R Assignment 1

Jonathan De Los Santos

Problem 1 (34 points)

The production of beer is a multibillion-dollar worldwide industry. The dataset in the following link include the alcohol per volume and calories of famous beer brands.

```
mydata <- read.csv("http://tiny.cc/isqs5347-beer")
```

A.

Compute the following statistics for variables "Alcohol" and "Calories": mean, median, variance, standard deviation, Q1 & Q3, and interquartile range. (14 points)

Let's look at some basic summary data of the Alcohol column. Mean, the average alcohol in each bear:

```
mean(mydata$Alcohol)
```

```
## [1] 4.955827
```

Median, the middle alcohol amount in each beer:

```
median(mydata$Alcohol)
```

```
## [1] 4.93
```

Variance, the average of the squared variations from the mean:

```
var(mydata$Alcohol)
```

```
## [1] 0.8045737
```

Standard deviation, the square root of the variance:

```
sd(mydata$Alcohol)
```

```
## [1] 0.8969803
```

Interquartile ranges, the 25th and 75th percentiles:

```
quantile(mydata$Alcohol, c(.25, .75))
```

```
##   25%   75%
## 4.510 5.265
```

Now we will examine the same data for calories: Mean:

```
mean(mydata$Calories)
```

```
## [1] 42.22047
```

Median:

```
median(mydata$Calories)
```

```
## [1] 43
```

Variance:

```
var(mydata$Calories)
```

```
## [1] 63.07799
```

Standard deviation:

```
sd(mydata$Calories)
```

```
## [1] 7.942165
```

Interquartile ranges:

```
quantile(mydata$Calories, c(.25, .75))
```

```
##  25%  75%
## 39.5 45.0
```

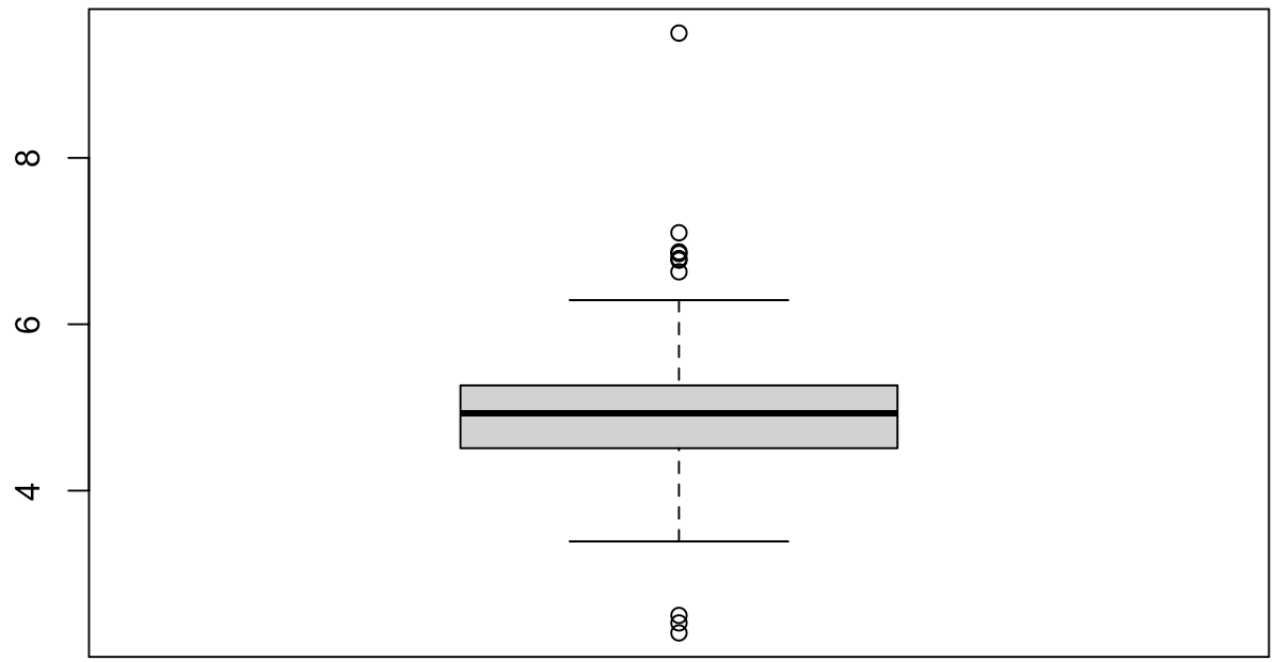
B.

Create separate boxplots for variables “Alcohol” and “Calories”. Are there any outliers for each variable (8 points).

Another good way to examine this dataset is with a boxplot which shows the middle of the dataset as well as it’s quartiles.

Here is the boxplot of Alcohol:

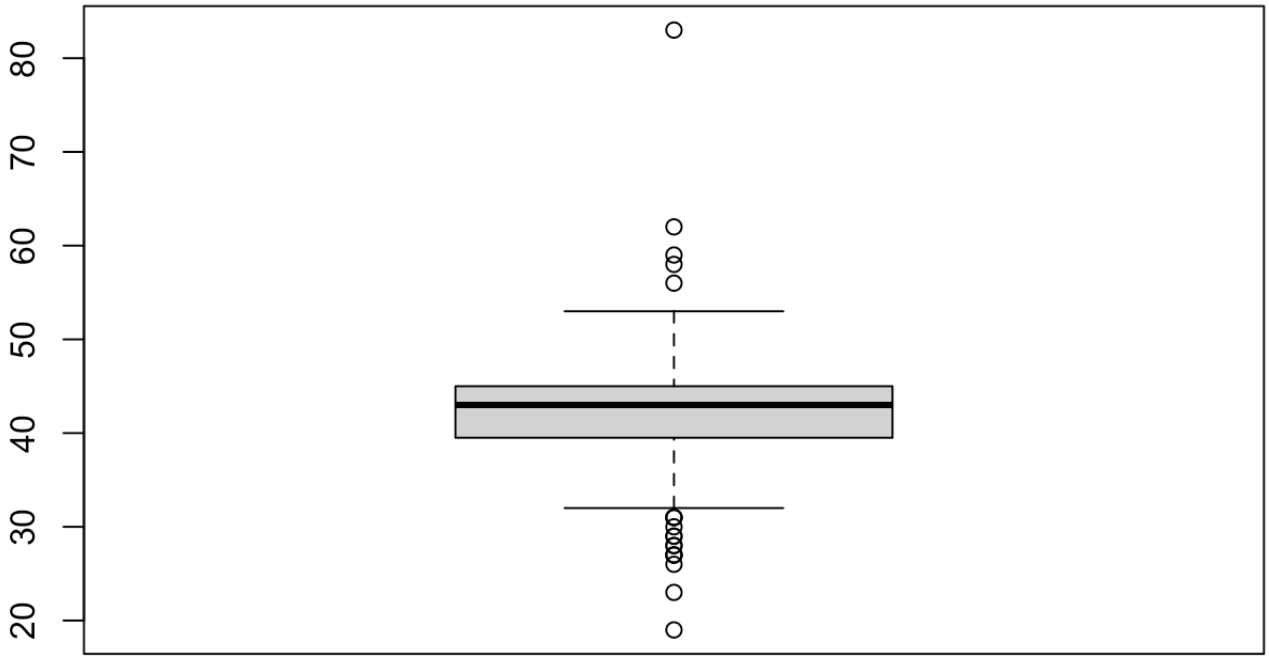
```
boxplot(mydata$Alcohol)
```



The dot you see at the very top is an extreme outlier in the data. Beyond that you can see there are several other points outside of the quartiles as well.

Here is the box plot for Calories:

```
boxplot(mydata$Calories)
```

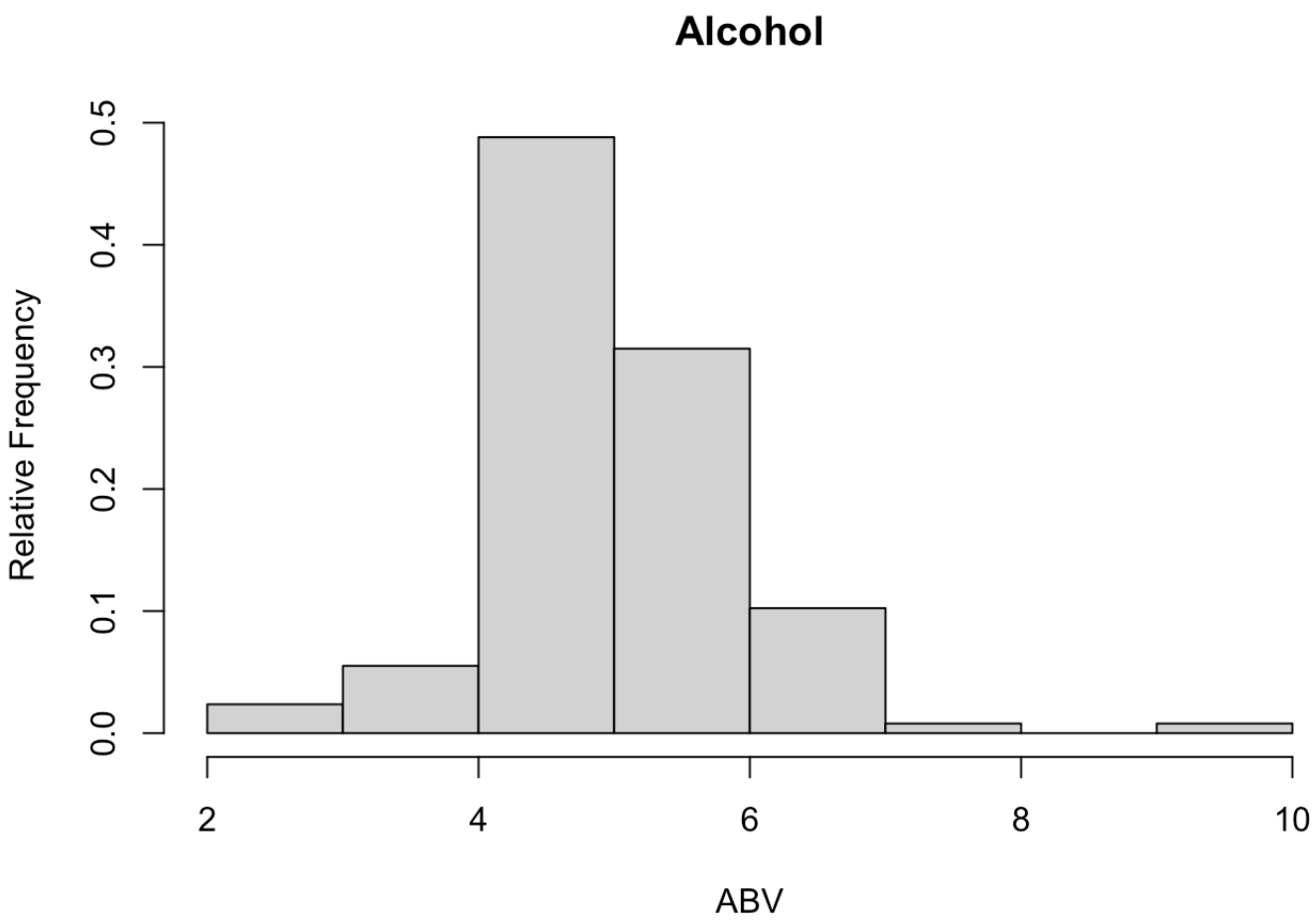


Again we see an extreme outlier at the top of the plot meaning we have a beer with an extraordinarly high amount of calories.

C.

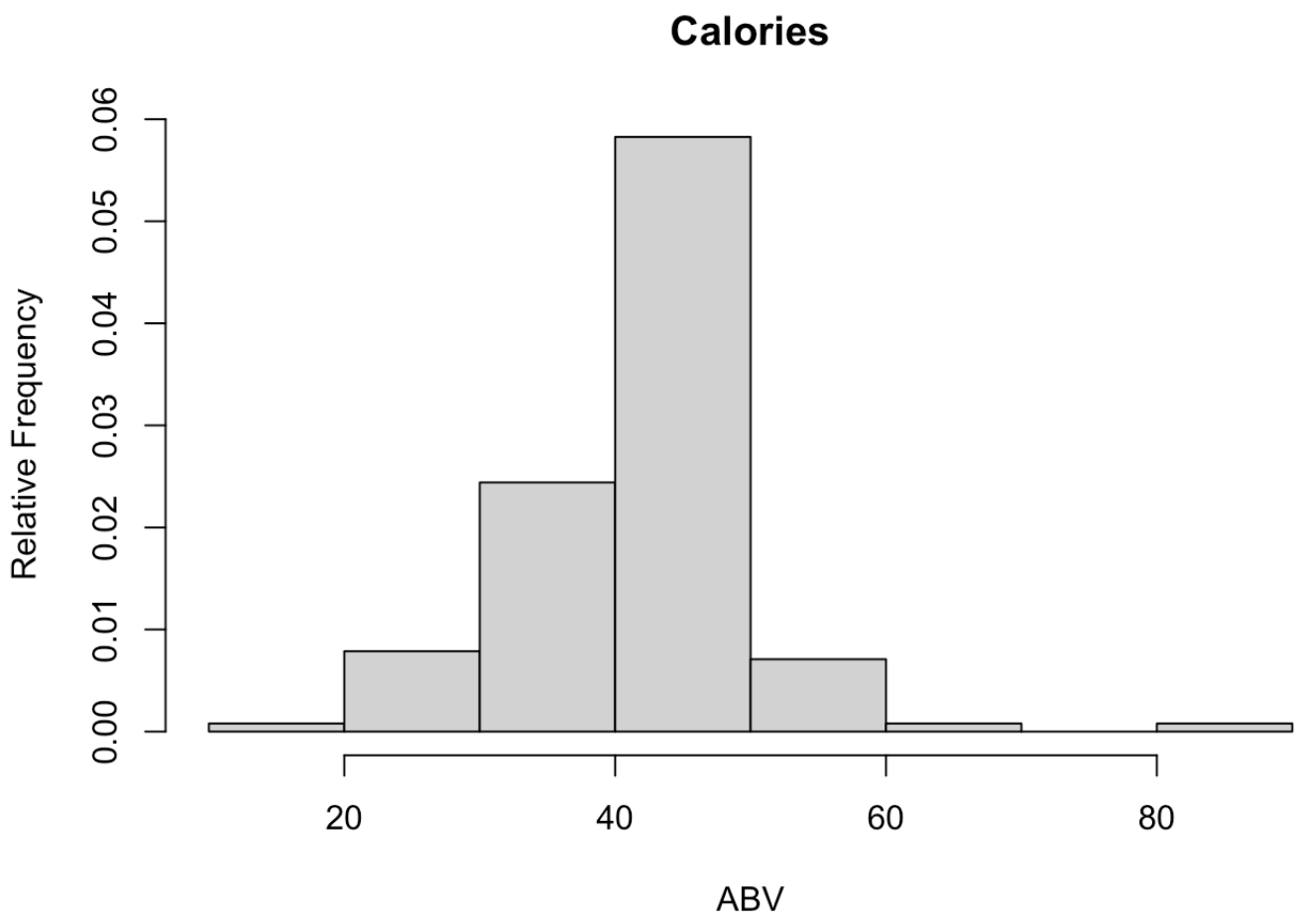
Create separate histograms for variables “Alcohol” and “Calories”. Make sure the y-axes presents the relative frequency (a value between 0 and 1). (6 points) This histogram will plot the Alcohol amounts with their relative frequencies

```
hist(mydata$Alcohol, freq = FALSE, main = "Alcohol", ylab = "Relative Frequency", xla
b = "ABV")
```



We can do the same for Calories:

```
hist(mydata$Calories, freq = FALSE, main = "Calories", ylab = "Relative Frequency", x
lab = "ABV")
```



D.

Are variables “Alcohol” and “Calories” skewed or symmetrical? If skewed, in which direction? What that means, explain. (6)

```
# hint: You can find skewness() function using "e1071" R library.
# install.packages("e1071") # if you are going to use it for the first time.
install.packages("e1071",  repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/1n/nvr79nb55tz9j4lsbrw6hsf80000gn/T//Rtmpu1iO54/downloaded_packages
```

```
library(e1071)
```

Skewness will tell us if the normal bell curve is distributed more in a certain direction from the median. Here, alcohol has a positive skew value so the distribution is right skewed. In other words, the mean is to the right of the median.

```
skewness(mydata$Alcohol)
```

```
## [1] 0.8563446
```

The same appears to be true for Calories, though it is not quite as skewed:

```
skewness(mydata$Calories)
```

```
## [1] 0.662515
```

Problem 2 (46 points)

Use the TTU graduate student exit survey data.

```
grad <- read.csv("http://westfall.ba.ttu.edu/isqs6348/Rdata/pgs.csv", header = T)
# Attach allows you to not have to type the dataset name to identify a column
attach(grad)
```

Two variables of interest are “FacTeaching”, a 1,2,3,4,5 rating of teaching at TTU by the student, and “COL”, the college from which the student graduated.

A.

Construct the contingency table showing counts of students in all combinations of these two variables. (10 points) The following table will show us a count of the number of students in each combination of college and the rating they gave to the faculty:

```
table(FacTeaching, COL)
```

```
##             COL
## FacTeaching  AG  AR  AS  BA DUAL  ED  EN  GR  HS  MC VPA
##           1   4   3  12   9    0   3   5   0   1   0   4
##           2  15   4  24  28    0   6  36   3   5   0   7
##           3  26   6 124  44    2  26  65   8  17   3  10
##           4  78  16 290 116    0 113 168  27  41  25  37
##           5  56   4 171  66    0  93  86  15  33   6  44
```

B.

Construct a contingency table showing the proportion (probability) of students in all combinations of these two variables (5 points). Round results by three decimals (1 points). We can view this as a probability table, showing the proportion of total students in each combined category:

```
t <- (table(FacTeaching, COL))/length(COL)

round(t,3)
```

```
##            COL
## FacTeaching    AG    AR    AS    BA  DUAL    ED    EN    GR    HS    MC   VPA
##           1 0.002 0.001 0.006 0.004 0.000 0.001 0.002 0.000 0.000 0.000 0.002
##           2 0.007 0.002 0.012 0.014 0.000 0.003 0.018 0.001 0.002 0.000 0.003
##           3 0.013 0.003 0.062 0.022 0.001 0.013 0.032 0.004 0.008 0.001 0.005
##           4 0.039 0.008 0.145 0.058 0.000 0.056 0.084 0.013 0.020 0.012 0.018
##           5 0.028 0.002 0.085 0.033 0.000 0.046 0.043 0.007 0.016 0.003 0.022
```

C.

What is the probability that a randomly selected student is from college of business administration (BA)? We call this the marginal probability, P(COL=BA). (5 points) If we sum the probabilities in each column, we can see the probability of each college. In this case there is a 13.1% chance a student picked at random would be from the college of Business Administration.

```
round(colSums(t),3)
```

```
##    AG    AR    AS    BA  DUAL    ED    EN    GR    HS    MC   VPA
## 0.089 0.016 0.310 0.131 0.001 0.120 0.180 0.026 0.048 0.017 0.051
```

D.

What is the probability that a randomly selected student is from BA and rates the teaching quality by 5? We call this the joint probability, P(COL = BA and FacTeaching = 5). (5 points) Looking back at our probability table, 3.3% of BA students rate the faculty a 5.

```
# Hint: You can find the answer in the table you made in part c. Just report the prob
ability value for the intersection of BA and 5. For example, the P(COL=AG and FacTeac
hing = 3) = 0.013.
```

E.

Given that a randomly selected student is from BA, what is the probability that he/she rates the teaching quality by 5? We call this the conditional probability, P(FacTeaching = 5 | COL=BA). (5 points) Knowing the probability of the student being from BA AND rating the teaching quality a 5 is 3.3%, we can calculate the conditional probability by dividing that by the probability that a student picked at random is from BA.

```
.033/.131
```

```
## [1] 0.2519084
```

We see that there is a 25.2% chance that a randomly selected student from BA rates the teaching quality a 5.

```
# Hint: Use the conditional probability formula. You can use the answer to part c and
d as inputs for conditional probability formula.
```

F.

Given that a randomly selected student is from college of education (ED), what is the probability that he/she rates the teaching quality by 5? In other words P(FacTeaching = 5 | COL=ED)? What is your conclusion about the difference between the quality of teaching in BA and ED. (5 points) From our previous tables, we know that the probability of the student being from ED and rating the teaching a 5 is 4.6%. We also know there is a 12% chance of a student being from ED. That gives us a conditional probability of:

```
.046/.12
```

```
## [1] 0.3833333
```

This is a 38% chance versus a 25% probability from the BA college. This implies that the quality of faculty is better in ED according to the students.

G.

What is the probability that a randomly selected student is fully happy about the teaching quality at TTU, hence rates FacTeaching = 5? We call this the marginal probability, P(FacTeaching=5). (5 points) Here is a table of probabilities for all of the rankings at TTU:

```
round(rowSums(t),3)
```

```
##     1     2     3     4     5
## 0.020 0.064 0.165 0.455 0.287
```

Which tells us 28.7% of students are fully happy about TTU faculty.

```
# Hint: For finding the P(FacTeaching = 5), you can perform colSums of table in part
b (if you defined FacTeaching as columns).
```

H.

Given that a randomly selected student rates the teaching quality by 5, what is the probability that he/she is graduated from BA? The P(COL=BA | FacTeaching = 5). (5 points) Knowing the probability of graduating from the BA AND rating the faculty a 5 is 3.3%, we can divide that by the 28.7% of students who rate the teaching quality a 5:

```
.033/.287
```

```
## [1] 0.1149826
```

Of all students who rate the teaching quality a 5, there is an 11.5% chance they graduated from the BA.

```
# You can use the answer to part g and d as inputs for conditional probability formul
a.
```