CIND 123 - Data Analytics: Basic Methods

Assignment 2 (10%)

[Qian (Jessie) Ma]

[CIND 123 Section D40 & student number: 501274167]

Instructions

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Review this website for more details on using R Markdown http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

Use RStudio for this assignment. Complete the assignment by inserting your R code wherever you see the string "#INSERT YOUR ANSWER HERE".

When you click the **Knit** button, a document (PDF, Word, or HTML format) will be generated that includes both the assignment content as well as the output of any embedded R code chunks.

Submit **both** the rmd and generated output files. Failing to submit both files will be subject to mark deduction.

Sample Question and Solution

Use `seq()` to create the vector $(100, 97 \ldots, 4)$.

```
seq(100, 3, -3)
```

```
##  [1] 100  97  94  91  88  85  82  79  76  73  70  67  64  61  58  55  52  49  46
## [20]  43  40  37  34  31  28  25  22  19  16  13  10   7   4
```

Question 1 (40 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 (https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv (https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv)) in several formats. Store the Titanic Data Set `titanic_train` using the following commands.

```
#install.packages("titanic")
library(titanic)
titanicDataset <- read.csv(file = "https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv", stringsAsFactors = F)
str(titanicDataset)
```

```
## 'data.frame':    1309 obs. of  14 variables:
##  $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : chr  "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. Helen Loraine" "Allison, Mr. Hudson Joshua Creighton" ...
##  $ sex      : chr  "female" "male" "female" "male" ...
##  $ age      : num  29 0.92 2 30 25 48 63 39 53 71 ...
##  $ sibsp    : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch    : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : chr  "24160" "113781" "113781" "113781" ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : chr  "B5" "C22 C26" "C22 C26" "C22 C26" ...
##  $ embarked : chr  "S" "S" "S" "S" ...
##  $ boat     : chr  "2" "11" "" "" ...
##  $ body     : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: chr  "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON" ...
```

a. Extract and show the columns `cabin`, `age`, `embarked` and `pclass` into a new data frame of the name 'titanicSubset'. (5 points)

```
#INSERT YOUR ANSWER HERE
titanicSubset<-titanicDataset[,c('cabin','age','embarked','pclass')]
head(titanicSubset)
```

```
##     cabin   age embarked pclass
## 1      B5 29.00        S      1
## 2 C22 C26  0.92        S      1
## 3 C22 C26  2.00        S      1
## 4 C22 C26 30.00        S      1
## 5 C22 C26 25.00        S      1
## 6     E12 48.00        S      1
```

b. Numerical data: Use the count() function from the `dplyr` package to display the total number of passengers that survived or not. (5 points) HINT: To count the occurrences of survived or not in the titanicDataset data frame using the `dplyr` package, you can use the pipe operator (%>%) to chain operations.

```
#INSERT YOUR ANSWER HERE
#install.packages("dplyr")
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
```

```
titanic_survival<-titanicDataset%>%
  count(survived)
titanic_survival
```

```
##   survived   n
## 1        0 809
## 2        1 500
```

```
#As the dataset did not come with a data dictionary, I will assume that 0 means died and 1 means survived - therefore 809 passengers died and 500 survived based on the output.
```

c. Categorical data: Use count() and group_by() functions from the `dplyr` package to calculate the number of passengers by `embarked` . (5 points) HINT: Use group_by() first then pipe the result to count() to calculate the number of passengers.

```
#INSERT YOUR ANSWER HERE
titanic_embarked<-titanicDataset%>%
  group_by(embarked)%>%
  count(embarked)
titanic_embarked
```

```
## # A tibble: 4 × 2
## # Groups:   embarked [4]
##   embarked     n
##   <chr>    <int>
## 1 ""           2
## 2 "C"        270
## 3 "Q"        123
## 4 "S"        914
```

```
#The output shows the number of passengers that embarked from each gate.
```

d. Find the passengers in data frame whose embarked information is an empty character (“”), and fill them by the most frequent embarked value. (3 points)

```
#INSERT YOUR ANSWER HERE
#Find number of records with empty embarked information, in this case there are 2 rows, at index
es 169 and 285
titanic_emptyembarked<-subset(titanicDataset,embarked=="")
titanic_emptyembarked
```

```
##     pclass survived                                          name    sex age sibsp
## 169      1        1                         Icard, Miss. Amelie female  38     0
## 285      1        1 Stone, Mrs. George Nelson (Martha Evelyn) female  62     0
##     parch ticket fare cabin embarked boat body      home.dest
## 169     0 113572   80   B28             6   NA
## 285     0 113572   80   B28             6   NA Cincinatti, OH
```

```
#Find the most frequent embarked value, which in this case is "S"
sort(table(titanicDataset$embarked),decreasing=T)
```

```
##
##   S   C   Q
## 914 270 123   2
```

```
#Fill in empty embarked information with "S" value
titanicDataset[c(169,285),'embarked']<-'S'
#Verify if the passenger embarked info is now filled
titanicDataset[c(169,285),]
```

```
##     pclass survived                                          name    sex age sibsp
## 169      1        1                         Icard, Miss. Amelie female  38     0
## 285      1        1 Stone, Mrs. George Nelson (Martha Evelyn) female  62     0
##     parch ticket fare cabin embarked boat body      home.dest
## 169     0 113572   80   B28        S    6   NA
## 285     0 113572   80   B28        S    6   NA Cincinatti, OH
```

```
#Verify that there are 2 additional passengers added to count of embarked 'S', you can see 2 wer
e added to the previous count of 914, now the total is 916
titanic_embarked<-titanicDataset%>%
  group_by(embarked)%>%
  count(embarked)
titanic_embarked
```

```
## # A tibble: 3 × 2
## # Groups:   embarked [3]
##   embarked     n
##   <chr>    <int>
## 1 C          270
## 2 Q          123
## 3 S          916
```

e. Use the aggregate() function to calculate the 'survivalCount' of each `embarked` and calculate the survival rate of each embarked. Then draw the conclusion on which embarked has the higher survival rate. (5 points)

```
#INSERT YOUR ANSWER HERE
#Calculate survival count of each embarked
survivalCount<-aggregate(survived~embarked,data=titanicDataset,FUN=sum)
survivalCount
```

```
##   embarked survived
## 1        C      150
## 2        Q       44
## 3        S      306
```

```
#Calculate survival rate of each embarked
survivalCount$survival_rate<-survivalCount$survived/table(titanicDataset$embarked)[survivalCount
$embarked]*100
survivalCount
```

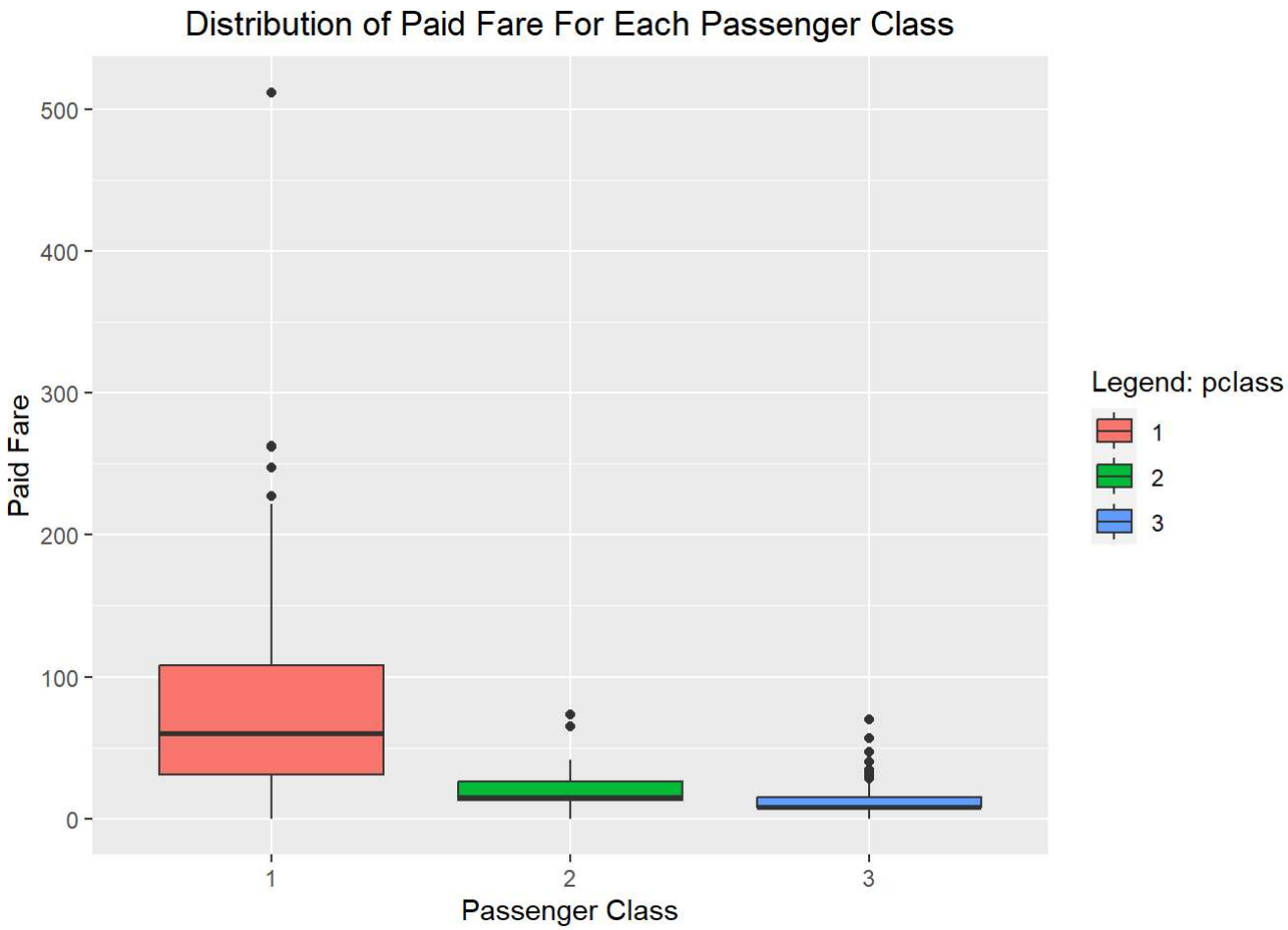
```
##   embarked survived survival_rate
## 1        C      150      55.55556
## 2        Q       44      35.77236
## 3        S      306      33.40611
```

```
#It is concluded that embarkation gate letter "C" has the highest survival rate.
```

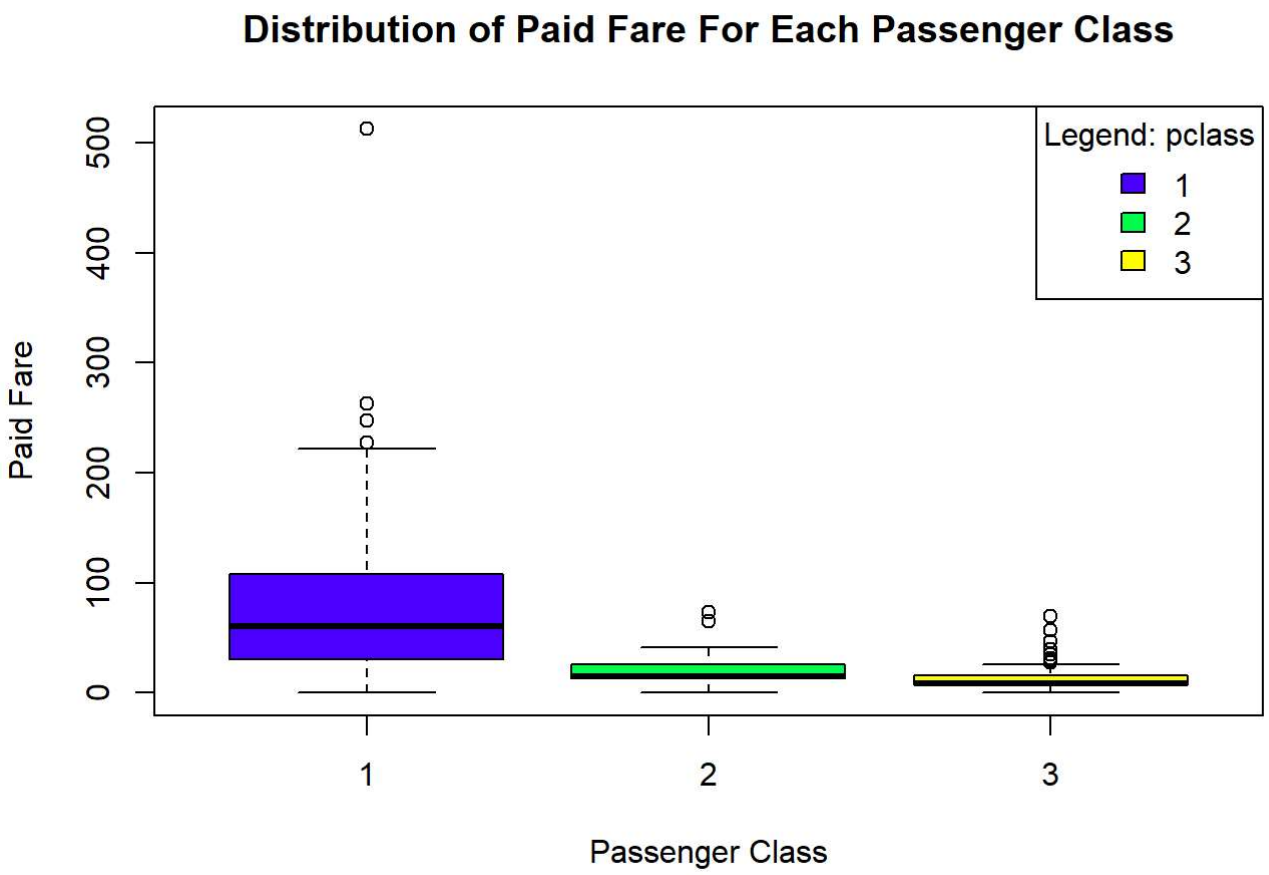
f. Use boxplot to display the distribution of fare for each pcalss and infer which passenger class is more expensive. (5 points)

```
#INSERT YOUR ANSWER HERE
#Boxplot method 1
#install.packages("ggplot2")
library(ggplot2)
ggplot(data=titanicDataset,aes(x=as.factor(pclass),y=fare,fill=as.factor(pclass)))+
  geom_boxplot()+
  labs(x="Passenger Class",y="Paid Fare")+
  ggtitle("Distribution of Paid Fare For Each Passenger Class")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_fill_discrete(name="Legend: pclass")
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```




```
#Boxplot method 2
boxplot(fare~pclass,data=titanicDataset,main="Distribution of Paid Fare For Each Passenger Clas
s",xlab="Passenger Class",ylab="Paid Fare",col=topo.colors((3)))
legend("topright",title="Legend: pclass",c("1","2","3"),fill=topo.colors(3))
```




```
#I can infer from the boxplots that Passenger Class 1 is more expensive than the rest because Q 1, Q3, the median, upper whisker and highest outliers are all higher than their counterparts as displayed in the other passenger classes on the boxplots.
```

g. Calculate the average fare for three pclass and describe if the calculation agrees with the box plot. (5 points)

```
#INSERT YOUR ANSWER HERE
average_fare<-titanicDataset%>%
  group_by(pclass)%>%
  summarize(average_fare=mean(fare))
average_fare
```

```
## # A tibble: 3 × 2
##   pclass average_fare
##    <int>        <dbl>
## 1      1         87.5
## 2      2         21.2
## 3      3         NA
```

```
#As the boxplots do not display the mean, we can compare the output values to the median which i s considered a type of average. Therefore, yes the calculation agrees with the boxplots because the values are very close to the median lines shown in the boxplot for the pclasses.
```

h. Use the for loop and if control statements to list the mens's non-empty home destinations, age 54 or less that embarked from Q (Queenstown), on the Titanic. (7 points)

```
#INSERT YOUR ANSWER HERE
#Display records with empty home destinations
titanic_emptyhomedest<-subset(titanicDataset,home.dest=="")
head(titanic_emptyhomedest)
```

```
##    pclass survived                            name    sex age sibsp parch
## 14      1        1    Barber, Miss. Ellen "Nellie" female  26     0     0
## 19      1        1           Bazzani, Miss. Albina female  32     0     0
## 24      1        1           Bidois, Miss. Rosalie female  42     0     0
## 25      1        1               Bird, Miss. Ellen female  29     0     0
## 29      1        1          Bissette, Miss. Amelia female  35     0     0
## 45      1        1 Burns, Miss. Elizabeth Margaret female  41     0     0
##      ticket     fare cabin embarked boat body home.dest
## 14    19877  78.8500              S    6   NA
## 19    11813  76.2917   D15        C    8   NA
## 24 PC 17757 227.5250              C    4   NA
## 25 PC 17483 221.7792   C97        S    8   NA
## 29 PC 17760 135.6333   C99        S    8   NA
## 45    16966 134.5000   E40        C    3   NA
```

```
#Assign NAs to empty values
titanicDataset[titanicDataset==""]<-NA
#Verify if there are still records with empty home destinations
nonemptyhomedest<-subset(titanicDataset,home.dest=="")
nonemptyhomedest
```

```
##  [1] pclass    survived  name      sex       age       sibsp     parch
##  [8] ticket    fare      cabin     embarked  boat      body      home.dest
## <0 rows> (or 0-length row.names)
```

```
#List men's non-empty home destinations, age 54 or less embarking from Q
destinations<-c()
for (i in 1:nrow(titanicDataset)){
  if (!is.na(titanicDataset$age[i])&&
      !is.na(titanicDataset$sex[i])&&
      !is.na(titanicDataset$home.dest[i])&&
      titanicDataset$sex[i]=="male"&&
      titanicDataset$age[i]<=54&&
      titanicDataset$embarked[i]=="Q"){
    destinations<-c(destinations,titanicDataset$home.dest[i])
  }
}
print(unique(destinations))
```

```
##  [1] "Fond du Lac, WI"
##  [2] "Ireland Chicago, IL"
##  [3] "Kingwilliamstown, Co Cork, Ireland New York, NY"
##  [4] "Co Cork, Ireland Charlestown, MA"
##  [5] "Ireland Philadelphia, PA"
##  [6] "Ireland New York, NY"
##  [7] "Co Limerick, Ireland Sherbrooke, PQ"
##  [8] "Philadelphia, PA"
##  [9] "Ireland Brooklyn, NY"
## [10] "Co Athlone, Ireland New York, NY"
## [11] "Aughnacliff, Co Longford, Ireland New York, NY"
## [12] "New York, NY"
```

Question 2 (15 points)

80 engines work together in a sequence. The historical data shows that the probability of each engine's failure is 0.05. We know that if one engine fails, the whole system stops.

a. What is the probability that the system operates without failure? (5 points)

```
#INSERT YOUR ANSWER HERE
dbinom(x=80,size=80,prob=0.95)
```

```
## [1] 0.01651537
```

b. Use the Binomial approximation to calculate the probability that at least 3 engines are defective? (5 points)

```
#INSERT YOUR ANSWER HERE
1-pbinom(q=2,size=80,prob=0.05)
```

```
## [1] 0.7693795
```

c. What is the probability that the second engine (B) is defective given the first engine (A) is not defective, i.e., P(B is defective|A is not defective), while we know that the first and second engines are independent. (5 points)

```
#INSERT YOUR ANSWER HERE
#When events A and B are indepedent (where event A has no effect on the probability of event B), the conditional probability of event B given event A is simply the probability of event B or P(B). The probability of a defective engine is 0.05.
P_B<-0.05
p_B_given_A=P_B
p_B_given_A
```

```
## [1] 0.05
```

Question 3 (25 points)

On average, John visits his parents 4 times a month

a. Find the probabilities that John visits his parents 1 to 6 times in a month? (5 points)

```
#INSERT YOUR ANSWER HERE
#Two ways to calculate the probability, first way:
ppois(6,lambda=4)-dpois(0,lambda=4)
```

```
## [1] 0.8710104
```

```
#Second way:
sum(dpois(1:6,lambda=4))
```

```
## [1] 0.8710104
```

b. Find the probability that John visits his parents 3 times or more in a month? (5 points)

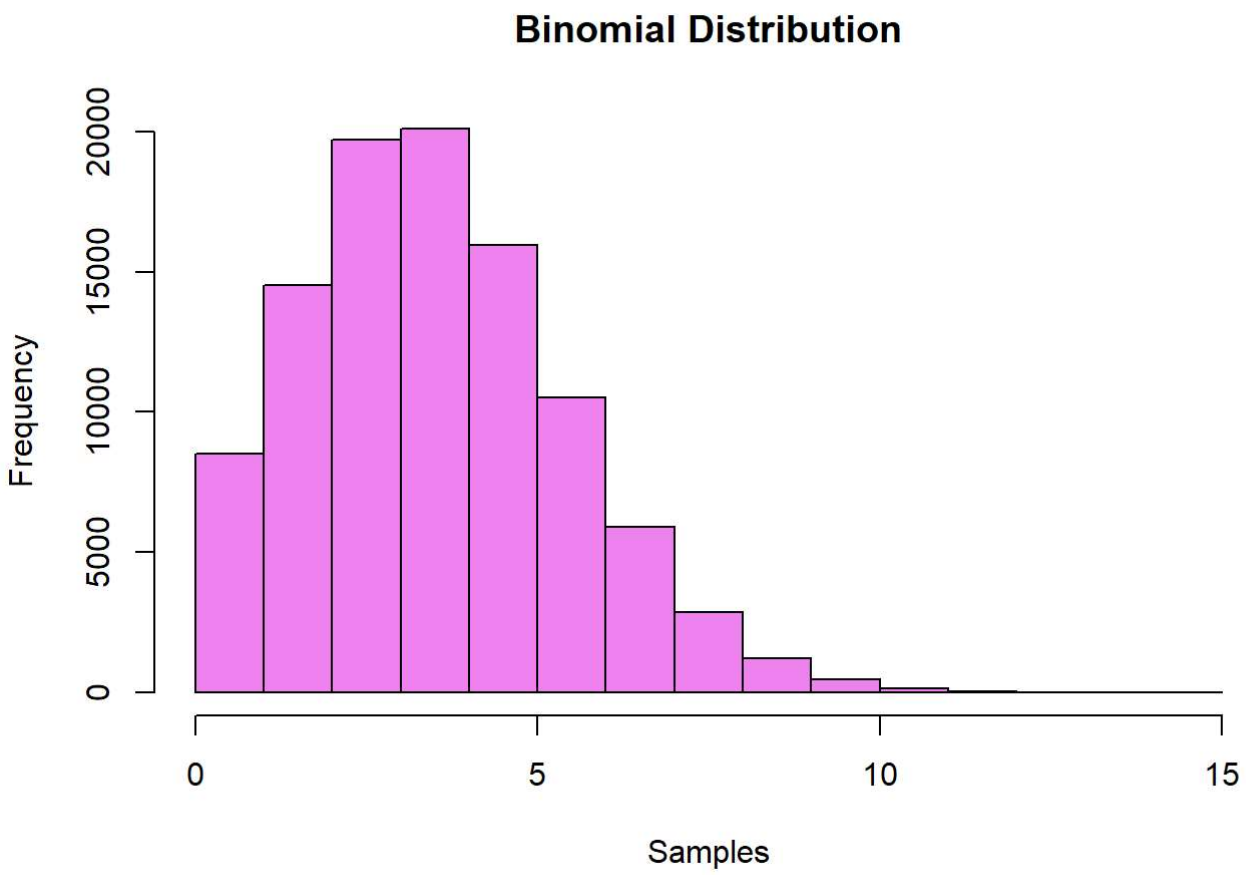
```
#INSERT YOUR ANSWER HERE
1-ppois(2,lambda=4)
```

```
## [1] 0.7618967
```

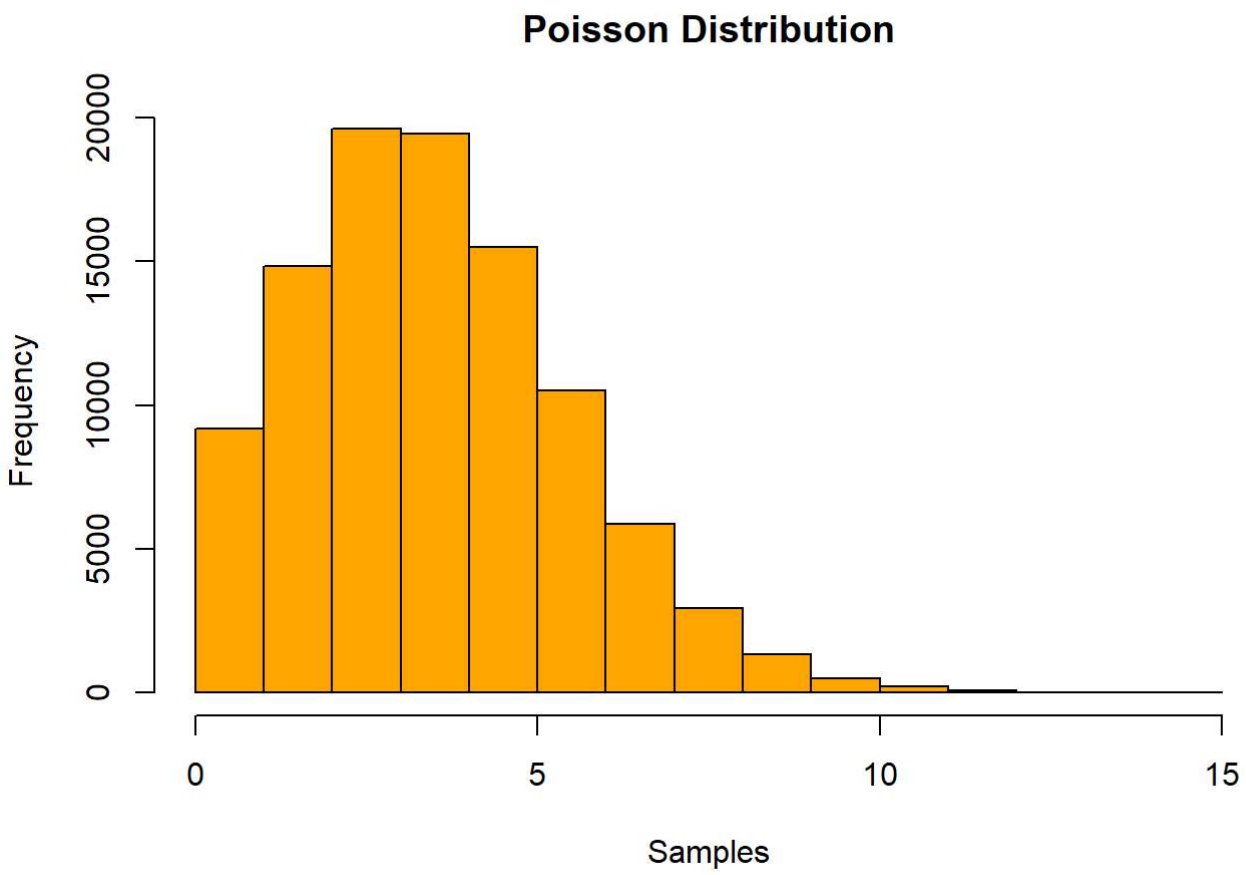
c. Compare the similarity between Binomial and Poisson distribution. (15 points @ 5 point each)

1. Create 100,000 samples for a Binomial random variable using parameters described in Question 2
2. Create 100,000 samples for a Poisson random variable using parameters described in Question 3
3. then illustrate on how well the Poisson probability distribution approximates the Binomial probability distribution. HINT: use multhist() from the 'plotrix' package

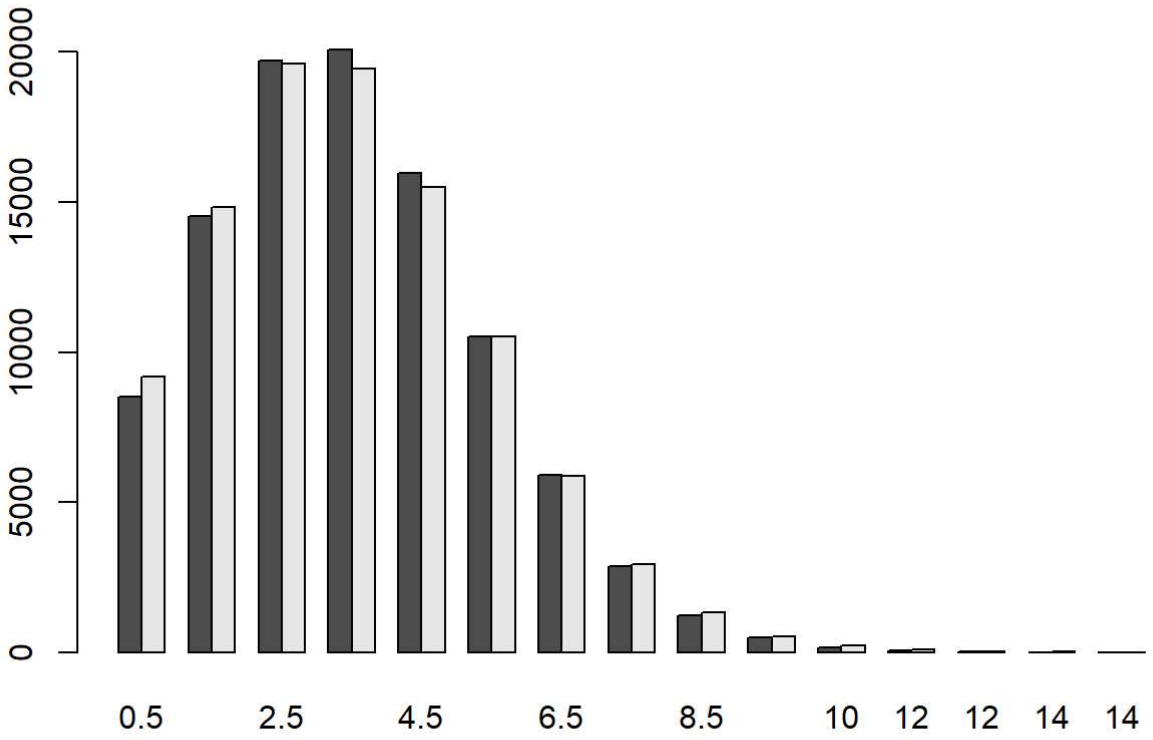
```
#INSERT YOUR ANSWER HERE
#1) Binomial samples
binomial_samples<-rbinom(100000,80,0.05)
#binomial_samples
#2) Poisson samples
poisson_samples<-rpois(100000,4)
#poisson_samples
#3) Illustrate approximation: Binomial Distribution
hist(binomial_samples,col="violet",xlab="Samples",ylab="Frequency",main="Binomial Distribution")
```




```
#3) Illustrate approximation: Poisson Distribution
hist(poisson_samples,col="orange",xlab="Samples",ylab="Frequency",main="Poisson Distribution")
```




```
#Combined histogram 1 using plotrix
#install.packages("plotrix")
library(plotrix)
multhist(list(binomial_samples,poisson_samples))
```




```
#Combined histogram 2 using plotly
#install.packages("plotly")
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

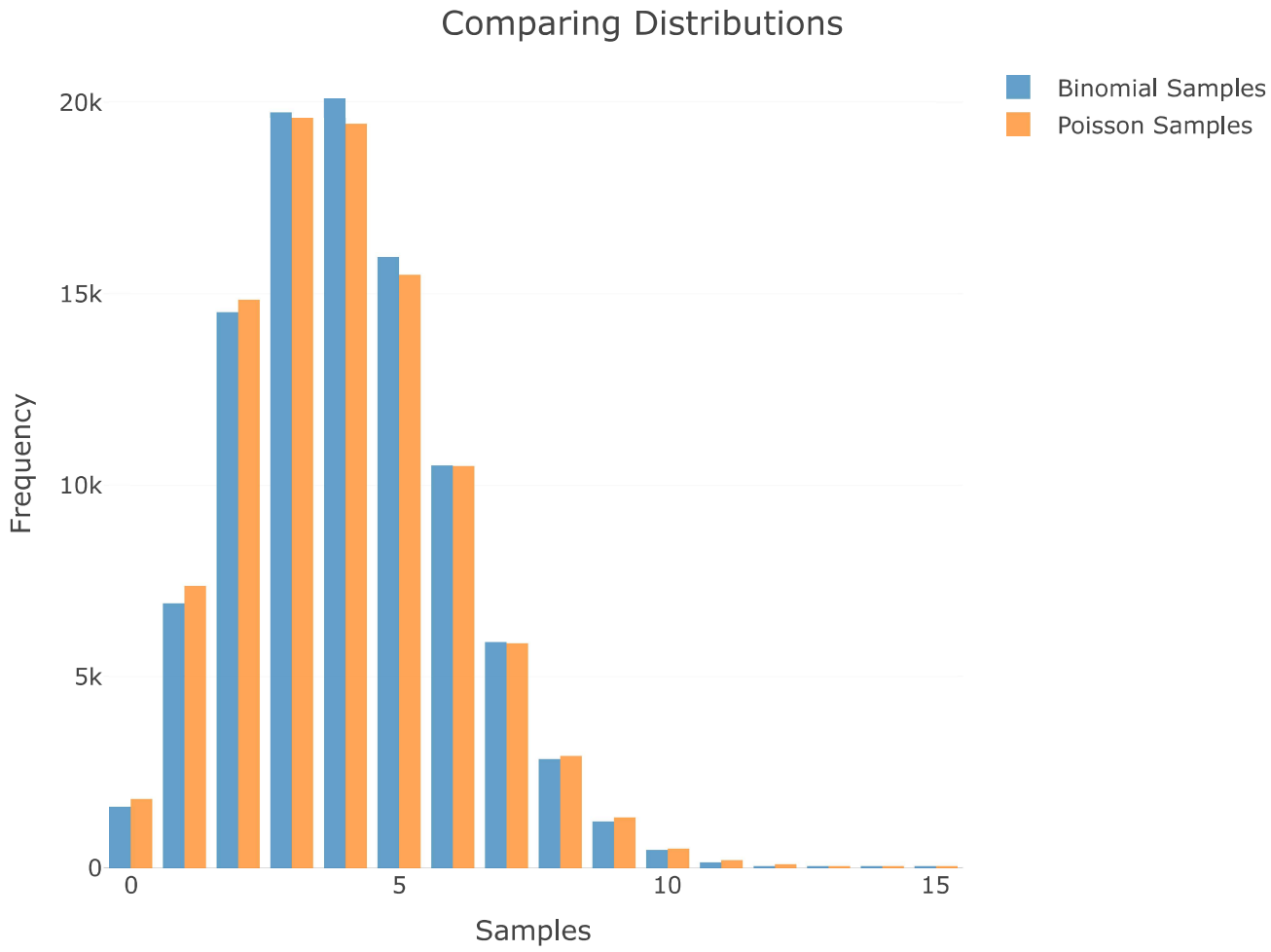
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(dplyr)
set.seed(123)

# Create Plotly figure with multiple histograms
fig <- plot_ly() %>%
  add_histogram(x = ~binomial_samples, name = "Binomial Samples", nbinsx = 30, opacity = 0.7) %
>%
  add_histogram(x = ~poisson_samples, name = "Poisson Samples", nbinsx = 30, opacity = 0.7) %>%
  layout(title = "Comparing Distributions",
         xaxis = list(title = "Samples"),
         yaxis = list(title = "Frequency"))
fig
```




```
#From observing the individual histograms and combined histograms above it is very apparent that the Poisson probability distribution approximates the Binomial probability distribution very wel
l.
```

Question 4 (20 points)

Write a script in R to compute the following probabilities of a normal random variable with mean 9 and variance 25

a. The probability that it lies between 8.2 and 17.3 (inclusive) (5 points)

```
#INSERT YOUR ANSWER HERE
pnorm(17.3,mean=9,sd=sqrt(25))-pnorm(8.2,mean=9,sd=sqrt(25))
```

```
## [1] 0.5151022
```

b. The probability that it is greater than 15.02 (5 points)

```
#INSERT YOUR ANSWER HERE
1-pnorm(15.02,mean=9,sd=sqrt(25))
```

```
## [1] 0.1142948
```

c. The probability that it is less than or equal to 11.8 (5 points)

```
#INSERT YOUR ANSWER HERE
pnorm(11.8,mean=9,sd=sqrt(25))
```

```
## [1] 0.7122603
```

d. The probability that it is less than 10 or greater than 13 (5 points)

```
#INSERT YOUR ANSWER HERE
pnorm(10,mean=9,sd=sqrt(25))+(1-pnorm(13,mean=9,sd=sqrt(25)))
```

```
## [1] 0.7911151
```

END of Assignment #2.