SDM Project: Establishing a Regression Model to Predict MLS Salaries

By: William Collins, Manuel Salgueiro Gentile, Raunak Ghosh, Koustav Dutta

NOVEMBER 7, 2021
UNIVERSITY OF SOUTH FLORIDA

Executive Summary

Based off the reported AIC and BIC values the reported "Best" model for predicting In-Field Player's salary was our glm model, which had a reported AIC of 3576.97 and a BIC of 3825.74. This was the model that best performed comparatively according to both AIC and BIC criteria. For Our base case for Season 2015 , on average players were paid 10.4% more in the year 2017 and 17.2% more in 2018 in comparison to 2015. While 2016 only showed a 4.1% increase in salary.

Our base Case for Club ATL , on average players were paid the highest in the following three clubs: MNUFC, NYCFC, and LA where the reported average salary was 33.5%,33.2%, and 31.5% respectively higher than Atlanta's average players salary. The lowest paid club on average was Chicago which reported a 5.8% lower salary than Atlanta.

Our Base Case for Position was W (Wildcard- multiple position players) , and on average players who specialized in one specific position were paid more compared to Wildcard players who played different positions in the field. Being a Midfielder and Forward were the best paid positions with 7.9% and 6.3% better salaries than Wildcard players. While defenders make on average 1.4% more than wildcard players. However, this was not found to be statistically significant

Our base case for DSP was not a DSP , on average players who were classified as designated players earned approximately 173.5% more than standard players.

Our base case for Conference was Eastern Conference , on average players who played for Western Conferences were found to make approximately 15.0% less than those who played in the eastern conference.

The goals effect for our model predicted that on average for every extra goal scored by a particular player his salary will increase by approximately 3.5%

The Red Card and Yellow Card Effect for our model predicted that on average players who have received additional red cards will get paid approximately 5% more. However, this is not statistically significant . On average players who have received additional yellow cards will get paid approximately 3.2% more . These effects may seem contradictory with the logic but when considering rc and yc received as a measure of aggressiveness, players who showcase higher aggressiveness in the field will get paid more. This claim is speculative.

The Fouls Committed Effect for our model on average each extra foul committed decreases the players salary by 0.5% .The Fouls Suffered Effect on average each extra foul received increases the players salary by .4% . With the reported average effect centered around 0 we can see how the effects of team assignment and season differs from the computational average. For example, seasonally 2015 had a lower random effect influence than that of the calculated average.

The QQ plot demonstrates reasonable normality within the data . A high VIF value was found for the Club variable however this IV was an essential predictor for this model and therefore deemed permissible Based on the residuals plot and the results of a Bartletts test (p-value < 2.2e-16) there appears to be clear evidence of heteroskedastic amongst the data. Since we are working with data tracked time(seasons) we expect there to be autocorrelation amongst the data.

Our base case for non-Designated Player Status, the designated player status has a 232% rise in player salary compared to non-designated players in our Mixed Effect Models for our goalkeeper model. The goals against average provides a 9% increment in goalkeeper salary in the mixed effect model. It is because the goalkeeper provides number of saves compared to average goals scored by the opposition . The loss factor provides a 5% increment in goalkeeper salary due to the number of saves provided per game even though the team may have lost the match. Western Conference goalkeepers have a 3.8% increment in salary compared to eastern conference goalkeepers. (We are interpreting the conference even though they are non-significant in our model.) The QQ plot demonstrates reasonable normality within the data . A high VIF value was found for the Club variable however this IV was an essential predictor for this model and therefore deemed permissible . Based on the residuals plot and the results of a Bartletts test (p-value < 2.2e-16) there appears to be clear evidence of heteroscedasticity amongst the data. Since we are working with data tracked time(seasons) we expect there to be autocorrelation amongst the data.

Problem Significance and Prior Literature

The interest in soccer—both in participation and popularity—is swelling in the United States. This fact is undeniable. According to a survey done by a Gallup poll in 2018, 7% of Americans say that soccer is their favorite sport to watch, while 9% said their favorite American sport was baseball[1]. As soccer bridges the gap in popularity within American sports enthusiasts its important to note that it was not always like this. In 1996, when the MLS soccer league first was established, it only had 10 teams, and was a failing business losing approximately 350 million dollars from 1996-2004.Of the ten teams that started in 1996 only 8 were left by 2001, two teams having been disbanded due to failing interest within the league. The inevitable disbandment of the entire league seemed to be all but assured as Major League Soccer was suffering from a lack of interest in all its teams during this early period. MLS unlike other American professional sports was unique in that one business entity, the league, owned all the teams that participated within it, leaving investors as the ones to 'run' the franchise teams. Therefore, poor attendance for one team was bad for all teams. Likewise, rules regarding team management that applied to one team also applied to all teams, which became detrimental early

[1] LoRé, M. (2019, April 30). Soccer's Growth In U.S. Has International Legends Buzzing. Retrieved from https://www.forbes.com/sites/michaellore/2019/04/26/soccers-growth-in-u-s-has-international-legends-buzzing/?sh=73b7020c17f1

on to this soccer league. It wasn't until 2007 that this approach to league-based management was changed when the MLS realized that the stifled popularity of the sport was not for a lack of interest within the American community towards soccer but rather a lack of interest towards the players on American teams. In the past the rules that the league set for all teams included things such as salary caps on how much each team will have to grant to each of their players. This helped the league limit it's expenses on players and theoretically promoted parity in the league as all teams have equal resources to acquire player talent (Coates, Frick, and Jewell 2016). But this had to change as the league was faced with an issue of recruitment, as desirable players didn't want to leave the European leagues to play for the lower-paying American leagues, so in 2007 the MLS instituted the Designated Player (DP) rule. This Designated Player rule allowed for teams to exclude three players from their team's standard salary cap restrictions. In 2007, David Beckham signed with the Los Angeles Galaxy, being one of the first all-star players to be recruited into MLS and one of the first players to have the DP rule applied. This change would signify a major shift in popularity and interest in MLS.

Nowadays there are 27 teams that participate within the American MLS with plans to further expand the number of teams to 30 by 2023.[2] In addition, The 24 MLS teams in existence during the 2019 season produced **$1.1 billion in revenue**, including distributions from the league. While in 2020 there was an estimated **$468 million** for 26 teams, the decline in revenue hypothesized to be primarily attributed to Covid.[3] Since instituting the Designated Player Rule, MLS has exploded and now a large portion of team budgets are dedicated to providing not only adequate but generous compensation to talented players either within the salary cap or beyond it using the DP rule. This report's primary purpose then is to explore how a player's stats will affect his level of salary, and more importantly identify which stats are most important in determining salary for a soccer player.

[2] Serrano, R. (2021, October 13). The 30th MLS expansion team will be announced next year. Retrieved from https://en.as.com/en/2021/10/14/soccer/1634162586_189727.html

[3] Badenhausen, K. (2021, July 20). Los Angeles FC Tops Sportico's 2021 MLS Valuations at $860 Million. Retrieved from https://www.sportico.com/valuations/teams/2021/mls-team-valuations-2021-1234634303/

Data Source

For our project we gathered our data from three sources. First, we collected the salary information from the Wolfram Data Repository. This dataset contains Major League Soccer players' salaries from 2007 to 2018. This file, besides containing the information regarding the players' first name, last name and position to uniquely identify each observation, also includes two columns for the Base Salary and the Guaranteed Compensation. These two variables are the core of our analysis depicting the compensation levels for each player measured in dollars ($). The file also shows a column depicting which club has the contractual responsibilities for each player and another column indicating the year for that data point.

The second dataset for our analysis focuses on data about the performance of MLS players and was found in Kaggle. The dataset includes multiple files, for our analysis we took only the all_players and all_goalkeepers files which included performance information corresponding to on-field players (strikers, midfielders and defenders) and another one which only focuses on performance indicators particular to goalkeepers. This has been done purposely since goals would be a good indicator of performance for a striker, but goalkeepers will rarely have any opportunity to score, so a much better indicator of their performance would be saves. The data spans from 1996 to 2020, showing the performance information for each player in the specific year indicated. Each specific performance indicator included is measured in different ways, most involved count of occurrences during that season such as goals or saves, other columns include a ratio or percentage as the type of measurement such as shot/conversion rate(goals/shots).

Our third data source involves the team data and was collected from American Soccer Analysis. It provides information about the collective performance of each team in the league with important attributes like total shots for the team, total goals, points in the league, etc. The purpose of adding this extra level of information is to better understand the performance of each team in the league and how each player impacts and contributes to the team. The team performance is measured through the count of different attributes such as the variables previously mentioned.

Date Preparation Methodology

Considering our analysis, we decided to subset for multiple years of these datasets which are available in both data sources. For this we selected the seasons of 2015-2018, as they are the most recent shared observations for our data sources. These four-year periods will provide some time notion on how the MLS changed season by season. Similarly, for the performance data we decided to subset the observations to only include regular season data, excluding from our scope the playoffs information, since this will create unbalanced data in terms of the number of games played for each team. If we consider the regular season and playoffs as well, some players from teams that classified for playoff games during those years will have higher goals and games won just because they had extra games compared to teams who just disputed the regular season. To avoid this and provide a fair and consistent judgement to evaluate performance across all players we have only selected the regular season games, which will give each player the same number of games than all of their conference competitors.

For our purpose of assessing how performance of MLS affects the salary component of these players, we focused on using the Base Salary as our dependent variable and the different individual performance indicators as well as team information for additional independent variables which would allow us to measure the impact of each of these components on the salary of the soccer players for the MLS League.

After downloading the corresponding files, we merged the salary dataset with the performance data. For each year in the Salary dataset, we appended first and last name to create the Full Name variable which would allow us to fetch and match each observation with its corresponding row in the Performance dataset. Although the process matched most of the observations, approximately 60 rows each year showed NA values for which we had to manually look at the observation and search for any difference in the syntax of the name like extra spaces, added middle name or any other difference in the Name that resulted in the Name column values not matching. These observations per each year, most of which were junior academy players and did not have significant salaries, were dropped accordingly, given the names mismatching to any observation in the performance dataset.

Similarly, the team data was merged to our complete dataset by matching the team for each player in the performance data with the team's attributes. We also performed the appropriate checks on the data to ensure a reasonable number of players for each season, club, position, etc. All these steps were performed for each year both for the goalkeeper data and the players data.

Variable Choice and Selection

Table of Effects (Predictors)(Players)

Predictor Variable	Expected Sign of effect	Rationale
Goals Scored	+	Having more goals leads to a positive effect on the player's salary as the club will want to keep the player as more goals means more matches won.
Assists	+	Having more assists leads to a positive effect on the player's salary as the club will want to keep the player as more assists means more goals which means more matches won.
Games Played	+	Playing more games leads to a positive effect on a player's salary as the club will want to keep the player as more games played means he is helping the club win more matches.
Minutes Played	+	Playing more minute's leads to a positive effect on a player's salary as the club will want to keep the player as more minutes played means he is helping the club win more matches.
Games Started	+	Starting more games leads to a positive effect on a player's salary as the club will want to keep the player as more games started means he is helping the club win more matches.
Shots on Goal	+	Having more shots on goal leads to a positive effect on the player's salary as the club will want to keep the player as more shots on goal means there is a higher possibility of more goals which means more matches won.
Goals Per 90 Minutes	+	Having more goals per 90 minutes leads to a positive effect on the player's salary as the club will want to keep the player as more goals per 90 minutes means more goals which means more matches won.
Shot Conversion Percentage	+	Having a better Shot Conversion Percentage leads to a positive effect on the player's salary as the club will want to keep the player as a better Shot Conversion Percentage means more goals which means more matches won.
Assists per 90 mins	+	Having more assists per 90 mins leads to a positive effect on the player's salary as the club will want to keep the player as more assists per 90 mins means more goals which means more matches won.
Fouls Committed	+/-	Having more fouls committed leads to a negative effect on salary as more fouls committed means more yellow and red cards which can cause a team to lose matches. This can also have a positive effect based on what position the player is playing as in some cases more fouls committed by defenders can lead to less goals scored by the opposing teams attackers.

Table of Effects (Predictors)(Player and GoalKeeper(GK))

Predictor Variable	Expected Sign of effect	Rationale
Yellow Card	-	Having more yellow cards leads to a negative effect on salary as more yellow cards means more red cards and worse performance of a team which can cause a team to lose matches.
Red Card	-	Having more red cards leads to a negative effect on salary as more red cards means worse performance of a team as the team would have to play one man down which can cause a team to lose matches.
Saves Made(GK)	+	Having more saves made leads to a positive effect on the player's salary as the club will want to keep the player as more saves made means more goals saved which means more matches won.
Goals Against Average	-	Having a higher Goals Against Average leads to a negative effect on the player's salary as the club will want to not keep the player with a higher Goals Against Average which means more matches lost.
Penalty kicks not saved/Penalty Kicks against	-	Having a higher Penalty kicks not saved leads to a negative effect on the player's salary as the club will want to not keep the player with a Penalty kicks not saved which means the goalkeeper with less saves will give away more goals which mean more matches lost.(This will not be included in our model because of inconsistent data)
Clean Sheets(GK)	+	Having more clean sheets leads to a positive effect on the player's salary as the club will want to keep the player with more clean sheets which means the goalkeeper with more games with no goals which means more matches won.
Save Percentage(GK)	+	Having a higher save percentage leads to a positive effect on the player's salary as the club will want to keep the player with a higher save percentage which means the goalkeeper with a higher save percentage will save more goals which means more matches won.
Designated Player	+	Being a Designated player will have a positive effect on the players salary as the designated players are the special players on each team and they are paid extra as they are usually a class above the rest.

Table of Effects (Predictors)(Teams)

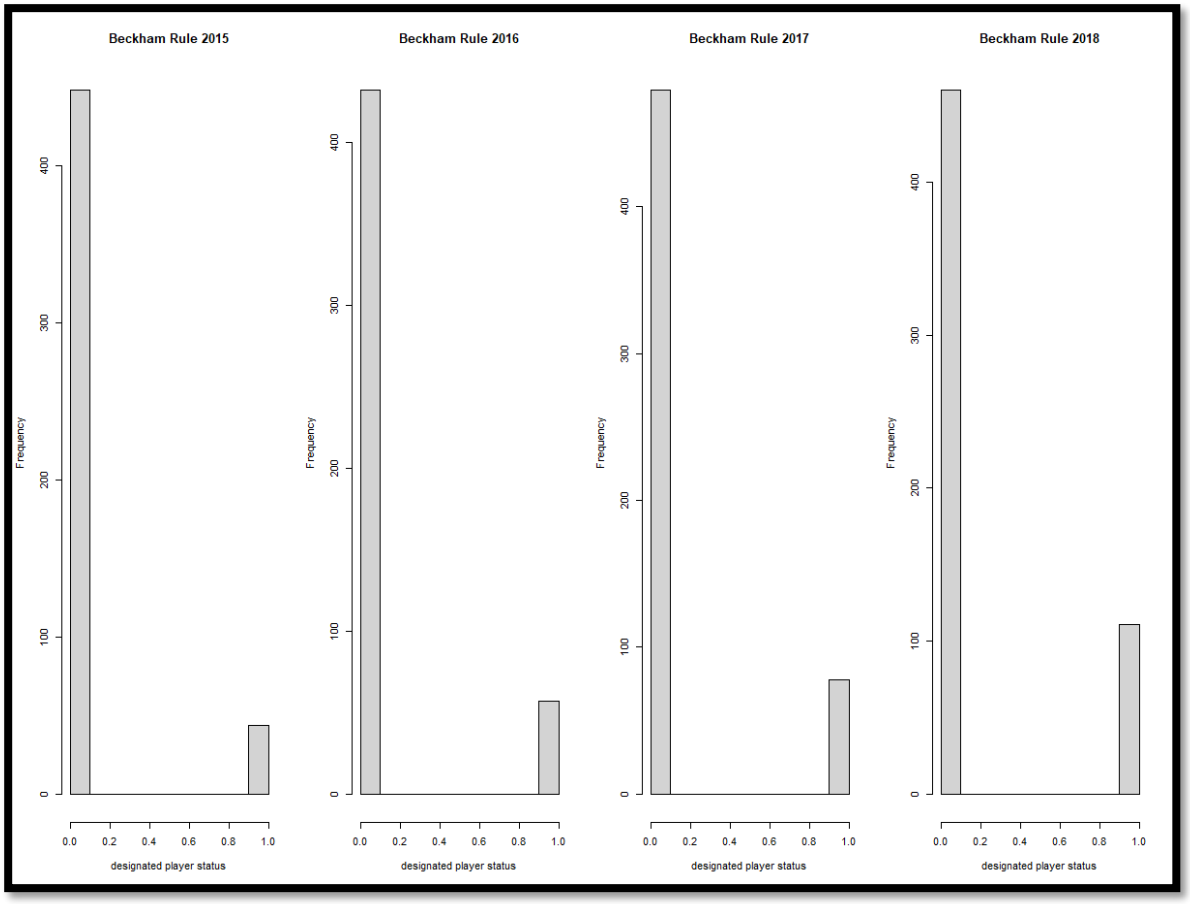
Predictor Variable	Expected Sign of effect	Rationale
Team (Team Name)	NA	We don't expect the team name to affect the salary of the players.
Season (Year of Season)	NA	We don't expect the season to affect the salary of the players.
Games (Total Games Played that Season)	+	Playing more games leads to a positive effect on a player's salary as the club will want to keep the player as more games played means he is helping the club win more matches.
ShtF (Shots For)	+	Having more shots for leads to a positive effect on the players salary as the club with more shots for will usually have more goals which leads to more matches won which in turn leads to a better finish in the standings which will lead to the club getting more money which means they will be able to afford to pay their players more
ShtA (Shots Against)	-	Having more shots against leads to a negative effect on the players salary as the club with more shots against will usually have more goals scored against them which leads to more matches lost which in turn leads to a worse finish in the standings which will lead to the club getting less money which means they will be not able to afford to pay their players more
GF (Total Goals For)	+	Having more Total Goals For leads to a positive effect on the players salary as the club with more Total Goals For will usually have more goals which leads to more matches won which in turn leads to a better finish in the standings which will lead to the club getting more money which means they will be able to afford to pay their players more
GA (Total Goals Against)	-	Having more Total Goals Against leads to a negative effect on the players salary as the club with more Total Goals Against will usually have more goals scored against them which leads to more matches lost which in turn leads to a worse finish in the standings which will lead to the club getting less money which means they will be not able to afford to pay their players more

Table of Effects (Predictors)(Cont.)(Teams)

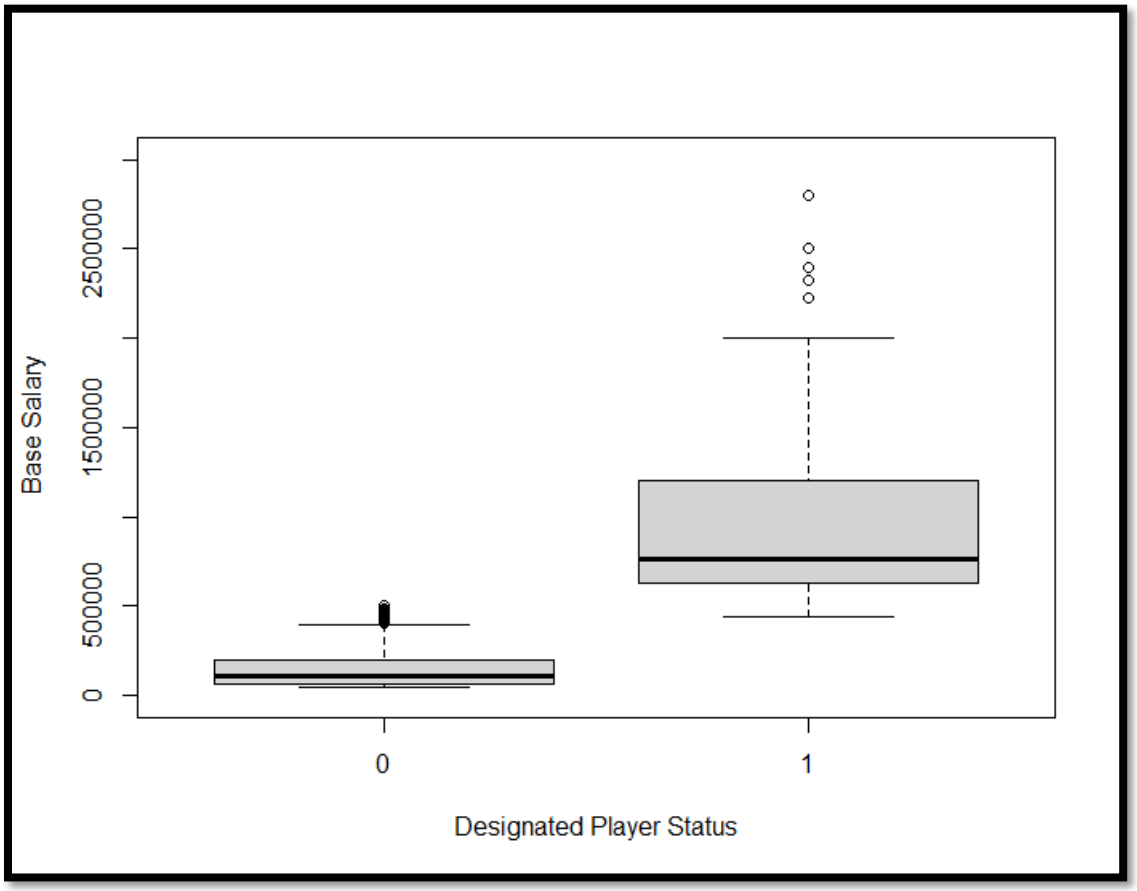
Predictor Variable	Expected Sign Of Effect	Rationale
xGF (Expected Total Goals For, given the same sample of shots over 1000 simulations)	N/A	The expected total goals for should not affect the actual salary of the players as the expected value won't determine the team's standings which determine their income and popularity. The variable that matters in total goals for not expected total goals for.
GD (Goal Difference)	+/-	Having a positive Goal difference will usually lead to a positive effect on the players salary, however, it is not guaranteed and likewise having a negative goal difference will usually lead to a positive effect on the players salary, however, it is not guaranteed as the goal difference might be positive because the club might have scored many goals in a few games but lost the majority of the games and vice versa for the negative goal difference
xGA (Expected Total Goals Against, given the same sample of shots over 1000 simulations)	N/A	The expected total goals against for should not affect the actual salary of the players as the expected value won't determine the team's standings which determine their income and popularity. The variable that matters in total goals against not expected total goals against.
xPts (Expected Total Points, given the same sample of shots over 1000 simulations)	N/A	The expected total points for should not affect the actual salary of the players as the expected value won't determine the team's standings which determine their income and popularity. The variable that matters in total points for not expected total points.
Pts (Total Points earned this Season)	+	Having more points leads to a positive effect on the players salary as the club with more points leads to a better finish in the standings which will lead to the club getting more money which means they will be able to afford to pay their players more

Descriptive Analysis and Data Visualization

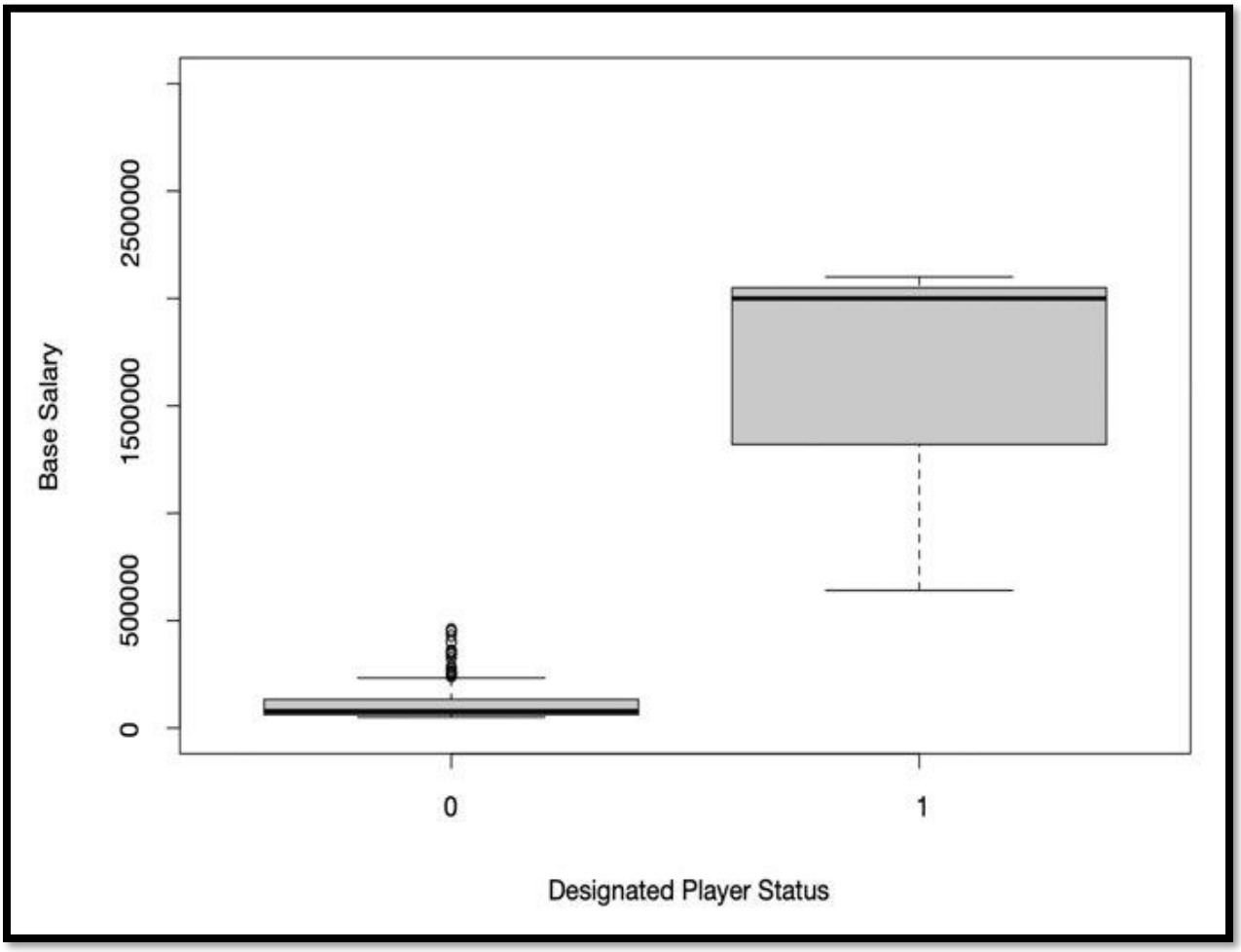
Designated Players vs. "Normal Players"



Seen in the above histogram plots there is an increasing trend in the number of designated players (coded as '1') present within the league. This is explained by two reasons: more teams are entering the league (ex. MNUFC started playing games in 2017 LAFC started playing games in 2018) and more teams are utilizing all of their Designated Player spots.



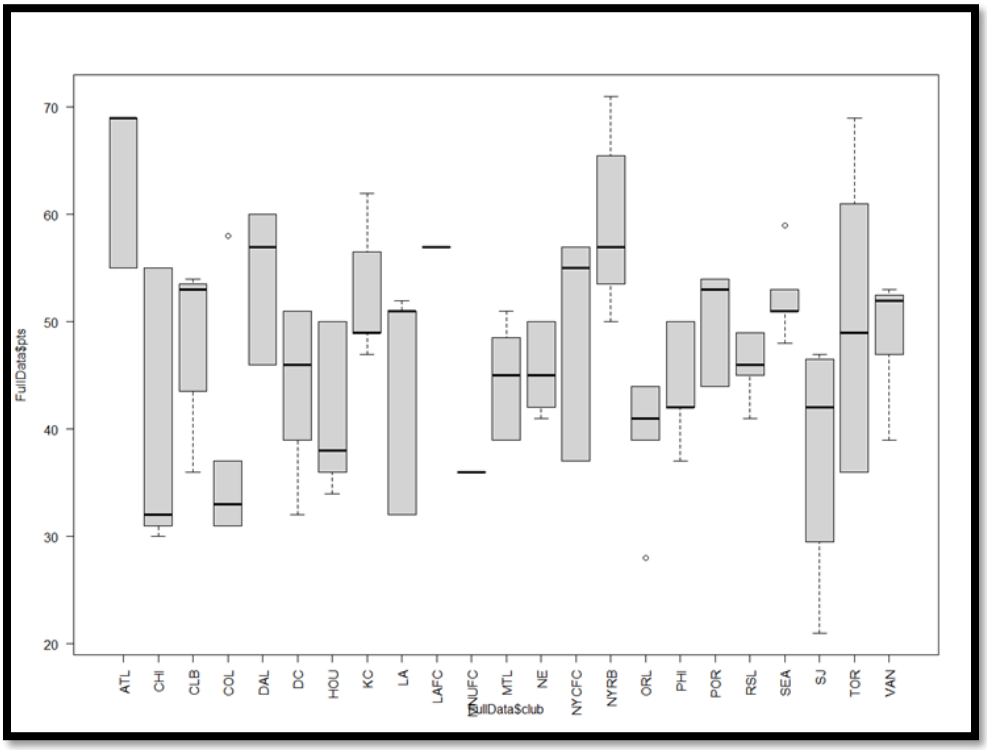
The featured Box-Plot above illustrates the significant salary disparity between Designated and Non-Designated players (goal keepers not included). Overall, the plot suggests that Designated Players had a higher IQR than regular players and that there is over a 1-million-dollar difference in average base salary between regular and designated players.



The featured Box-Plot above illustrates the significant salary disparity between Designated and Non-Designated players within goalkeepers alone. Overall there was a greater reported IQR for Designated Players than regular players, and this effect was even more pronounced in goalkeepers than in the previous plot which included in-field players.The current plot suggests that there is over a 2-million-dollar difference in average base salary between regular and designated players. However, this data is not a good representation as there is only 2 designated players (Tim Howard and Brad Guzan) for goalkeeper and they make significantly more than an average goalkeeper in the MLS.

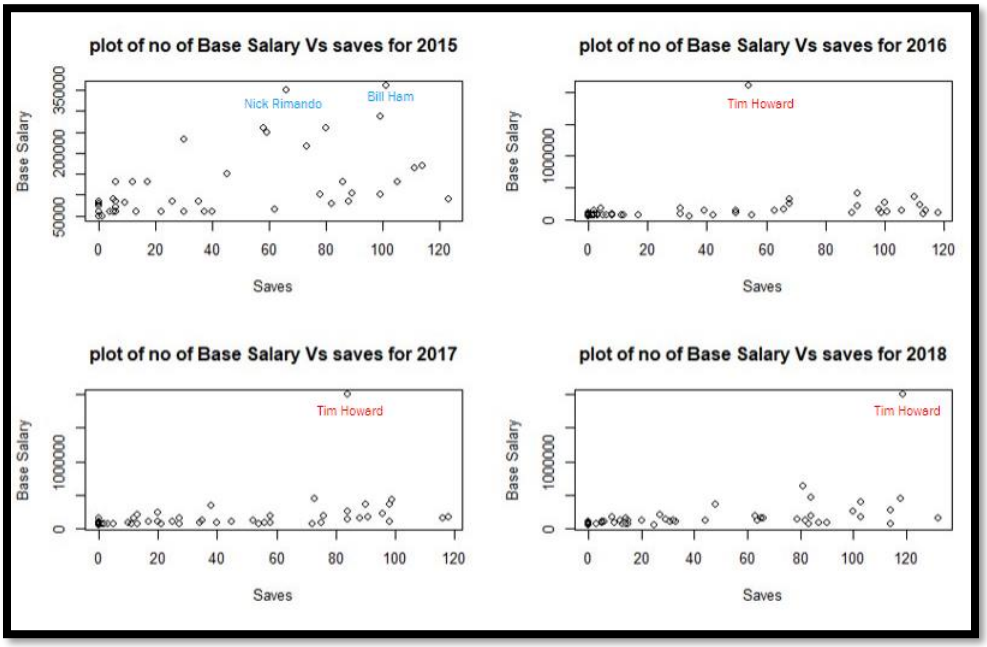
Players and Team Performances Visualized

Points Distribution By Club



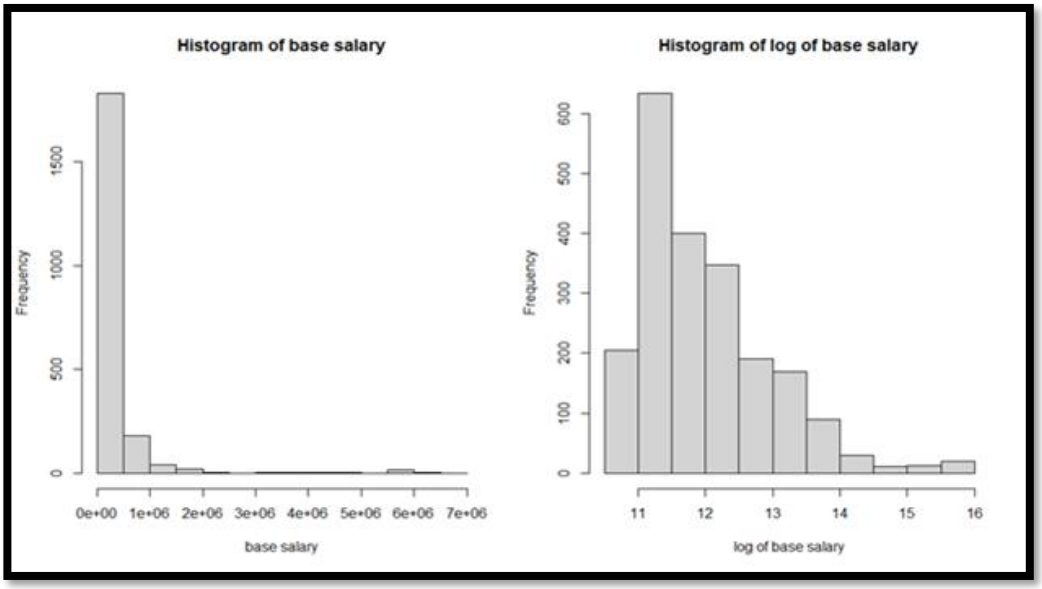
Based on the above boxplot of Club vs. Total Team Points, Atlanta's Club had the highest reported average points out of any team.Chicago on average had the lowest reported average points out of any team. An additional note is that LAFC and MNUFC had one and two playing seasons, respectively which explains the low variation within the data points.

Plot of Base Salary Vs Saves (Goalkeeper Data)



The above plots are of Base Salary vs Saves within our goalkeeper subset across the seasons. It is illustrated that from 2016 till 2018, the designated player, Tim Howard, was the highest paid goalkeeper by a significant amount. In 2015, there were no designated players.

Base Salary vs. Log of Base Salary



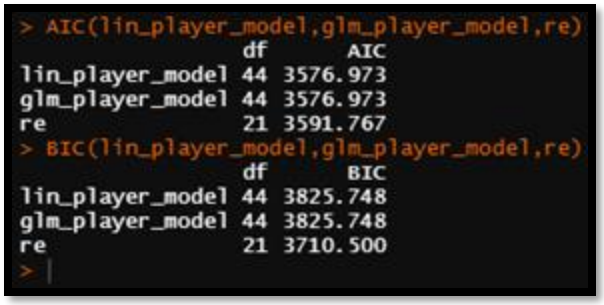
The histogram of base salary is on the left and the histogram of log of base salary is on the right. From the left plot its evident that the data is heavily right-skewed however when we apply the log transformation to base salary it becomes more normalized. When model building the log of base salary will be used.

Models

In our analysis for in-field players, we tried three different models as showcased below by the stargazer output including an OLS regression, a generalized linear model (GLM) and Mixed Effects model.

	Dependent variable:		
	log(base.salary)		
	OLS	*normal*	*linear mixed-effects*
	(1)	(2)	(3)
season2016	0.041 (0.037)	0.041 (0.037)	
season2017	0.104*** (0.037)	0.104*** (0.037)	
season2018	0.172*** (0.041)	0.172*** (0.041)	
clubCHI	-0.058 (0.108)	-0.058 (0.108)	
clubCLB	0.192* (0.106)	0.192* (0.106)	
clubCOL	-0.020 (0.118)	-0.020 (0.118)	
clubDAL	0.203* (0.106)	0.203* (0.106)	
clubDC	0.104 (0.107)	0.104 (0.107)	
clubHOU	0.215* (0.110)	0.215* (0.110)	
clubKC	0.120 (0.108)	0.120 (0.108)	
clubLA	0.315*** (0.108)	0.315*** (0.108)	
clubLAFC	0.222 (0.146)	0.222 (0.146)	
clubMNUFC	0.335*** (0.129)	0.335*** (0.129)	
clubMTL	0.042 (0.103)	0.042 (0.103)	
clubNE	-0.023 (0.106)	-0.023 (0.106)	
clubNYCFC	0.332*** (0.106)	0.332*** (0.106)	
clubNYRB	0.065 (0.104)	0.065 (0.104)	
clubORL	-0.008 (0.108)	-0.008 (0.108)	
clubPHI	-0.024 (0.105)	-0.024 (0.105)	
clubPOR	0.175 (0.107)	0.175 (0.107)	
clubRSL	0.141 (0.111)	0.141 (0.111)	
clubSEA	0.246** (0.109)	0.246** (0.109)	
clubSJ	0.279** (0.117)	0.279** (0.117)	
clubTOR	0.243** (0.098)	0.243** (0.098)	
clubVAN	0.195* (0.108)	0.195* (0.108)	
positionD	0.014 (0.052)	0.014 (0.052)	0.012 (0.052)
positionF	0.063 (0.055)	0.063 (0.055)	0.069 (0.054)
positionM	0.079 (0.050)	0.079 (0.050)	0.081 (0.050)
g	0.035* (0.020)	0.035* (0.020)	0.037* (0.020)
a	0.003 (0.006)	0.003 (0.006)	0.003 (0.006)
mins	0.0002*** (0.00003)	0.0002*** (0.00003)	0.0002*** (0.00003)
fc	-0.005*** (0.002)	-0.005*** (0.002)	-0.006*** (0.002)
fs	0.004*** (0.001)	0.004*** (0.001)	0.004*** (0.001)
yc	0.032*** (0.009)	0.032*** (0.009)	0.032*** (0.009)
rc	0.050 (0.032)	0.050 (0.032)	0.050 (0.032)
sc.	0.002 (0.001)	0.002 (0.001)	0.002 (0.001)
Designatedplayerstatus1	1.735*** (0.040)	1.735*** (0.040)	1.734*** (0.040)
ConferenceID1	-0.150*** (0.045)	-0.150*** (0.045)	-0.048 (0.034)
shtf	0.0002 (0.0004)	0.0002 (0.0004)	0.0005 (0.0004)
pts	-0.003 (0.002)	-0.003 (0.002)	-0.003* (0.002)
tsc	0.006 (0.012)	0.006 (0.012)	0.014 (0.010)
pgtr	-0.004 (0.010)	-0.004 (0.010)	-0.004 (0.010)
Constant	11.101*** (0.293)	11.101*** (0.293)	11.097*** (0.219)
Observations	2,109	2,109	2,109
R^2	0.682		
Adjusted R^2	0.676		
Log Likelihood		-1,745.487	-1,774.883
Akaike Inf. Crit.		3,576.973	3,591.767
Bayesian Inf. Crit.			3,710.500
Residual Std. Error	0.559 (df = 2066)		
F Statistic	105.692*** (df = 42; 2066)		

We selected our GLM model as our best model, considering it resulted in good values both for the AIC and BIC comparison among models as referred to in the image below:



Considering our best model as the GLM model, the following interpretations for the coefficients shown above are presented:

For the season effects, we must consider that the baseline is season 2015, which is the first season of data in our analysis. Knowing this, the coefficients are 0.041, 0.104 and 0.172 for years 2016, 2017 and 2018 respectively. This means that on average players were paid 10.4% more in the year 2017 and 17.2% more in 2018 in comparison to 2015. While 2016 only showed a 4.1% increase in salary compared to 2015 baseline. This illustrates the fact that the base salary for MLS outfield players has been continuously increasing season by season.

As far as the club effects on salary, we have selected Atlanta FC as our baseline level, considering they do have average numbers regarding their base salaries when compared to other MLS teams. The above shown coefficients highlight that on average players were paid the highest in MNUF, NYCFC and LA where the reported average salary for these clubs was respectively 33.5%, 33.2% and 31.5% higher than Atlanta's average players salary. Similarly, we can note that the lowest paid club on average was Chicago which reported a 5.8% lower salary than Atlanta.

For the position, we have selected Wildcard as our baseline level. This level was created referring to all the different players who have multiple positions listed as their position values. For instance, we have players on our data that can perform both as Defender and Midfielder or Midfielders that can adopt Forward roles in the game. At first instance, this flexibility could have been interpreted as a positive aspect for salary, but the model shows that on average players who specialized in one specific position were paid more compared to Wildcard players who played different positions in the field. Being a Midfielder and Forward were the best paid positions with 7.9% and 6.3% better salaries than Wildcard players who did not have specific defined position. Similarly, defenders make on average 1.4% more than Wildcard players. However, this was not found to be statistically significant, so our interpretation would not be considered as much valid.

For Designated Players effect on salary, we expected and found significant relationship between these two components. Our base case for Designated Players is 0, which means not being identified as a Designated Player. The coefficient was 1.175, which means that on average players who were classified as designated players earned approximately 173.5% more than standard players. This makes sense as being classified as Designated Player depends on how the salary of that player is.

As far as each Conference effect on salary, the results demonstrate that there is a difference between salaries in each MLS Conference. Considering the base case or baseline level as 0 being Eastern Conference, we can assert that on average players who played for Western Conference were found to make approximately 15.0% less than those who played in the Eastern conference.

Another important effect to consider is the Goals effect on salary. As shown by the 0.035 coefficient, on average for every extra goal scored by a particular player, will increase it salary by approximately 3.5%.

The Fouls Committed effect on salary presented by the model indicated that on average each extra foul committed by the player decreases the players salary by 0.5%. Similarly, the Fouls Suffered effect on salary displayed that on average each extra foul received by a player increases the players salary by 0.4%. These two effects proved to be remarkably interesting as they indicate how players should behave in the field to maximize their compensation with detailed actions such as fouls.

Lastly, some interesting coefficients to consider are the Red Card and Yellow Card effect on salary. As showed by the results, on average players who have received an additional red card will get paid approximately 5% more. Similarly, on average for each additional yellow card a player receives, he will get paid approximately 3.2% more. These effects may seem contradictory with the logic but when considering red cards and yellow cards received as a measure of aggressiveness, players who showcase higher aggressiveness in the field will get paid more according to the model. This claim is speculative, as the values show no statistical significance for these estimates.

GoalKeeper Models:

In our analysis for GoalKeepers, we tried two different models as showcased below by the stargazer output including a generalized linear model (GLM) and Mixed Effects model.

	Dependent variable:	
	log(base.salary)	
	linear mixed-effects	*normal*
	(1)	(2)
season2016		0.036 (0.077)
season2017		0.108 (0.074)
season2018		0.170** (0.076)
clubCHI		0.010 (0.205)
clubCLB		-0.043 (0.207)
clubCOL		0.250 (0.202)
clubDAL		0.075 (0.205)
clubDC		0.191 (0.194)
clubHOU		0.031 (0.211)
clubKC		-0.109 (0.225)
clubLA		-0.002 (0.210)
clubLAFC		-0.307 (0.271)
clubMNUFC		0.023 (0.240)
clubMTL		0.047 (0.199)
clubNE		-0.093 (0.206)
clubNYCFC		-0.024 (0.203)
clubNYRB		0.230 (0.212)
clubORL		0.023 (0.195)
clubPHI		-0.026 (0.204)
clubPOR		-0.072 (0.211)
clubRSL		0.262 (0.208)
clubSEA		0.008 (0.212)
clubSJ		-0.089 (0.214)
clubTOR		0.017 (0.195)
clubVAN		0.193 (0.209)
ga	-0.011 (0.009)	-0.011 (0.010)
sv	0.001 (0.003)	0.001 (0.004)
gaa	0.090*** (0.030)	0.090*** (0.031)
w	0.017 (0.021)	0.013 (0.022)
l	0.048* (0.027)	0.046 (0.030)
t	0.025 (0.021)	0.029 (0.023)
sho	0.039 (0.027)	0.043 (0.028)
Designatedplayerstatus1	2.326*** (0.205)	2.147*** (0.237)
ConferenceID1	0.038 (0.051)	0.044 (0.081)
Constant	11.050*** (0.052)	10.927*** (0.173)
Observations	256	256
Log Likelihood	-125.660	-112.122
Akaike Inf. Crit.	277.320	294.243
Bayesian Inf. Crit.	323.407	

Base Case for non-Designated Player Status :

The designated player status has a 232% rise in player salary compared to non-designated players in our Mixed Effect Models

The goals against average provides a 9% increment in goalkeeper salary in the mixed effect model. It is because the goalkeeper provides number of saves compared to average goals scored by the opposition

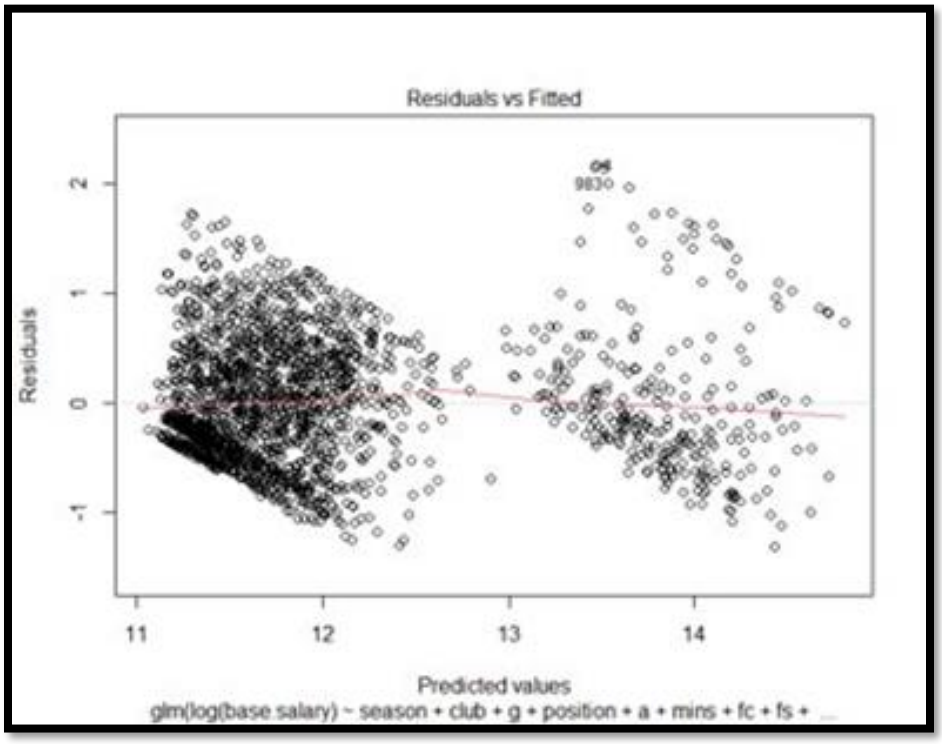
The loss factor provides a 5% increment in goalkeeper salary due to the number of saves provided per game even though the team may have lost the match.

Western Conference goalkeepers have a 3.8% increment in salary compared to eastern conference goalkeepers.

Note: We are interpreting the conference even though they are non-significant in our model.

Quality Checks

Player Model:



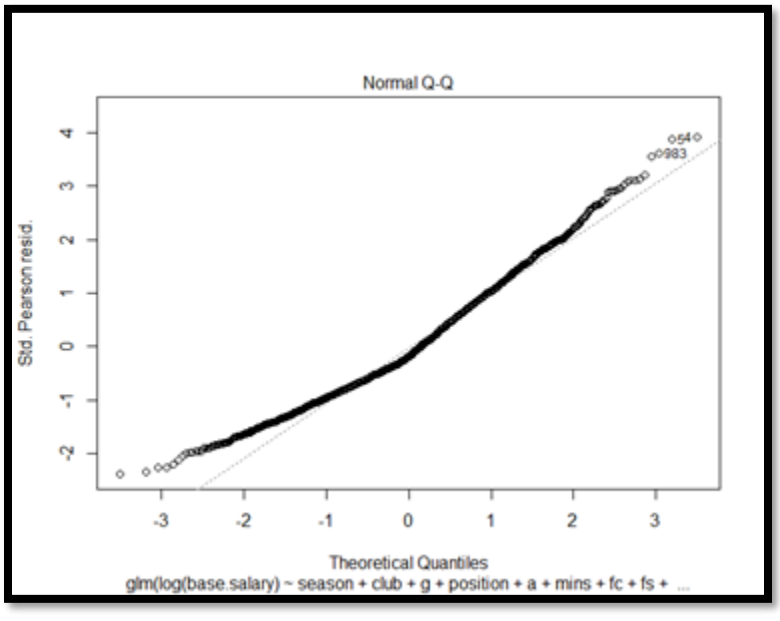
Heteroscedasity and Linearity:

Based on the residuals plot and the results of a Bartletts test (p-value < 2.2e-16) there appears to be clear evidence of heteroscedasity amongst the data. In addition, it seems that the data follows a linear relationship.

	GVIF	Df	GVIF^(1/(2*Df))
season	1.504918	3	1.070497
club	21.698333	22	1.072441
g	2.568016	1	1.602503
position	1.741524	3	1.096869
a	2.209877	1	1.486566
mins	4.337228	1	2.082601
fc	4.827743	1	2.197213
fs	3.252071	1	1.803350
yc	2.782497	1	1.668082
rc	1.125054	1	1.060686
sc.	1.356705	1	1.164777
Designatedplayerstatus	1.308190	1	1.143761
ConferenceID	3.255210	1	1.804220
shtf	2.997495	1	1.731328
pts	2.708827	1	1.645852
tsc	3.027802	1	1.740058

Multicollinearity

A high VIF value was found for the Club variable however this IV was an essential predictor for this model and therefore deemed permissible. No other issues were identified in multicollinearity.



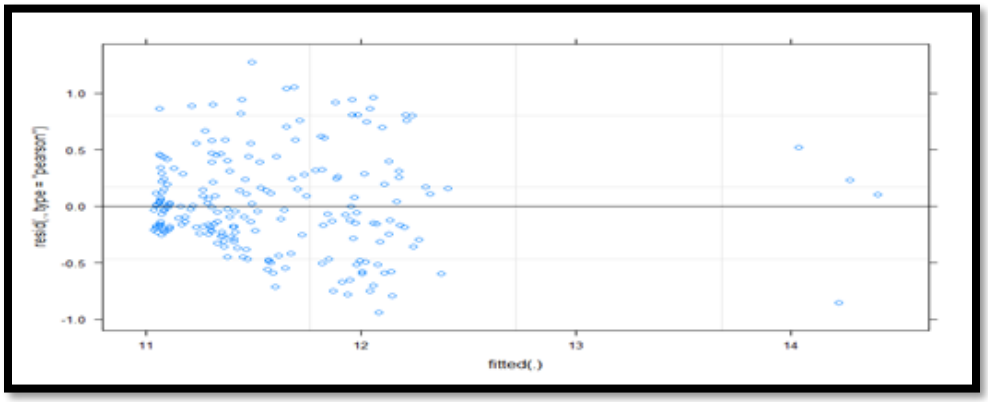
Normality

The QQ plot demonstrates that there is a reasonable level of normality within the data. There is a small exception close to the left tail end but a majority of the points fall on the QQ line.

GoalKeeper Models:

Linearity:

We can observe that the model does exhibit linearity as the residual vs fitted values are centered around the 0 line.



Multicollinearity:

The High VIF values found for this model suggests the model has high multicollinearity because GA,SV,Sho,W, and L all had reported VIF values higher than 5.

VIF Values	
Ga-46.6	sv-29.8
T-8.8	Sho-14.6
Designatedplayer -1.05	Gaa-1.3
W-19.82	ConferenceID-1.04
L-28.17	

Normality:

The Kurtosis demonstrates reasonable normality within the data because the kurtosis value is centered around 3.

```
> kurtosis(residual)
[1] 3.471274
```

Heteroscedasity:

Based on the residuals plot and the results of a Bartletts test (p-value < 1.11e-05) there appears to be clear evidence of heteroscedasity amongst the data.

```
        Bartlett test of homogeneity of variances

data:  list(residual, fitted_re)
Bartlett's K-squared = 19.312, df = 1, p-value = 1.11e-05
```

Autocorrelation:

Since we are working with data tracked time(seasons) we expect there to be autocorrelation amongst the data.

Recommendations

After reviewing the models and our analysis, some actionable recommendations for players were generated based on the results. For the best potential salary in MLS, players should follow the below guidelines to maximize their earning potentials:

•Join the Eastern Conference.

•Become a Designated Player.

•Specialize as a Mid-Fielder.

•Play for one of the three clubs: MNUFC, NYCFC, or LA

•Earn a few yellow cards for aggressive team contributions.

•Score as many goals as possible.

References

Iehl, Aaron Michael Anderson, "An empirical analysis on major league soccer player earnings" (2020). Honors Program Theses. 418. https://scholarworks.uni.edu/hpt/418

Olsen, D. (2015, January 29). Visualizing MLS Salaries Compared to Other U.S. Leagues. Retrieved from https://www.americansocceranalysis.com/home/2015/1/26/visualizingmlssalaries

MLS Players' Salaries Retrieved from: https://datarepository.wolframcloud.com/resources/MLS-Players-Salaries

Major League Soccer Stats Database - Goalkeepers' data. Retrieved from https://www.kaggle.com/josephvm/major-league-soccer-ataset?select=all_goalkeepers.csv

Major League Soccer Stats Dataset - Outfield players data. Retrieved from:https://www.kaggle.com/josephvm/major-league-soccer-dataset?select=all_players.csv

Major League Soccer Stats Dataset – Team. Retrieved from: https://app.americansocceranalysis.com/#!/mls/xgoals/teams

Appendix

Players Model Code

```
library(readxl)

library(vioplot)

library(corrplot)

library(PerformanceAnalytics)

library(ggplot2)

library(stargazer)

library(lme4)

library(car)
```

```
library(lmtest)

library(moments)

library(psych)

library(dplyr)

setwd("C:\\Users\\Willi/Desktop\\Statistical Data Mining")

soccerdata = read_excel("MLS Team Data_update.xlsb.xlsx", sheet = "updated_player_team")

View(soccerdata)

#Preprocessing

colnames(soccerdata) = tolower(make.names(colnames(soccerdata)))

soccerdata_new = subset(soccerdata,select = -
        c(club..grouped.,first.name,last.name,pkg.a,season...34))

colnames(soccerdata_new)[1] = "season"

View(soccerdata_new)

colSums(is.na(soccerdata_new))

which(! complete.cases(soccerdata_new))

colSums(is.na(soccerdata_new))

str(soccerdata_new)

soccerdata_new <- soccerdata_new[-c( 272 , 301, 363, 382, 441, 498, 507, 509, 574, 612, 634,
        667, 673, 704, 729, 738, 746, 798, 860, 874, 884, 885 , 888 , 911 , 934,

                944, 1021, 1022, 1050, 1068, 1078, 1106, 1187, 1191, 1210, 1233, 1267, 1281,
        1285, 1398, 1410, 1483, 1489, 1494, 1526, 1532, 1534, 1564, 1573, 1597, 1598, 1614
        ,1766 ,1843,

                1853, 1889, 1893, 1970, 1982, 2032 ,2043, 2053 ,2062, 2073 ,2128, 2141, 2142,
        2170, 2174, 2184, 2185, 2192, 2195),]



#cleaning teams which are not significant such as USA,PAN ,CIV, JAM, CAN, HON

soccerdata_new_clean = soccerdata_new[!(soccerdata_new$team == "USA" |
        soccerdata_new$team == "PAN" | soccerdata_new$team == "CIV"|
        soccerdata_new$team == "JAM"| soccerdata_new$team == "CAN"|
        soccerdata_new$team == "HON" ),]
```

```
table(soccerdata_new_clean$team)
table(soccerdata_new_clean$club)

#Factor the variables
soccerdata_new_clean$name = as.character(soccerdata_new_clean$name)
soccerdata_new_clean$team = as.factor(soccerdata_new_clean$team)
soccerdata_new_clean$club = as.factor(soccerdata_new_clean$club)
soccerdata_new_clean$position = as.factor(soccerdata_new_clean$position)
soccerdata_new_clean$season = as.factor(soccerdata_new_clean$season)


Data2015<-subset(soccerdata_new_clean,soccerdata_new_clean$season==2015)
Data2016<-subset(soccerdata_new_clean,soccerdata_new_clean$season==2016)
Data2017<-subset(soccerdata_new_clean,soccerdata_new_clean$season==2017)
Data2018<-subset(soccerdata_new_clean,soccerdata_new_clean$season==2018)


#Creating the Designated Player Status.The mls salary cap for any individual player cannot
      exceed
#its Maximum Budget Charge for that year therefore players who have a salary
#greater than that can be considered Designated Players,logically
Data2018$Designatedplayerstatus<- ifelse(Data2018$base.salary>504375, 1,0)
Data2017$Designatedplayerstatus<- ifelse(Data2017$base.salary>480625, 1,0)
Data2016$Designatedplayerstatus<- ifelse(Data2016$base.salary>457500, 1,0)
Data2015$Designatedplayerstatus<- ifelse(Data2015$base.salary>436250, 1,0)

FullData<-rbind(Data2015,Data2016,Data2017,Data2018)

#Teams shot conversion
```

```
FullData$tsc = (FullData$gf/FullData$shtf)*100

#Goal Remainder after subtracting Player contribution (Total goals- player goals)
FullData$pgtr = (FullData$gf-FullData$g)

#Player Position Wild card entry
table(FullData$position)

FullData$position = ifelse(FullData$position == "D","D",ifelse(FullData$position ==
        "F","F",ifelse( FullData$position == "M","M","W")))


View(FullData)
str(FullData)


table(FullData$club)
#segregating data for east coast and west coast teams where east coast is 0 and west coast is 1
FullData$ConferenceID<-ifelse(FullData$team=="NYRB"| FullData$team=="ATL"|
        FullData$team=="NYCFC"| FullData$team=="DC"| FullData$team=="COL"|
        FullData$club=="PHI"| FullData$team=="MTL"| FullData$club=="NE"|
        FullData$team=="TOR"| FullData$team=="ORL"| FullData$team=="CHI",0,1)
table(FullData$ConferenceID)
#Factor the variables
FullData$Designatedplayerstatus = as.factor(FullData$Designatedplayerstatus)
FullData$ConferenceID = as.factor(FullData$ConferenceID)
FullData$position = as.factor(FullData$position)

FullData$position = relevel(FullData$position,"W")
```

```
#Data visualization


hist(FullData$base.salary, main = "Histogram of base salary", xlab = "base salary")

hist(log(FullData$base.salary), main = "Histogram of log of base salary",xlab = "log of base
    salary")

plot(FullData$g,FullData$mins,main = "Plot of minutes played vs goals scored",xlab = "goals
    scored",ylab = "minutes played")

plot(FullData$g,log(FullData$base.salary),cex = 1.2,main = "Plot of log of base salary vs goals
    scored",xlab = "goals scored",ylab = "log of base salary")


plot(FullData$club,log(FullData$base.salary), las = 2, main = "BoxPlot of base salary of each
    team in MLS from 2015 till 2018",xlab = "clubs",ylab = "log of base salary")

par(mfrow = c(1,2))

plot(FullData$season,log(FullData$total.compensation),main = "BoxPlot of total compensation
    in MLS from 2015 till 2018",xlab = "seasons",ylab = "log of total compensation")

plot(FullData$season,log(FullData$base.salary),main = "BoxPlot of base salary in MLS from
    2015 till 2018",xlab = "seasons",ylab = "log of base salary")


par(mfrow=c(1,4))


hist(Data2015$Designatedplayerstatus,main = "Beckham Rule 2015",xlab = "designated player
    status")

hist(Data2016$Designatedplayerstatus,main = "Beckham Rule 2016",xlab = "designated player
    status")

hist(Data2017$Designatedplayerstatus,main = "Beckham Rule 2017",xlab = "designated player
    status")

hist(Data2018$Designatedplayerstatus,main = "Beckham Rule 2018",xlab = "designated player
    status")


plot(FullData$position,log(FullData$base.salary),main = "plot of base salary vs player position",
    xlab = "Position", ylab="log of base salary")
```

```
table(Data2015$club)
table(Data2017$club)



boxplot(Data2018$pts~Data2018$club,las=2)
boxplot(FullData$gf~FullData$club,las=2)

### There are >18 players above the 3 million dollar limit in this plot which are not shown due
    to the cutoff.
boxplot(FullData$base.salary~FullData$Designatedplayerstatus,ylab = " Base
    Salary",xlab="Designated Player Status",ylim=c(0,3000000))
?axis
#barplot(FullData$pts,FullData$team)

table(Data2015$pts)

# Grouped Bar Plot of the amount of Players per club
counts <- table(Data2015$pts, Data2015$club)
counts
barplot(counts, main="Players per Club",
    xlab="CLUB NAMES",ylim=c(0,70),las=2)

table(Data2015$pts,Data2015$team)

hist(Data2015$pts)
hist(Data2016$pts)
hist(Data2017$pts)
hist(Data2018$pts)
```

```
table(FullData$team)


#Violin Plot
ggplot(FullData, aes(name, base.salary)) +geom_violin()




#Correlation
par(mfrow=c(1,1))
x = select_if(FullData, is.numeric)
cor_player = round(cor(x),2)
cor_player
corrplot(cor_player,cex = 1.2,las = 2)


cor()


#models
#players: season,club,position,g,a,mins,fc,fs,yc,rc,sc.,designatedplayerstatus,conferenceID,tsc
#Teams:
lin_player_model =
        lm(log(base.salary)~season+club+position+g+a+mins+fc+fs+yc+rc+sc.+Designatedplaye
        rstatus+ConferenceID+shtf+mins+pts+tsc+pgtr,data = FullData)
summary(lin_player_model)
glm_player_model =
        glm(log(base.salary)~season+club+g+position+a+mins+fc+fs+yc+rc+sc.+Designatedpla
        yerstatus+ConferenceID+shtf+mins+pts+tsc,data = FullData)
summary(glm_player_model)
#Applying pooled effects to playermodels
```

```
#pooled <- plm(log(base.salary) ~
        season+club+position+g+a+mins+fc+fs+yc+rc+sc.+DesignatedplayerstatusConferenceI
        D+shtf+mins+pts+tsc, data=d,index=c('season','club'), model="pooling")

#summary(pooled)                          # OLS model

#plmtest(pooled)                          # LM test of pooled model

#plmtest(pooled, effect="twoways", type="bp")        # Data shows panel effect


#Applying  fixed and random effects to player models

#fixed1 <-  plm(log(base.salary) ~
        season+club+position+g+a+mins+fc+fs+yc+rc+sc.+DesignatedplayerstatusConferenceI
        D+shtf+mins+pts+tsc, data=FullData, model="within")

#summary(fixed1)                          # Fixed effects model

#fixef(fixed1)

#summary(fixef(fixed1))


#random <- plm(log(base.salary) ~
        season+club+position+g+a+mins+fc+fs+yc+rc+sc.+DesignatedplayerstatusConferenceI
        D+shtf+mins+pts+tsc, data=FullData, model="random")

#summary(random)                          # Random effects model

#ranef(random)

#summary(ranef(random))


#pFtest(fixed1, pooled)                   # F test for nested models: FE is better

#phtest(fixed1, random)                   # Hausman test: FE is better


#Applying lmer models


re <- lmer(log(base.salary) ~
        (1|season)+(1|club)+position+g+a+mins+fc+fs+yc+rc+sc.+DesignatedplayerstatusConf
        erenceID+shtf+mins+pts+tsc+pgtr, data=FullData,REML=FALSE)
```

```
summary(re)
confint(re)
AIC(re)
fixef(re)                          # Magnitude of fixed effects
ranef(re)                          # Magnitude of random effects
coef(re)                           # Magnitude of total effects
```

```
AIC(lin_player_model,glm_player_model,re)
BIC(lin_player_model,glm_player_model,re)
stargazer(lin_player_model,glm_player_model, re, type="text",
          single.row=TRUE,out="models.html")

#Assumptions
plot(glm_player_model)

#' MV normality: Shapiro-Wilks test
shapiro.test(glm_player_model$res)

kurtosis(glm_player_model$residuals)
#' Homoskedasticity: Bartlett's test for normal distribution model
bartlett.test(list(glm_player_model$res, glm_player_model$fit))
plot(glm_player_model$res ~ glm_player_model$fit)

# Multi-collinearity:
vif(glm_player_model)
```

```
# Independence/autocorrelation:
dwtest(glm_player_model)

plot(glm_player_model$res ~ glm_player_model$fit)
```

GoalKeeper Model Code:

```
library(readxl)
library(vioplot)
library(corrplot)
library(PerformanceAnalytics)
library(ggplot2)
library(stargazer)
library(plm)
library(lme4)
library(car)
library(lmtest)
library(moments)
library(psych)
library(tidyverse)
library(dplyr)
setwd("C:/Users/rauna/Downloads/")
gkdata = read_excel("MLSCompleteData.xlsx", sheet = "GKCompletedata")
View(gkdata)
str(gkdata)
#Preprocessing
colnames(gkdata) = tolower(make.names(colnames(gkdata)))
View(gkdata)
```

```
colSums(is.na(gkdata))

which(! complete.cases(gkdata))

colSums(is.na(gkdata))

str(gkdata)

gkdata_new <- gkdata[-c(4,55,88,104,121,158,242,253),]



#cleaning teams which are not significant such as USA,PAN ,CIV, JAM, CAN, HON


gkdata_new_clean = gkdata_new[!(gkdata_new$team == "USA" | gkdata_new$team == "PAN" |
          gkdata_new$team == "CIV"| gkdata_new$team == "JAM"| gkdata_new$team == "CAN"|
          gkdata_new$team == "HON" ),]

gkdata_new_clean = subset(gkdata_new_clean,select = -c(ga...28,sv.,w.,pkg.a,first.name,last.name))

table(gkdata_new_clean$team)

table(gkdata_new_clean$club)

View(gkdata_new_clean)

names(gkdata_new_clean)[13]  <- "ga"

str(gkdata_new_clean)

#Factor the variables

gkdata_new_clean$name = as.character(gkdata_new_clean$name)

gkdata_new_clean$team = as.factor(gkdata_new_clean$team)

gkdata_new_clean$club = as.factor(gkdata_new_clean$club)

gkdata_new_clean$position = as.factor(gkdata_new_clean$position)

gkdata_new_clean$season = as.factor(gkdata_new_clean$season)



Data2015<-subset(gkdata_new_clean,gkdata_new_clean$season==2015)

Data2016<-subset(gkdata_new_clean,gkdata_new_clean$season==2016)

Data2017<-subset(gkdata_new_clean,gkdata_new_clean$season==2017)

Data2018<-subset(gkdata_new_clean,gkdata_new_clean$season==2018)
```

```
#Creating the Designated Player Status.The mls salary cap for any individual player cannot exceed
#its Maximum Budget Charge for that year therefore players who have a salary
#greater than that can be considered Designated Players,logically
Data2018$Designatedplayerstatus<- ifelse(Data2018$base.salary>504375, 1,0)
Data2017$Designatedplayerstatus<- ifelse(Data2017$base.salary>480625, 1,0)
Data2016$Designatedplayerstatus<- ifelse(Data2016$base.salary>457500, 1,0)
Data2015$Designatedplayerstatus<- ifelse(Data2015$base.salary>436250, 1,0)

FullData<-rbind(Data2015,Data2016,Data2017,Data2018)

#Teams shot conversion
FullData$tsc = (FullData$gf/FullData$shtf)*100

table(FullData$position)



View(FullData)
str(FullData)

#segregating data for east coast and west coast teams where east coast is 0 and west coast is 1
FullData$ConferenceID<-ifelse(FullData$team=="NYRB"| FullData$team=="ATL"|
        FullData$team=="NYCFC"| FullData$team=="DC"| FullData$team=="COL"|
        FullData$club=="PHI"| FullData$team=="MTL"| FullData$club=="NE"|
        FullData$team=="TOR"| FullData$team=="ORL"| FullData$team=="CHI",0,1)
table(FullData$ConferenceID)
#Factor the variables
FullData$Designatedplayerstatus = as.factor(FullData$Designatedplayerstatus)
FullData$ConferenceID = as.factor(FullData$ConferenceID)
FullData$position = as.factor(FullData$position)
```

```
#Data visualization


par(mfrow = c(2,2))

plot(Data2015$sv,Data2015$base.salary, main = "plot of no of Base Salary Vs saves for 2015", xlab =
        "Saves", ylab = "Base Salary")

text(Data2015$sv, Data2015$base.salary,Data2015$name , cex = 1.2,pos=4, col="red")


plot(Data2016$sv,Data2016$base.salary, main = "plot of no of Base Salary Vs saves for 2016", xlab =
        "Saves", ylab = "Base Salary")

text(Data2016$sv, Data2016$base.salary,Data2016$name , cex = 1.2,pos=4, col="red")


plot(Data2017$sv,Data2017$base.salary, main = "plot of no of Base Salary Vs saves for 2017", xlab =
        "Saves", ylab = "Base Salary")

text(Data2017$sv, Data2017$base.salary,Data2017$name , cex = 0.6,pos=4, col="red")


plot(Data2018$sv,Data2018$base.salary, main = "plot of no of Base Salary Vs saves for 2018", xlab =
        "Saves", ylab = "Base Salary")

text(Data2018$sv, Data2018$base.salary,Data2018$name , cex = 0.6,pos=4, col="red")




#Correlation

x = select_if(FullData, is.numeric)

cor_player = round(cor(x),2)

cor_player

corPlot(cor_player,cex = 1.2,las = 2)


#Applying lmer models
```

```
re <- lmer(log(base.salary) ~
        (1|season)+(1|club)+ga+sv+gaa+w+l+t+sho+w+Designatedplayerstatus+ConferenceID,
        data=FullData,REML=FALSE)



summary(re)

confint(re)


fixef(re)                       # Magnitude of fixed effects

ranef(re)                       # Magnitude of random effects

coef(re)                        # Magnitude of total effects


AIC(re)

BIC(re)

stargazer( re, type="text", single.row=TRUE)
```