MATH1324 Assignment 2

Code

Modelling the Distribution of Football Goals

Group Details

- Casey-Ann Charlesworth (s3132392)

Problem Statement

Football goals scored within a game are discrete events and therefore should fit within a Poisson distribution. However, different leagues will have different scoring rates and in order to determine whether the Poisson distribution is accurate across the entire sport, we will analyse a number of leagues in Europe, reporting on both home and away goals separately, in order to determine whether there are underlying factors that distort an expected consistently Poisson distribution.

Load Packages

Hide

```
library(dplyr)
library(readr)
library(magrittr)
library(ggplot2)
```

Data

Data was collected from a number of disparate leagues in Europe, with data covering the 2016/17 season (incomplete) and 2015/16 season (complete). These two datasets were then combined (per league). Data for five leagues was initially collected, however, due to space restrictions, we will only report on three of them, being:

Scottish Premier League, link (http://www.football-data.co.uk/scotlandm.php), accessed 6 April 2017

Bundesliga 1 (Germany), link (http://www.football-data.co.uk/germanym.php), accessed 6 April 2017

Ethniki Katigoria (Greece), link (http://www.football-data.co.uk/greecem.php), accessed 6 April 2017

The CSV files were inspected and combined using Python, returning only the variables of Division, Full Time Home Goals, and Full Time Away Goals. All other data/variables were removed for this investigation.

The data was checked for outliers and any missing values (totalling <0.0015% of data collected) were removed.

Hide

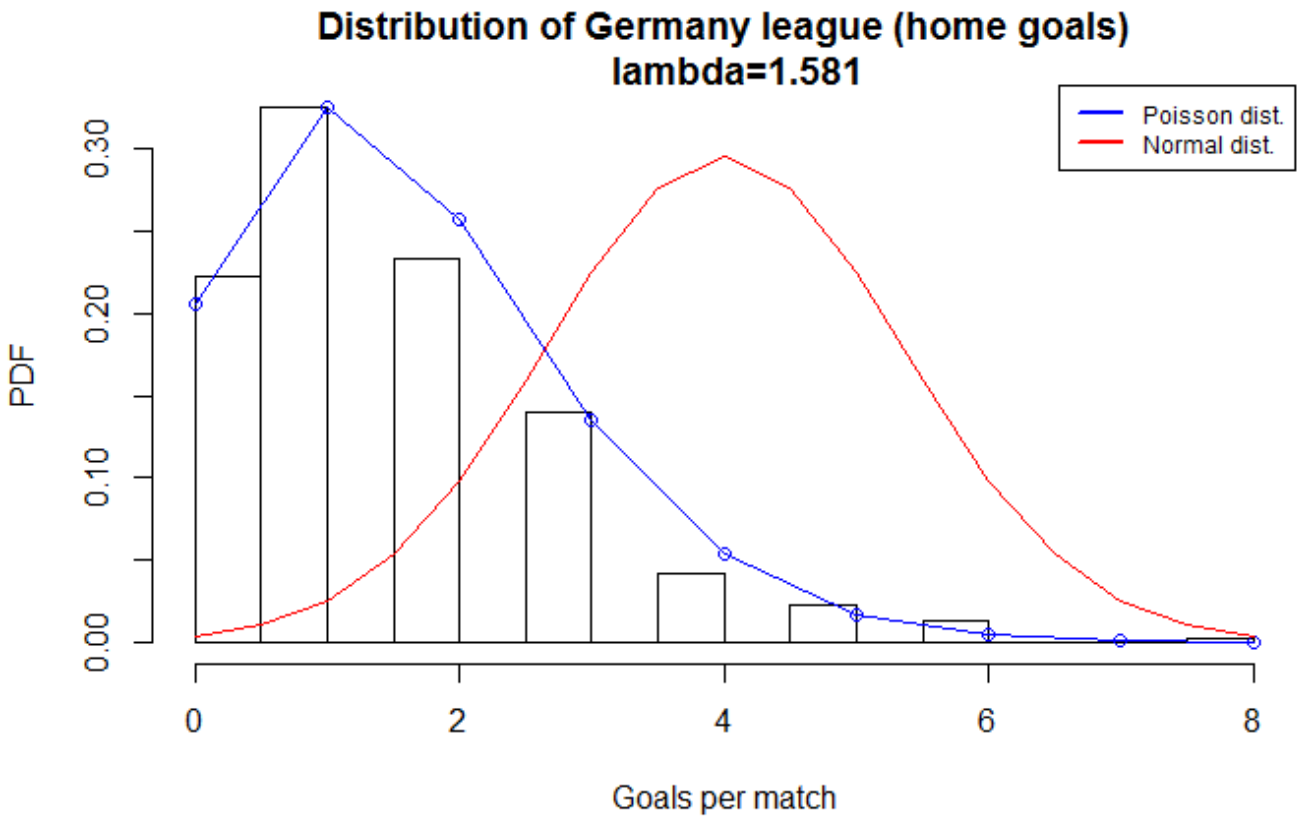
```
#Read data into R
Germany <- read_csv("~/Studies/RMIT - Master of Data Science/1710/IntroToStatistics/Assignmen
t2/Germany.csv")
Greece <- read_csv("~/Studies/RMIT - Master of Data Science/1710/IntroToStatistics/Assignment
2/Greece.csv")
Scotland <- read_csv("~/Studies/RMIT - Master of Data Science/1710/IntroToStatistics/Assignme
nt2/Scotland.csv")

#Update variables to correct types
Greece$Greece_home <- as.integer(Greece$Greece_home)
Greece$Greece_away <- as.integer(Greece$Greece_away)
```

Distribution Fitting

Below, each league has been visualised into a histogram with a Poisson distribution overlay. For comparison, a normal distribution has also been included to highlight the fit of the Poisson model.

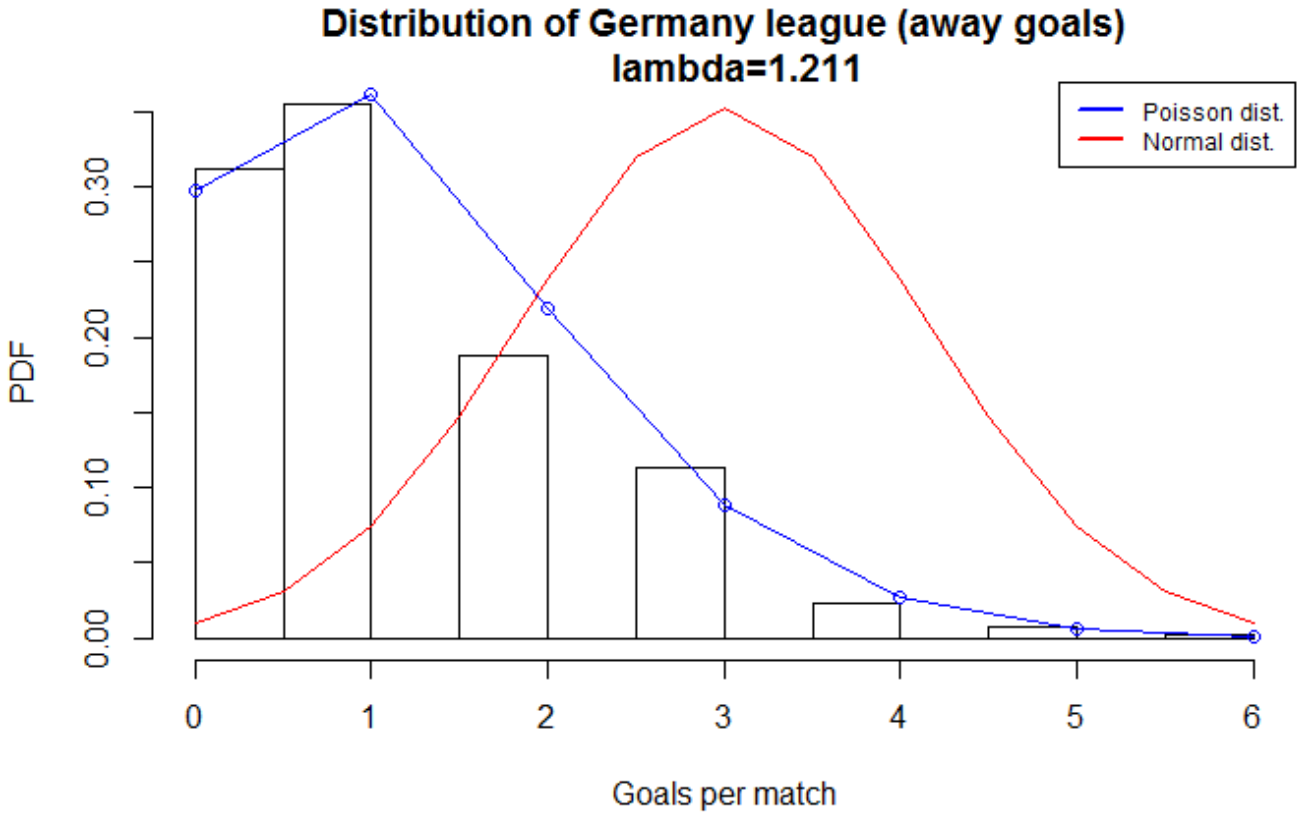
Hide

```
#Function to create histogram
createHist <- function(df, country, ha) {
  h <- hist(df, breaks=(max(df)*2), plot=FALSE)
  h$counts <- h$counts/sum(h$counts)
  plot(h, main=paste("Distribution of ", country, " league (", ha, " goals)\
   lambda=",round(mean(df),3), sep=""),
      xlab="Goals per match", ylab="PDF")
  points(c(0:max(df)), dpois(x=c(0:max(df)), lambda=mean(df)), type = "p", col = "blue")
  lines(c(0:max(df)), dpois(x=c(0:max(df)), lambda=mean(df)), type = "l",col = "blue")
  #Add in a normal distribution line for comparison
  lines(seq(0, max(df), by=.5), dnorm(seq(0, max(df), by=.5),
      max(df)/2, sd(df)), col="red")
  #Add legend
  legend("topright", c("Poisson dist.", "Normal dist."), lty=c(1,1), lwd=c(2.5,2.5),
         col=c("blue", "red"), cex=0.75)
  }
#Pass each variable to the function
createHist(Germany$Germany_home, "Germany", "home")
```



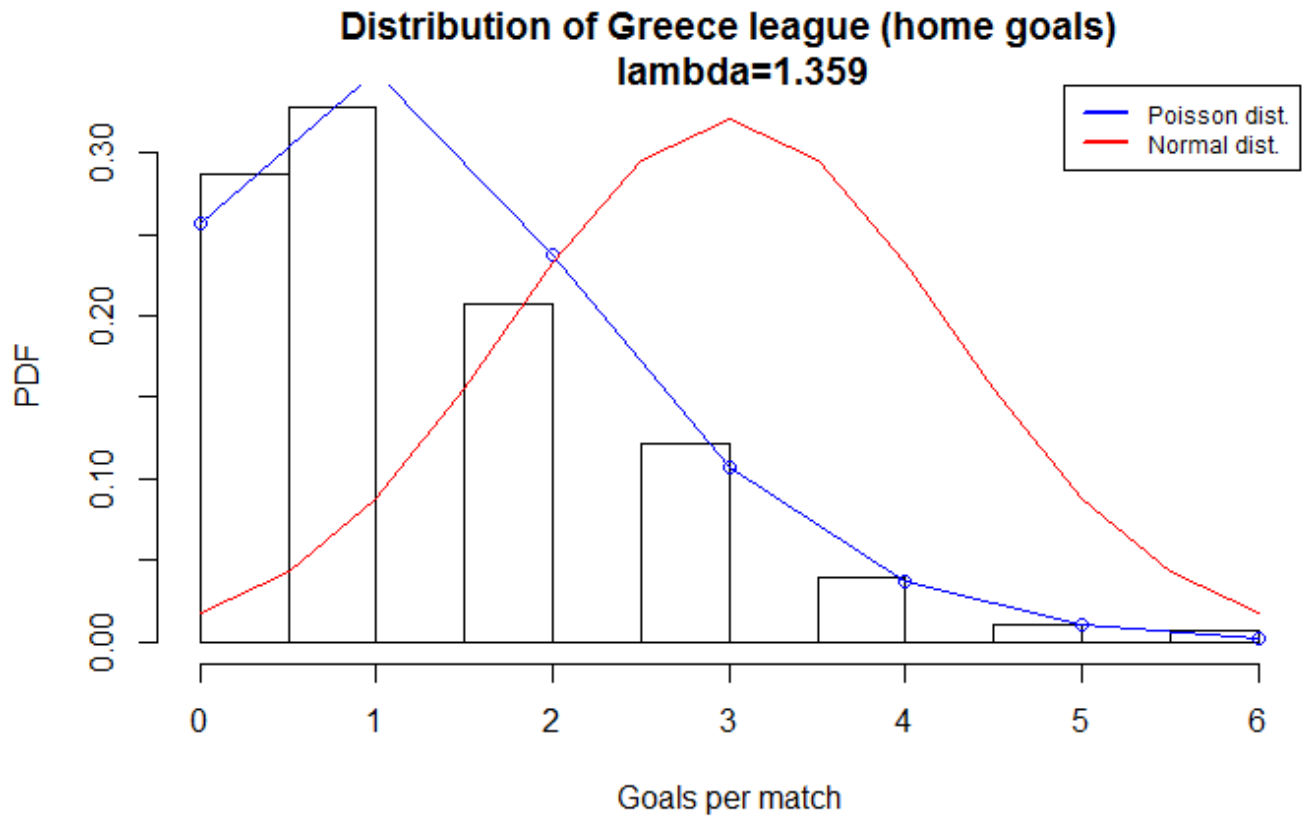
Hide

```
createHist(Germany$Germany_away, "Germany", "away")
```



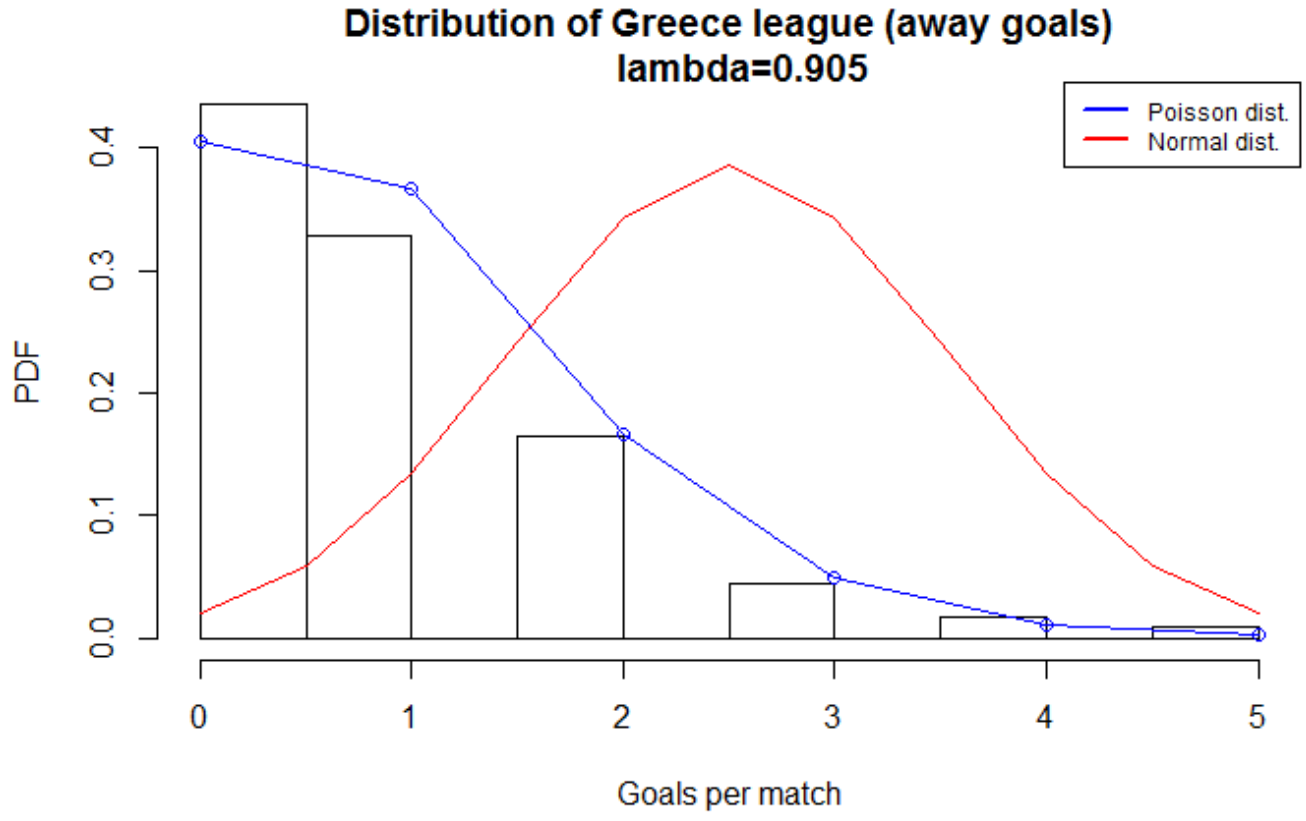
Hide

```
createHist(Greece$Greece_home, "Greece", "home")
```



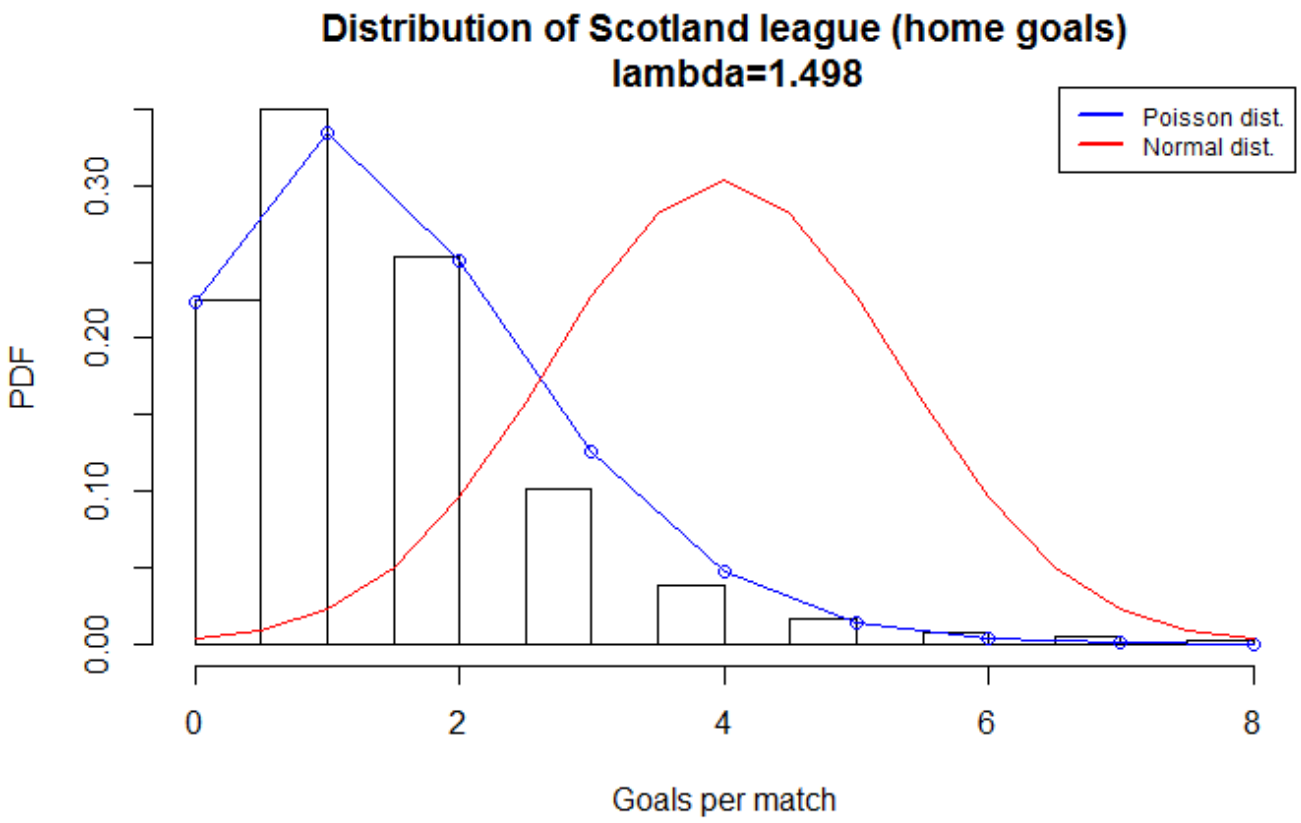
Hide

```
createHist(Greece$Greece_away, "Greece", "away")
```



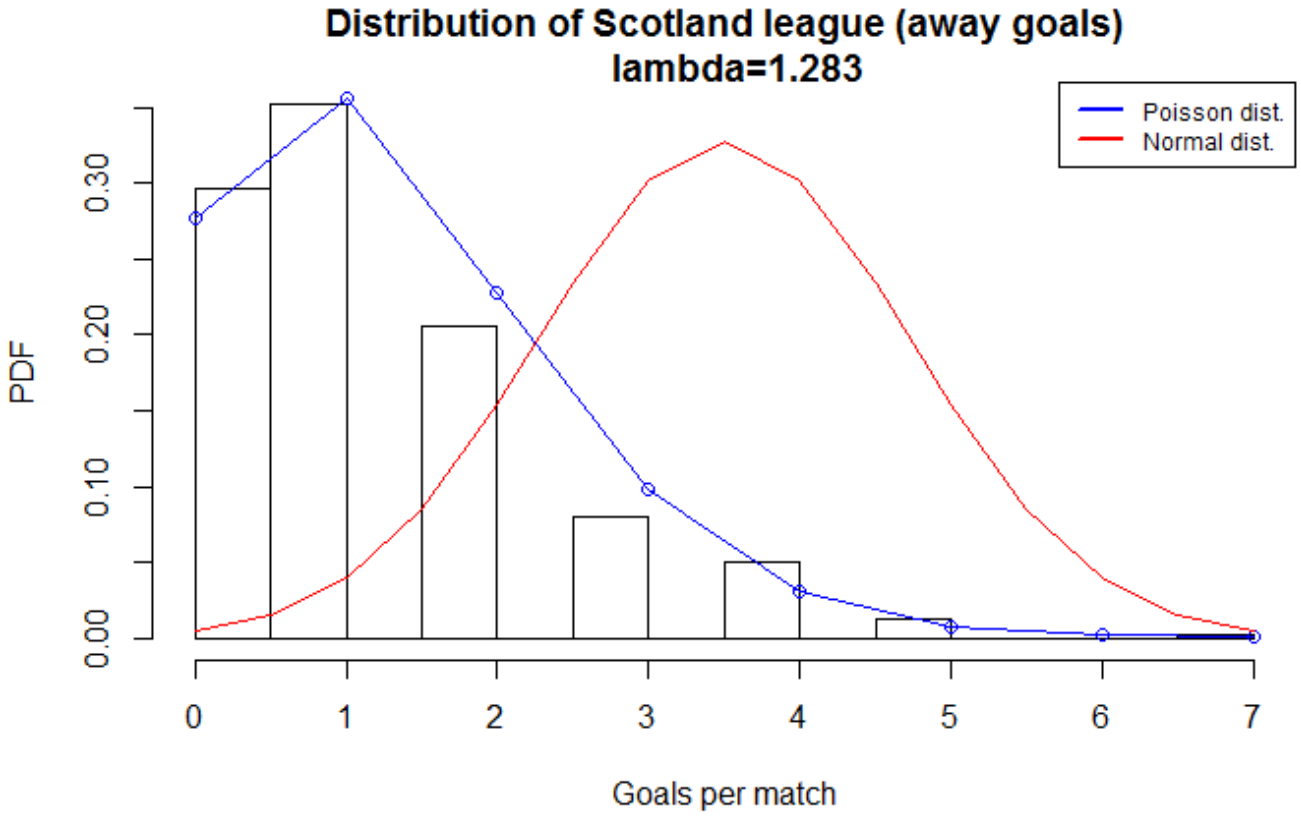
Hide

```
createHist(Scotland$Scotland_home, "Scotland", "home")
```



Hide

```
createHist(Scotland$Scotland_away, "Scotland", "away")
```



The Poisson distribution relies on a right-skew distribution (as shown above). To further highlight this, the table below has been created to show that a home team in each league has a probability of >0.9 of scoring between 0-3 goals per game, whereas away teams have a probability of >0.85 of scoring between 0-2 goals per game. Compare this with a team scoring 5 goals in a game (home team <.02; away team <.01). This indicates a right-skew distribution, and thus models the Poisson distribution.

Hide

```
#Function to create probabilities and join into a data frame
createProb <- function(df, df2) {
  c(round(ppois(3, mean(df))-ppois(-1, mean(df)),3),
       round(ppois(2, mean(df2))-ppois(-1, mean(df2)), 3),
       round(dpois(5, mean(df)), 3),
       round(dpois(5, mean(df2)), 3))
}
Germany_Pr <- createProb(Germany$Germany_home, Germany$Germany_away)
Greece_Pr <- createProb(Greece$Greece_home, Greece$Greece_away)
Scotland_Pr <- createProb(Scotland$Scotland_home, Scotland$Scotland_away)
df <- data.frame(Germany_Pr, Greece_Pr, Scotland_Pr)
row.names(df) <- c("Home (Pr(0<=X<=3))","Away (Pr(0<=X<=2))","Home (Pr(X=5))", "Away (Pr(X=
5))")
df
```

	Germany_Pr <dbl>	Greece_Pr <dbl>	Scotland_Pr <dbl>
Home (Pr(0<=X<=3))	0.924	0.951	0.935
Away (Pr(0<=X<=2))	0.877	0.936	0.861
Home (Pr(X=5))	0.017	0.010	0.014
Away (Pr(X=5))	0.006	0.002	0.008

4 rows

Interpretation

As predicted, Football scores are modelled on the Poisson distribution. Although different data was collected, and different probabilities of scoring were output, it is clear that the game of Football follows the predicted right-skewed pattern. Even the variances in scoring between leagues had no effect on the distribution, due to the sport having such a low scoring rate.

By overlaying the normal distribution, it was clear that this was not a fit. Higher scoring games that would not often see zero goals scored would be more likely to fit this distribution. However, in Football, it is quite common to see a zero scoreline. Therefore, this would naturally lead to a right-skew distribution.

It was interesting to note that the Greek league had a higher rate of zero goals scored for away teams. The datasets above were specifically chosen based on their locations and for their differing seasonal weather. It was initially thought that warmer climates would lead to higher scores, however, in the above analysis, this was not true at all. Therefore, it would be interesting to analyse this aspect further to see whether colder or warmer climates contribute toward higher scoring games and whether there are other weather factors that may have an effect on goal scoring.