Homework 1

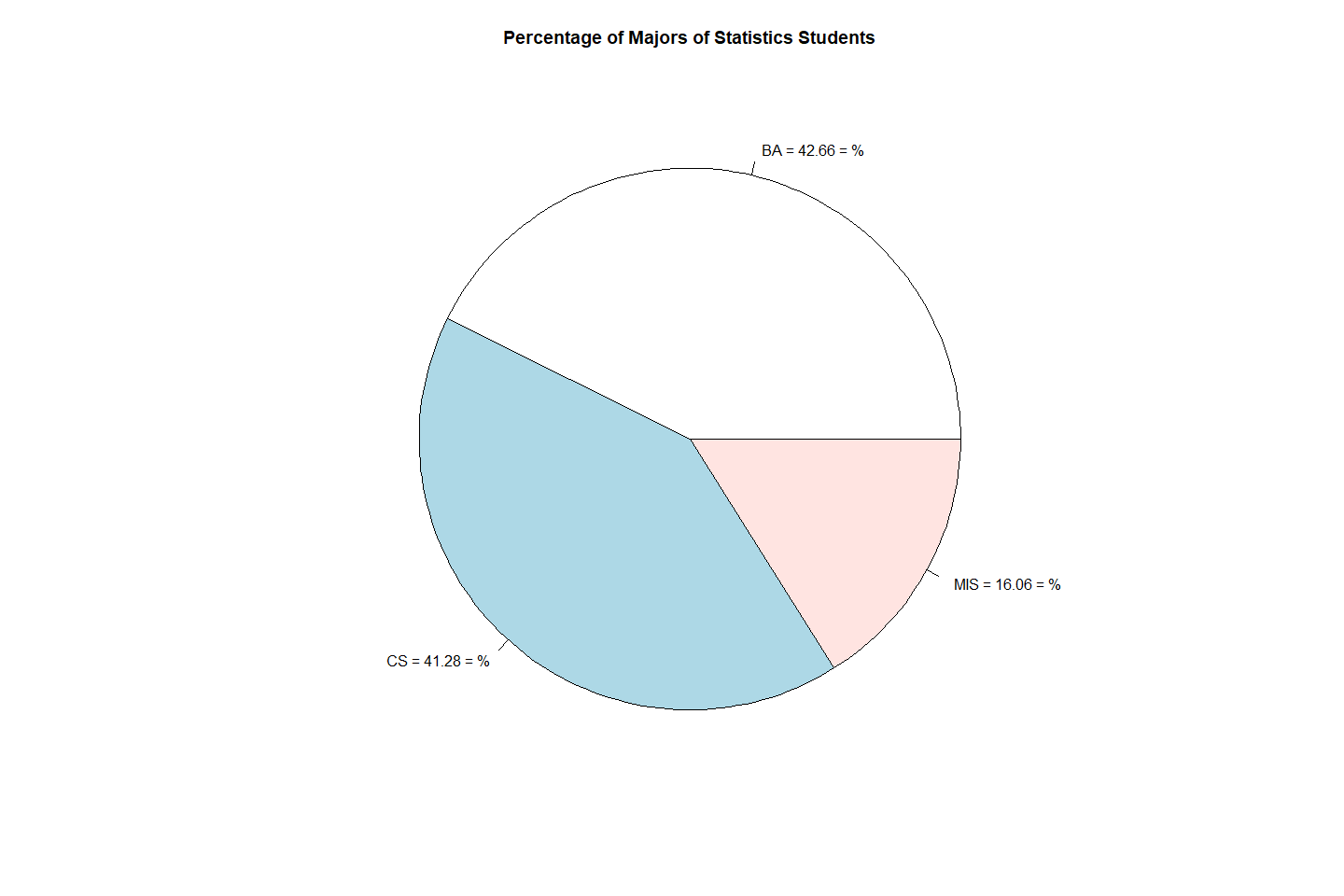
Elikem Asudo Tsatsu Gale-Zoyiku

2023-01-25

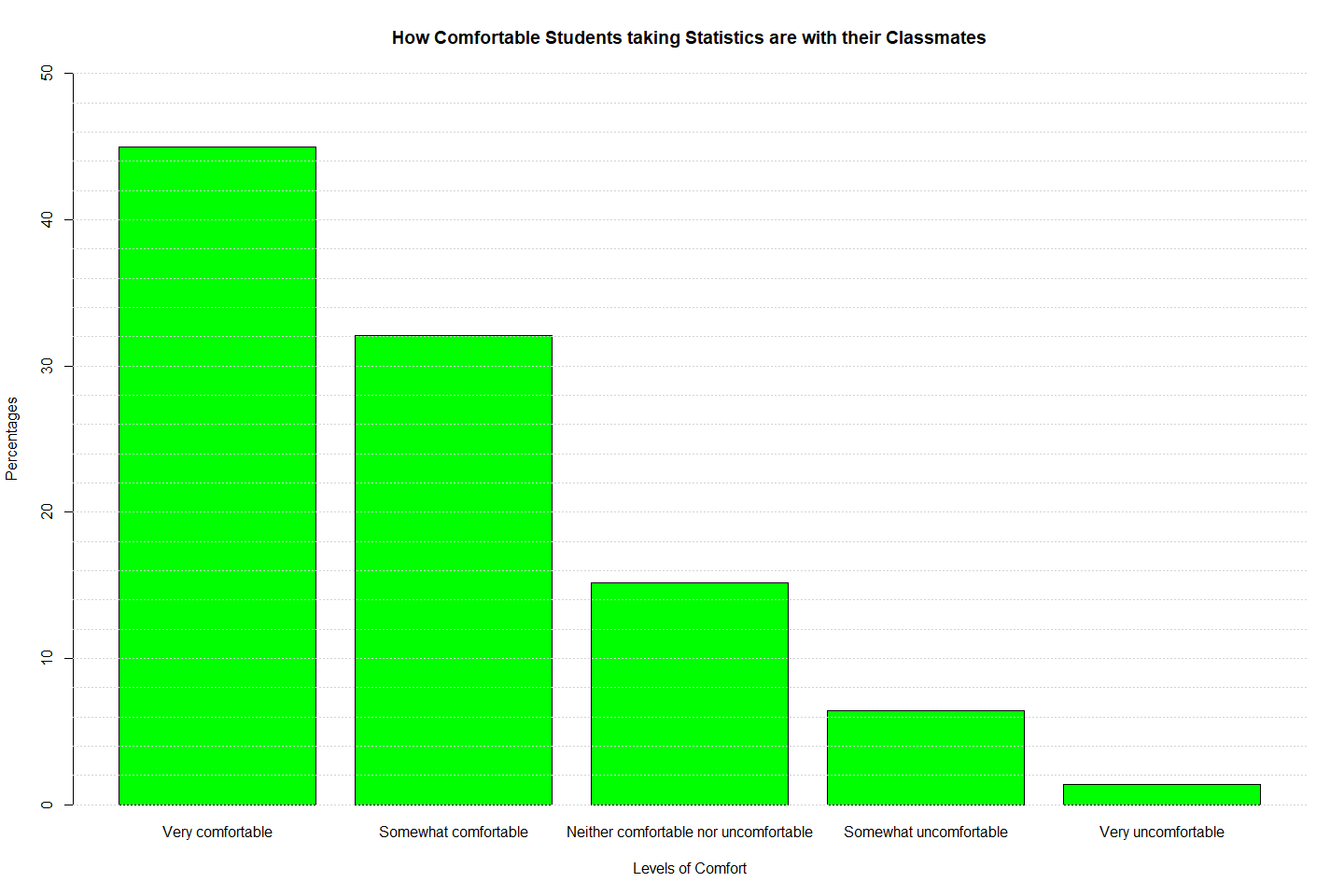
library(readr)  
StudentsData <- read.delim("C:/Users/egale/OneDrive - Ashesi University/Desktop/Statistics with Probability/StudentsData.txt")

### ###Question 1

#a.  
t = table(StudentsData$Major)  
prop = prop.table(t)  
percent = prop.table(t)\*100  
pf = round(percent,2)  
lbl = paste(c(" BA"," CS"," MIS"),pf,"%",sep=" = ")  
ttl = paste(c('Percentage of Majors of Statistics Students'))  
pie(t,label=lbl,main=ttl)



#c  
t1 <- table(StudentsData$Comfortable\_classmates)  
comfort <- factor(StudentsData$Comfortable\_classmates, levels = c("Very comfortable","Somewhat comfortable","Neither comfortable nor uncomfortable","Somewhat uncomfortable","Very uncomfortable"))  
t2<-table(comfort)  
barplot(100\*t2/length(comfort), main = "How Comfortable Students taking Statistics are with their Classmates", col ="green", ylab = "Percentages",xlab='Levels of Comfort', ylim=c(0,50))  
abline(h=(seq(0,50,2)), col="lightgray", lty="dotted")



#a.

**218** students are taking Statistics this semester.

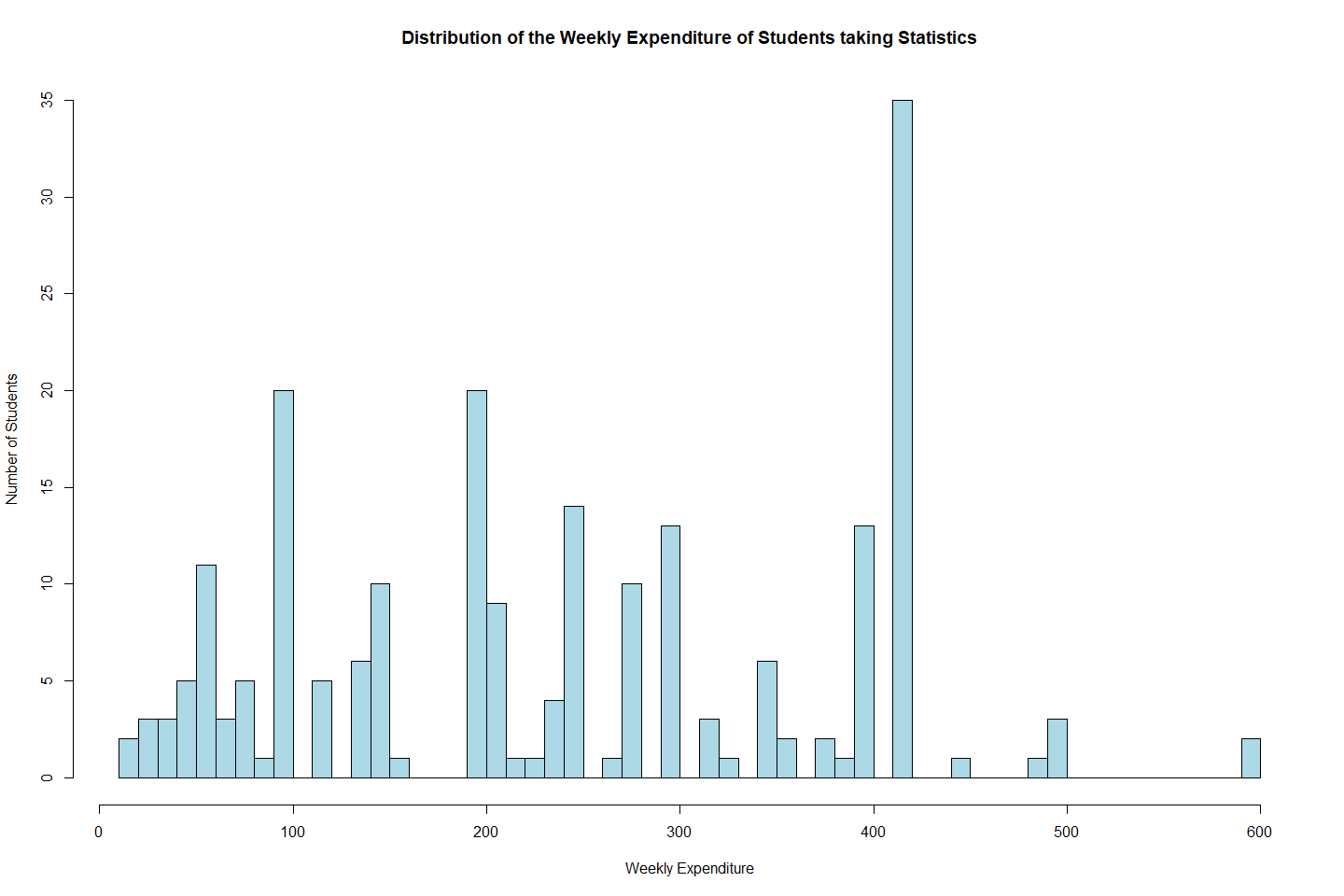
**42.66%** of students in the Statistics class are **BA** students.

**41.28%** of students taking Statistics are **CS** students.

**16.06%** of students taking Statistics are **MIS** majors.

### ###Question 2

hist(StudentsData$Expenditure, col='lightblue', main='Distribution of the Weekly Expenditure of Students taking Statistics', xlab='Weekly Expenditure', ylab='Number of Students',breaks=50)



summary(StudentsData$Expenditure)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.0 120.0 240.0 245.1 396.2 600.0

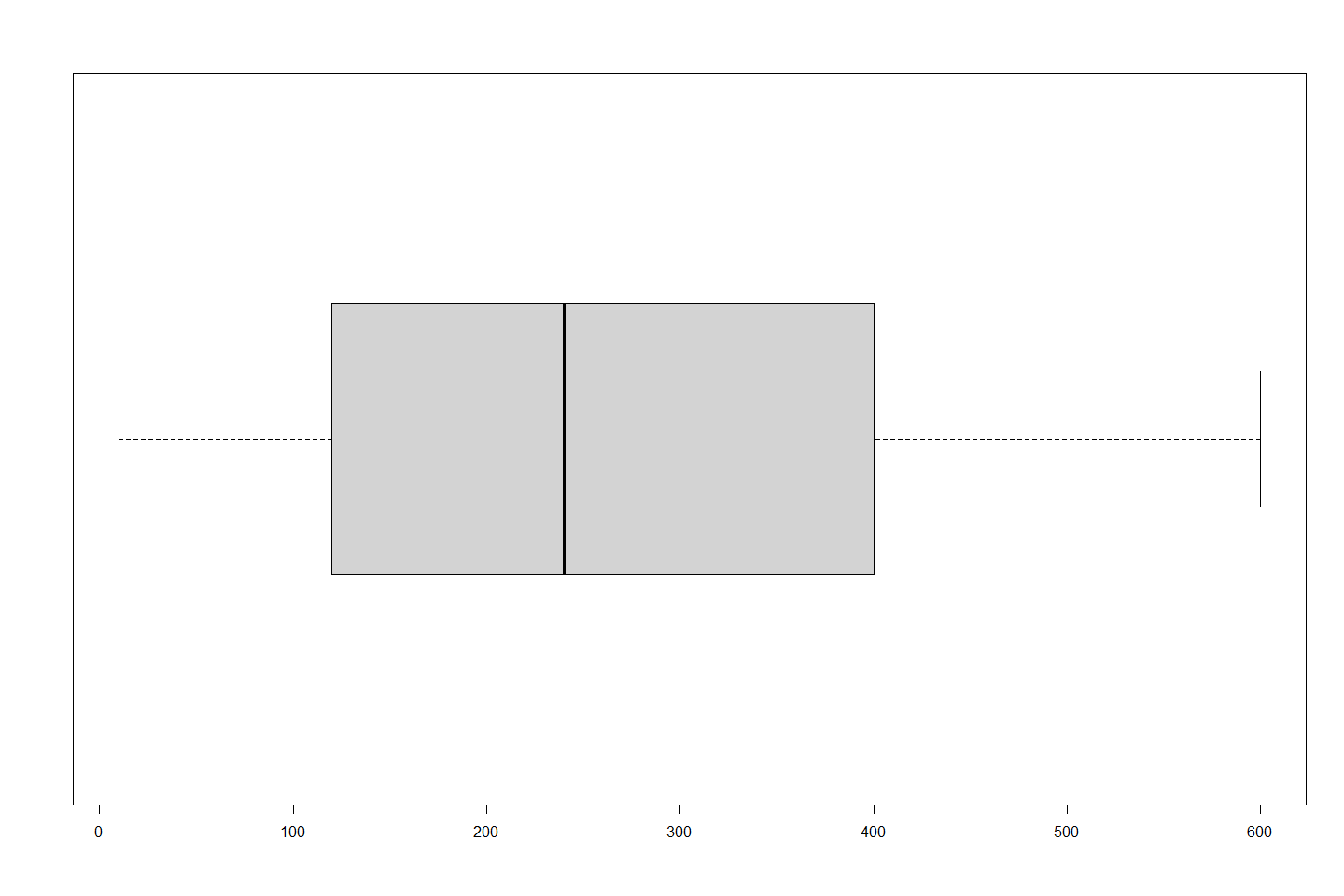
IQR(StudentsData$Expenditure)

## [1] 276.25

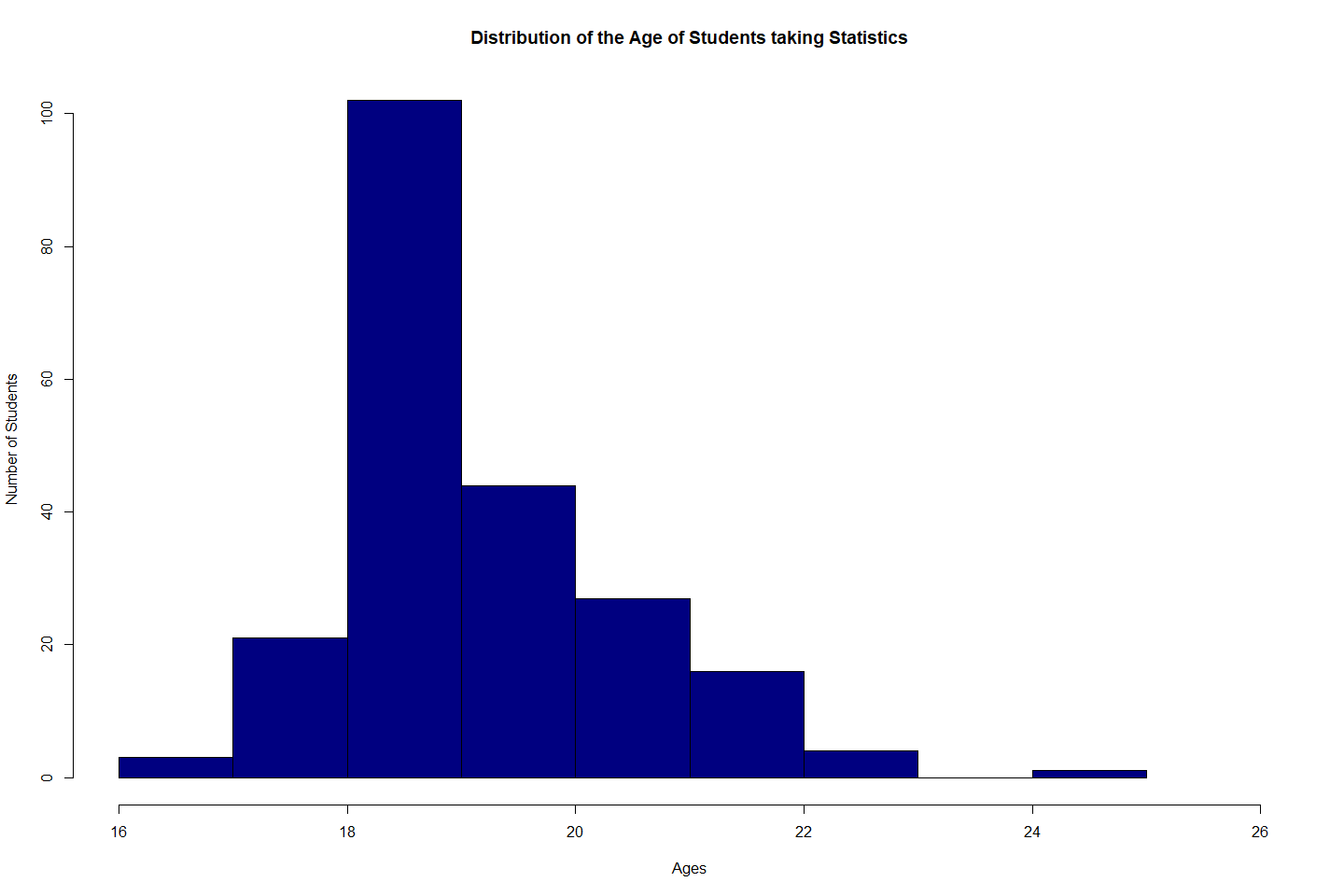
sd(StudentsData$Expenditure)

## [1] 136.8793

#USING a boxplot to check for outliers in the distribution  
boxplot(StudentsData$Expenditure,horizontal = TRUE)



hist(StudentsData$Age, col='navyblue', main='Distribution of the Age of Students taking Statistics', xlab='Ages', ylab='Number of Students',xlim=c(16,26))



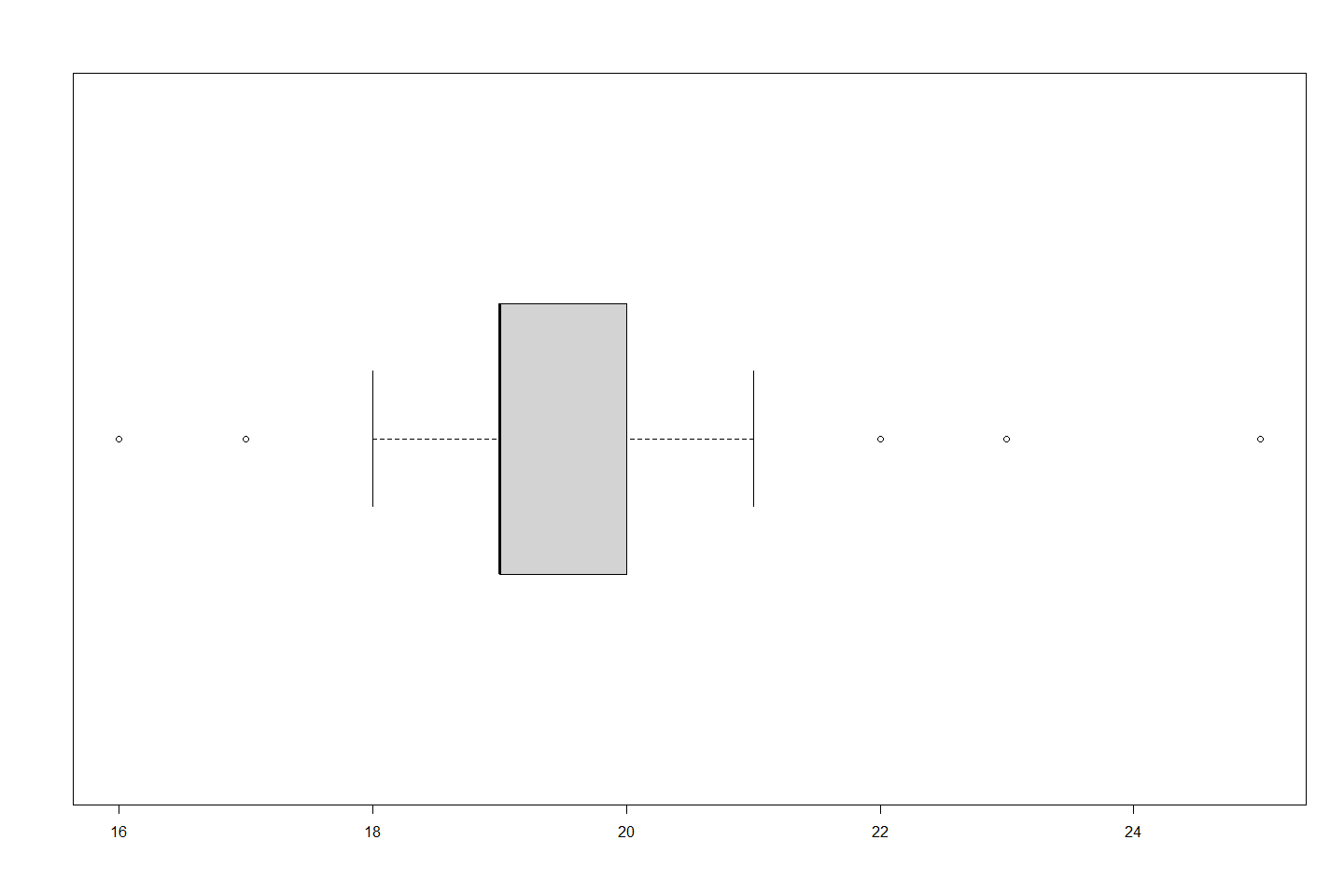
summary(StudentsData$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 19.00 19.00 19.64 20.00 25.00

IQR(StudentsData$Age)

## [1] 1

#USING a boxplot to check for outliers in the distribution  
boxplot(StudentsData$Age,horizontal = TRUE)



The distribution of the weekly expenditure of students taking statistics is skewed to the left while the distribution of their ages is skewed right. This is because the distribution of the weekly expenditure has its mean<medianmedian and mode. The weekly expenditure distribtion has an IQR of 276.25 cedis whiles the age distribution has an IQR of 1. The weekly expenditure distribtion has a median of 240 cedis while the age distribution has a median of 19 years. The age distribution has approximately 5 outliers while the weekly expenditure distribution has no outliers.

### ###Question 3

var=c(91,94,97,100,102,102,103,108,111,112,115,115,116,116,117,117,117,122,122,123,124,128,129,130,132)  
a = summary(var);a

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 91.0 103.0 116.0 113.7 122.0 132.0

IQR(var)

## [1] 19

l1 <- (quantile(var,0.25))  
l2 <- (quantile(var,0.75))  
ir <- IQR(var)  
lower <- l1 - 1.5\*(ir);lower

## 25%   
## 74.5

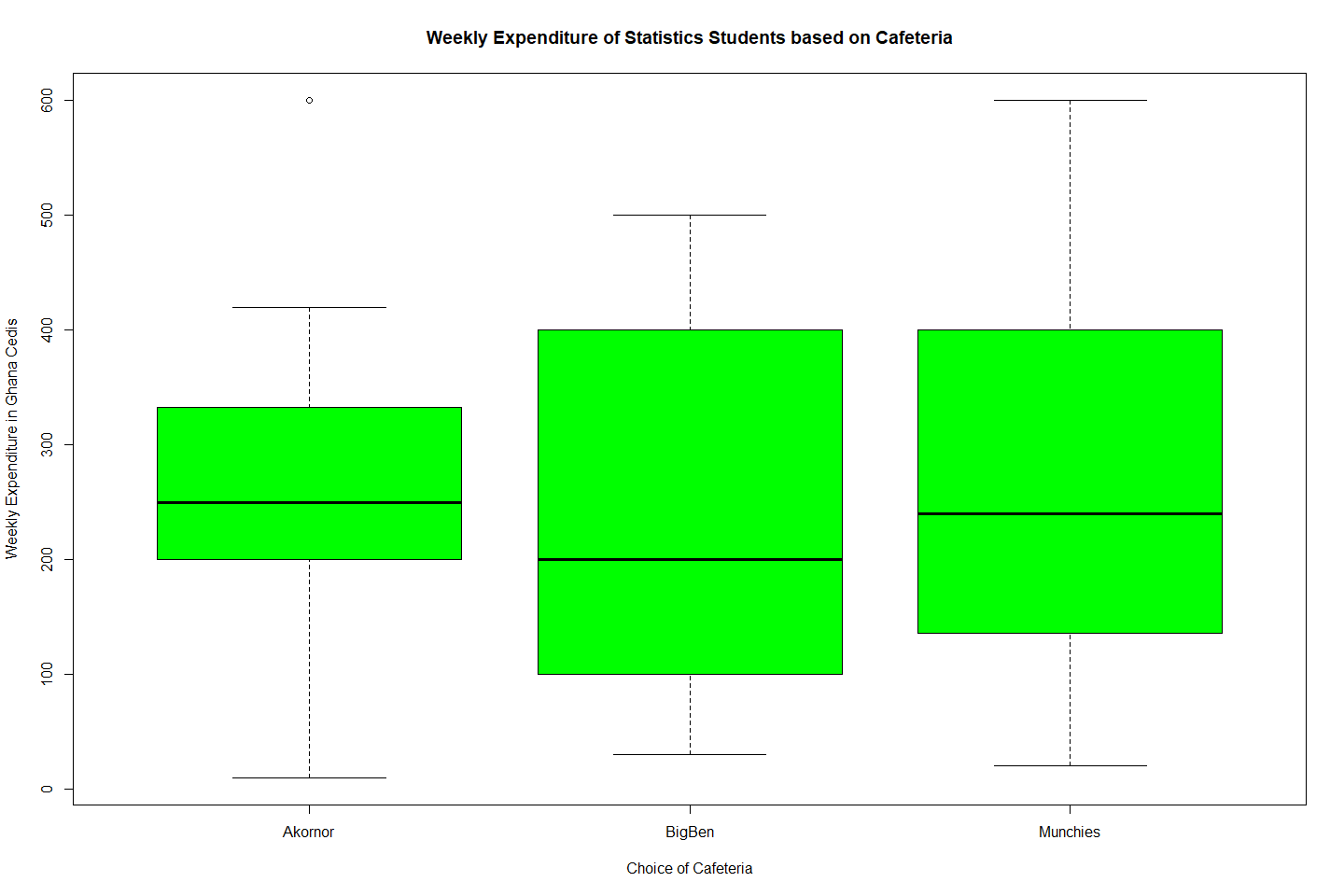
upper <- l2 + 1.5\*(ir);upper

## 75%   
## 150.5

1. 91, 94, 97, 100, 102
2. 116 Cedis
3. 107 cedis, because it appears thrice in the dataset, more than any other value.
4. The middle 50% of customers paid between 103 cedis and 122 cedis as fees.
5. 19
6. There are no outliers in this dataset, as calculated in the code chunk above. There is no value below l1 (74.50 cedis) or l2 (150.50 cedis).

### ###Question 4

#a.  
boxplot(StudentsData$Expenditure ~ StudentsData$Cafeteria,main="Weekly Expenditure of Statistics Students based on Cafeteria",ylab="Weekly Expenditure in Ghana Cedis",xlab='Choice of Cafeteria',col='green')



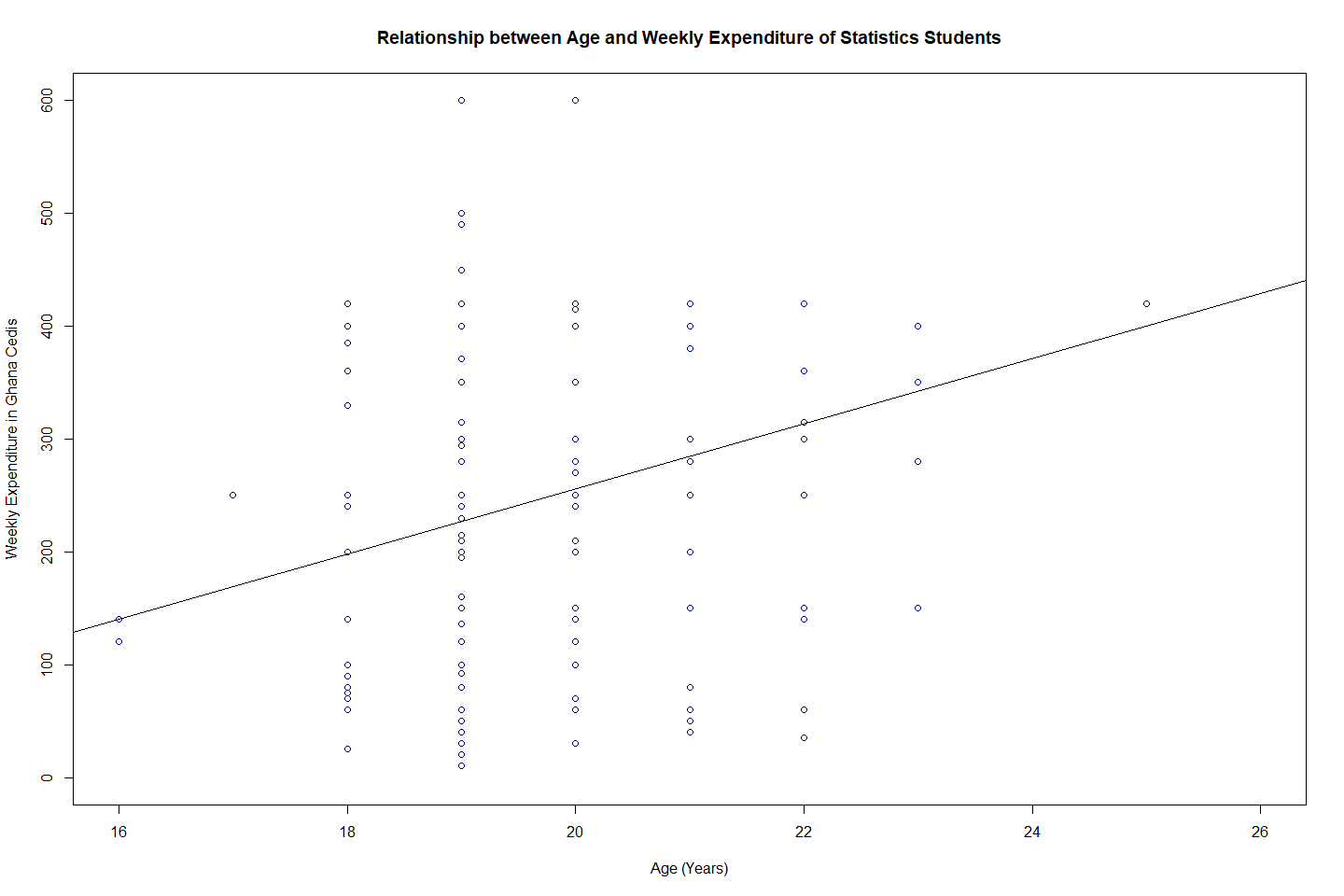
#b.   
tapply(StudentsData$Expenditure,StudentsData$Cafeteria,fivenum)

## $Akornor  
## [1] 10.0 200.0 250.0 332.5 600.0  
##   
## $BigBen  
## [1] 30 100 200 400 500  
##   
## $Munchies  
## [1] 20 136 240 400 600

1. In this box and whisker plot, the relationship between students’ choice of cafeteria, and their weekly expenditure is explored. Choice of cafeteria is the categorical and explanatory variable, and weekly expenditure is the quantitative and response variable. The question ‘Can a student’s weekly expenditure be explained by their prefered cafteria?’ is answered. The median weekly expenditure of students who prefer Akornor is 250 cedis, 200 cedis for BigBen, and 240 cedis for Muchies. However, 25% of students who prefer Munchies or BigBen generally spend more than students who prefer Akornor (except for one outlier who spends more than every other student in the dataset). The weekly expenditure of students who prefer Munchies vary a lot more than the students who prefer the other cafeterias.

### ###Question 5

#a,b.   
plot(col='navyblue',x = StudentsData$Age,  
y = StudentsData$Expenditure,   
xlab = "Age (Years)",  
ylab = "Weekly Expenditure in Ghana Cedis",  
xlim = c(16, 26),  
ylim = c(0, max(StudentsData$Expenditure)),main = "Relationship between Age and Weekly Expenditure of Statistics Students") + abline(lm(StudentsData$Expenditure ~ StudentsData$Age, data= StudentsData))



## integer(0)

#c.  
cor(StudentsData$Age,StudentsData$Expenditure)

## [1] 0.2679785

1. Based on the pattern observed in the scatter plot and the value for the correlation coefficient (0.2679), there does appear to be a linear relationship between the two variables, age and expenditure. The direction of the relationship is positive, meaning that as one variable increases, the other variable also tends to increase. However, the strength of the relationship is weak, as indicated by the low correlation coefficient value. Additionally, there are outliers present in the data.

### ###Question 6

#a.  
GradeA= c(588, 579, 585, 580, 583)  
meanA = mean(GradeA);meanA

## [1] 583

medianA = median(GradeA);medianA

## [1] 583

IQR(GradeA)

## [1] 5

standardDA = sd(GradeA);standardDA

## [1] 3.674235

GradeB = c( 570, 574, 572, 571, 578)  
meanB = mean(GradeB);meanB

## [1] 573

medianB = median(GradeB);medianB

## [1] 572

IQR(GradeB)

## [1] 3

standardDB = sd(GradeB);standardDB

## [1] 3.162278

#c.  
GradeB = c( 570, 574, 572, 571, 588)  
meanB = mean(GradeB);meanB

## [1] 575

medianB = median(GradeB);medianB

## [1] 572

a = IQR(GradeB)  
standardDB = sd(GradeB);standardDB

## [1] 7.416198

b = quantile(GradeB,0.75)  
c = 1.5\*(a) + b

1. Grade B. The mean thickness of the Grade B tires is 573mm, and the median thickness is 572 mm. These values are closer to the expected 575mm than the Grade A tires, which have a mean and median of 583mm. Additionally, the standard deviation of Grade B tires is 3.16mm, which is less than the standard deviation of Grade A tires, 3.67mm. This means that the Grade B tires deviate less from the mean than the Grade A tires do. Also, the inter-quartile range of Grade A is larger than the inter-quartile range of Grade B. Hence, Grade B tires have higher consistency and are closer to the expected 575mm than Grade A.
2. My responses for (a) will change in that the mean and standard deviation for Grade B increase to 575mm and 7.42 mm. The median does not change. My responses for (b) will not change in that because the presence of 588mm in the dataset is an outlier. This can be confirmed by computing the inter-quartile range of Grade B, and using the interquartile rule.The operations manager of Hankook would be better served by ignoring the outlier. In the case where the outlier is kept, the operations manager would be best served by using comparing the median values.