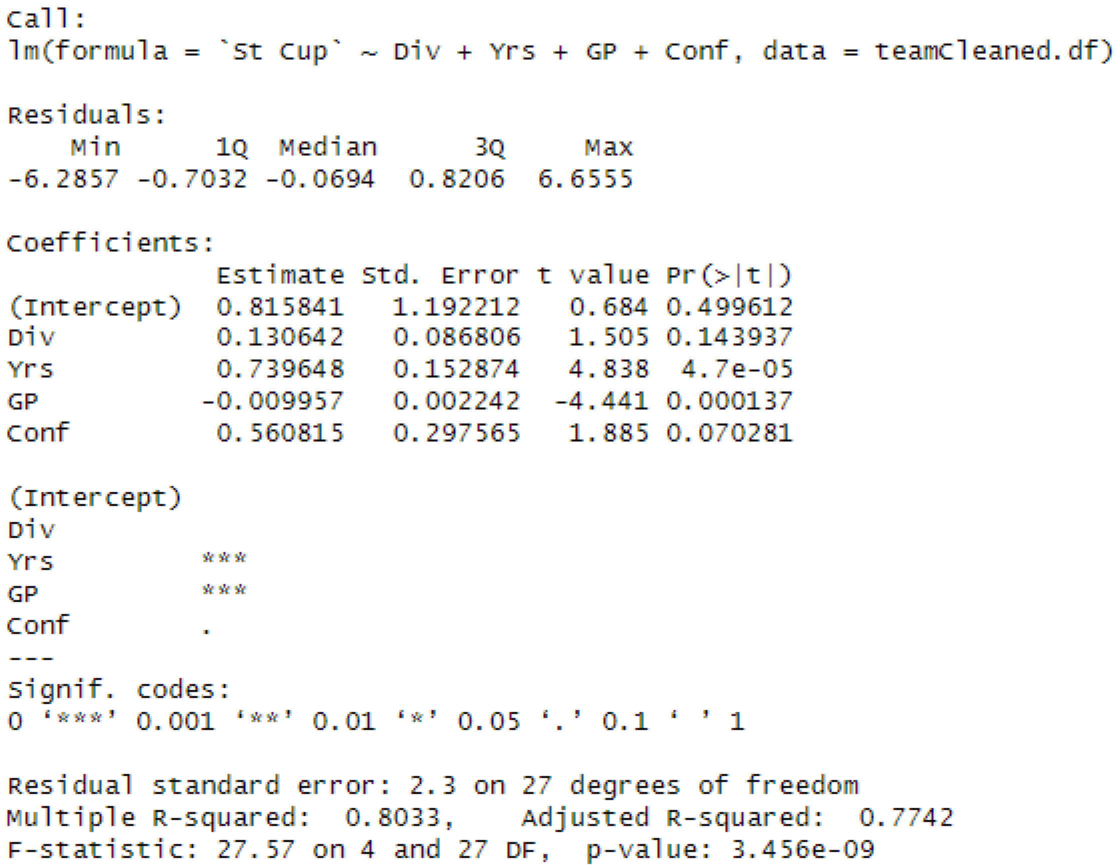
Summary of step.model1



```
Call:
lm(formula = `St Cup` ~ T + Div + Yrs + GP + Conf, data = teamCleanedStep.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4636 -0.9572 -0.0604  0.9634  6.2942 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.370731   1.809064  -0.758 0.455442    
T           -0.015319   0.009721  -1.576 0.127157    
Div          0.149099   0.085323   1.747 0.092362 .  
Yrs          0.829292   0.159342   5.204 1.96e-05 ***
GP          -0.009203   0.002235  -4.118 0.000343 ***
Conf         0.569245   0.289764   1.965 0.060243 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.239 on 26 degrees of freedom
Multiple R-squared:  0.8205,	Adjusted R-squared:  0.786 
F-statistic: 23.77 on 5 and 26 DF,  p-value: 6.185e-09
```

Appendix P

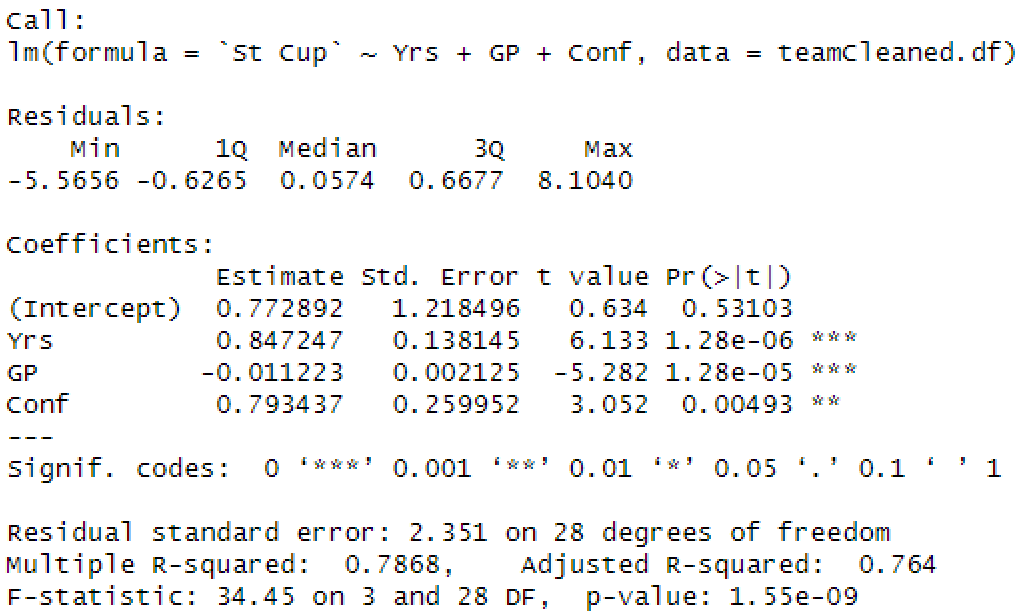
Summary of step.model2



```
Call:
lm(formula = `St Cup` ~ Div + Yrs + GP + Conf, data = teamCleaned.df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2857 -0.7032 -0.0694  0.8206  6.6555

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.815841   1.192212   0.684 0.499612
Div          0.130642   0.086806   1.505 0.143937
Yrs          0.739648   0.152874   4.838  4.7e-05
GP          -0.009957   0.002242  -4.441 0.000137
Conf         0.560815   0.297565   1.885 0.070281

(Intercept)
Div
Yrs         ***
GP          ***
Conf        .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.3 on 27 degrees of freedom
Multiple R-squared:  0.8033,    Adjusted R-squared:  0.7742
F-statistic: 27.57 on 4 and 27 DF,  p-value: 3.456e-09
```

Appendix Q

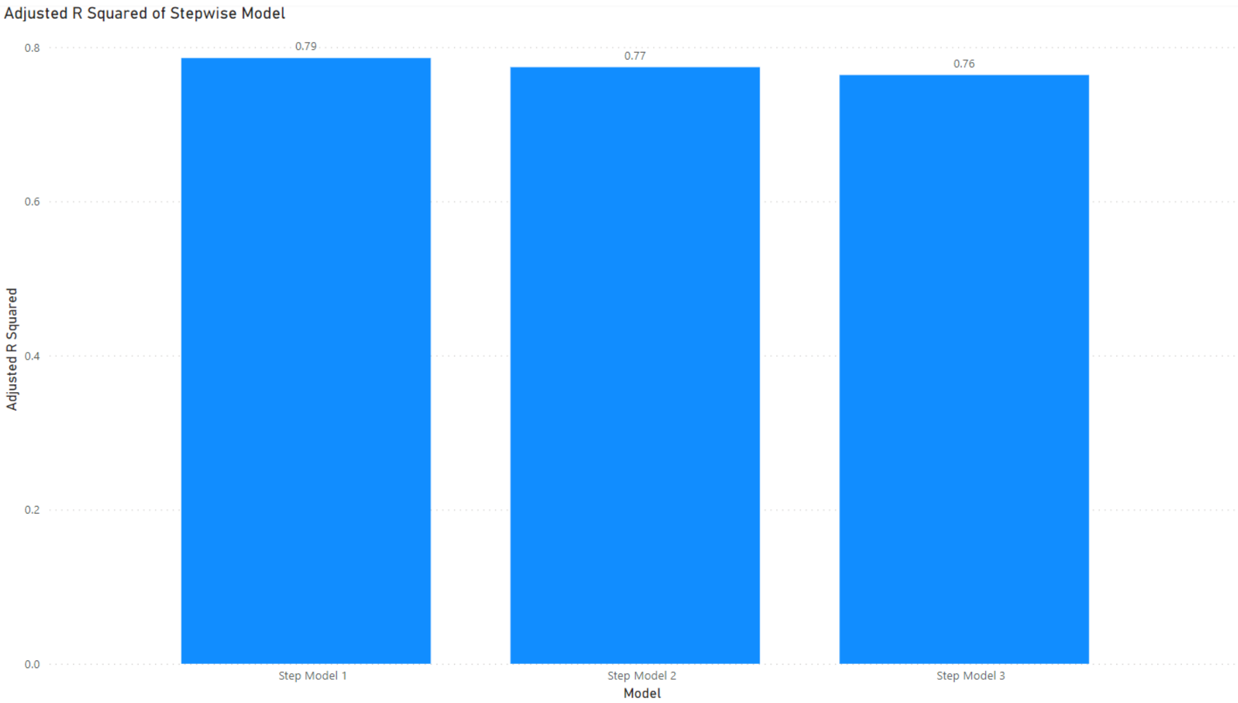
Summary of step.model3



```
Call:
lm(formula = `St Cup` ~ Yrs + GP + Conf, data = teamCleaned.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5656 -0.6265  0.0574  0.6677  8.1040 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.772892   1.218496   0.634  0.53103    
Yrs          0.847247   0.138145   6.133 1.28e-06 ***
GP          -0.011223   0.002125  -5.282 1.28e-05 ***
Conf         0.793437   0.259952   3.052  0.00493 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.351 on 28 degrees of freedom
Multiple R-squared:  0.7868,	Adjusted R-squared:  0.764 
F-statistic: 34.45 on 3 and 28 DF,  p-value: 1.55e-09
```

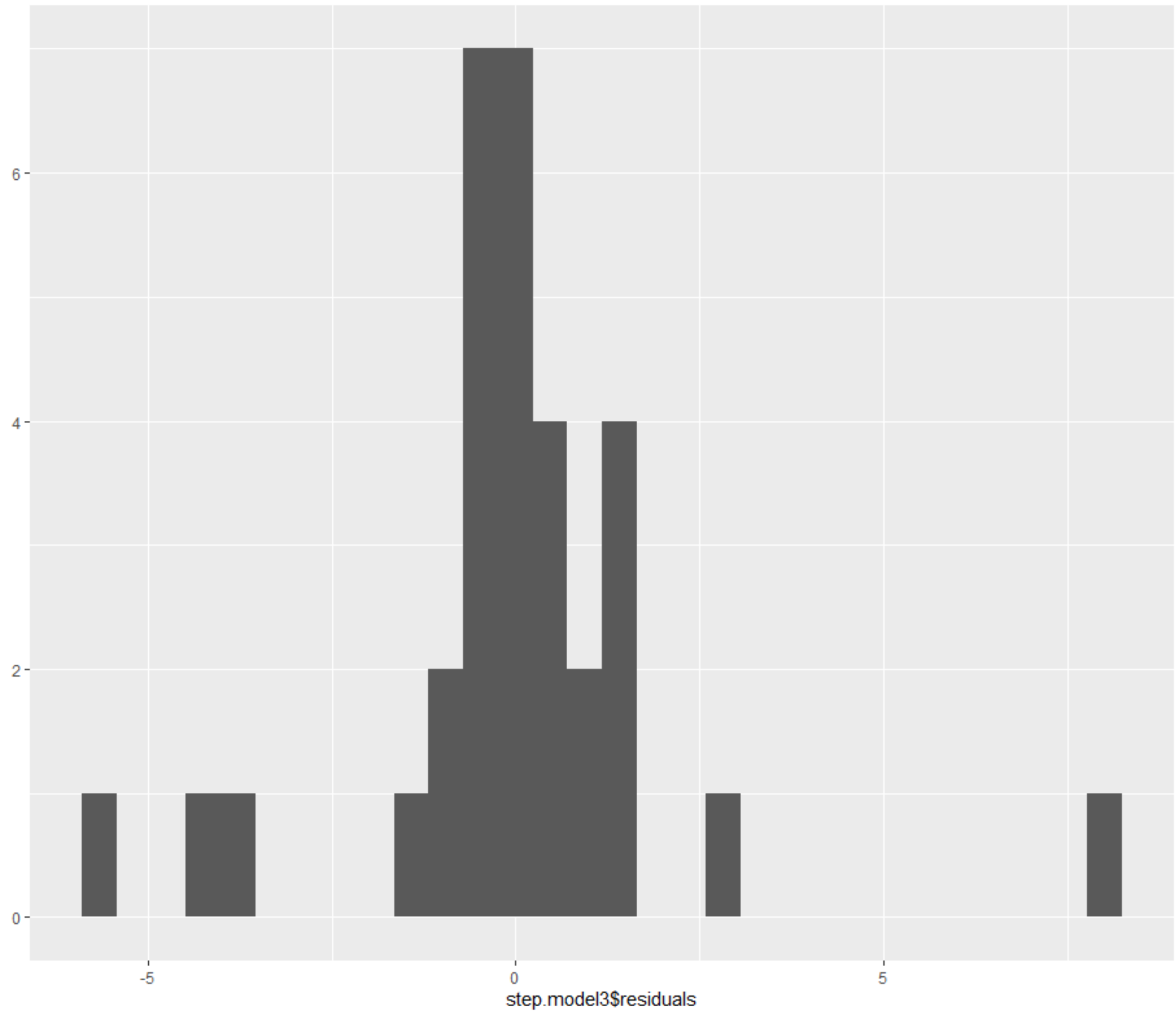
Appendix R

Visual of Adjusted R Squared of Machine Learning Based Models



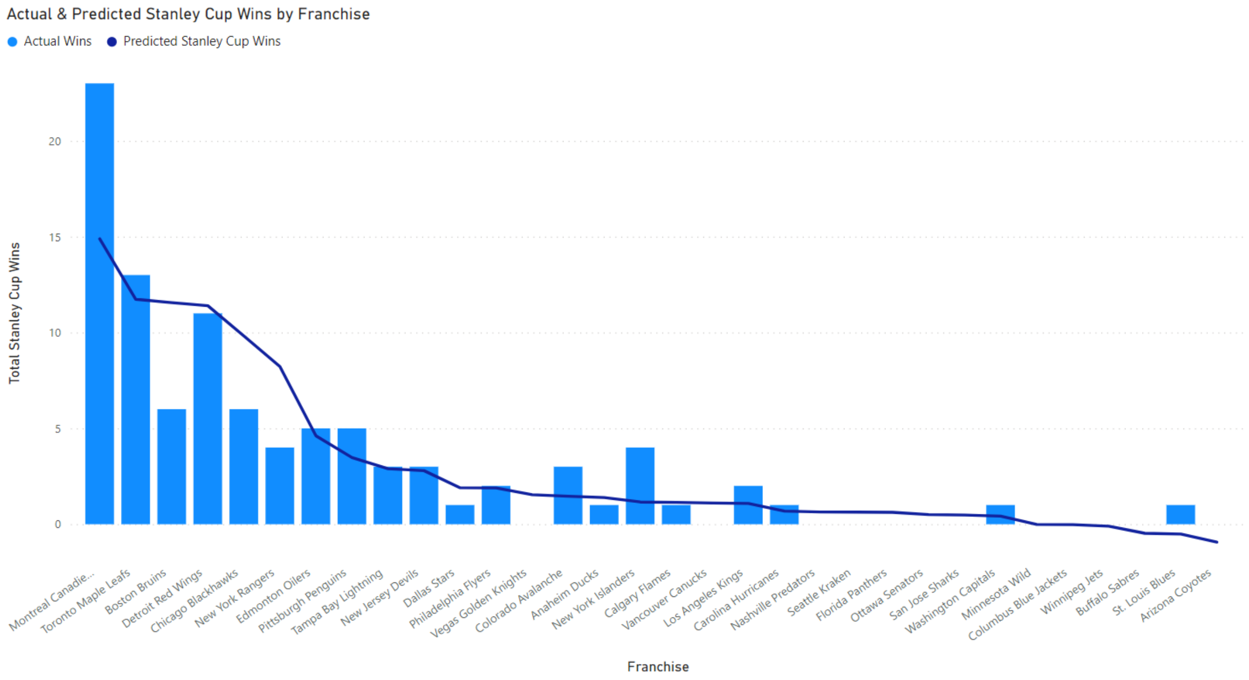
Appendix S

Residual Distribution from step.model3



Appendix T

Visualization of Actual Stanley Cup Wins Versus Predicted Stanley Cup Wins by our step.model3



References

2022-23 NHL skater statistics. Hockey. (n.d.). Retrieved April 14, 2023, from https://www.hockey-reference.com/leagues/NHL_2023_skaters.html

List of all the NHL teams &amp; other hockey teams. Hockey. (n.d.). Retrieved April 14, 2023, from https://www.hockey-reference.com/teams/#active_franchises

NHL. (2013). *Alex Ovechkin stats and news*. NHL.com. Retrieved April 14, 2023, from https://www.nhl.com/player/alex-ovechkin-8471214

NHL stats. NHL.com. (n.d.). Retrieved April 14, 2023, from https://www.nhl.com/stats/teams

Predict NHL player salaries. Kaggle. (n.d.). Retrieved April 14, 2023, from https://www.kaggle.com/datasets/camnugent/predict-nhl-player-salaries?select=train.csv