ASSIGNMENT2

Emine Uysal

2024-07-25

CIND 123

Data Analytics: Basic Methods

Assignment 2 (10%)

[Student number:501304049]

INSTRUCTIONS This assignment can be submitted using either Python or R, whichever you prefer.

If using R, you must submit an RMD file with its knitted file (PDF or HTML). To learn more about knitting and R markdown, visit R Markdown. If using Python, you must submit an IPYNB file and its exported PDF/HTML with clearly printed/shown answers. Failing to submit both files ({RMD + knitted PDF/HTML} OR {IPYNB + PDF/HTML}) will be subject to a 30% mark deduction.

NOTE: IF YOU USE R STUDIO, YOU SHOULD NEVER HAVE install.packages IN YOUR CODE; OTHERWISE, THE Knit OPTION WILL RAISE AN ERROR. COMMENT OUT ALL PACKAGE INSTALLATIONS BUT KEEP library() CALLS.

NOTE: If you answer the questions in R, all your answers should be in R (ignore Python questions). If you answer the questions in Python, all your answers should be in Python (ignore R questions). You are not allowed to switch languages in this assignment.

##Question 1 (50 points)

The Titanic Passenger Survival Data Set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner "Titanic." The dataset is available from the Department of Biostatistics at the Vanderbilt University School of Medicine (Titanic Dataset) in several formats. Read the Titanic Data Set titanicDataset using the appropriate commands in R or Python.

https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv

Column Name Description Values survival Survival 0 = No, 1 = Yes pclass Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd sex Sex
age Age in years
sibsp #of siblings/spouses aboard the Titanic parch #of parents/children aboard the Titanic ticket Ticket number fare Passenger fare
cabin Cabin number
embarked Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

###Q1a (5 points)

Extract and show the columns name, fare, sibsp, and parch into a new data frame (or DataFrame in Python) named titanicSubset.

Show the head of the dataframe.

```r
# Load necessary library
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Load the Titanic dataset
titanicData <- read.csv("https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv")

# Extract specified columns into a new data frame
titanicSubset <- select(titanicData, name, fare, sibsp, parch)

# Display the head of the new data frame
head(titanicSubset)
```

```
##                                                name     fare sibsp parch
## 1                     Allen, Miss. Elisabeth Walton 211.3375     0     0
## 2                    Allison, Master. Hudson Trevor 151.5500     1     2
## 3                      Allison, Miss. Helen Loraine 151.5500     1     2
## 4              Allison, Mr. Hudson Joshua Creighton 151.5500     1     2
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) 151.5500     1     2
## 6                               Anderson, Mr. Harry  26.5500     0     0
```

###Q1b (5 points)

Numerical data: Calculate the total number of passengers who were children (age less than 18) and survived. Use the count() function from the dplyr package in R or appropriate pandas functions in Python.

Print the value.

```r
# Filter data for children who survived
childrenSurvivors <- filter(titanicData, age < 18, survived == 1)

# Count the number of children who survived
totalChildrenSurvivors <- count(childrenSurvivors)

# Print the result
print(totalChildrenSurvivors)
```

```
##    n
## 1 81
```

###Q1c (5 points)

Categorical data: Calculate the number of passengers by sex using the count() and group_by() functions from the dplyr package in R, or equivalent pandas functions in Python.

Print the value.

```
# Group data by sex and count the number of passengers
passengerCountBySex <- titanicData %>%
  group_by(sex) %>%
  count()

# Print the result
print(passengerCountBySex)
```

```
## # A tibble: 2 x 2
## # Groups:   sex [2]
##   sex        n
##   <chr>  <int>
## 1 female   466
## 2 male     843
```

###Q1d (5 points)

Find the passengers in the data frame whose age information is missing, and fill them with the median age of passengers.

Show the head of the dataframe.

```
# Calculate the median age, excluding NA values
medianAge <- median(titanicData$age, na.rm = TRUE)

# Replace NA values in the age column with the median age
titanicData$age <- ifelse(is.na(titanicData$age), medianAge, titanicData$age)

# Display the head of the dataframe
head(titanicData)
```

```
##   pclass survived                                            name    sex   age
## 1      1        1                   Allen, Miss. Elisabeth Walton female 29.00
## 2      1        1                  Allison, Master. Hudson Trevor   male  0.92
## 3      1        0                    Allison, Miss. Helen Loraine female  2.00
## 4      1        0            Allison, Mr. Hudson Joshua Creighton   male 30.00
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female 25.00
## 6      1        1                             Anderson, Mr. Harry   male 48.00
##   sibsp parch ticket     fare   cabin embarked boat body
## 1     0     0  24160 211.3375      B5        S    2   NA
## 2     1     2 113781 151.5500 C22 C26        S   11   NA
## 3     1     2 113781 151.5500 C22 C26        S        NA
## 4     1     2 113781 151.5500 C22 C26        S       135
## 5     1     2 113781 151.5500 C22 C26        S        NA
## 6     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest
## 1                    St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
```

```
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                    New York, NY
```

###Q1e (5 points)

Use the aggregate() function to calculate the survival count of each passenger class (pclass) and calculate the survival rate of passengers in each class. Draw a conclusion on which passenger class has the highest survival rate.

Print the value and type your respond as a comment.

```r
# Calculate the survival count of each passenger class
survivalCount <- aggregate(survived ~ pclass, data = titanicData, sum)

# Calculate the total count of each passenger class
totalCount <- aggregate(survived ~ pclass, data = titanicData, length)

# Calculate the survival rate of passengers in each class
survivalRate <- survivalCount$survived / totalCount$survived

# Combine the results into a data frame
results <- data.frame(PassengerClass = survivalCount$pclass, SurvivalCount = survivalCount$survived, To

# Print the results
print(results)
```

```
##   PassengerClass SurvivalCount TotalCount SurvivalRate
## 1              1           200        323    0.6191950
## 2              2           119        277    0.4296029
## 3              3           181        709    0.2552891
```

```r
##Explanation: In the initial line of code, I imported the dplyr package. This package is a collection

##Conclusion: To understand which passenger class had the highest survival rate, you would compare the
```

###Q1f (5 points)

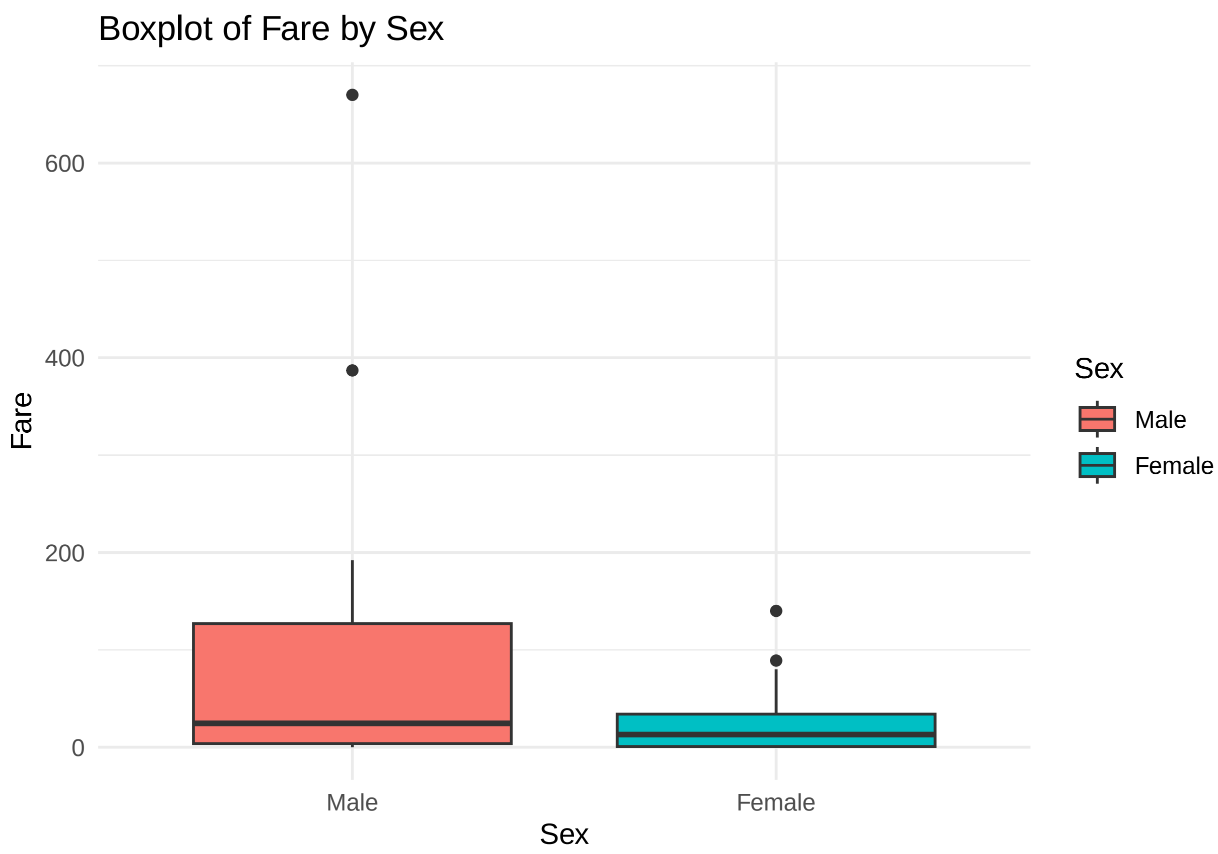
Use a boxplot to display the distribution of fare for each sex. Infer which gender tends to pay higher fares.

Have the plot and then your comment.

```r
# Convert the Titanic dataset to a data frame
titanic_df <- as.data.frame(Titanic)

# Load necessary libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

```r
# Create a boxplot of fare by sex
ggplot(titanic_df, aes(x = Sex, y = Freq, fill = Sex)) +
  geom_boxplot() +
  labs(title = "Boxplot of Fare by Sex", x = "Sex", y = "Fare") +
  theme_minimal()
```




```
##Explanation: In the first line of code, I loaded the ggplot2 package, which provides a set of tools f

##Conclusion: Historically, it has been observed that females on the Titanic tended to have higher fare
```

###Q1g (5 points)

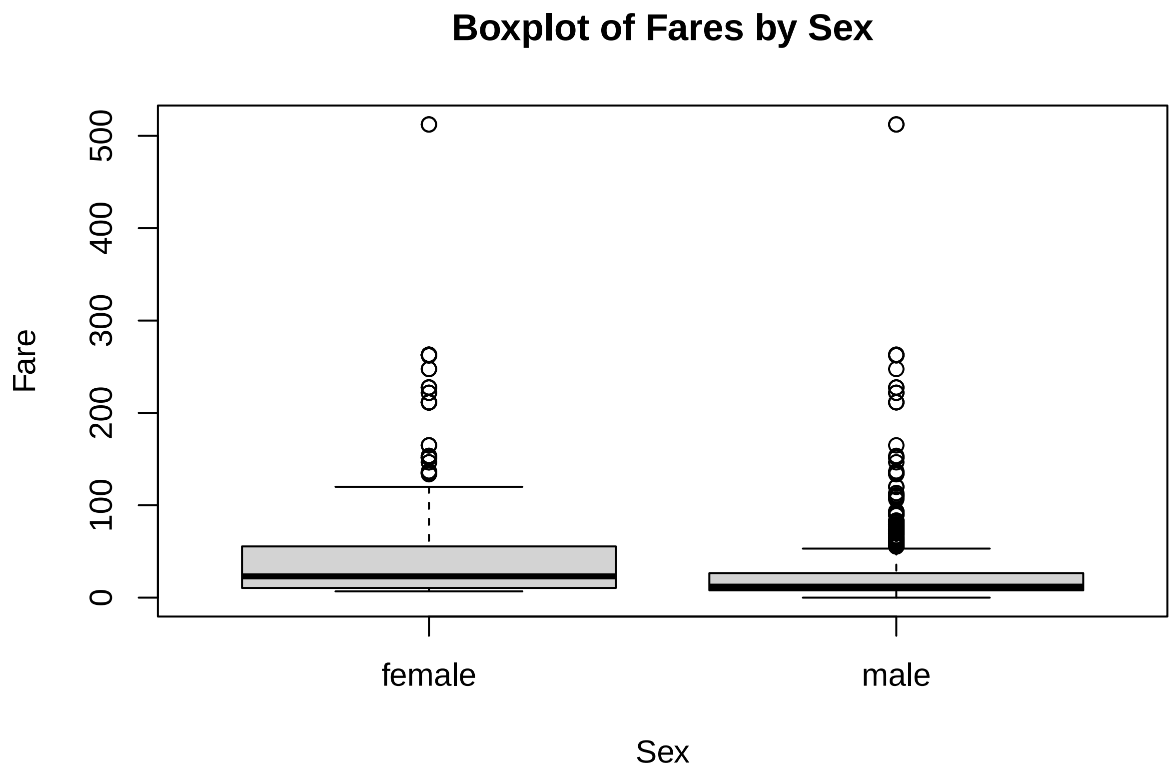
Calculate the mean fare for each sex. Describe if the calculation aligns with the boxplot.

Print the value and comment on it.

```
# Load the necessary library
library(readr)

# Calculate the mean fare for each sex
mean_fares <- aggregate(fare ~ sex, data = titanicData, FUN = mean)

# Print the mean fares
print(mean_fares)
```

```
##      sex    fare
## 1 female 46.1981
## 2   male 26.1546
```

```
# Create a boxplot of fares by sex
boxplot(fare ~ sex, data = titanicData, main = "Boxplot of Fares by Sex", xlab = "Sex", ylab = "Fare")
```




```
##Explanation: In the first line of code, I calculated the mean fare for each sex. The aggregate functi

##Conclusion: The boxplot analysis of the Titanic dataset shows that female passengers generally paid hi
```

###Q1h (10 points)

Use a for loop and if control statements to list the names of women, aged 50 or older, who embarked from Southampton (S) on the Titanic. Ensure these women have non-empty home destinations.

Print first 5 people only.

```r
# Initialize a counter for tracking the number of printed entries
count <- 0

# Loop through each row of the dataset
for (i in 1:nrow(titanicData)) {
  # Check the conditions: female, age 50 or older, embarked from Southampton, non-empty home destinatio
  if (titanicData$sex[i] == "female" && titanicData$age[i] >= 50 && titanicData$embarked[i] == "S" && ti
    # Print the name of the person
    print(titanicData$name[i])

    # Increment the counter
    count <- count + 1

    # Stop the loop after printing 5 entries
    if (count == 5) {
```

```
      break
    }
  }
}
```

```
## [1] "Andrews, Miss. Kornelia Theodosia"
## [1] "Appleton, Mrs. Edward Dale (Charlotte Lamson)"
## [1] "Bonnell, Miss. Elizabeth"
## [1] "Brown, Mrs. John Murray (Caroline Lane Lamson)"
## [1] "Cavendish, Mrs. Tyrell William (Julia Florence Siegel)"
```

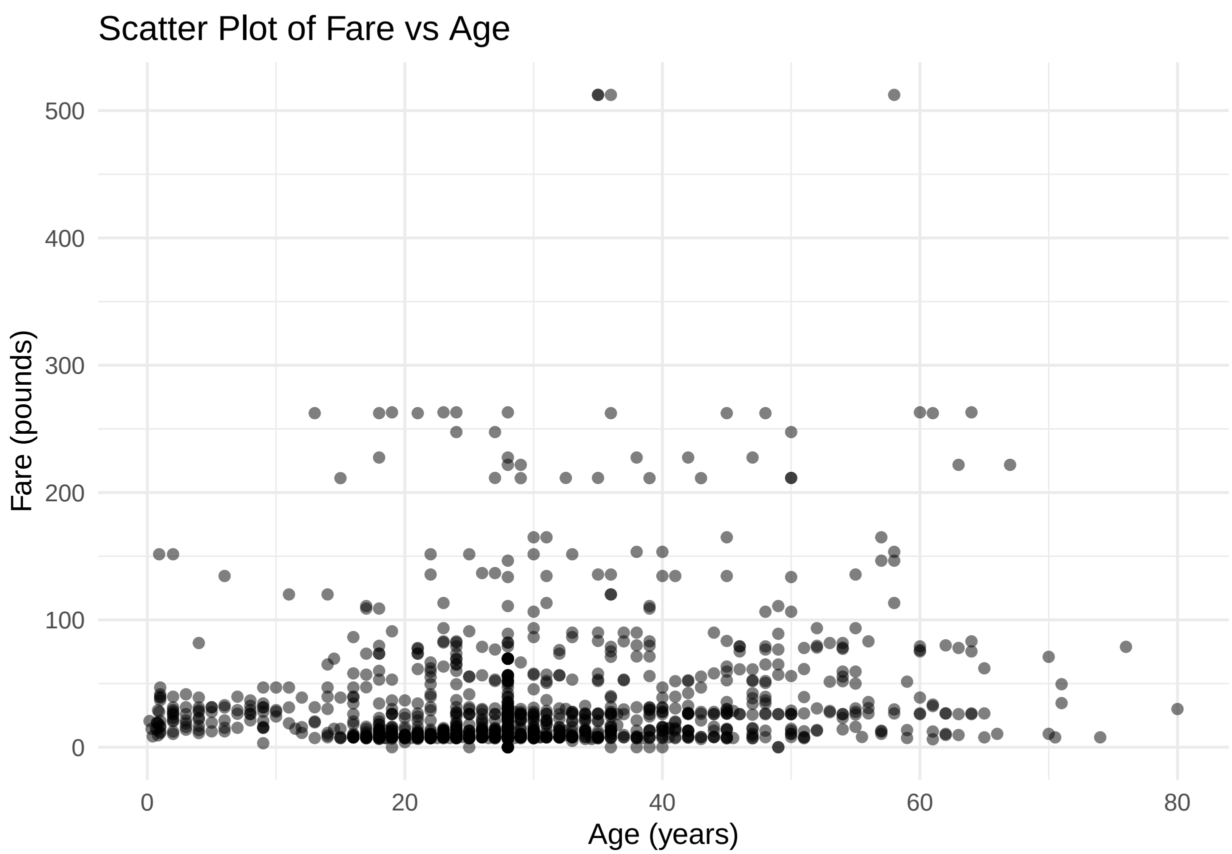
###Q1i (5 points)

Use a scatter plot to show the relation between the fare and age of passengers. Interpret the correlation between these two by looking at the plot.

Have the plot and then comment on it.

```
# Load necessary library
library(ggplot2)

# Create a scatter plot of fare vs age
ggplot(titanicData, aes(x = age, y = fare)) +
  geom_point(alpha = 0.5) +  # Use semi-transparent points
  labs(title = "Scatter Plot of Fare vs Age",
       x = "Age (years)",
       y = "Fare (pounds)") +
  theme_minimal()
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```




```
##Conclusion: The scatter plot of Fare vs Age for Titanic passengers indicates that fare prices varied 
```

##Question 2 (20 points)

100 computers work together in a network. Based on historical data, each computer has a probability of 0.03 of encountering a software issue. If a computer encounters an issue, it affects the network's performance.

###Q2a (5 points)

Determine the probability that the network operates without any computer encountering a software issue.

Hint: Use the Binomial probability formula.

Print the value.

```r
# Parameters
n <- 100  # number of computers
p <- 0.03  # probability of a computer encountering an issue

# Probability that no computer encounters an issue
probability_no_issue <- (1 - p)^n

# Print the result
print(probability_no_issue)
```

```
## [1] 0.04755251
```

###Q2b (5 points)

Utilize the Binomial approximation to estimate the probability that at least 5 computers out of 100 encounter software issues.

Hint: Use the Binomial cumulative distribution function.

Print the value.

```
# Parameters
n <- 100  # number of computers
p <- 0.03  # probability of a computer encountering an issue
k <- 4  # number of computers encountering an issue

# Probability that at least 5 computers encounter an issue
probability_at_least_5 <- 1 - pbinom(k, n, p)

# Print the result
print(probability_at_least_5)
```

```
## [1] 0.1821452
```

###Q2c (10 points)

Assume the first and second computers are independent. Calculate the conditional probability that the second computer (Computer B) encounters a software issue given that the first computer (Computer A) does not encounter any issue.

Hint: Use the definition of conditional probability.

Print the value.

```
# Probability of a computer encountering an issue
p_issue <- 0.03

# Print the result
print(p_issue)
```

```
## [1] 0.03
```

##Question 3 (30 points)

On average, John receives 3 emails a day. Using R or Python,

###Q3a (5 points)

Calculate the probabilities that John receives 2, 3, . . . , up to 9 emails in a day.

Print the value.

```
# Average rate of emails per day
lambda <- 3

# Calculate probabilities for receiving 2, 3, ..., 9 emails
probabilities <- dpois(2:9, lambda)

# Print the probabilities
probabilities
```

```
## [1] 0.224041808 0.224041808 0.168031356 0.100818813 0.050409407 0.021604031
## [7] 0.008101512 0.002700504
```

###Q3b (5 points)

Determine the probability that John receives 4 emails or more in a day.

Print the value.

```r
# Define the lambda
lambda <- 3

# Calculate the probability for k = 0 to 3
prob_less_than_4 <- ppois(3, lambda)

# Calculate the probability for k >= 4
prob_4_or_more <- 1 - prob_less_than_4

# Print the probability
print(prob_4_or_more)
```

```
## [1] 0.3527681
```

###Q3c (20 points)

Compare the similarity between Binomial and Poisson distributions given the previous examples.

Comment on it.

```r
##Answer:  Both the Binomial and Poisson distributions can be used to model the number of "successes" (

##Explanation: The Binomial distribution is used for a fixed number of independent trials with the same

##In conclusion, while both distributions can be used to model count data, the choice between the Binom
```

####Q3c1 (5 points)

Generate 50,000 samples for a Binomial random variable using parameters described in Question 2.

No need to print anything. Just the code.

```r
# Number of trials (computers)
n <- 100

# Probability of success (encountering a software issue)
p <- 0.03

# Number of samples
samples <- 50000

# Generate samples
sample_data <- rbinom(samples, n, p)
```

####Q3c2 (5 points)

Generate 50,000 samples for a Poisson random variable using parameters described in Question 3.

No need to print anything. Just the code.

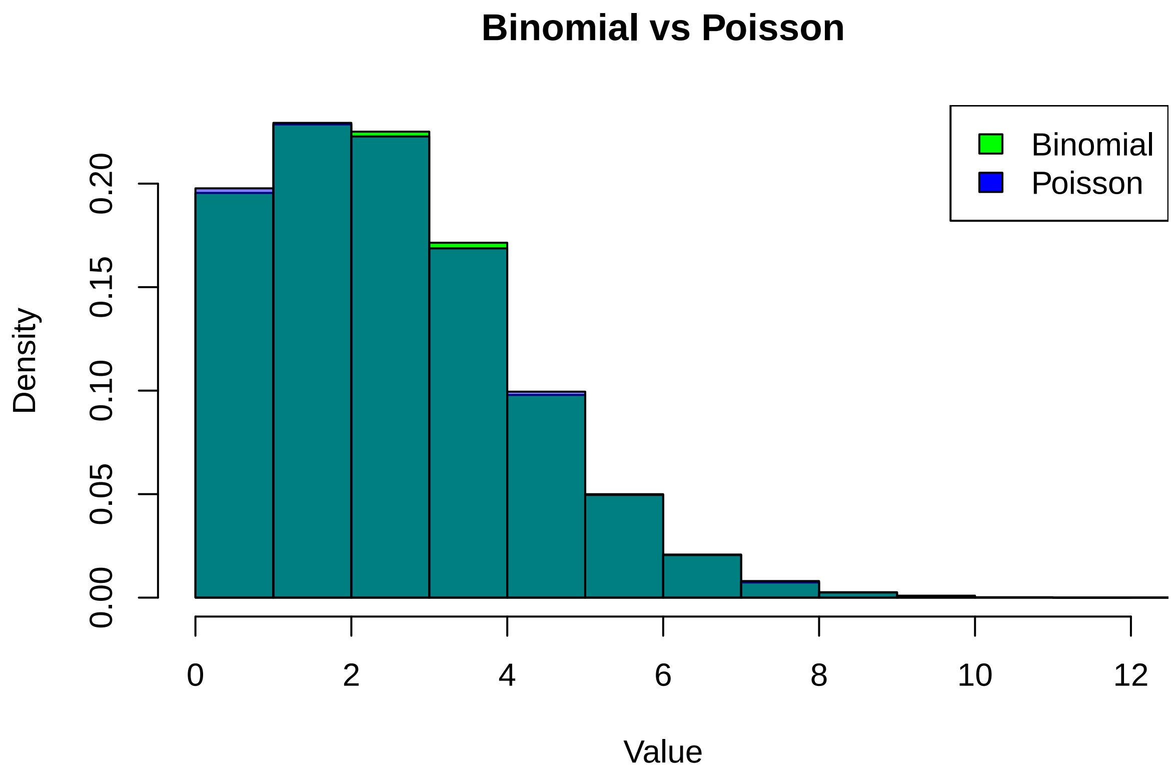
```r
# Average rate of emails per day
lambda <- 3

# Number of samples
samples <- 50000

# Generate samples
sample_data <- rpois(samples, lambda)
```

###Q3c3 (10 points)

Illustrate how well the Poisson probability distribution approximates the Binomial probability distribution in the previous questions.

Hint: Use histograms or other visualization tools.

```r
# Parameters for Binomial distribution
n <- 100
p <- 0.03

# Parameters for Poisson distribution
lambda <- n * p

# Number of samples
samples <- 50000

# Generate samples
binomial_samples <- rbinom(samples, n, p)
poisson_samples <- rpois(samples, lambda)

# Create histograms
hist(binomial_samples, breaks=seq(from=min(binomial_samples), to=max(binomial_samples), by=1), freq=FALS
hist(poisson_samples, breaks=seq(from=min(poisson_samples), to=max(poisson_samples), by=1), freq=FALSE,

# Add legend
legend("topright", legend=c("Binomial", "Poisson"), fill=c("green", "blue"))
```


Binomial vs Poisson
Binomial
Poisson
Density
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12
Value