Homework Assignment 1

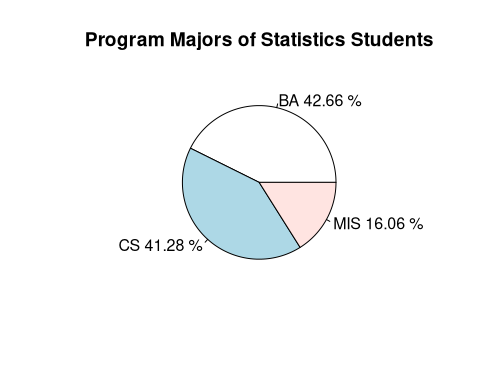
David Abeiku Saah

29/01/2023

library(readr)  
  
StudentsData <- read\_delim("StudentsData.txt",   
 delim = "\t", escape\_double = FALSE,   
 trim\_ws = TRUE)

### Question 1a

major = table(StudentsData$Major);  
prop = prop.table(major);  
percent = prop.table(major)\*100;  
pf = round(percent,2);  
lbl = paste(c("BA", "CS", "MIS"), pf, "%", sep=" ");  
pie(major, labels = lbl, main = "Program Majors of Statistics Students");



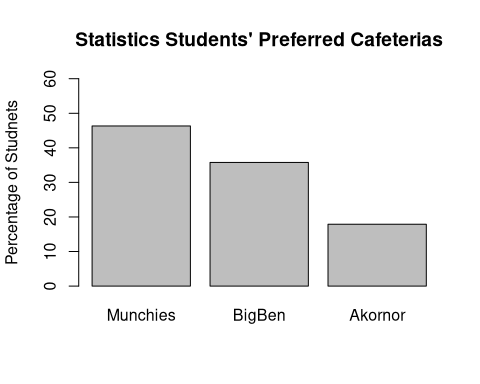
The major of Statistics students are not equally distributed among the 3 categories. However, it appears that a majority of students offer either Business Administration (42.66%) or Computer Science (41.28%). The remaining students (16.06%) offer Management Information Systems. This means that if one were to pick a student at random from the Statistics class, the student would most likely be a Computer Science or a Business Administration Student.

### Question 1b

# tabulate view counts for each factor  
caf\_1 <- table(StudentsData$Cafeteria);caf\_1

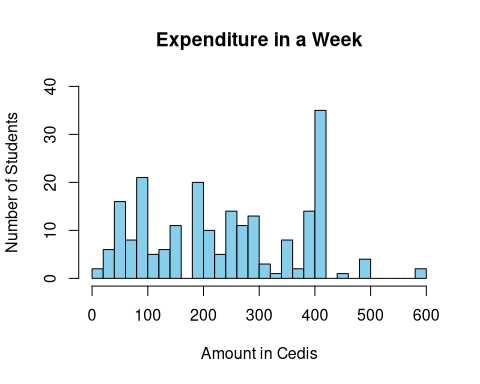
##   
## Akornor BigBen Munchies   
## 39 78 101

# reorder factor levels from highest to lowest based on observed counts in caf\_1  
caf\_2 <- factor(StudentsData$Cafeteria, levels = c("Munchies", "BigBen", "Akornor"));  
  
# tabulate view counts for each factor now in descending order  
caf\_table <- table(caf\_2);  
  
# generate bar chart  
barplot((100\*caf\_table) / length(caf\_2), main = "Statistics Students' Preferred Cafeterias", ylab = "Percentage of Studnets", ylim = c(0, 60));

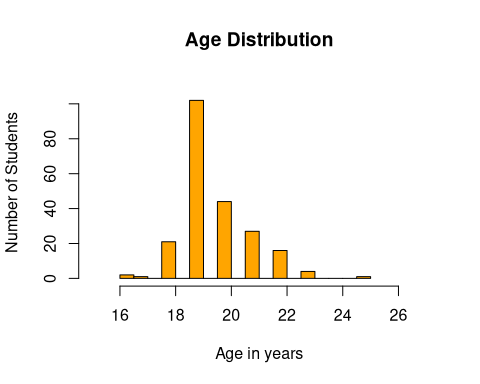


### Question 2a

hist(StudentsData$Expenditure, breaks = 40, col = "skyblue", ylim = c(0, 40), xlim = c(0, 600), main="Expenditure in a Week", xlab = "Amount in Cedis", ylab = "Number of Students");



hist(StudentsData$Age, breaks = 20, col = "orange", ylim = c(0, 110), xlim = c(15, 27), main="Age Distribution", xlab = "Age in years", ylab = "Number of Students");



### Question 2b

summary(StudentsData$Expenditure);

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

summary(StudentsData$Age);

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

# Calculations for expenditure.  
# Expenditure is right skewed since mean > median.  
median(StudentsData$Expenditure);

## [1] 240

IQR(StudentsData$Expenditure);

## [1] 276.25

# Calculations for age.  
# Age is roughly symmetrical since its mean and median are approximately the same.  
mean(StudentsData$Age);

## [1] 19.63761

sd(StudentsData$Age);

## [1] 1.270675

### Question 2c

The distribution for students’ age is roughly symmetrical while that of students’ expenditure is right skewed. This means that most students are either 19 or 20 years of age (since center = 19.64 years) and most students spend below GHS 240. The age distribution of Statistics students is more consistent since it has a small measure of spread (1.27) while the amount students spend per week varies greatly because it has 276.25 as a measure of spread. There also appear to be outliers in both distributions.

### Question 3

1. The first 5 values: 91, 94, 97, 100, 102.
2. The median fees paid is GHS 116.
3. Yes. 117 is the mode because it appears most frequently in the data.
4. The middle 50% of customers paid between 122.5 and 102.5 cedis as fees.

**Workings:**  
The middle 50% is between the third quartile(Q3) and first quartile(Q1).

Q1 = 1/4 \* (n + 1)  
Q1 = 1/4 \* (25 + 1)  
Q1 = 6.5  
Q1 = Between the 6th and 7th observation.  
Q1 = 102.5

Q3 = 3/4 \* (n + 1)  
Q3 = 3/4 \* (25 + 1)  
Q3 = 19.5  
Q3 = Between the 19th and 20th observation.  
Q3 = 122.5

1. IQR = Q3 - Q1 = 122.5 - 102.5 = 20

lower\_limit = 102.5 - (1.5 \* 20);lower\_limit

## [1] 72.5

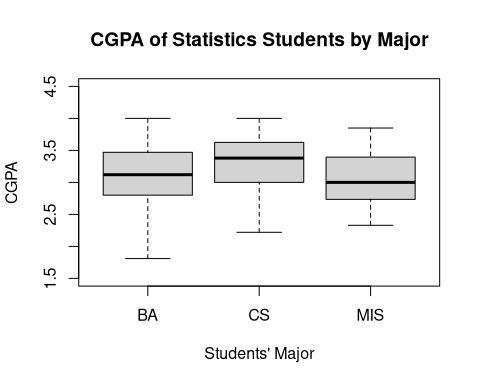
upper\_limit = 117.5 + (1.5 \* 20);upper\_limit

## [1] 147.5

1. There are no outliers in the distribution.

### Question 4a

boxplot(StudentsData$CGPA ~ StudentsData$Major, main="CGPA of Statistics Students by Major", ylab="CGPA", xlab = "Students' Major", ylim = c(1.5, 4.5));



### Question 4b

# Five number summary  
# min, Q1, Median, Q3, max  
tapply(StudentsData$CGPA, StudentsData$Major, fivenum);

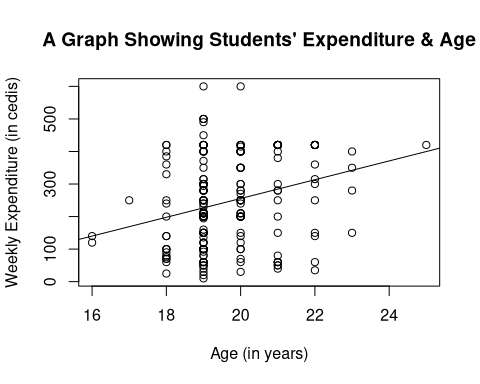
## $BA  
## [1] 1.81 2.80 3.12 3.47 12.00  
##   
## $CS  
## [1] 2.220 3.000 3.380 3.625 4.000  
##   
## $MIS  
## [1] 2.330 2.735 3.000 3.395 3.850

### Question 4c

The median CGPA for Computer Science (3.38) is greater than the other two distributions (3.12 for BA and 3.00 for MIS). The CGPA spread for CS students is 0.625, 0.67 for BA students, and 0.66 for MIS students. This shows that CS Students have a consistent CGPA compared to the students in other majors. When a Statistics student is selected at random, if he/she offers CS, they will most likely have a high CGPA.

### Question 5a

plot(StudentsData$Age, StudentsData$Expenditure, main = "A Graph Showing Students' Expenditure & Age", xlab = "Age (in years)", ylab = "Weekly Expenditure (in cedis)");  
  
model = lm(StudentsData$Expenditure ~ StudentsData$Age);  
abline(model);



cor(StudentsData$Expenditure, StudentsData$Age, use = "complete.obs")

## [1] 0.2679785

### Question 5d

From the graph, the distribution neither shows a distinct shape/pattern nor a form of relationship between a student’s age and their weekly expenditure. The correlation coefficient of the distribution is 0.27, which suggests no linear relationship between a student’s age and the amount of money they spend within a week. There are possible outliers in the distribution. In conclusion, a Statistics student’s weekly expenditure does not depend on their age.

### Question 6a

gradeA = c(588, 579, 585, 580, 583);  
gradeB = c(570, 574, 572, 571, 578);  
  
# Mean, median, standard deviation for Grade A  
mean(gradeA);

## [1] 583

median(gradeA);

## [1] 583

sd(gradeA);

## [1] 3.674235

# Mean, median, standard deviation for Grade B  
mean(gradeB);

## [1] 573

median(gradeB);

## [1] 572

sd(gradeB)

## [1] 3.162278

### Question 6b

From the question, a grade of tire with an inner diameter of 575 millimeters is of high quality. Grade B tires provides a better quality. Though the median and mean of Grade A tires are symmetric (583 millimeters for both), their deviation from the actual value is very high compared to Grade B tires which only deviates 2-3 points from the quality value (575 mm). To further support this claim, the standard deviation of the inner diameter of Grade B tires are smaller (3.16) and therefore more consistent with the value provided by the mean (573). On the other hand, the standard deviation for Grade A tires are large, hence, the inner diameter of the tires varies.

### Question 6c

gradeA = c(588, 579, 585, 580, 583);  
gradeB = c(570, 574, 572, 571, 588);  
  
# Mean, median, standard deviation for Grade A  
mean(gradeA);

## [1] 583

median(gradeA);

## [1] 583

sd(gradeA);

## [1] 3.674235

# Mean, median, standard deviation for Grade B  
mean(gradeB);

## [1] 575

median(gradeB);

## [1] 572

sd(gradeB)

## [1] 7.416198

My answer still does not change. The median (572) for Grade B tires stays the same while the mean (from 573 to 575) and standard deviation (from 3.16 to 7.42) change since they are influenced by the new value. Nonetheless, I will still choose Grade B tires because the measure of center is still close to the standard quality measurement compared to Grade A tires.