Introduction

As an avid MLS fan, I like many others was excited to see the arrival of Messi to Inter Miami, along with famous Spanish pair Jordi Alba and Sergio Busquets. It made me curious that spending big money on these players actually impacts regular season performance because there are many examples through MLS history where a big star player on huge wages, like Lorenzo Inisgne, Steven Gerrard, or Ezequiel Barco, anecdotally did not help the team achieve regular season success.

Data Collection

To explore this hunch further, I took data from FBref to analyze the relationship between player wages, general performance, minutes played and team standings in the 2022 season. I merged the datasets in Excel. The only transformation I had was to group the weekly wages by team.

Initial Data Exploration

Creating a bar chart with a best-fit line to visualize how much money per week each team was spending and annotating which conference and position the team ended up in. The bar chart illustrates that spending a lot of money had no clear relationship with overall standings as Toronto FC with the biggest weekly wage ended up near the bottom of the Eastern Conference; whereas New York Red Bulls were closer to the top.

To find controls for my future OLS I ran a covariance matrix because there was a risk of multicollinearity. The multicollinearity risk came from the fact that a player's performance (xG+/-) means that on average they are more likely to earn more money and play more minutes (Min) on the field. Also, older players (age) can demand a higher wage due to their experience. Therefore, ensuring that there was not a strong correlation between minutes played, player's performance and age was necessary to reduce the risk of multicollinearity.

There is a weak positive correlation between age, minutes played, and weekly wages. It makes sense that older players get paid more; however, MLS has publicly stated that they want to be a league where they develop talents to sell them to the bigger clubs in Europe. There are success stories like Alphonso Davies and Miguel Almiron to name a few; however, the correlation is not as strong as MLS would like.

However, age is minimally correlated with xG+/-. This suggests that older players are not necessarily delivering on their experience on the field. Therefore, investing in older players with a bigger pay packet might not be the best investment for a league with stringent squad rules like the MLS.

OLS Regression

My dependent variable is position and the outcome variable is weekly wages, utilizing controls such as minutes played, age, and xG+/-. Indeed age, minutes played, and weekly wages are positively correlated; however, it is necessary to include minutes played because one does not want to include players on a huge wage, but did not play due to injury or not being selected by the manager. Age is necessary because MLS has a stereotype of purchasing players beyond their peak. As xG+/- is minimally correlated with wage, minutes played, and age, it would serve as an effective control variable in our analysis, ensuring that the observed relationships with other variables are not confounded by the effects of these factors. This allows for a more accurate understanding and interpretation of how other predictors influence the outcome without the undue influence of wage, minutes, or age.

The output suggests that weekly wage is extremely weakly correlated with position. Moreover, xG+/- is positively correlated with position as well. Both are statistically significant. Age also has a positive relationship; however, it does not have a statistical significance.

I do not want to come to any conclusions with these findings as I still need to rethink the model to accurately predict anything as it still has strong multicollinearity issues as seen with the high Jaque Berra score. Also since standings are not normally distributed, I need to come up with a better model to come up with some findings that actually can be actionable. Please advise