Enze Cai

BUSN 323

Dr. Brandon Ware

Apr. 28, 2022

Tableau Project

**Introduction**

The topic I selected for this project is to discover the relationship between soccer players’ salaries and their skills. It sounds obvious that a high-skill player must have a high salary, but the situation is different for China National Football Team (I’ll use CNFT as an abbreviation unofficially since China National Football Team doesn’t have an official abbreviation). I’m Chinese, and every Chinese hope CNFT has a chance to participate in the world cup, but it’s even hard for them to get the qualification. The reason that CNFT is being ridiculed in China, especially on social media, is that they cannot get good results even though they have the best training equipment, system, and nutrition structure in the world. Therefore, I decided to analyze the relationship between soccer players’ salaries and skills, and I use SOG (shots on goal, number of shots on target) for quantifying players’ skills.

**Data Collection**

The datasets I used for this project are America’s MLS (Major League Soccer) basic stats and MLS players’ salaries, and both datasets are from [www.sportsvizsunday.com/americansports](http://www.sportsvizsunday.com/americansports). Though data is way easier to find online than it used to be, it’s still hard to find a useful dataset for a pre-specific topic. I forgot how I found this dataset, but it took me a lot of time. From the description of the website team, it says “Sports Viz Sunday is an initiative to encourage the creation and sharing of sports-themed visualizations”, so it’s a public sports-themed database. We can download a variety of sports datasets on the website by just clicking the link. The website provides datasets about American football, baseball, basketball, ice hockey, soccer, NCAA, etc., so it’s a powerful database website for sports fans. For the purpose of analyzing the relationship between players’ salaries and skills, I use the data from two datasets. One of them is about the statistics of MLS soccer players from 1996 to 2019, including SOG (shots on goal), SHTS (number of shots), GP (game played), etc. Another dataset is about the salaries of players from 2004 to 2018. For this project, I will use the datasets for the year 2018 to ensure time consistency.

**New Statistic**

For the new statistic, I did linear regression to check if there is a strong linear relationship between soccer players’ salaries and SOG. There are several options for quantifying soccer players’ skills, such as SOG, SHTS, SOG%, etc. The reason I chose SOG as the index for quantifying players’ skills is that a player with high SHTS doesn’t mean that he has a high number of goals; same for SOG%, if a player only shot once and he goals in a season, he would get 100% for SOG% then, but it didn’t mean that he has high skills since he only shot for once. Before I did linear regression, I organized and cleaned the data. I use the “lookup” function in excel to import the value of SOG from another spreadsheet to make sure the value has corresponded to the correct player. And I use a scatterplot to conduct linear regression experiments and use the regression function under data analysis to draw a summary. Here are the results:

Table

Description automatically generated

From the graph, we can see the trendline has a positive coefficient, and it shows a positive relation between base salary and SOG, which means that with the increase in base salary of a soccer player, his SOG will also increase. Both graph and summary output provided R-square statistic equals 0.18, which means 18% of the variation of the dependent variable is explained by the predictor variable. 18% is quite high for real-world data, so we can see some relationship between players’ salaries and SOG. We can also notice that there are some outliers in the graph, which is acceptable since we have a very large sample size.

**Visualization**

I have created four visualizations in tableau, which are scatterplot, stacked bubbles, treemap, and a line graph reflecting linear regression. I have merged SOG and salaries into one spreadsheet so that they can be imported conveniently, but they are from two different datasets.

Chart, scatter chart

Description automatically generated

A scatterplot is used to present the values of two numerical variables by dots, and we can easily check the distribution of trends between these two variables from the scatterplot. Also, a scatterplot is necessary to reflect linear regression. In this case, each dot represents a soccer player, and the position of the dot is determined by the player’s salary and SOG. I have added the trendline to the plot, including the linear equation and r-squared statistic, which are the same as the results of the new statistic. Therefore, with a p-value < 0.0001, we can draw the same result as the new statistic, that the change in salary is related to the change of SOG, and the positive relationship is reflected by the plot.

Chart, bubble chart

Description automatically generated

Then I have included a stacked bubble chart in my visualization to reflect the relationship between SOG and the salary of a player by size and color. In this case, size is measured by the salary of a player, which means if a player has a higher salary, he will have a bigger bubble than others. And the color is measured by SOG, which means if a player has high SOG, he will have darker bubble color than others. Therefore, from the graph, we can see most of the small bubbles have a light color, which implies players with low SOG has low salary than others. However, there exists an exception. For example, Bastian, Michael Bradley, and Jozy Altidore have large bubbles without dark color, which are considered outliers in statistics, and I’ll provide some possible reasons in the conclusion.

Chart, treemap chart

Description automatically generated

The third visualization is the treemap reflecting the relationship between SOG and salaries. It would be better if we use the categorical variables as a color for the treemap so that we can easily see the difference between variables. In this case, I use the same setting from the stacked bubble to the treemap, and the graph shows more neat and clear. We can see some evident small squares with dark colors. The positive linear pattern is hard to see from the treemap, it’s probably because of the low r-squared value, though it’s relevantly high in a real-world case.

Chart, line chart

Description automatically generated

The last visualization I have created is the line graph between SOG and the average salary corresponding to SOG. I created this visualization by improving the first scatterplot linear regression. The difference between these two linear regressions is that the first one is in terms of every single player, and the second one is using the average salary for each increment of SOG. The advantage of the second one is that we use the average as a variable to avoid a lot of uncertainty, and we can see the r-squared value is 0.31 which means the trendline fits the data better than the first plot with the r-squared value equal to 0.18.

**Conclusion**

By conducting all these visualizations and new statistics, I can conclude that there is a positive relationship between soccer players’ SOG and salaries at the significance level at 0.05 (conclusion drawn from the p-value<0.05), which means soccer players’ salaries have a positive relationship with their skills. However, we can also see many outliers in the dataset, and there are several possible reasons. Since a sports player cannot always stay in peak physical condition, they can have a high salary but didn’t have a good performance in a particular season. The reason can also be an emergency, like an injury.

I can improve this project in several parts. The first one is that I can reduce the sample size, to make the data cleaner and more intuitive. Also, I’m not a soccer fan, but a workable option is that I can visualize the dataset for different positions of players since I personally believe the salaries vary from each different position (we can also test this hypothesis).