**Determining Which Variables Contribute to Shot Success in Soccer**

**Chapter 1: Introduction**

According to the Nielsen Company, soccer is the uncontestably most watched sport in the world, and is a business for teams worth hundreds of millions of dollars. In fact, 43% of the sports viewers in the World are interested in watching soccer, compared to 37% from Basketball viewers; and the FIFA World Cup Media rights income budgeted a whopping total of $3.0 billion USD for this event alone (World Football Report, 2018, www.nielsen.com/wp-content/uploads/sites/3/2019/04/world-football-report-2018.pdf). As a result, clubs put a great deal of time and resources towards achieving success **(example and source)**. This includes collecting ever increasing amounts of data about their team’s performances and those of others. Unfortunately, most of these data sets are not publicly available. There are still small data sets that have become publicly available, such as this one obtained by Luca Pappalardo on Figshare. From this small data set we hope to gain some insight into the game and an understanding of how data is coming to be used in soccer.

**Chapter 2: Data Set Overview**

2.1 Data Source

The dataset contains all matches played in Europe’s top five leagues during the 2017/2018 season (380 matches per league for a total of 1900 matches, producing 643,149 match events), the 2016 European Championship (51 matches, producing 78,139 match events), and the 2018 World Cup (64 matches, 101,758 match events). The data was imported by league and then compiled into a single data frame, giving a total of 3,251,294 events and 19 variables. This data was originally released by Wyscout.

2.2 Calculating and Adding Event Distance Variable

One variable we are interested in looking at is the distance of assists and key passes and whether this plays a role in the success of the pass and whether it results in a successful shot on goal or not. We have to calculate distance for each event using the x,y-coordinates given in the data set. **(What if we used the typical dimensions of FIFA as an assumption, 120x75 yards, to give an estimate of distance?)**

2.3 Selecting Variables and Subsetting Data

With so many variables available we had to start by narrowing the number to look at. Since we’re interested in what variables effect the success rate of a shot on goal, we selected the variables immediately preceding a goal. These include shots, assists, and key passes, and the x,y-coordinates and event distance for each. Separate data frames were created for each of the following events of interest:

1. Key passes: pass leading to unsuccessful scoring opportunity
2. Assists: pass leading to a goal
3. Shots: unsuccessful shots on goal
4. Goals: successful shots on goal

**Chapter 3: Exploratory Data Analysis: Understanding Selected Independent Variables**

3.1 Understanding the Field

To understand our data, we needed to first understand how the field is set up. The field is set up so that the length is given by the x-coordinate, and the width is given by the y-coordinate. Each coordinate is given as a percentage. Events are always oriented so that the attacking team is going from left to right.

3.2 Event Visualization

Using the ggsoccer library to plot event data on an image of a pitch and bordering the x and y axis with event density plots we were quickly able to understand how each event is distributed on the pitch, shown in Figure 1.

**Please note that the dimension on a soccer field, per ruling of the FIFA, are NEVER squared, and although dimension may vary, the typical layout is of 120 by 75 yards (excluding? the limit lines) (source: backyardsidekick.com/soccer-field-size-layout-and-dimensions/).**

Several observations can be made from this visualization. First, the x & y-coordinates for both key passes and assists are multimodal and appear to be symmetric across the y=50 line. This makes sense as teams try to attack down both sides of the field. With this data being an average over so many games getting a symmetric split between the two sides of the field is expected. This symmetry could be a factor that varies from team to team, however. Second, the x & y-coordinate distributions for shots and goals is much closer to normal. The exception being the x-coordinate for shots which is bimodal. This is interesting because it indicates that a significant portion of shots are take farther away from goal, however, very few goals are scored from these shots. Despite this, players still seem to take these shots.

The distance distributions for key passes and assists are displayed in Figure 2. Both show fairly normal distributions, despite the fact that neither the x,y-coordinates for these two events show normal distributions.

3.3 Differences in Event Distributions

From looking at Figures 1 and 2, we already see several points that are outliers. For example, when looking at the figure showing shot events several shots can be seen to occur in the defensive half of the field. When looking at the bordering density plot, we see that these events, while they do occur, are rare compared with the majority of shots taken. From this we felt safe cleaning our events to remove these sorts of outliers before taking a deep dive into the distributions.

3.3.1 Fitting Event Distributions

After cleaning the data we plotted histograms for each variable, comparing key passes with assists and shots with goals by plotting them on the same plots. We then attempted to fit each distribution with a normal distribution. As seen in Figure 4, not all variables are well fitted by a normal distribution, such as the x & y-coordinate distribution of key passes and assists, and the x-coordinate shot distribution. The rest appear to be reasonably well fit with a normal distribution.

3.3.2 T-tests: Comparing Assists vs Key Passes and Shots vs Goals

T-tests were performed to check for differences between x & y coordinates between key passes and assists and between shots and goals. The pass distance between key passes and assists was also checked for differences using the t-test. With population sizes for each of our events on the order of several thousand for assists and goals to greater than ten thousand for key passes and shots, t-tests can still be applied to determine differences in the sample means, thanks to the central limit theorem which states that the distribution of sample means of a population still show a normal distribution, with large enough sample size, even if the population shows a non-normal distribution.

Assists vs Key Passes:

T-tests with a 95% confidence level were run on the assist and key pass variables, x-coordinate, y-coordinate, and distance. For each of these variables the null hypothesis was set as no difference between the variable means for assists and key passes.

Below is the result for the x-coordinate. The resulting p-value is on the order of 1e-07 signifying a statistically significant difference in the x-coordinate between assists and key passes. This result shows that, on average, final passes made further up the field are more likely to result in an assist than a key pass.

When looking at the y-coordinate a p-value of 0.5 is obtained which would indicate no difference between assists and key pass. Since these distributions are not actually normal and symmetric around the y=50 line, when we assume a normal distribution, we get an inaccurate result.

The last variable to be looked at was the distance variable. The t-test here resulted in a p-value < 2e-16. So there is a significant difference between the distance covered by an assist vs a key pass, with assists having an average distance of 29.2 with a confidence interval from 28.7 - 29.8, and key passes having a mean distance of 33.2. This shows that final passes that have to cover a shorter distance are more likely to result in an assist.

Goals vs Shots:

Next, t-tests with a 95% confidence level were run on the goals and shots variables, x & y-coordinates. For these variables the null hypothesis was again set as no difference between the variable means for assists and key passes.

Below is the result for the x-coordinate. The resulting p-value is < 2e-16, meaning there is a statistically significant difference between goals and shots. With the shot x-coordinate mean equal to 84.1 and lying outside the confidence interval for goals, 90.7 - 90.9, we can conclude that attempts on goal take further up the field are more likely to result in a goal.

Looking now at the results for the y-coordinate, a p-value of 0.005 is obtained, still indicating a statistically significant difference of this variable between goals and shots. The shot mean y-coordinate is 49.2 and sits just outside the confidence interval of 49.3 - 49.8 for the goals y-coordinate. This is an interesting result as we expected for there not to be a difference in the y-coordinate. This is because shots are taken from both sides of the field in about equal measure, yet it seems that perhaps the shot distribution favors the right side more than the left, but more goals are scored from a more central location. It could be that while this result is statistically significant, in reality though it probably isn't. For example, an average professional soccer pitch is about 70 meters wide. So, a difference of 0.1% in the y-coordinate is a difference of 0.07 meters, which is not really a significant difference.

**Chapter 4: Smart Questions**

4.1 SMART Question 1: Given the variables Key Pass, Shot, Assist, and Goal within the data from the 2016 European Championship, the 2017/2018 Europe’s top five leagues, and the 2018 World Cup, which configuration of the variables result in a higher probability of a goal being scored?

The t-tests done in the previous section demonstrate that there are differences in our selected variables between key passes and assists and between shots and goals. However, this does not answer the question: which variables result in the highest probability on goal? In other words, in a match, what field positions should teams target to improve the likelihood of scoring a goal? To answer this, we calculated and plotted the success probabilities of the final pass (successful = assist = goal) and shots (successful = goal) as a function of their starting x-coordinate, y-coordinate and int the case of the final pass the pass distance. The results are shown below in Figure 6.

From these plots we can make some conclusions about each variables role in determining a successful shot on goal. First, goal probability varies with respect to all selected variables. The one exception to this is the final pass x-coordinate which remains relatively constant. Second, goal probability shows the greatest change as functions of final pass distance and shot x-coordinate. In the case of final pass distance, the probability of a goal approaches 0.4 as the length of the pass decreases. In the case of shot x-coordinate, the goal probability also approaches 0.4 as the x-coordinate approaches 100, gets closer to the goal. The goal probability of the final pass and shot y-coordinates also shows peaks, in these two cases the peak is symmetric around the y=50 line. However, the difference between goal probability at the peak and at the baseline is slightly lower, about 0.25, compared with the final pass distance and shot x-coordinate variables.

4.2 SMART Question 2: How do these variables and configuration vary with different teams?

Based on the t-test results, we now know that there is some evidence that there are differences between means of x & y coordinates and distance of key passes vs assists, and separately, x & y coordinates of shots and goals. An interesting follow-up question is whether or not teams build strategies around these variables and if so, do these strategies differ across teams? We can use boxplots to visualize and ANOVA to test for any differences.

For now, our team demonstrates this extension of our analyses using World Cup data for 5 teams with the highest frequency of shots + goals and assists + key passes. These 5 teams are Belgium, England, Croatia, Brazil, and France. This makes sense as, with the exception of Brazil, these teams were all semifinalists, and Brazil still making it to the quarterfinals of the 2018 World cup.

For each of these teams, we plot the x and y coordinates and distance of their final passes (not distinguishing between assists and key passes because strategy is made prior to the end result of a goal or non-goal). We also plot the x and y coordinates of their shots (once again, not distinguishing between goals and non-goals). See Figure 7 below.

With a p-value of 0.02, the ANOVA results show that there is some statistical evidence of differences across the five teams with respect to the x-coordinates of their final passes.

Similarly, with a p-value of 0.007, the ANOVA results show that there is some statistical evidence of differences across the five teams with respect to the x-coordinates of their shots.

These results indicate that the 5 teams do have differences in strategies in the World Cup when it comes to the x-coordinates of their shots and final passes. This matters because, from the t-tests already presented, differences in x-coordinate means of successful and unsuccessful events were statistically significantly different.

What the ANOVA results do not tell us is which among the 5 teams are different from each other. The natural next step here is to conduct some sort of pairwise comparisons that can identify which pairs reveal significant differences - something our team can consider in the future. Our team also hasn't reported the ANOVA results for variables that did not show statistically significant differences. This is due to the limited sample we used for demonstration purposes. We can consider running ANOVA over all teams in the World Cup and see from there if we see any differences. For now, it is enough to say that among the 5 teams chosen for this project, there is no evidence of different strategies when it comes to y-coordinates and distance.

**Chapter 5: Modelling Shot Success**

**Chapter 6: Conclusions**

1. Using t-tests to compare configurations between assists and key passes, and separately, goals and shots, the team found statistically significant differences in means of the following:

a. Starting x-coordinate of assists vs key passes. In particular, this result shows that, on average, final passes made further up the field are more likely to result in an assist than a key pass.

b. Distance covered by assists vs key passes. In particular, this shows that final passes that have to cover a shorter distance are more likely to result in an assist.

c. Starting x-coordinate of shots vs goals. In particular, we can conclude that attempts on goal taken further up the field are more likely to result in a goal.

2. From the probability plots on the same variables of interest, the team concludes the following:

a. First, goal probability varies with respect to all but one of the variables, which is the start x-coordinate of the final pass.

b. Second, goal probability shows the greatest change as functions of final pass distance and shot x-coordinate. That is, the probability of a goal appears most sensitive to these two variables.

3. Using ANOVA to test for differences in strategies among 5 teams in the World Cup (Belgium, England, Croatia, Brazil, and France), the team found some statistically significant differences among these teams with respect to the following:

a. Starting x-coordinate of final passes

b. Starting x-coordinate of shots

c. ANOVA does not give information as to which pairs of teams are different from each other, but we can conclude that at least some of the 5 teams, when compared to others, have different strategies when it comes to their starting points for passes and shots.

4. [PLACEHOLDER FOR MODELLING CONCLUSIONS]