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

**Chapter 1: Introduction**

Soccer is the most watched sport in the world and is a business for teams worth hundreds of millions of dollars. As a result, clubs put a great deal of time and resources towards achieving success. 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.

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.

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 At a glance

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. 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

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

The t-tests were done assuming a normal distribution of the event variables despite most of the distributions not being normal. This was quick and easy, even though there's room for error it still allows us to get a sense of the differences in variables the determine success or failure.

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: Which match variables result in the highest probability of a shot on goal being successful?

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

**Chapter 5: Modelling Shot Success**

**Chapter 6: Conclusions**