Homework 01

Name

Task 1 - Poisson Distribution, Football Goals

This RMarkdown document provides a structured task for analyzing the dataset representing the number of goals scored in the last four World Cup football tournaments. It includes R code chunks that perform various analyses, including maximum likelihood estimation, plotting histograms and probability distributions, and assessing the goodness of fit for the Poisson and normal distributions.

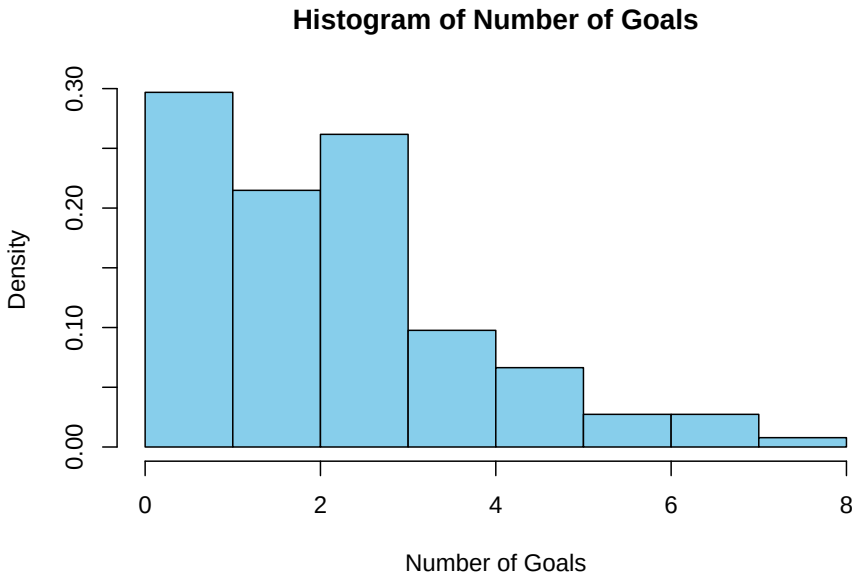
Background

You are provided with a dataset representing the number of goals scored in each game in the last four World Cup football tournaments. We are interested in understanding the distribution of these goals and comparing it with the theoretical Poisson distribution.

Data Exploration

Load the dataset and perform exploratory data analysis (EDA) to understand the distribution of the number of goals scored. Visualize the data using a histogram.

```
football_data <- read.csv("football_data.csv", header = FALSE)
data<- football_data[,1]
hist(data, freq = FALSE, main = "Histogram of Number of Goals",
     xlab = "Number of Goals", ylab = "Density", col = "skyblue", border = "black")
```



Maximum Likelihood Estimation (MLE)

We will use the maximum likelihood estimation method to estimate the parameter λ of the Poisson distribution for the football data.

Define the negative log-likelihood function for the Poisson distribution

```
negloglik_poisson <- function(par, x) {
  return(-sum(log(dpois(x, lambda = par))))
}
```

```
# Find the maximum likelihood estimate of lambda using optimization
op_poisson <- optimize(negloglik_poisson, interval = c(0, 10), x = data)

# Display the estimated parameter lambda
op_poisson$minimum
```

```
## [1] 2.566391
```

The estimated parameter λ for the Poisson distribution is approximately op_poisson$minimum.

Visualizing Poisson Distribution

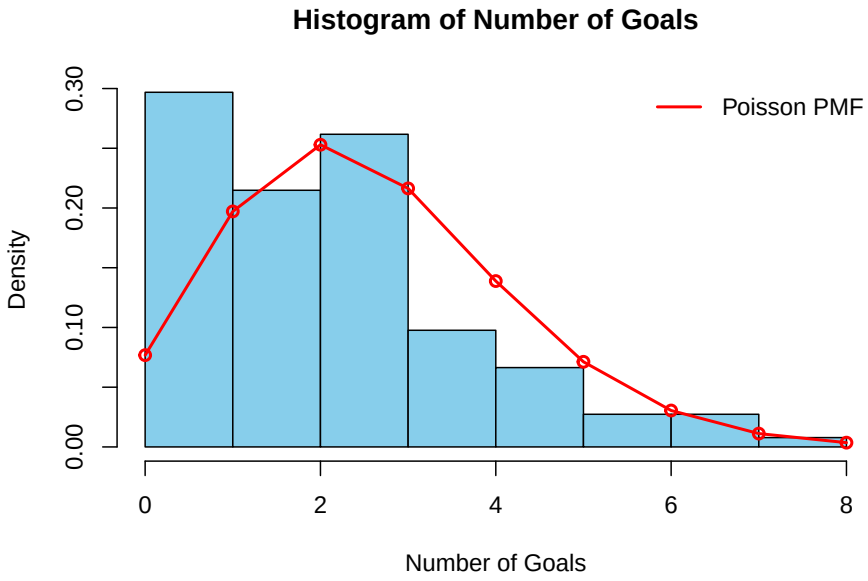
Plot the Poisson probability mass function (PMF) with the estimated λ alongside the histogram of the football data.

```
# Generate x values for PMF
x_values <- 0:max(data)

# Calculate PMF for estimated lambda
pmf <- dpois(x_values, lambda = op_poisson$minimum)


hist(data, freq = FALSE, main = "Histogram of Number of Goals",
     xlab = "Number of Goals", ylab = "Density", col = "skyblue", border = "black")

# Overlay PMF on histogram
lines(x_values, pmf, type = "o", col = "red", lwd = 2)
legend("topright", legend = "Poisson PMF", col = "red", lwd = 2, bty = "n")
```

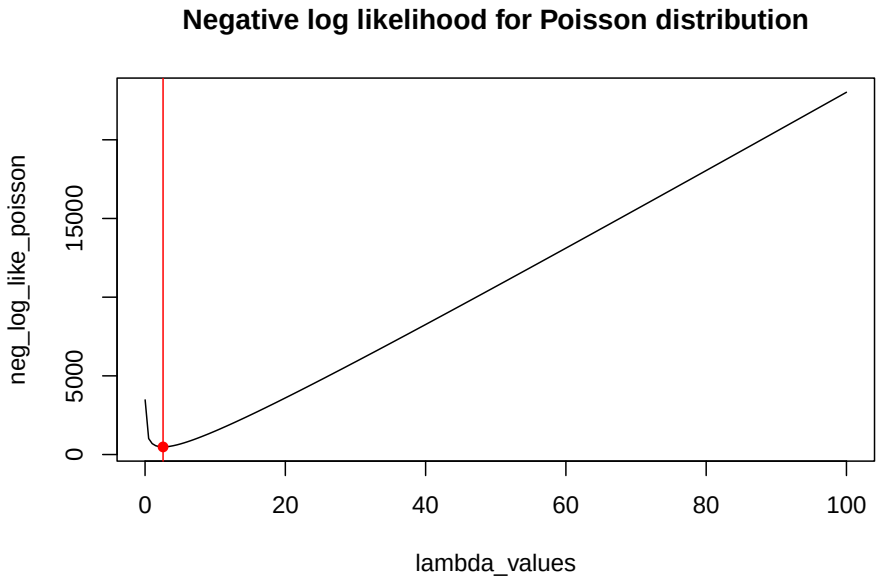


The red line represents the Poisson probability mass function with $\lambda \approx 2.5663905$, overlaid on the histogram of the football data.

```r
# Calculate the negative log likelihood for a range of lambda values
lambda_values <- seq(0.01, 100, length.out = 200)
neg_log_like_poisson <- rep(0, length(lambda_values))

for (i in 1:length(lambda_values)) {
  neg_log_like_poisson[i] <- negloglik_poisson(lambda_values[i], x = data)
}

# Plot the negative log likelihood function for lambda
plot(lambda_values, neg_log_like_poisson, type = "l", main = "Negative log likelihood for Poisson distri
points(op_poisson$minimum, op_poisson$objective, col = "red", pch = 16)
abline(v = op_poisson$minimum, col = "red")
```



Normal Distribution Approximation

Estimate Parameters

Estimate the parameters (mean and standard deviation) of the normal distribution that best approximate the Poisson distribution.

```
# Estimate parameters of normal distribution
mu_hat <- mean(data)
sigma_hat <- sqrt(var(data) * (length(data) - 1) / length(data))

# Print the estimated parameters
mu_hat
```

```
## [1] 2.566406
```

```
sigma_hat
```

```
## [1] 1.698886
```

Plot the histogram of the data along with the probability density function (PDF) of both the estimated Poisson distribution and the normal distribution.

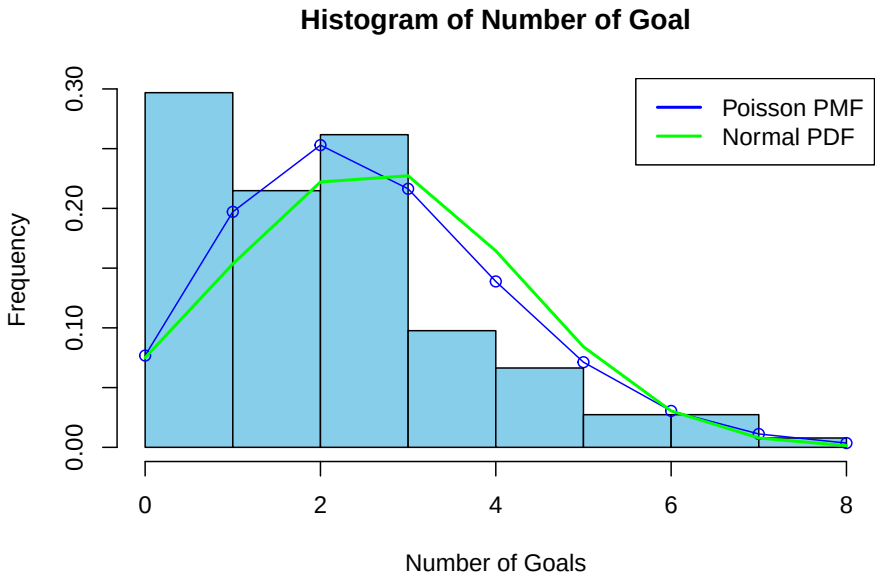
```
# Plot histogram of data
hist(data, freq = FALSE, main = "Histogram of Number of Goal",
     xlab = "Number of Goals", ylab = "Frequency", col = "skyblue")

xx <- 0:max(data)
# Plot PMF of estimated Poisson distribution
points(xx, pmf, type = "o", col = "blue")

# Plot PDF of estimated normal distribution
pdf_normal <- dnorm(xx, mean = mu_hat, sd = sigma_hat)
lines(xx, pdf_normal, col = "green", lwd = 2)

# Add legend
```

```r
legend("topright", legend = c("Poisson PMF", "Normal PDF"),
       col = c("blue", "green"), lty = 1, lwd = 2)
```



Conclusion

In this task, we explored the distribution of the number of goals scored in the last four World Cup football tournaments. We estimated the parameters of both the Poisson and normal distributions and assessed their goodness of fit to the data.

Task 2 - Principal component analysis, correlations

In this task, you'll be handling the dataset stored in the file named `food_data.RData`, located within the interactive outline vault labeled as `Homework 1`. Inside this file, you'll find a data table named `data.short`. This table contains information about the significance of various products sold in shops within a particular city. Each value within the table represents the proportion of items sold in different shops relative to the total number of items sold during different times of the day. The columns of the table correspond to different shop products, while the rows represent various time intervals throughout the day. Your goal is to investigate the relationships of different products importances during the day using PCA and correlation analysis.

a)

Compute the sample correlation coefficient for each pair of variables (shop products) and **interpret** it:

Variables	Results	Interpretation
`example`	0	The correlation is zero, which means...
`alcohol, snacks`	`0.68`	`insert`
`alcohol, babies`	`-0.73`	`insert`
`breakfast, dairy eggs`	`0.95`	`insert`

b)

Perform a principal component analysis of provided data table. State which variable (shop types) has the most influence on the first three components:

PC1	PC2	PC3
babies	produce	snacks

Perform PCA:

```
data.pca <- prcomp(data.short, scale = TRUE)
```

Code for determining the most important variables for the first 3 principal components:

```
s <- summary(data.pca)
score <- s$rotation
```

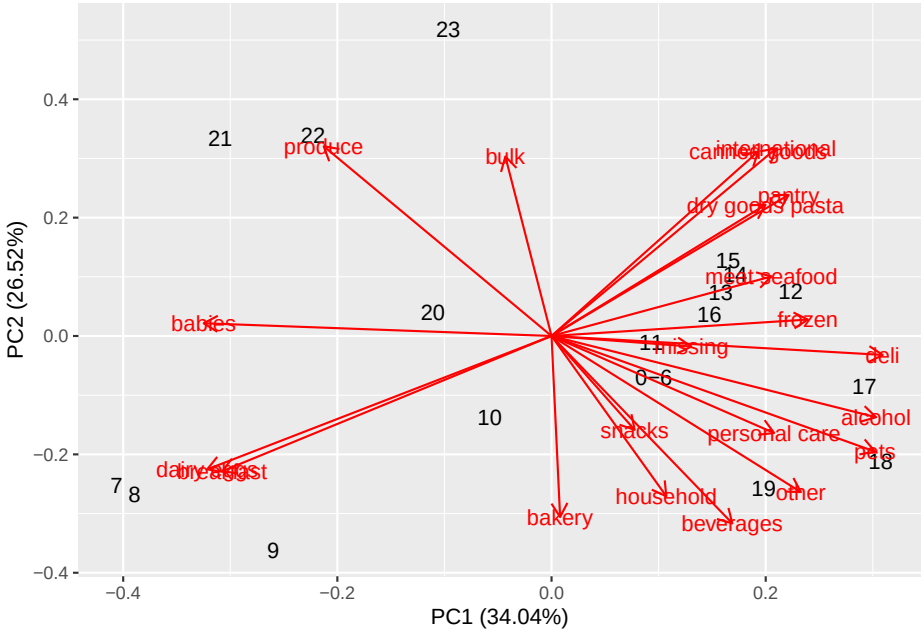
c)

How many of components do You need to explain **at least 90%** of the total amount of variance in Your data?

Number of components
6

d)

Create a scatter plot of data points using the first two components. What is your evaluation of the final plot? Can you decide if there is any division of the observations (different daytimes)? What is the typical product selled in the morning hours?



Question	Answer
Interpret the result	`insert`
Are there any division?	`yes, ...`
Typical product for the morning	`breakfast`