Section Assignment 5

Liam Mueller

1/29/2022

Please Read!

This week, your assignment is to work with the tidyverse package to rearrange and manipulate data tables.

Since we have already downloaded and installed the tidyverse, this week, simply load the tidyverse functions into your working library:

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

This line of code above will need to be run each time you start a new R session.

One of the traits of a great programmer is their ability to solve a problem they haven't seen before. One of the best ways to solve a problem you have not seen before is to see if anyone else has. Great programmers are experts of web searching. Do not be afraid to find code on an online help page and run it!

Copying from the internet is one of the foundations of learning how to program, but it only works as learning if you reflect on why the code you used works. For that reason, this week and in future weeks, you will need to annotate your code or answer a specific question about the process. Use the lines that look like this at the end of each question to input your explanation of the code:

```
note<-'your note here'
print(note)
```

You will be graded on not just your code, but your explanation. Remember, in this class, it is okay to copy code, but you still need to demonstrate independent thought.

Questions: This week we will be working with plant chemistry data. The data set is from 382 unique plants that were sampled for both their saponin and tannin concentration. We will be coming back to this data set again in future weeks, so the the work you put in this week will save you time a few weeks down the road!

Question 1. (1 point) Read in the associated .csv file for this assignment "PlantChemistry.csv".

```
ChemData<-read.csv(file="PlantChemistry.csv")
str(ChemData)
```

```
## 'data.frame':    382 obs. of  3 variables:
##  $ Plant_ID: chr  "Plant_Sample_1" "Plant_Sample_2" "Plant_Sample_3" "Plant_Sample_4" ...
##  $ saponins: num  0.2072 0.1818 0.1231 0.0461 0.0632 ...
##  $ tannins : num  8.64 8.58 6.78 8.2 8.58 6.73 8.84 9.28 6.88 8.53 ...
```

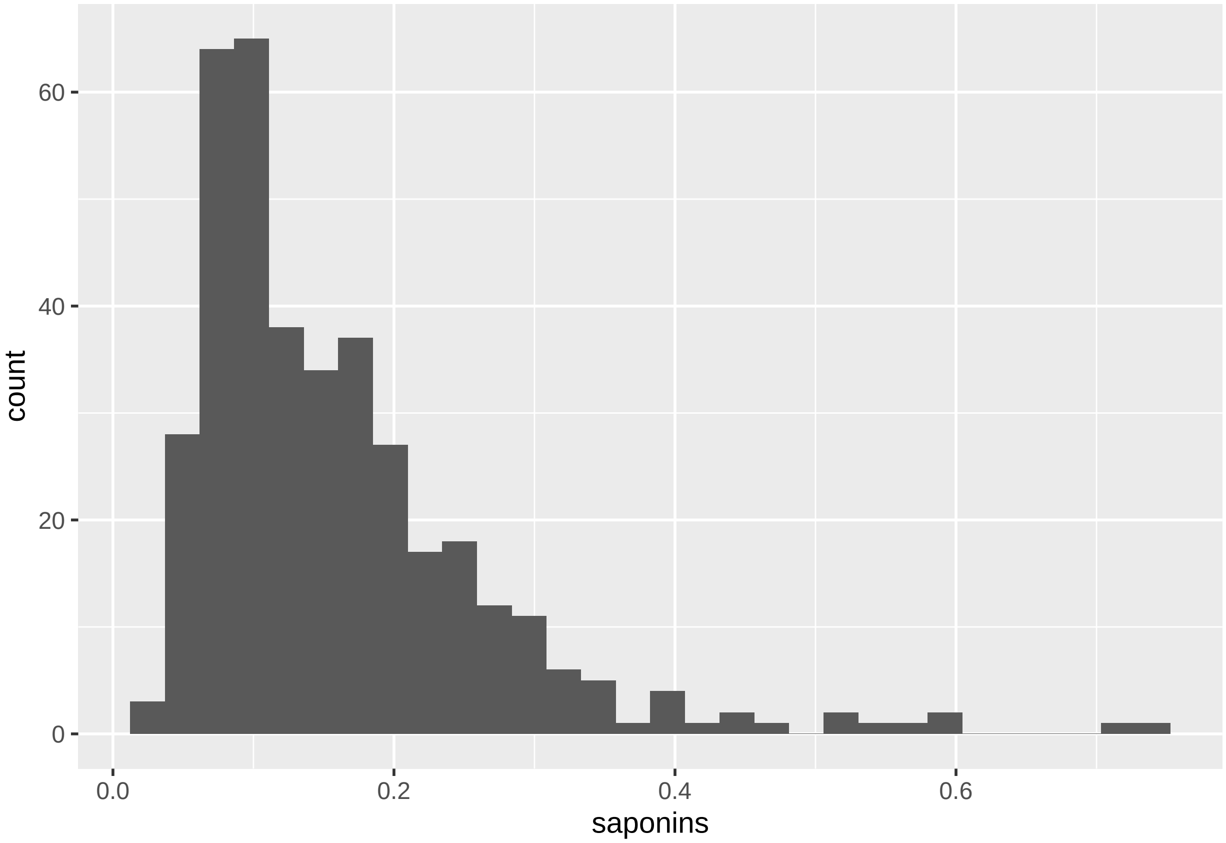
What does the `str()` function tell you about the shape of the object ChemData? Answer in the space provided below. [1] "Test, Test,Test"

Question 2.(2 point): One of the two numeric data columns in "ChemData" is not normally distributed. Use ggplot to build a histogram to visualize the distribution of both the "saponins" and "tannins" columns. Hint: In week 3 you used the `geom_bar()` or `geom_col()` function to build your plot. This week, try `geom_histogram()`.

```
Hist_saponins<-ggplot(data = ChemData,mapping = aes(x = saponins))+
  geom_histogram()

Hist_saponins
```

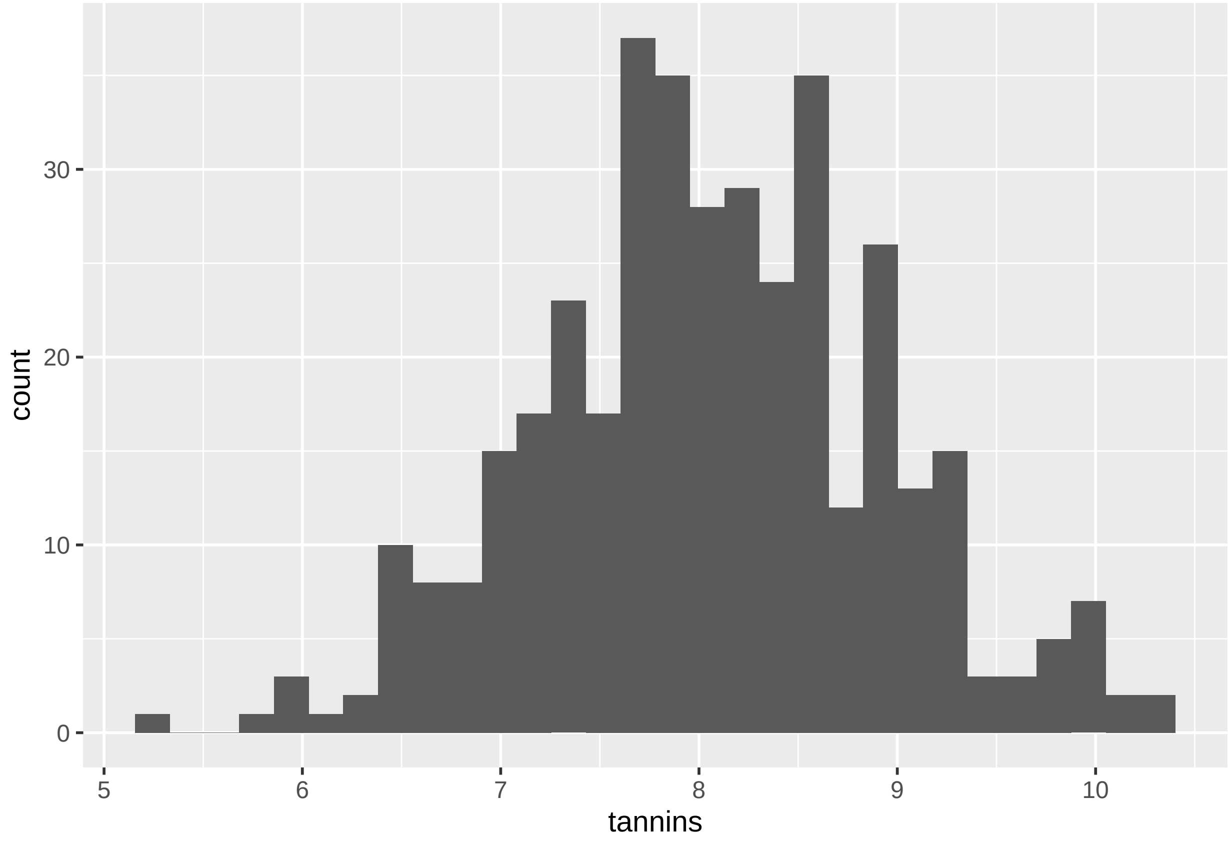
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Hist_tannins<-Hist_saponins<-ggplot(data = ChemData,mapping = aes(x = tannins))+
  geom_histogram()

Hist_tannins
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Which of the two columns of data seem more non-normal to you based on the visual test of examining the histograms? [1] ""

Question 3. (2 points): While the "eye test" is good, lets use a statistical test to determine the probability that our data are random variables drawn from a normal distribution. While it may not be the best test for normality, perform the Kolmogorov–Smirnov test on both the saponins and tannins data. Hint: `y="pnorm"`.

```
#saponins
ks.test(x = ChemData$saponins,y = "pnorm",mean(ChemData$saponins),sd(ChemData$saponins))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ChemData$saponins
## D = 0.12617, p-value = 1.045e-05
## alternative hypothesis: two-sided
```

Interpret the results for the ks.test you performed on the saponins data. Your answer should include what the null and alternate hypothesis for the test are and whether you choose to reject the null hypothesis or not. [1] ""

```
#tannins
ks.test(x = ChemData$tannins,y = "pnorm",mean(ChemData$tannins),sd(ChemData$tannins))
```

```
## Warning in ks.test(x = ChemData$tannins, y = "pnorm", mean(ChemData$tannins), :
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ChemData$tannins
## D = 0.02698, p-value = 0.9437
## alternative hypothesis: two-sided
```

Interpret the results for the ks.test you performed on the tannins data. Your answer should include what the null and alternate hypothesis for the test are and whether you choose to reject the null hypothesis or not. [1] ""

Question 4 (1 point): By now, you have identified that one of the columns of data in ChemData is not normally distributed. Using the tidyverse, we can easily create a new column in our data set that contains log transformed data of the column that we identified is non-normal. Hint: Check out the “Manipulate Variables” column on the dplyr cheat-sheet from last week, specifically the `mutate()` function.

```r
ChemData_with_log<-mutate(ChemData,logsaps=log(saponins))
str(ChemData_with_log)
```

```
## 'data.frame':    382 obs. of  4 variables:
##  $ Plant_ID: chr  "Plant_Sample_1" "Plant_Sample_2" "Plant_Sample_3" "Plant_Sample_4" ...
##  $ saponins: num  0.2072 0.1818 0.1231 0.0461 0.0632 ...
##  $ tannins : num  8.64 8.58 6.78 8.2 8.58 6.73 8.84 9.28 6.88 8.53 ...
##  $ logsaps : num  -1.57 -1.7 -2.09 -3.08 -2.76 ...
```

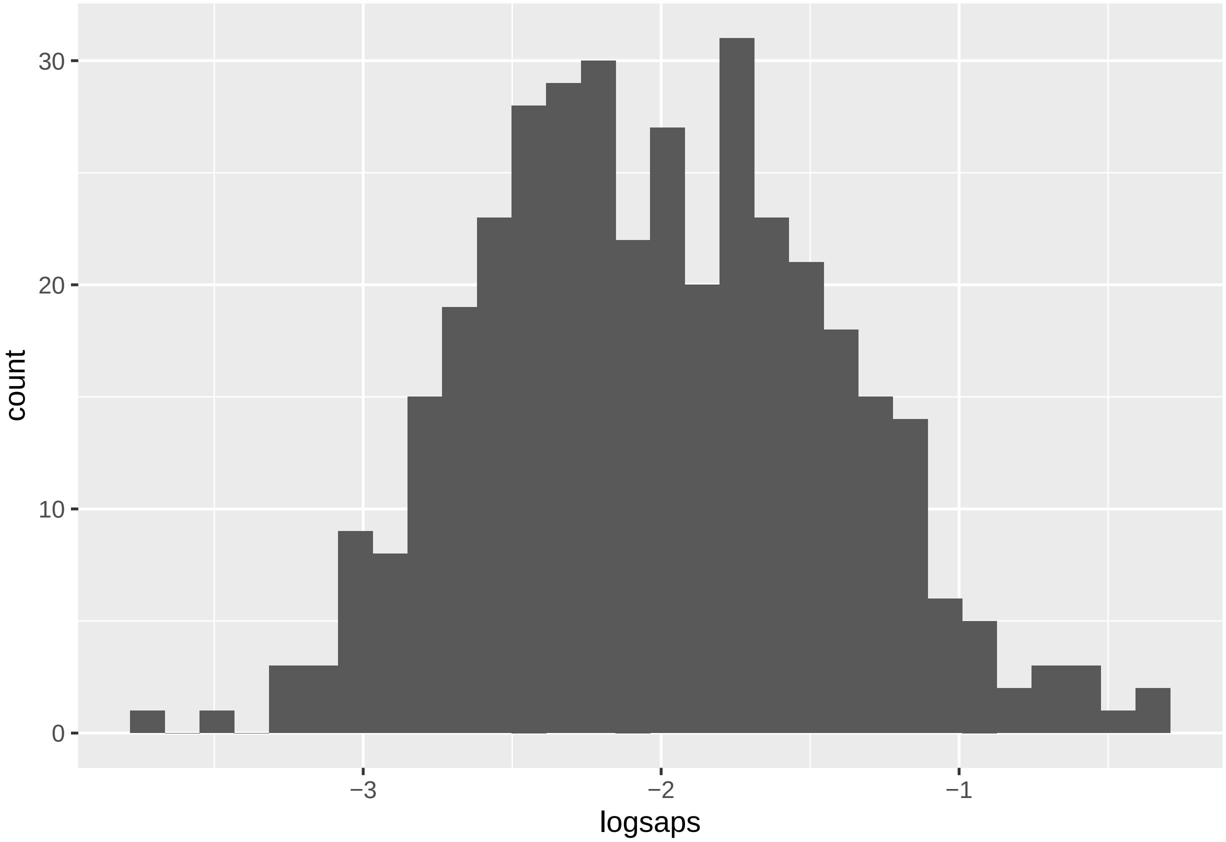
Add your notes for Q4 below, explain which column you chose to mutate and what the mutate function has done: [1] ""

Question 5 (1 point): Lets test and see if the data transformation has worked. In this question we will use a histogram to determine if our transformed data meet the assumptions of a normal distribution.

```
#Histogram of the transformed data:
Hist_Transformed<-ggplot(data =ChemData_with_log,mapping = aes(x = logsaps)) +
  geom_histogram()

Hist_Transformed
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



Do the transformed data appear normally distributed? Compare to the appropriate histogram you made in question 2. [1] ""

Question 6 (1 point): Lets test and see if the data transformation has worked. In this question we will use a Kolmogorov–Smirnov test to determine if our transformed data meet the assumptions of a normal distribution.

```
#Kolmogorov-Smirnov test on the transformed data
ks.test(x = ChemData_with_log$logsaps,y = "pnorm",mean(ChemData_with_log$logsaps),sd(ChemData_with_log$l
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ChemData_with_log$logsaps
## D = 0.049186, p-value = 0.3138
## alternative hypothesis: two-sided
```

Interpret the results for the ks.test you performed on the transformed data. Your answer should include what the null and alternate hypothesis for the test are and whether you choose to reject the null hypothesis or not. [1] ""

Once you are done, click the “Knit” button above(It looks like a blue ball of yarn). Save the file with your name and the week number in the file name:

(for example: “Liam_Mueller_Section_5”).

Then upload the pdf file to canvas/gradescope under the Week 5 Section Assignment before the deadline.

And that is it for this week!