Homework 01

Name

Task 1 - Poisson Distribution Analysis for Football Data

In this task, you will analyze the dataset provided in the file `football_data.csv` (available in the interactive outline vault `Homework 1`). This dataset contains the number of goals scored in each game during the last four World Cup football tournaments. Your goal is to explore the distribution of these goals and compare it with the theoretical Poisson distribution.

A)

Using the provided dataset `football_data.csv`, create a histogram to visualize the distribution of the number of goals scored in World Cup football tournaments.

```
# INSERT your code here (histogram)
```

B)

Define a negative log-likelihood function for the Poisson distribution, given a parameter λ and a dataset `x`.

```
# INSERT your code here (function definition)
```

C)

Using optimization techniques, find the maximum likelihood estimate of the parameter λ for the Poisson distribution with the given dataset, and then display (**print**) the estimated value of λ.

```
# INSERT your code here (optimalization + displaying the values)
```

D)

Visualize the Poisson probability mass function (PMF) with the estimated parameter λ alongside the histogram of the football data. What insights can you derive from this visualization?

```
# INSERT your code here (histogram + PMF)
```

Your interpretation:

E)

Plot the negative log-likelihood function for the Poisson distribution with varying λ values and your numerical estimation (red point). Describe the importance of this plot and how it helps in estimating the parameter lambda for the Poisson distribution.

```
# INSERT your code here (plot)
```

F)

Estimate the parameters (mean and standard deviation) of the normal distribution that best approximate the Poisson distribution (using MLE). Then, plot the histogram of the data along with the probability

density functions (PDFs) of both the estimated Poisson distribution and the normal distribution. Discuss the differences between the two distributions and how the parameters affect their shapes; which model is better for our data?

```
# INSERT your code here (MLE for normal distribution + histogram + density)
```

Your interpretation:

Conclusion

Please **conclude** your analysis by summarizing the exploration of the distribution of the number of goals scored in the last four World Cup football tournaments, including the estimation of parameters for both the Poisson and normal distributions.

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`	`insert`	`insert`
`alcohol`, `babies`	`insert`	`insert`
`breakfast`, `dairy eggs`	`insert`	`insert`

B)

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

Perform PCA:

```
# INSERT your code for here (PCA)
```

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

```
# INSERT your code for here (importances of variables)
```

PC1	PC2	PC3
`insert variable name`	`insert variable name`	`insert variable name`

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
`insert number`

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?

Task	Answer
Interpret the result	`insert`
Are there any division?	`insert`
Typical product for the morning	`insert`