**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

3.3.1 Fitting Event Distributions

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

**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**