**STAT 2600 *Exam1 from 4pm to 5:30pm on 2/8* NAME: Bryan Greener**

**( ) out of 19 points + 1 bonus point is your total score.**

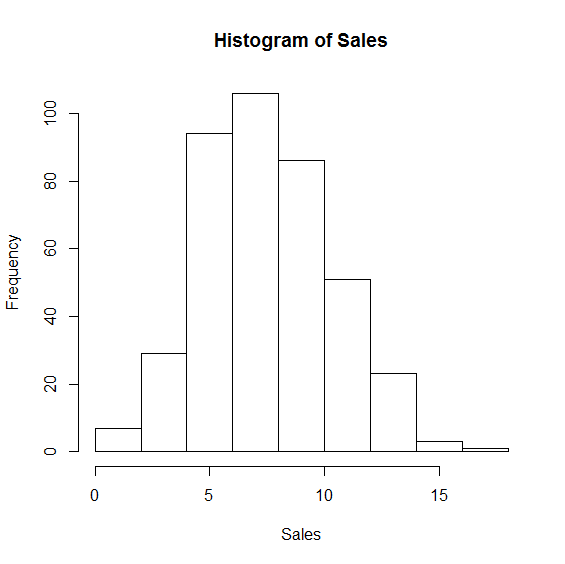
**In order to get full credit, you must use R to answer the problems along with the R codes used. Download ‘data.csv’ file. Write your answer right after each subproblem. Do not forget to make a copy of figures and paste it in this file.**

Data contains sales of child car seats at 400 different stores and consists of the following 11 variables:

|  |
| --- |
| ***Sales***: Unit sales (in thousands) at each location  ***CompPrice***: Price charged by competitor at each location  ***Advertising***: Local advertising budget for company at each location (in thousands of dollars)  ***Population***: Population size in region (in thousands)  ***Price***: Price company charges for car seats at each site  ***Age***: Average age of the local population  ***Education***: Education level at each location  ***ShelveLoc***: A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site  ***Urban***: A factor with levels No and Yes to indicate whether the store is in an urban or rural location  ***US***: A factor with levels No and Yes to indicate whether the store is in the US or not  ***Income***: A factor with levels low, middle and high indicating community income level |

1. Inference about ***Sales*** variable.

(a) Make a histogram and describe the distribution briefly. 2pts



The distribution of this dataset is pretty close to symmetric. As the mean and median values, as shown in part b and c are very close to each other.

(b) Find the mean and standard deviation. 1pt

Mean=7.496325

Standard Deviation=2.824115

(c) What is the five number summary? 1pt

0.00 5.38 7.49 9.32 16.27

(d) Check whether there are outliers or not. **Must use a 1.5IQR rule and show your work in R**. 2pts

**Above Q3**

> which(Sales>(9.32+outlier))

[1] 8 12 18 19 22 26 31 47 57 69 74 83 88 99 122 140 150 159 170

[20] 172 190 194 213 233 242 271 273 293 295 305 317 326 340 349 353 358 368 377

[39] 385 396

**Below Q1**

> which(Sales<(5.38-outlier))

[1] 29 35 40 41 51 58 63 85 107 144 162 166 175 198 204 210 216 281 298

[20] 325 329 360

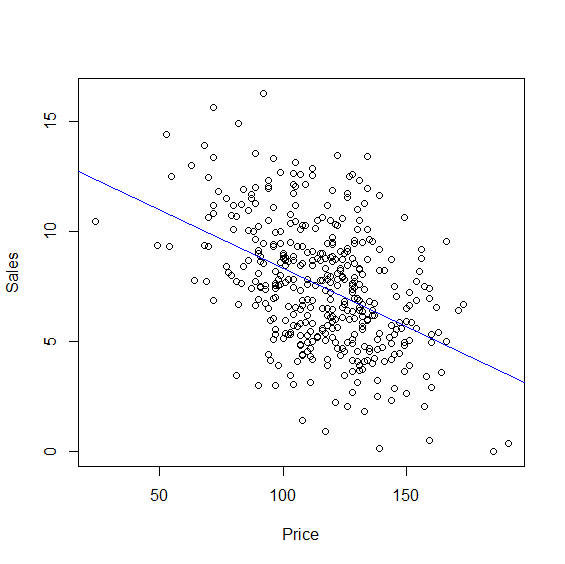
2. We will attempt to predict ***Sales*** based on ***Price*** by using a least-squares regression line, i.e., we treat ***Price*** and ***Sales*** as an explanatory and response variable, respectively.

(a) What are the intercept and the slope of the least-squares regression line? 2pts

Intercept: 143.7589

Slope: -3.730354

(b) Make a scatterplot and add the line to the plot. 1pt



(c) Comment on the form, direction, and strength of the relationship in the scatterplot (b). 1pt

Form: If this has any form at all, I would say it is clustered however the form is very loose.

Direction: Negative

Strength: The points don’t fit this line too well however the closer you get to the LSR line the more the points seem to clump up and form a tightly packed line. That being said, a good majority of the points aren’t anywhere near the LSR line.

(d) Predict a value of ***Sales*** at ***Price***=150. 1pt

When Price=150, sales will be roughly 5.5

(e) What proportion of the variation in ***Sales*** is explained by ***Price***? 1pts

You can see on the graph that as price rises, the amount of sales drops faster and faster. Near the lower price points, we had many more outliers on the high side of sales as well.

3. Investigate the association between ***Income*** and ***US*** variables:

(a) Fill out the actual counts in the following table. 2pts

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Income | | |
| Low | Middle | High |
| US | Yes | 60 | 131 | 67 |
| No | 41 | 70 | 31 |

(b) Give the joint distribution of ***US*** and ***Income***. 1pt

Low Mid High

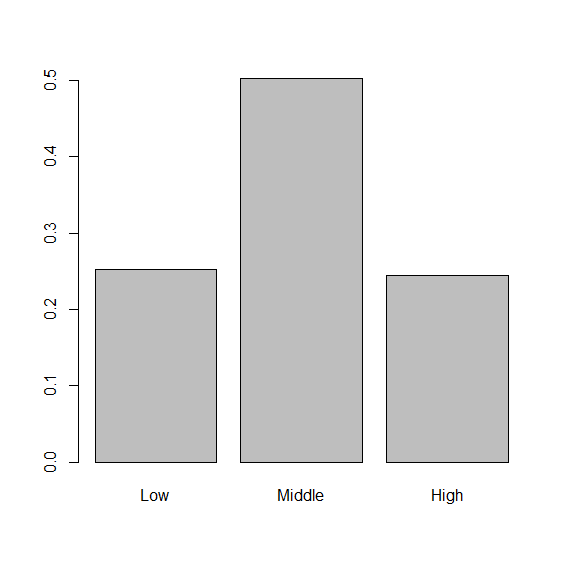
US 0.1500 0.3275 0.1675

NOT US 0.1025 0.1750 0.0775

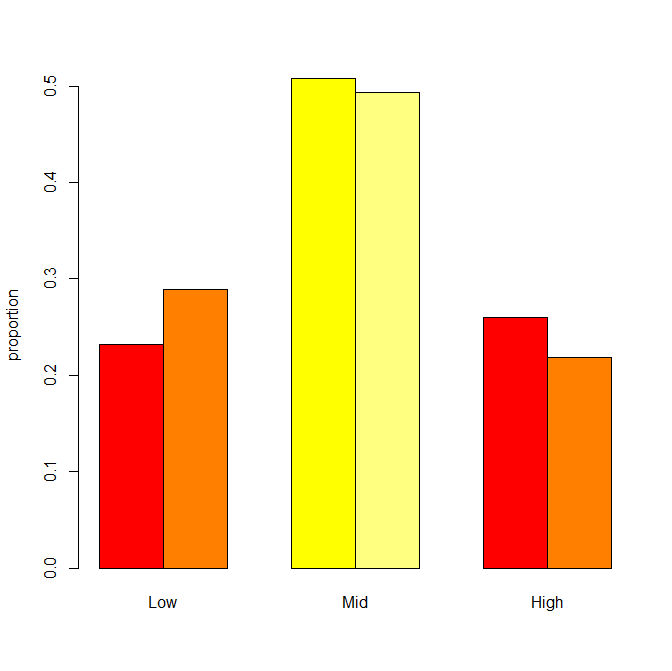
(c) What is the marginal distribution of ***Income***? Display the result graphically. 1pt

Low Middle High

0.2525 0.5025 0.2450



(d) What is the conditional distribution of ***Income*** for Yes and for No? Display the result graphically. 2pt



(e) Summarize the relationship between ***US*** and ***Income*** using the result of (c). 1pt

Those that are in the US tend to have a higher income than those that are not in the us as long as they are in the Middle and Upper classes.

Attached below is the code:

# Import/Attach Dataset

data=read.csv("data.csv")

attach(data)

# Show Variables

names(data)

# #1

# a)

hist(Sales)

# b)

mean(Sales)

sd(Sales)

# c)

fivenum(Sales)

# d)

IQR(Sales)

# find how far from Q1 and Q3 outlier must be

outlier=(IQR(Sales)\*1.5)-IQR(Sales)

# Outliers above Q3

which(Sales>(9.32+outlier))

# Outliers below Q1

which(Sales<(5.38-outlier))

# #2

# a and b)

# Calculate Regression Line

r=cor(Price,Sales)

sdy=sd(Price)

sdx=sd(Sales)

ybar=mean(Price)

xbar=mean(Sales)

b1=r\*sdy/sdx

# Display b1

b1

b0=ybar-b1\*xbar

# Display b2

b0

yhat=b0+b1\*xbar

# Plot points

plot(Price,Sales)

# Create line and add it to scatterplot

fit=lm(Sales~Price)

abline(fit,col='blue')

# #3

data2=c(60,131,67,41,70,31)

datamatrix=matrix(data2,nrow=2,ncol=3,byrow=TRUE)

rownames(datamatrix)=c('US','NOT US')

colnames(datamatrix)=c('Low','Mid','High')

datamatrix

row1=sum(datamatrix[1,])

row2=sum(datamatrix[2,])

col1=sum(datamatrix[,1])

col2=sum(datamatrix[,2])

col3=sum(datamatrix[,3])

total=sum(data2)

prop.joint=datamatrix/total

prop.joint

# Generate marginal distribution for row/col

prop.margin.row=c(row1/total,row2/total)

names(prop.margin.row)=c('US','NOT US')

prop.margin.col=c(col1/total,col2/total,col3/total)

names(prop.margin.col)=c('Low','Middle','High')

barplot(prop.margin.row)

barplot(prop.margin.col)

prop.margin.col

prop.margin.row

# Generate conditional distribution

prop1=datamatrix[1,]/row1

prop2=datamatrix[2,]/row2

names(prop1)=c('Low','Middle','High')

names(prop2)=c('Low','Middle','High')

table1=prop.table(datamatrix,margin=1)

table2=prop.table(datamatrix,margin=2)

barplot(table1,beside=TRUE,col=heat.colors(4),ylab="proportion")