DATA SCIENCE AND MACHINE LEARNING (MSC)

DAMA51: Foundations in Computer Science

Academic Year: 2022–2023

#2 Written Assignment	
Submission Deadline	**Wednesday, 1 February 2023, 23:59:59 EET**
Student Name	**Kritikos Antonios**

Guidelines

The deadline is definitive.
An indicative solution will be posted online along with the returning of the graded assignments.
The assignment is due via the STUDY submission system. **You are expected to deliver a document (.DOC, .ODT, .PDF) and a compressed (.ZIP, .RAR) file containing all your work:**

- 1 document file (this document) with the answers to all the topics, along with the R code and the results of the execution of the code
- 1 compressed file with 3 R scripts that correspond to topics 3, 4 and 5.

You should not make any changes in the written assignment file other than providing your own answers. You should also type all of your answers into Word and not attach any handwritten notes as pictures into your work otherwise a 5% reduction of your final grade will be applied. Make sure to name all the files (ZIP file, DOC file and R script files) with **your last name first followed by a dash and the names of each component at the end**. For example for the student with last name Aggelou the files should be named as follows: Aggelou-HW1.zip, Aggelou-HW1.doc, Aggelou-Topic3.R, Aggelou-Topic4.R and Aggelou-Topic5.R. The R script files should automatically run with the **source** command and generate the correct results. Also, please include comments before each command to explain the functionality of the command that follows. Unless otherwise stated in the question, all numerical answers should be given to **three decimal places**.

Topic	Points	Grades
1. Online QUIZ	40	
2. Article Review	10	
3. Principal Component Analysis	20	
4. Confusion Matrix	10	
5. Hypothesis Testing – χ^2 Test	20	
TOTAL	**100**	**/100**

Topic 1: Online QUIZ

Complete the corresponding online quiz available at:

https://study.eap.gr/mod/quiz/view.php?id=24566

You have one effort and unlimited time to complete the quiz, up to the submission deadline. **(40 points)**

Topic 2: Article Review

The review article by Shamina Ahmed et al. entitled "Artificial intelligence and machine learning in finance: A bibliometric review" (available at https://doi.org/10.1016/j.ribaf.2022.101646) uses a bibliometric approach to review the artificial intelligence and machine learning literature in the finance field highlighting the main application areas. Summarize the main research streams reviewed as well as the methodologies used in each of these areas. Finally give your personal view on the main research challenges and how to tackle these based on AI and ML methodologies (you may use a personal business perspective, if available, beyond the finance field).

Note: You should write up your answer to a maximum of 300 words. Any text in excess of 300 words will not be taken into consideration.

(10 points)

In the review article entitled "Artificial intelligence and machine learning in finance: A bibliometric review" are the findings of a conducted research which had as basis 348 articles published between 2011 and 2021 from the Q1 and Q2 finance journals of Scopus. The six main research streams were: How AI and ML can be implemented and provide insights in the fields of: 1)Bankruptcy prediction and credit-risk assessment. 2)Stock price prediction, portfolio management, volatility, and liquidity. 3)Prediction of the prices of oil, gold, and agriculture products. 4)Anti-money laundering, anti-fraud detection, and risk management. 5)Behavioral finance. 6)Big data analytics, blockchain, and data mining. Obviously all the previous researches and articles are finance-adjacent but it holds a good basis for the time being.

Even though the articles are recent (considering that AI and ML blew up as a field in the past ten years) and extremely informing, we still lack the necessary knowledge on how can we expand and utilize the usage of AI and ML both in finance and in other fields. To manage that, we must tackle some problems. The most important problem is the data that is needed to do the research, so governments and universities must provide funds to implement AI and ML in studies and businesses, so the researchers can gather data to expand their observations in the field, and provide more and accurate insights in the utilization of AI. By tackling that problem, all other possible problems will be either eliminated or minimized, because for any research of this kind, the most important part is the data. Having more data means that the possibilities of error in any research will be massively avoided, so the researchers will have strong foundation on targeting more accurately the aspects of this kind of new technology, hence, helping people understand and utilize it properly.

Topic 3: Principal Components Analysis

For this topic, you'll use the built-in data set **USArrests** of R to answer the following points. Within each answer frame below you should include the R code as well as the results **(20 points)**

a. Using R, first review each attribute included in the **USArrests** dataset (write the name and type) and calculate the mean and standard deviation. Include your answers in the Tables provided **(4 points)**

(2 points) Name and type of variables

Variable Name	Variable Type
Murder	Num
Assault	Int
UrbanPop	Int
Rape	Num

R-code and Results:

```
myDataFrame <- as.data.frame(USArrests)
myDataFrame
```

```
> myDataFrame <- as.data.frame(USArrests)
> myDataFrame
               Murder Assault UrbanPop Rape
Alabama          13.2     236       58 21.2
Alaska           10.0     263       48 44.5
Arizona           8.1     294       80 31.0
Arkansas          8.8     190       50 19.5
California        9.0     276       91 40.6
Colorado          7.9     204       78 38.7
Connecticut       3.3     110       77 11.1
Delaware          5.9     238       72 15.8
Florida          15.4     335       80 31.9
Georgia          17.4     211       60 25.8
Hawaii            5.3      46       83 20.2
Idaho             2.6     120       54 14.2
Illinois         10.4     249       83 24.0
Indiana           7.2     113       65 21.0
Iowa              2.2      56       57 11.3
Kansas            6.0     115       66 18.0
Kentucky          9.7     109       52 16.3
Louisiana        15.4     249       66 22.2
Maine             2.1      83       51  7.8
Maryland         11.3     300       67 27.8
Massachusetts     4.4     149       85 16.3
Michigan         12.1     255       74 35.1
Minnesota         2.7      72       66 14.9
Mississippi      16.1     259       44 17.1
Missouri          9.0     178       70 28.2
Montana           6.0     109       53 16.4
Nebraska          4.3     102       62 16.5
Nevada           12.2     252       81 46.0
New Hampshire     2.1      57       56  9.5
New Jersey        7.4     159       89 18.8
New Mexico       11.4     285       70 32.1
New York         11.1     254       86 26.1
North Carolina   13.0     337       45 16.1
North Dakota      0.8      45       44  7.3
Ohio              7.3     120       75 21.4
Oklahoma          6.6     151       68 20.0
Oregon            4.9     159       67 29.3
Pennsylvania      6.3     106       72 14.9
Rhode Island      3.4     174       87  8.3
South Carolina   14.4     279       48 22.5
South Dakota      3.8      86       45 12.8
Tennessee        13.2     188       59 26.9
Texas            12.7     201       80 25.5
Utah              3.2     120       80 22.9
Vermont           2.2      48       32 11.2
Virginia          8.5     156       63 20.7
Washington        4.0     145       73 26.2
West Virginia     5.7      81       39  9.3
Wisconsin         2.6      53       66 10.8
Wyoming           6.8     161       60 15.6
```

```
> str(USArrests)
'data.frame':   50 obs. of  4 variables:
 $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
 $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
 $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
 $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

(2 points) Mean and standard deviation for each variable

Variable Name	Mean	Standard Deviation
Murder	7.788	4.35551
Assault	170.76	83.33766
UrbanPop	65.54	14.47476
Rape	21.232	9.366385

R-code and Results:

str(USArrests)

```
> mean(USArrests$Murder)
[1] 7.788
> sd(USArrests$Murder)
[1] 4.35551
> mean(USArrests$Assault)
[1] 170.76
> sd(USArrests$Assault)
[1] 83.33766
> mean(USArrests$UrbanPop)
[1] 65.54
> sd(USArrests$UrbanPop)
[1] 14.47476
> mean(USArrests$Rape)
[1] 21.232
> sd(USArrests$Rape)
[1] 9.366385
```

b. Using R, based on question (a), compute the principal components, ensuring that PCA is applied on scaled variables. **(4 points)**

prcomp(USArrests, scale=TRUE)

```
> prcomp(USArrests, scale=TRUE)
Standard deviations (1, .., p=4):
[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation (n x k) = (4 x 4):
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

c. Using R and your results from question (b), write in the Table below, for each principal component, what is the percentage of the total variance that is explained. **(4 points)**

Percentage of the total variance that is explained (1 point/PC)

	PC1	PC2	PC3	PC4
Percentage of the total variance explained	0.6201	0.2474	0.8914	0.04336

R-code and Results:

pca <- prcomp(myDataFrame, scale=TRUE)

summary(pca)

```
> pca <- prcomp(myDataFrame, scale=TRUE)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

d. Using R and based on the previous results, write the coefficients of the linear combination of the original variables from which the principal components (PCs) are constructed (PCA loadings) in the Table below. **(4 points)**

PCA loadings

	PC1	PC2	PC3	PC4
Murder	-0.5358995	0.4181809	-0.3412327	0.64922780
Assault	-0.5831836	0.1879856	-0.2681484	-0.74340748
UrbanPop	-0.2781909	-0.8728062	-0.3780158	0.13387773
Rape	-0.5434321	-0.1673186	0.8177779	0.08902432

R-code:

pca$rotation

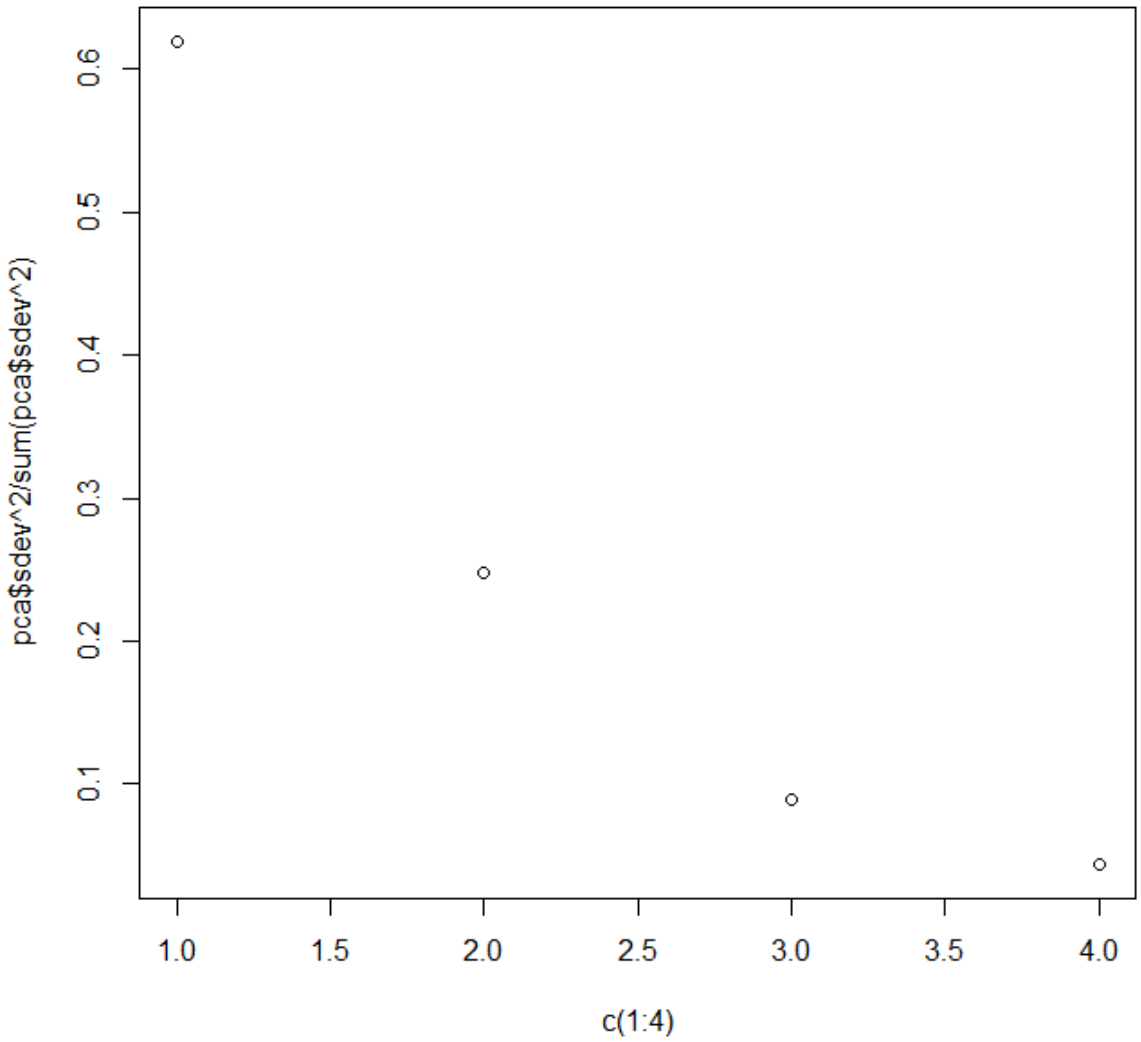
```
> pca$rotation
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

e. Using R, and the previous results, create a scree plot. Based on the scree plot, how many principal components would you retain? **(4 points)**

(3 point) Create a scree plot

R-code and Plot:

```
plot(c(1:4), pca$sdev^2 / sum(pca$sdev^2))
```



(1 point) How many principal components would you retain?

After the third principal component we see a pronounced drop off so we'll decide to retain the first three principal components.

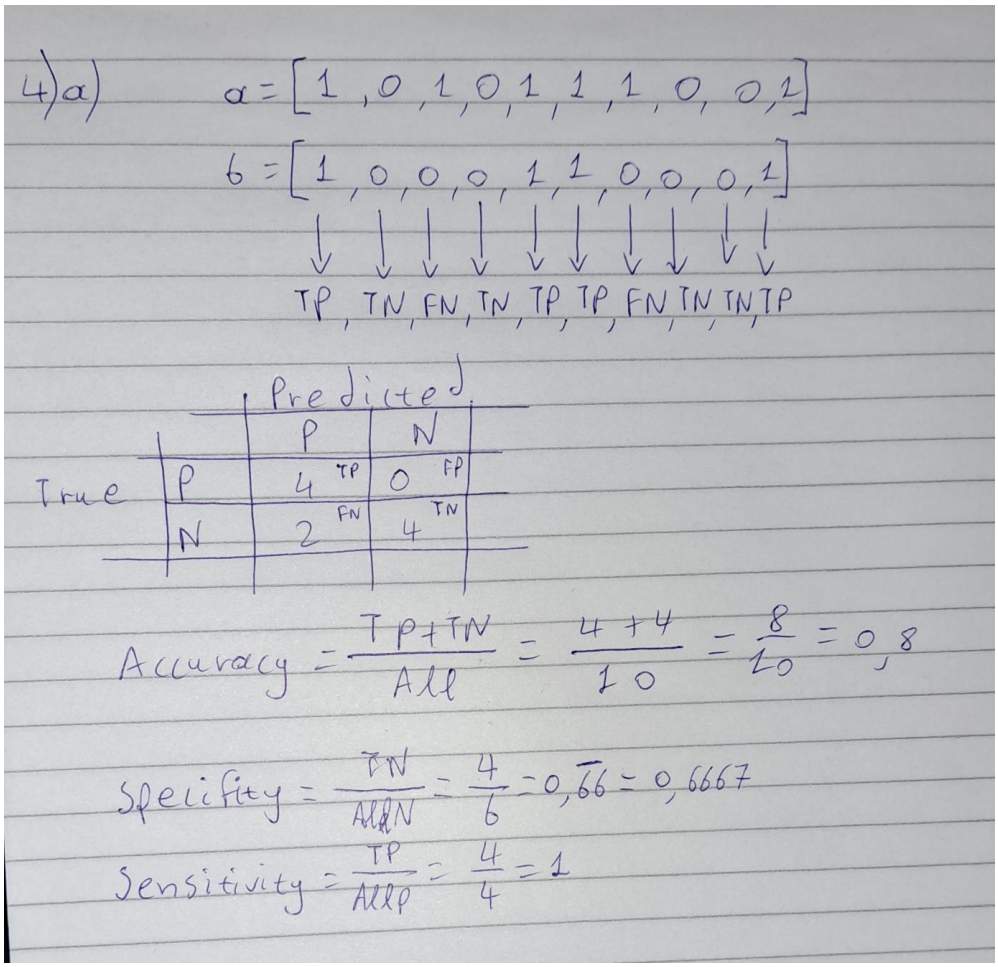
Topic 4: Confusion Matrix

You are given the following vectors a = [1,0,1,0,1,1,1,0,0,1] and b = [1,0,0,0,1,1,0,0,0,1], which represent the True and predicted values, respectively, of a diagnostic test being positive (0) or negative (1) for a specific disease. **(10 points)**

a. Using "pen and paper", fill in the confusion matrix below and then calculate the following statistics: sensitivity, specificity and accuracy (type the formulas, calculations and results in the answer frame below). **(5 points)**

(2 points) Confusion Matrix

True Class	Predicted Class	
	Positive (0)	**Negative (1)**
Positive (0)	4	0
Negative (1)	2	4



(1 point) Sensitivity = 1.0000

(1 point) Specificity = 0.6667

(1 point) Accuracy = 0.8

b. Using R, create the confusion matrix (fill in the one provided below) and calculate the following statistics: sensitivity, specificity and accuracy to verify your results from question (a). **(5 points)**

(2 point) Confusion Matrix

True Class	Predicted Class	
	Positive (0)	Negative (1)
Positive (0)	4	0
Negative (1)	2	4

R-code and Results:

(1 point) Sensitivity = 1.0000

(1 point) Specificity = 0.6667

(1 point) Accuracy = 0.8

```
install.packages('caret')
library(caret)
true_value <- factor(c(1,0,1,0,1,1,1,0,0,1))
predicted_value <- factor(c(1,0,0,0,1,1,0,0,0,1))
confusionMatrix(data=predicted_value,reference=true_value)
```

```
> library(caret)
Loading required package: ggplot2
Loading required package: lattice
> true_value <- factor(c(1,0,1,0,1,1,1,0,0,1))
> predicted_value <- factor(c(1,0,0,0,1,1,0,0,0,1))
> confusionMatrix(data=predicted_value,reference=true_value)
Confusion Matrix and Statistics

          Reference
Prediction 0 1
         0 4 2
         1 0 4

               Accuracy : 0.8
                 95% CI : (0.4439, 0.9748)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.1673

                  Kappa : 0.6154

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.6667
         Pos Pred Value : 0.6667
         Neg Pred Value : 1.0000
             Prevalence : 0.4000
         Detection Rate : 0.4000
   Detection Prevalence : 0.6000
      Balanced Accuracy : 0.8333

       'Positive' Class : 0
```

Topic 5: χ^2 Test

For this topic, you will use the built-in data set **iris** of R in order to answer the following points. **(20 points)**

a. Using R, first find the variables included in the dataset **iris** and write their names and types in the Table below. Then add a new variable, `Size_sepal`, which is "small", if the length of the sepal is smaller than the median sepal length of all flowers, or "big" otherwise. **(5 points)**

(2 points) Variable names and type

Variable Name	Variable Type
Sepal.Length	num
Sepal.Width	num
Petal.Length	num
Petal.Width	num
Species	Factor

R-code and Results:

str(myDataSet)

```
> str(myDataSet)
'data.frame':   150 obs. of  6 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1
 $ Size_sepal  : Factor w/ 2 levels "big","small": 2 2 2 2 2 2 2 2 2 2 ...
```

(3 points) add variable "Size_sepal"

R-code:

myDataSet$Size_sepal <- myDataSet$Size_sepal <- as.factor(ifelse(myDataSet$Sepal.Length <median(myDataSet$Sepal.Length), 'small', 'big'))

str(myDataSet)

```
> myDataSet$Size_sepal <- myDataSet$Size_sepal <- as.factor(ifelse(myDataSet$Sepal.Length <median(myDataSet$Sepal.Length), 'small
> str(myDataSet)
'data.frame':   150 obs. of  6 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Size_sepal  : Factor w/ 2 levels "big","small": 2 2 2 2 2 2 2 2 2 2 ...
```

b. Using R, create a contingency table, including sums, for the variables `Species` and `Size_sepal`. Create a stacked barplot in R to visualize `Size_sepal` (y-axis) for each of the `Species` (x-axis). **(5 points)**

(3 points) contingency table for the variables Species and Size_sepal

	Size_sepal		
Species	***Big***	***Small***	***Sum***
Setosa	1	49	50
versicolor	29	21	50
virginica	47	3	50
Sum	77	73	150

R-code and Results:

addmargins(table(myDataSet$Species,myDataSet$Size_sepal),c(1,2))

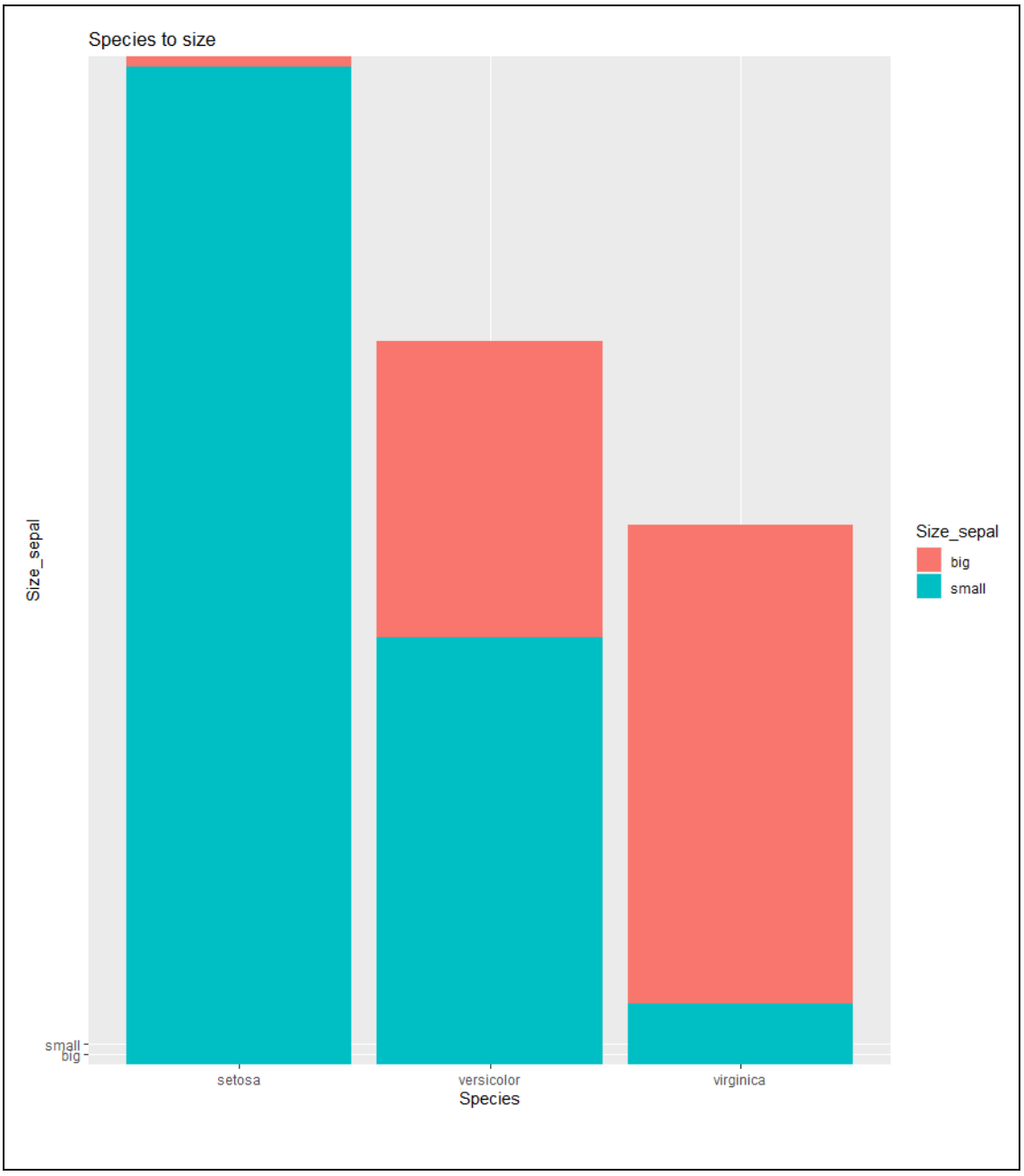
```
> addmargins(table(myDataSet$Species,myDataSet$Size_sepal),c(1,2))

            big small Sum
setosa        1    49  50
versicolor   29    21  50
virginica    47     3  50
Sum          77    73 150
```

(2 points) Create a stacked barplot

R-code and Plot:

ggplot(myDataSet, aes(x=Species, y=Size_sepal, fill=Size_sepal)) +
geom_bar(position="stack", stat="identity") + ggtitle("Species to size")



c. Define two hypotheses (null and alternative) in the Table below, to test whether the `Size_sepal` variable is independent of the `Species` variable. Using R, perform the Chi-Squared Test of Independence (using default values) and write if you can reject or not the null hypothesis at a significance level of a=0.05. **(5 points)**

(2 points) Define the two hypotheses below:

For Size_sepal and Species	
Ho (null Hypothesis):	*The Size_sepal variable is independent of the Species variable*
H1 (alternative Hypothesis):	*The Size_sepal variable is dependent to the Species variable*

(3 points) Perform the Chi-Squared Test of Independence and write if you can reject or not the null hypothesis for the variables Size_sepal and Species

R-code:

chisq.test(myDataSet$Size_sepal, myDataSet$Species)

```
> chisq.test(myDataSet$Size_sepal, myDataSet$Species)

        Pearson's Chi-squared test

data:  myDataSet$Size_sepal and myDataSet$Species
X-squared = 86.035, df = 2, p-value < 2.2e-16
```

p-value:

p-value < 2.2e-16

Decision:

The p-value is less than the alpha value(0.05) so we reject the null hypothesis and we accept the alternative hypothesis that the size is dependent to the species.

d. Using R, based on the results from question (c), calculate the degrees of freedom (df) and fill in the Tables below with the observed and expected values for variables `Species` and `Size_sepal`. **(5 points)**

(3 points) Observed and expected values for the variables Species and Size_sepal

Observed values for the variables Species and Size_sepal

Observed Values	*Size_sepal*	
Species	*big*	*Small*
setosa	1	49
versicolor	29	21
virginica	47	3

Expected values for the variables Species and Size_sepal

Expected Values	*Size_sepal*	
Species	*big*	*small*
Setosa	25.667	24.333
Versicolor	25.667	24.333
Virginica	25.667	24.333

R-code and Results:

(chisq.test(myDataSet$Species, myDataSet$Size_sepal))$observed

(chisq.test(myDataSet$Species, myDataSet$Size_sepal))$expected

```
> (chisq.test(myDataSet$Species, myDataSet$Size_sepal))$observed
                 myDataSet$Size_sepal
myDataSet$Species big small
       setosa       1    49
       versicolor  29    21
       virginica   47     3
> (chisq.test(myDataSet$Species, myDataSet$Size_sepal))$expected
                 myDataSet$Size_sepal
myDataSet$Species      big    small
       setosa     25.66667 24.33333
       versicolor 25.66667 24.33333
       virginica  25.66667 24.33333
```

(2 points) Calculate the degrees of freedom (df) and write the result below:

df(Species - Size_sepal) = (r – 1)(c - 1) = (3 – 1)(2 -1) = 2