MTH 324, Lab #8, Fall 2022 Name _____

Instructions: Follow along with the tutorial portion of the lab. Replicate the code examples in R on your own, along with the demonstration. Then use those examples as a model to answer the questions/perform the tasks that follow. Copy and paste the results of your code to answer questions where directed. Submit your response file and the code used (both for the tutorial and part two). Your code file and your lab response file should each include your name inside.

ANOVA

We have seen some ANOVA output from R in class lectures, so in this lab we are going to provide some examples of running the ANOVAs ourselves. One of the main issues we are going to run into is if the data is not already in a spreadsheet, the data for ANOVA common in textbooks are not necessarily set up to easily calculate the test in R. You may find it easier to set the data up in Excel and then import it into R for analysis. If you've done ANOVA tests in Excel, we actually need to reverse the common Excel format.

Consider the data shown from Excel:

	A	B	C
2			
3	Flavor 1	Flavor 2	Flavor 3
4	13	12	7
5	17	8	19
6	19	6	15
7	11	16	14
8	20	12	10
9	15	14	16
10	18	10	18
11	9	18	11
12	12	4	14
13	16	11	11

In R, what we will want is two columns of data. One column that indicates the flavor factor variable, and one column that indicates the measurement.

A sample of the data in correct format is shown here.

	A	B	C
1	Flavor	Rating	
2	1	13	
3	2	12	
4	3	7	
5	1	17	
6	2	8	
7	3	19	
8	1	19	
9	2	6	
10	3	15	
11	1	11	
12	2	16	
13	3	14	

The data we will work with for our examples will be the mtcars data (we will replicate many of the examples we did in class, or variations of them), so it is already in the appropriate dataframe format. In a two-way ANOVA, there will be two columns of factor variables, etc. We will encounter problems in the wrong format particularly in the homework (and maybe in the examples for this lab) because the display in a compact table is better for a textbook, and easier to work with if calculated by hand. In the world of real data, the dataframe format will be more common.

The ANOVA function in R will not work if both variables are numeric, one must be treated as a factor. Sometimes when we import data, this will happen when we don't want it to, but we can make a variables in a dataframe into a factor (string) variable by using the function factor() or as.factor(). If the data is already in the form of a string then this step should not be necessary. We encountered a similar issue when we created comparative boxplots.

To run the ANOVA, we use the function aov(). We need to specify which numerical variable is being modeled by which factor variable. We can do this by putting a ~ between the column names, and specifying the dataframe. Then we use the summary() function to print the results.

```
3  data(mtcars)
4  mtcars$cyl <- factor(mtcars$cyl)
5  one_way <- aov(mpg~cyl,data=mtcars)
6  summary(one_way)
```

The variable being modeled goes in front of the ~ and the factor variable after it. Recall that our hypothesis test we are conducting is that all the means are the same in the null hypothesis, and the alternative is that one mean is different.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
cyl           2  824.8   412.4    39.7 4.98e-09 ***
Residuals    29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our P-value is much less than the standard significance level of 0.05, and so we know that at least one mean is different.

To determine which one is different, or if they are all different, then we need to apply Tukey's method to obtain confidence intervals for each pair. (Alternatively, you could run pairs of two-sample t-tests, but we would need to separate our data out into multiple vectors because the t-test function takes data in a different format.)

We can look at a comparative boxplot to see what that suggests.

```
8  library(ggplot2)
9  ggplot(data=mtcars,aes(x=mpg,y=cyl))+geom_boxplot()+
10   labs(title="Boxplot of MPG by Cylinders",xlab="Miles Per Gallon",ylab="Number of Cylinders")
11
```

We can apply the TukeyHSD() function to obtain the confidence intervals. If the confidence intervals for pairwise comparisons. If the interval contains 0, the factors can be grouped together. If they do not contain zero, then they are grouped separately.

```
11
12  library(multcompView)
13  tukey_one <- TukeyHSD(one_way)
14
```

You will need to download the package if you have not already.

If you print the results of our tukey_one model, it will print the endpoints of the confidence intervals for each interval.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ cyl, data = mtcars)

$cyl
          diff        lwr        upr     p adj
6-4  -6.920779 -10.769350 -3.0722086 0.0003424
8-4 -11.563636 -14.770779 -8.3564942 0.0000000
8-6  -4.642857  -8.327583 -0.9581313 0.0112287
```

We can visualize the intervals using the plot function.

```
15  plot(tukey_one , las=1 , col="brown")
16
```

As we view the plots, keep the proper interpretation of the intervals in mind. It's not the overlap of the intervals we care about, but the inclusion of zero.

Let's look at our two-way ANOVA model. To additional variables for main effects only, in our model statement include a + sign between the factor variables. This variable, too, will have to be converted to a factor. Let's add the number of gears to our model and see what happens.

```
21  mtcars$gear <- factor(mtcars$gear)
22  two_way <- aov(mpg~cyl+gear,data=mtcars)
23  summary(two_way)
24
```

Do the number of gears also contribute to our gas mileage model?

```
             Df Sum Sq Mean Sq F value   Pr(>F)
cyl           2  824.8   412.4   38.00 1.41e-08 ***
gear          2    8.3     4.1    0.38    0.687
Residuals    27  293.0    10.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we saw in lecture, gears does not contribute to the model when cylinders are also included.

These are the main effects only, if we want to include interaction terms, use a * sign instead of the + sign.

```
two_way <- aov(mpg~cyl*gear,data=mtcars)
summary(two_way)
```

Since the main effect of gear turned out not to be significant, it's not surprising that the interaction effect is also not significant.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
cyl           2  824.8   412.4  36.777 4.92e-08 ***
gear          2    8.3     4.1   0.368    0.696
cyl:gear      3   23.9     8.0   0.710    0.555
Residuals    24  269.1    11.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Three-way ANOVA or N-way ANOVA continues in this same vein. Let's see if we can find a combination of variables that is significant.

```
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am)
three_way <- aov(mpg~gear+am+carb,data=mtcars)
summary(three_way)
```

Let's look at our summary results.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
gear          2  483.2  241.62  27.294 8.46e-07 ***
am            1   72.8   72.80   8.224 0.008699 **
carb          5  366.4   73.28   8.278 0.000136 ***
Residuals    23  203.6    8.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, all the main effects are significant. We could add interaction terms, but their presence in the model will depend on the data that we have. If we don't have the right kind of combinations of data, we won't be able to detect the interactions. We have only 32 observations total here, it turns out we are missing all the possible combinations to test all 4 interaction types.

As with the one-way, we can apply Tukey's method.

```
tukey_three <- TukeyHSD(three_way)

plot(tukey_three , las=1 , col="brown")
```

Based on the confidence interval data.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ gear + am + carb, data = mtcars)

$gear
         diff       lwr        upr     p adj
4-3  8.426667  5.540837 11.3124968 0.0000006
5-3  5.273333  1.425560  9.1211068 0.0061917
5-4 -3.153333 -7.119527  0.8128608 0.1370163

$am
        diff        lwr      upr     p adj
1-0 1.805128 -0.4102485 4.020505 0.1053983


$carb
          diff        lwr         upr     p adj
2-1 -1.209381  -5.759160  3.34039765 0.9598072
3-1 -3.232381  -9.603345  3.13858341 0.6224240
4-1 -7.292048 -11.841826 -2.74226902 0.0006325
6-1 -5.105714 -14.975570  4.76414126 0.6034924
8-1 -9.805714 -19.675570  0.06414126 0.0521887
3-2 -2.023000  -8.100513  4.05451265 0.9019076
4-2 -6.082667 -10.211523 -1.95380987 0.0016711
6-2 -3.896333 -13.579361  5.78669417 0.8086052
8-2 -8.596333 -18.279361  1.08669417 0.1021235
4-3 -4.059667 -10.137179  2.01784599 0.3351497
6-3 -1.873333 -12.533996  8.78732909 0.9935329
8-3 -6.573333 -17.233996  4.08732909 0.4195936
6-4  2.186333  -7.496694 11.86936084 0.9799919
8-4 -2.513667 -12.196694  7.16936084 0.9635910
8-6 -4.700000 -17.756592  8.35659162 0.8693595
```

Since there are three factor variables, there are three different plots output. You'll have to scroll back through the plots to see the first two. If we look at just the carb graph, we can see that some are grouped together and some separately because some of the intervals include 0 and some don't. Another method of visualizing the relationships is to list the factor levels for a variable in order of their means, and then use the intervals to underline those that group together. An example from the Devore text is shown.

75%	87%	62%	50%	37%	25%
5.37	5.44	5.98	6.59	6.77	7.68

To interpret this, we'd say that the levels 75%, 87% and 62% group together (behave similarly), 62% and 50% are similar and group together, 50% and 37% also form a group. While 25% is not like any of the other levels and is in a group by itself.

For levels that group together, a two-sample t-test or ANOVA on just those groups should fail to reject the null hypothesis for anything in the group. If you choose two levels that are not grouped together, the test should be able to reject the null hypothesis. It's a good idea to look at a boxplot broken out by levels to confirm these results intuitively.

Tasks

1. Use the mtcars data set and run an ANOVA test on mpg and the vs variable. Clearly state your hypothesis test. There are only two levels, so break out your variables into vectors by level and run a two-sample t-test. Do the results of the two methods agree? Plot the Tukey interval, and boxplots. Do the visuals appear to agree with the analysis? Paste all graphs and summary tables here.

2. Use the data at the top of the lab on flavor ratings. Finish converting the data to a dataframe and then conduct the one-way ANOVA. If the ANOVA test rejects the null, create your Tukey intervals, and a boxplot. Check your normality assumptions with a normal probability plot. Paste your analyses and graphs below.

3. Import the data from the lab 8 data file. There are three factor variables and two numerical variables. Conduct a two-way ANOVA of either of the numerical variables, using any two of the factor variables. Test for an interaction. Explain the results of the test. Support your analysis with Tukey's method and boxplots. Paste all graphs and analysis here.

4. Conduct a three-way ANOVA on the same numerical variable as above. Describe the results. Include supporting documentation here. Check that the numerical variable you chose is approximately normally distributed.

References:

1. Discovering Statistics Using R. Andy Field, Jeremy Miles, Zoe Field. (2012)
2. https://book.stat420.org/applied_statistics.pdf
3. https://scholarworks.montana.edu/xmlui/handle/1/2999
4. https://www.rstudio.com/resources/cheatsheets/
5. https://data-flair.training/blogs/hypothesis-testing-in-r/
6. https://www.scribbr.com/statistics/anova-in-r/
7. https://r-graph-gallery.com/84-tukey-test.html
8. https://www.real-statistics.com/one-way-analysis-of-variance-anova/basic-concepts-anova/
9. https://statisticsglobe.com/combine-two-vectors-into-data-frame-in-r