514 Lab 2

Julio Pagan, Joseph Fulkerson

Due Date 9/27

```
# Import needed packages
library(ggplot2)
```

1. Changing the author field and file name. (5 points)

(a) Change the `author:` field on the Rmd document

(b) Rename this file to “HW2_YourGroupNumberHere.Rmd”, where YourGroupNumberHere is changed to your group number (e.g. Group1).

2. Measure of location and variability (20 points)

A certain polymer is used for evacuation systems for aircraft. It is important that the polymer be resistant to the aging process. Twenty specimens of the polymer were used in an experiment. Ten were assigned randomly to be exposed to an accelerated batch aging process that involved exposure to high temperatures for 10 days. Measurements of tensile strength of the specimens were made, and the following data were recorded on tensile strength in psi:

```
No aging: 227 222 218 216 218 217 225 229 228 221
Aging: 219 214 218 203 215 211 209 204 201 205
```

```
# You can use the following code to create a data frame
strength = c( 227 ,222, 218, 216, 218, 217, 225, 229, 228,221,219,214,218,203,215,211,20
9,204,201,205)
aging<-as.factor(c(rep(0,10),rep(1,10)))
polymerData<-data.frame(strength,aging)
```

(a) Calculate the sample mean tensile strength of the two samples. (5 points)

```
mean(polymerData$strength[polymerData$aging==0])
```

```
## [1] 222.1
```

```
mean(polymerData$strength[polymerData$aging==1])
```

```
## [1] 209.9
```

> Sample mean of sample with no aging: 222.1 Sample mean of sample with aging: 209.9 #### (b) Calculate the median for both. Discuss your observation with the mean and median of each group. (5 points)

```
median(polymerData$strength[polymerData$aging==0])
```

```
## [1] 221.5
```

```
median(polymerData$strength[polymerData$aging==1])
```

```
## [1] 210
```

> Median of sample with no aging: 221.5 Median of sample with aging: 210

(c) Calculate the sample variance as well as standard deviation in tensile strength for both samples. (5 points)

```
var(polymerData$strength[polymerData$aging==0])
```

```
## [1] 23.65556
```

```
sd(polymerData$strength[polymerData$aging==0])
```

```
## [1] 4.863698
```

```
var(polymerData$strength[polymerData$aging==1])
```

```
## [1] 42.1
```

```
sd(polymerData$strength[polymerData$aging==1])
```

```
## [1] 6.488451
```

> Sample variance of sample with no aging: 23.65556 Sample standard dev. of sample with no aging: 4.863698 Sample variance of sample with aging: 42.1 Sample standard dev. of sample with aging: 6.488451

(d) Does there appear to be any evidence that aging affects the variability in tensile strength? (5 points)

> Yes, the group that ages has an increased variation

3.Normal Distribution with `qnorm` (15 points)

(a) Please use `qnorm` to find the critical values $z_{0.025}$. (5 points)

```
qnorm(1-0.025)
```

```
## [1] 1.959964
```

(b) Please use `qnorm` to find the critical values $z_{0.005}$. (5 points)

```
qnorm(1-0.005)
```

```
## [1] 2.575829
```

(c) Which value is larger? Why? (5 points)

> $z_{0.005}$ is larger at 2.575829 while the `qnorm` for $z_{0.025}$ is 1.959964

4. Working With Data (60 points)

- Obtaining the wine quality dataset (http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/)

(a) Import the `winequality-red` data set and name it `winequalRed`. (5 points)

```
# here is a hint for the r version
# -- change these commands as needed and delete these comments before submitting your wo
rk --
# if you downloaded the data set as a .csv file then you can read it in as follows:
# winequalRed <- read.csv("~/Documents/datasets/winequality-red.csv", sep=";")
# To view the data set
#   View(winequalRed)
```

```
winequalRed <-read.csv("winequality-red.csv", sep=";")
head(winequalRed)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

(b) Create a table of the `quality` and `alcohol` attributes from the `winequalRed` data set. (5 points)

Do not save the output from the code.

```
# hint: if you have two data columns named X and Y in your data frame, you can use code
like this to create  a table:
table(my.data.set$X, my.data.set$Y)
```

```
table(winequalRed$quality,winequalRed$alcohol)
```

```
##    
##     8.4 8.5 8.7 8.8  9 9.05 9.1 9.2 9.23333333333333 9.25 9.3 9.4 9.5 9.55
##   3   1   0   0   0  1    0   0   0                0    0   0   0   0    0
##   4   0   0   0   0  2    1   2   3                0    0   2   2   0    0
##   5   0   1   0   2 11    0  14  50                0    0  44  79  97    1
##   6   1   0   2   0 16    0   7  17                1    1  13  22  40    1
##   7   0   0   0   0  0    0   0   2                0    0   0   0   2    0
##   8   0   0   0   0  0    0   0   0                0    0   0   0   0    0
##    
##     9.56666666666667 9.6 9.7 9.8 9.9 9.95 10 10.0333333333333 10.1 10.2 10.3
##   3                0   0   1   1   1    1  0                0    0    1    0
##   4                0   6   2   3   1    0  4                0    1    0    1
##   5                0  38  35  49  25    0 29                0   23   21   13
##   6                1  15  14  23  18    0 25                2   21   20   18
##   7                0   0   2   1   4    0  8                0    2    4    1
##   8                0   0   0   1   0    0  1                0    0    0    0
##    
##     10.4 10.5 10.55 10.6 10.7 10.75 10.8 10.9 11 11.0666666666667 11.1 11.2
##   3    0    0     0    0    1     0    0    1  1                0    0    0
##   4    3    1     0    0    0     0    0    3  4                0    1    3
##   5   12   31     0    8    7     0    9   13 19                0    7    8
##   6   25   25     1   14   18     1   22   27 22                1   15   15
##   7    1   10     1    6    1     0   11    5 11                0    4   10
##   8    0    0     0    0    0     0    0    0  2                0    0    0
##    
##     11.3 11.4 11.5 11.6 11.7 11.8 11.9 11.95 12 12.1 12.2 12.3 12.4 12.5 12.6
##   3    0    0    0    0    0    0    0     0  0    0    0    0    0    0    0
##   4    1    2    2    0    0    1    0     0  1    0    0    0    0    0    0
##   5    5    9    3    1    1    2    1     1  1    1    1    0    0    0    1
##   6   18   18   19    8    9   15   14     0 10    4    7    5    7   11    2
##   7    7    2    6    6   11   10    5     0  9    8    4    7    6    9    2
##   8    1    1    0    0    2    1    0     0  0    0    0    0    0    1    1
##    
##     12.7 12.8 12.9 13 13.1 13.2 13.3 13.4 13.5 13.5666666666667 13.6 14 14.9
##   3    0    0    0  0    0    0    0    0    0                0    0  0    0
##   4    0    0    0  0    1    0    0    0    0                0    0  0    0
##   5    0    1    2  4    0    0    0    0    0                0    0  0    1
##   6    6    8    3  0    0    1    2    1    1                0    1  4    0
##   7    2    7    3  2    0    0    1    1    0                1    3  1    0
##   8    1    1    1  0    1    0    0    1    0                0    0  2    0
```

(c) Save the first twenty records of the `winequalRed` data set as a data frame with name `winequalRed20Rec` and show summary of it. (5 points)

```
winequalRed20Rec<-as.data.frame(winequalRed[1:20,])
head(winequalRed20Rec)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

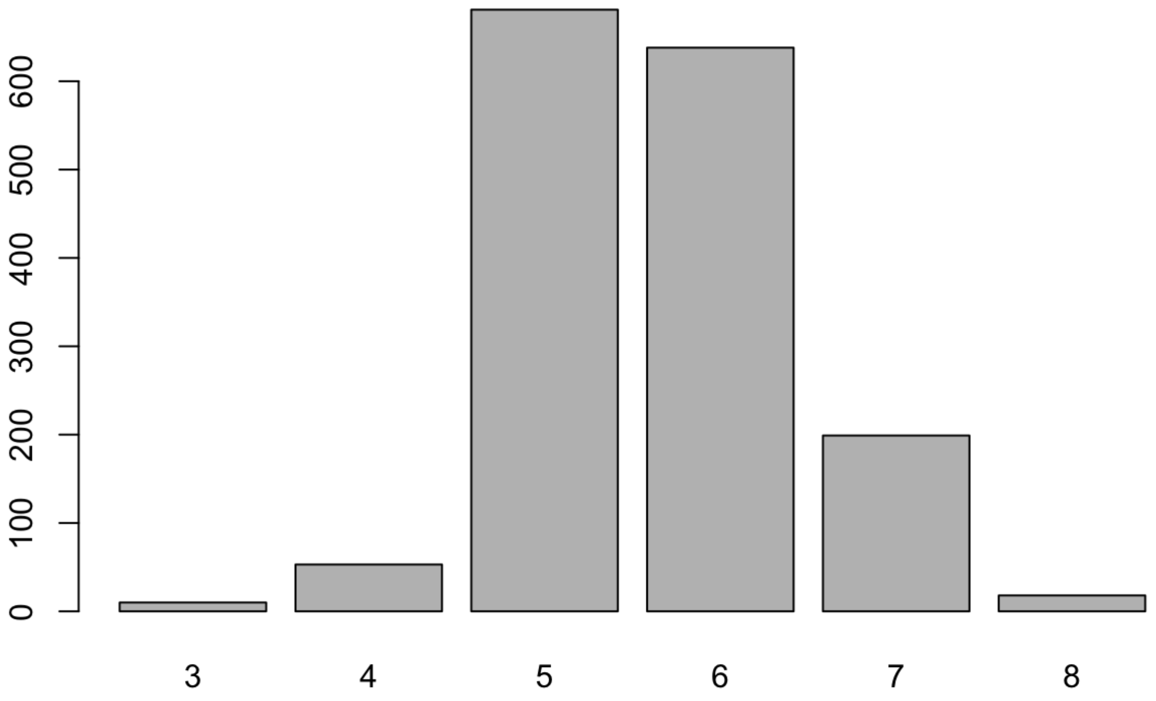
(d) Save the `density` and `pH` records of the `winequalRed` data set as as a data frame with name `winequalRedDePH` and show summary of it. (5 points)

```
winequalRedDePH <-as.data.frame(winequalRed[,8:9])
head(winequalRedDePH)
```

```
##   density   pH
## 1  0.9978 3.51
## 2  0.9968 3.20
## 3  0.9970 3.26
## 4  0.9980 3.16
## 5  0.9978 3.51
## 6  0.9978 3.51
```

(e) Use `barplot` to investigate `quality` attribute. Show your result. (5 points)

```
barplot(table(winequalRed$quality))
```



> Most have a quality of 5 & 6

(f) Create a new dataframe of the wine data set with a new column `highlowqual` with `low` when `quality` ≤ 5 and `high` when `quality` > 5. Find the mean and standard deviation for the attributes `alcohol` for the two classes. Based on the statistical information, describe if there exists difference for `alcohol` between the low quality and high quality red wines. (15 points)

```
winequalRedWithColumn <- transform(winequalRed, highlowqual=ifelse(quality > 5, 'high',
'low'))
head(winequalRedWithColumn)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality highlowqual
## 1       5         low
## 2       5         low
## 3       5         low
## 4       6        high
## 5       5         low
## 6       5         low
```

```
mean(winequalRedWithColumn[winequalRedWithColumn$highlowqual == 'low', 'alcohol'])
```

```
## [1] 9.926478
```

```
sd(winequalRedWithColumn[winequalRedWithColumn$highlowqual == 'low', 'alcohol'])
```

```
## [1] 0.7580065
```

```
mean(winequalRedWithColumn[winequalRedWithColumn$highlowqual == 'high', 'alcohol'])
```

```
## [1] 10.85503
```

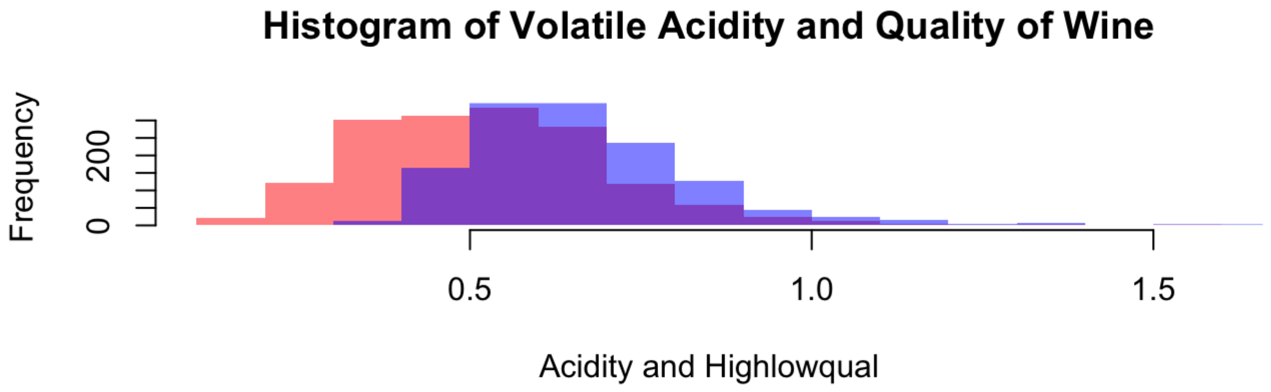
```
sd(winequalRedWithColumn[winequalRedWithColumn$highlowqual == 'high', 'alcohol'])
```

```
## [1] 1.106109
```

> The mean for low quality wine is 9.926478 and the standard deviation for low quality wine is 0.7580065. The mean for high quality wine is 10.85503 and the standard deviation for high quality wine is 1.106109

(g) Select any numeric attribute and show an overlay histogram of it with `highlowqual`. What conclusion can you draw from the plot? (10 points)

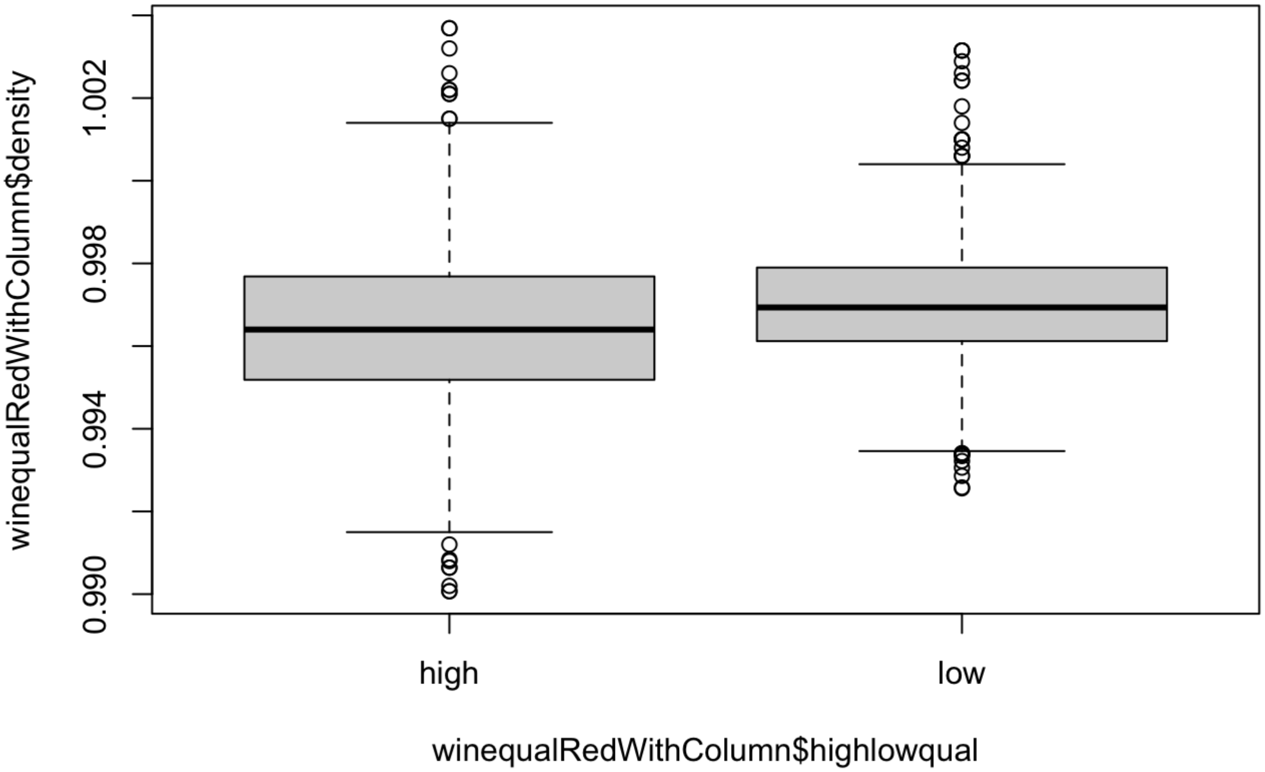
```
layout(1:2)
hist(winequalRedWithColumn[,2], main="Histogram of Volatile Acidity and Quality of Wine",
     xlab="Acidity and Highlowqual", col=rgb(1,0,0,.5), border=NA)
hist(winequalRedWithColumn[,10],col=rgb(0,0,1,.5), border=NA, add=TRUE)
```



> We can see the distribution of acidic levels

(h) Select any numeric attribute and show an overlay boxplot of it with a `highlowqual`. What conclusion can you draw from the plot? (10 points)

```
boxplot(winequalRedWithColumn$density~winequalRedWithColumn$highlowqual)
```



> High quality wine covers a higher range of density vs low quuality wine that has a more narrow range of density