Research Proposal on MLS Player Annual Earning Prediction

Introduction:

Soccer is one of the most popular sports in the world, and some soccer players have extremely high salaries. In 2021, Ronaldo earns roughly $125 million, which makes him the world's highest-paid soccer player(Gastelum, 2021). As a soccer fan, I want to have a better understanding of the high-salary phenomenon in the sports field. Meanwhile, MLS is a men's professional soccer league in the United States(US Soccer, 2021). By analyzing data about MLS players in 2018, I would like to answer my research question: What is the relationship between the annual earning of a MLS player and his performance in the year?

The importance of this research is that soccer players know their earnings relate to which statistics, so they can have a direction for improvement. Also, soccer clubs can better manage their budget to optimize their benefit, without choosing overvalued players.

Background/Literature:

Through a search on the UofT library, I found a paper about soccer clubs' choices in Spanish and English leagues from 1994 to 2004 that shows that win maximization better approximates the choices of soccer clubs than profit maximization(Garcia-del-Barrio & Szymanski, 2009). It shows the importance of my research question because we need to determine whether buying a player with a high salary really relates to victory. Also, a study analyzed the productivity and salary structure in MLS using the Gini coefficient and coefficient of variation. The researcher concluded that as salary increases, production declines(Coates et al., 2016). It verifies the inflation of salary among MLS players, which helps me anticipate the distribution of salary among MLS players: They will be highly skewed. It is important because to fit a linear regression model on earnings, I probably need to adjust my response variable.

Another study using positive assortative matching based on productivity showed that the high performance in MLS is related to wages(Scarfe et al., 2018). The study used data from previous years, and I expect I can reproduce and agree with the result by using a different approach. In addition, an analysis of the wage pattern in MLS is done using generalized least squares estimation. One of the most influential determinants they found is the number of games the player started with(Celik & Ince-Yenilmez, 2017). The approach they used is similar to mine, which means that I can include this variable as one of the variables in my analysis, although I have a different focus on predictors from them.

Data source:

I am using 2 datasets for datas of their common MLS players' statistic in 2018. In my analysis, I will determine which predictors should be kept.

Main variables:

- **Total.Compensation**: Average annual guaranteed compensation of a MLS player.
- **GP**: Number of game a MLS player in.
- **GS**: Number of game a MLS player in when the game started.
- **G**: Number of goals scored by a MLS player.
- **A**: Number of on target scoring attempts of a MLS player.
- **SHTS**: Total scoring attempts of a MLS player.
- **SOG**: On target scoring attempts of a MLS player.
- **FS**: Number of fouls(unfair act) a MLS player suffered from.
- **OFF**: Number of times a MLS player is in an offside position, that is when he is nearer to his opponents's goal line than both the ball and the second last opponent(Rothschild, 2016).

The **Total.Compensation** data is scraped by Ben Jones:
https://data.world/dataremixed/mls-player-salaries-2010-2018
The original source of the data is: https://mlsplayers.org/resources/salary-guide
The **rest of main variables** are scraped by Joseph Mohr:
https://www.kaggle.com/josephvm/major-league-soccer-dataset
The original source of the data is: https://www.mlssoccer.com/stats

Exploratory Data Analysis

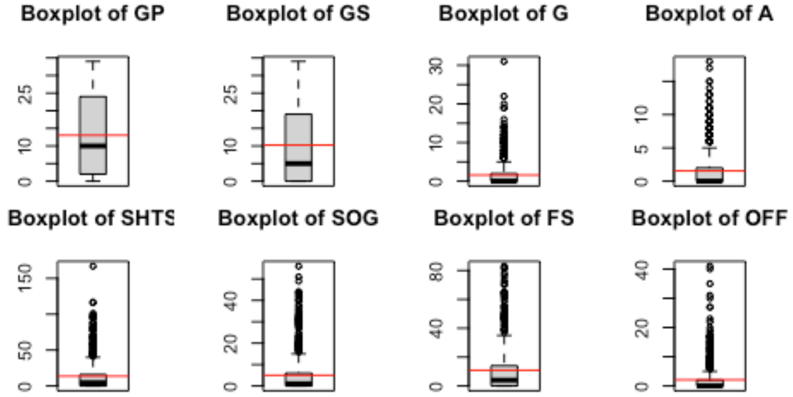
Sample Reliability, Accuracy and Sampling Bias

The original source of data is MLSPA and MLS. MLSPA is an organization that served as the collective bargaining representative for MLS players(MLSPA, 2021), while MLS is the official website of the league. Therefore, I anticipate the data is reliable and accurate. Data is directly scraped by from the source so the sampling bias that is most likely to occur is selection bias.

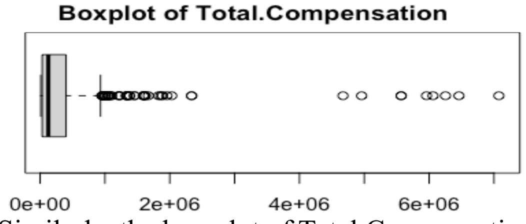
Generalization and Confounding Factors

After a natural join, I ended up with 690 different observations in my dataset. Generally, MLS has 776 players participating each year (Transfermarkt, 2021), so I expect that the result can be generalized to other MLS players. However, I failed to collect some confounding variables, and they may affect the generalizability of the result to a larger population. Possible confounding variables are the level of the soccer league, the budget of the team, popularity, age, and cultural differences.

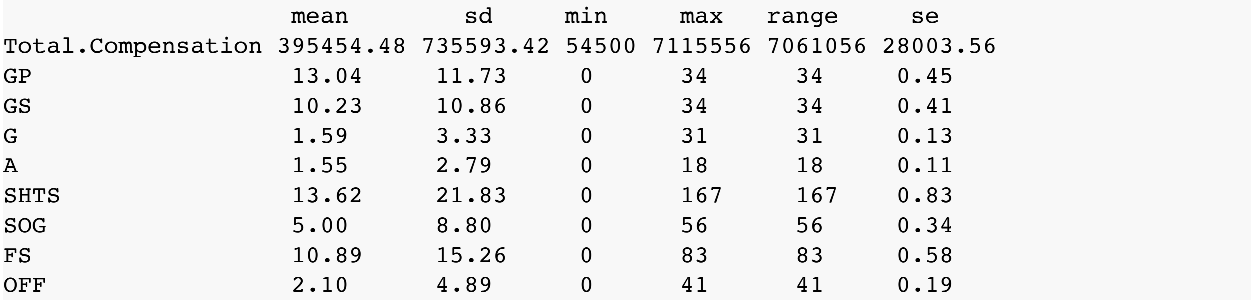
By exploring I found no missing information for any variables in my sample.
The following are box plots of GP, GS, G, A, SHTS, SOG, FS, and OFF.



The redline represents the mean of each variable. In the plots, GP and GS don't have any outliers. Both of them range from 35 to 0, and the interquartile range covers a wide range of variables. From the position of the median and mean, we can see that the median is lower than the mean, which means the data is slightly right-skewed. On the other hand, we can see from the box plots that the interquartile ranges for the rest of the variables are very narrow. Also, the mean is close to the higher quartile and the median is close to the lower quartile. Compared with other variables, the distribution of FS is slightly less skewed because the median is still visible. Visibly, G, A, SHTS, SOG, FS, and OFF all have many outliers.



Similarly, the box plot of Total.Compensation shows that this variable is right-skewed as well. Most of its data range from 0 to 1 million, but the highest outlier reaches 7 million.

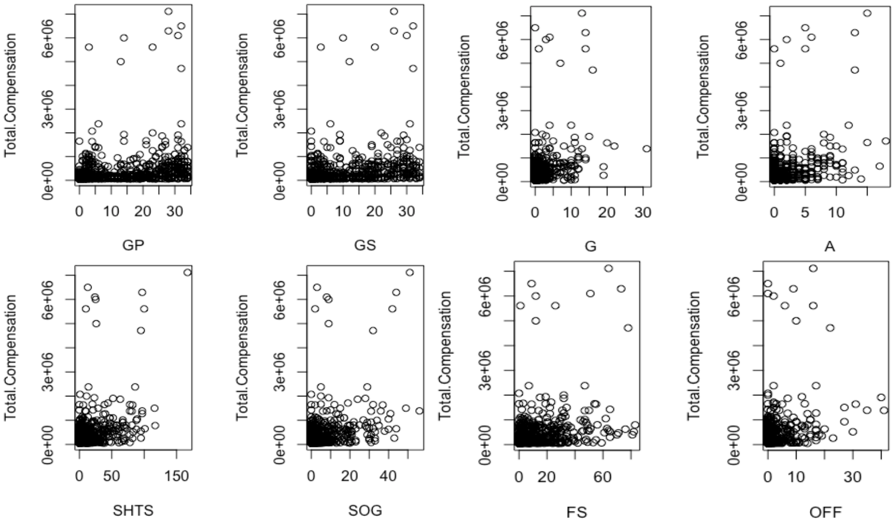

```
                        mean         sd     min     max   range      se
Total.Compensation 395454.48 735593.42 54500 7115556 7061056 28003.56
GP                     13.04      11.73      0       34      34     0.45
GS                     10.23      10.86      0       34      34     0.41
G                       1.59       3.33      0       31      31     0.13
A                       1.55       2.79      0       18      18     0.11
SHTS                   13.62      21.83      0      167     167     0.83
SOG                     5.00       8.80      0       56      56     0.34
FS                     10.89      15.26      0       83      83     0.58
OFF                     2.10       4.89      0       41      41     0.19
```

In the numeric table, the standard deviation is affected by the skewness and outliers. Currently, we have a large spread for each variable. The large standard error of Total.Compensation means the mean of Total.Compensation doesn't provide much useful information for the data.

Overall, all the variables are highly right-skewed. Skewed distribution of response may cause violation of normality assumption of linear regression, while skewed distribution of predictors may cause violation of linearity assumption. They will both effect common variance assumptions. It means that without adjustment, the result of my analysis may have a high variance or be biased. Additionally, abnormal data points are presented. They may affect the generalization of my data, depends on

Linear Model

My research question can be answered using a linear model. First, fitting a linear model is meaningful because I am looking for a quantitative relationship and possibly using performance to predict earnings. Also, previous research shows a relationship between performance and salary, although in a different approach. Finally, scatter plots of predictors and totals shows some noisy linear relationships are shown below.



I will use Total.Compensation as my response variable, because Total.Compensation represents the total annual earnings of an MLS player, which is what I am interested in. I will use GP, GS, G, A, SHTS, SOG, FS, and OFF as my predictors because they are statistics about an MLS player in different directions. Right now, GP, GS doesn't look linearly related to Total.Compensation, but I use them as the main variables because previous work has shown that GS is a factor that affects salary (Celik and Ince-Yenilmez, 2017). After adjusting the skewness problem, I hope that there is a clearer linear relationship between my predictor variables and the response.

A linear model is likely appropriate for my variables. First, the annual earning of a player along with other predictors are all numeric variables that can be fitted to a linear model. Then, regarding assumptions of linear regressions, it probably satisfies the uncorrelated errors assumption. My observations are all from different players, and I expect them to be independent variables. Hence, they should have uncorrelated errors. The assumption of normality may be violated, because Total.Compensation is right-skewed. Also, the assumption of linearity may be violated, because my predictors are right-skewed too. Due to the skewness, when predicting using a lower value, the variance is likely to be lower than using a higher value. That means the assumption of linearity may be violated. However, these issues are anticipated and I plan to normalize my data before fitting the model, possibly by applying the Central Limit Theorem or using a transformation function.

Another anticipated problem is the influence of outliers. If they come from designated players, removing them may make my model more representative for the majority of the MLS players. I plan to keep an eye on useless predictors, and whether the outliers should be included when working on my data analysis. Without losing information about the response, I am aiming for a simpler model that has fewer predictors.

References

Celik, O. B., & Ince-Yenilmez, M. (2017). Salary differences under the salary cap in Major League Soccer. *International Journal of Sports Science & Coaching*, *12*(5), 623–634. https://doi.org/10.1177/1747954117727809

Coates, D., Frick, B., & Jewell, T. (2016). Superstar salaries and soccer success. *Journal of Sports Economics*, *17*(7), 716–735. https://doi.org/10.1177/1527002514547297

Garcia-del-Barrio, P., & Szymanski, S. (2009). Goal! profit maximization versus win maximization in soccer. *Review of Industrial Organization*, *34*(1), 45–68. https://doi.org/10.1007/s11151-009-9203-6

Gastelum, A. (2021, September 21). *Ronaldo passes Messi as Forbes' top-paid soccer player*. Sports Illustrated. Retrieved October 20, 2021, from https://www.si.com/soccer/2021/09/21/cristiano-ronaldo-highest-paid-soccer-player-forbes-messi-neymar.

Learn about the MLSPA. MLS Players Association. (n.d.). Retrieved October 21, 2021, from https://mlsplayers.org/about-us.

Major League Soccer 2021. Transfermarkt. (n.d.). Retrieved October 20, 2021, from https://www.transfermarkt.us/major-league-soccer/startseite/wettbewerb/MLS1.

Professional Leagues. US Soccer. Sponsored by Volkswagen. (n.d.). Retrieved October 20, 2021, from http://www.ussoccer.com/about/affiliates/professional-soccer.

Rothschild, T. (2016, January 21). *Soccer 101: Explaining the offside rule: Orlando City*. orlandocitysc. Retrieved October 20, 2021, from https://www.orlandocitysc.com/news/soccer-101-explaining-offside-rule.

Scarfe, R., Singleton, C., & Telemo, P. (2018). Do high wage footballers play for high wage teams? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3297644