*We are in different sections, so each of us will submit one for our corresponding section

STA 106

Group members:

Aria Hamidi

Siddarth Vinnakota

Data Set: loseit.csv (QUESTION 2)

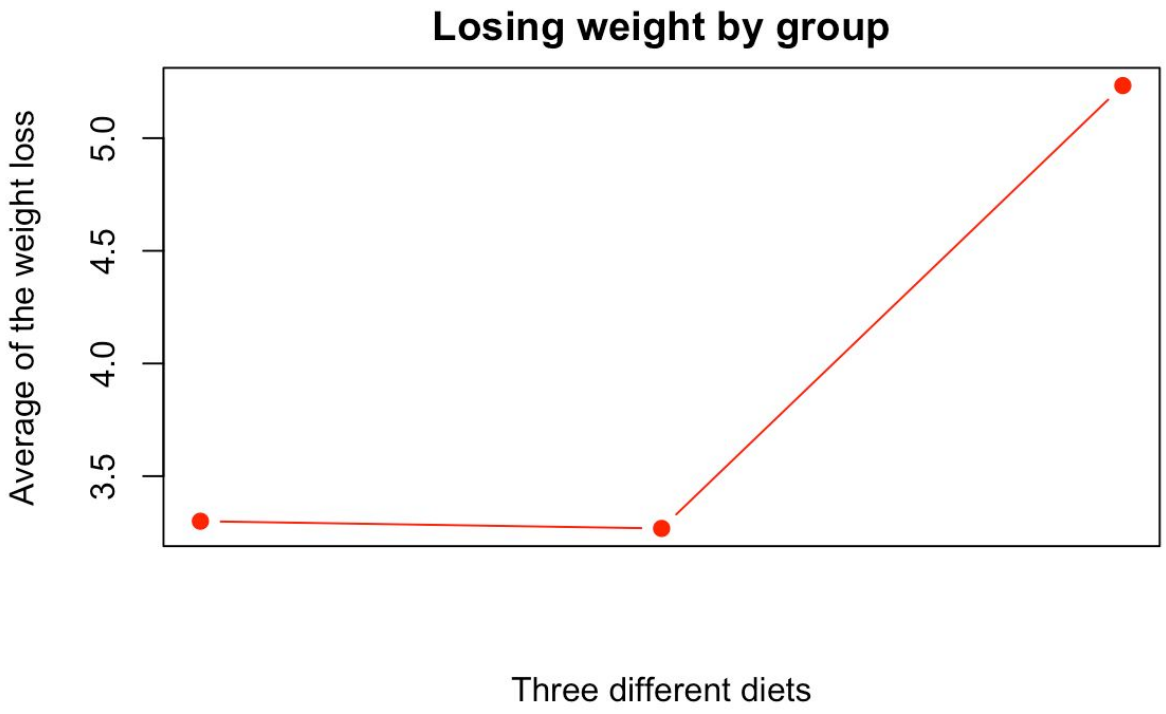
Exam 1 Project

I. Introduction

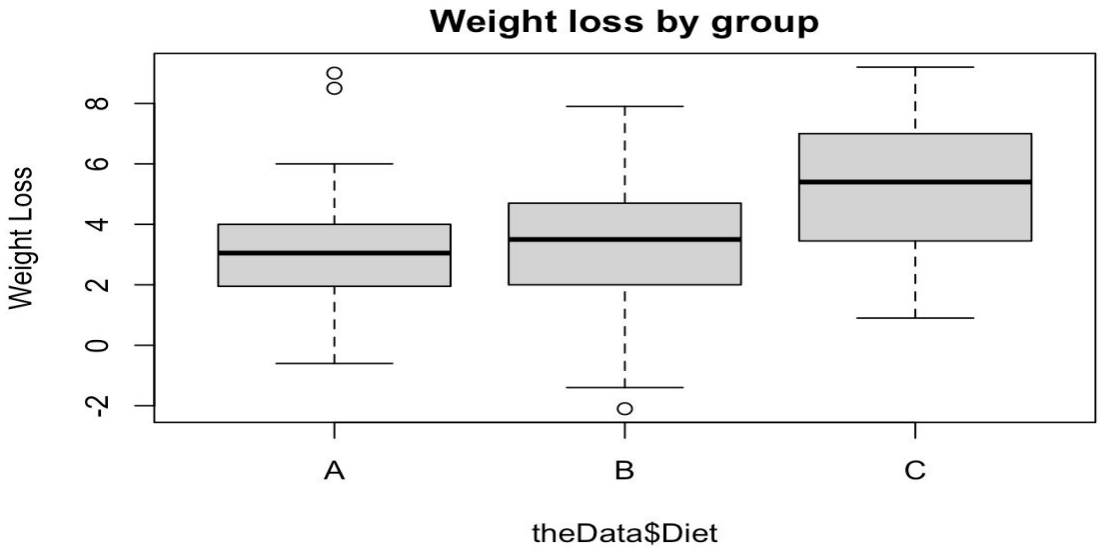
The question we are trying to answer for this data set is which of the diets, yields the best results in losing weight? We might be interested in this answer to find out the best diet plan to lose weight. We will be taking a Single Factor ANOVA approach in order to answer this question.

II. Summary

One of the first things that we did for this part was calculating the mean of each different diet. The sample mean for Diet A is 3.3000, for B is 3.2680, and for C it is 5.2333. From the means themselves, it looks like Diets A and B would share similar data, while C would be a more effective dieting method, as it is a larger mean. By using the sample means that we calculated, we created a scatter plot which potentially gives an idea how big or small the average weight loss of each group is and how different their means are from each other. Here's the scatter plot we created:



As the plot illustrates, the sample mean for diet A and B are close to each other, whereas the sample mean for diet C is larger than them. Furthermore, for better and more detailed data visualization, we decided to create a boxplot that showed us the outliers of each group of diet. Here's the boxplot:



As you can observe, there are a few outliers in diet A and diet B. Sketching the boxplot led us one step forward in the diagnostics; for example, by observing the plot, we already knew that there's no outliers for diet C; therefore, we didn't need to identify any limit values to cut off any data from diet C.

Another thing that we did to have an idea about the trend of the data was including the summary of each diet group:

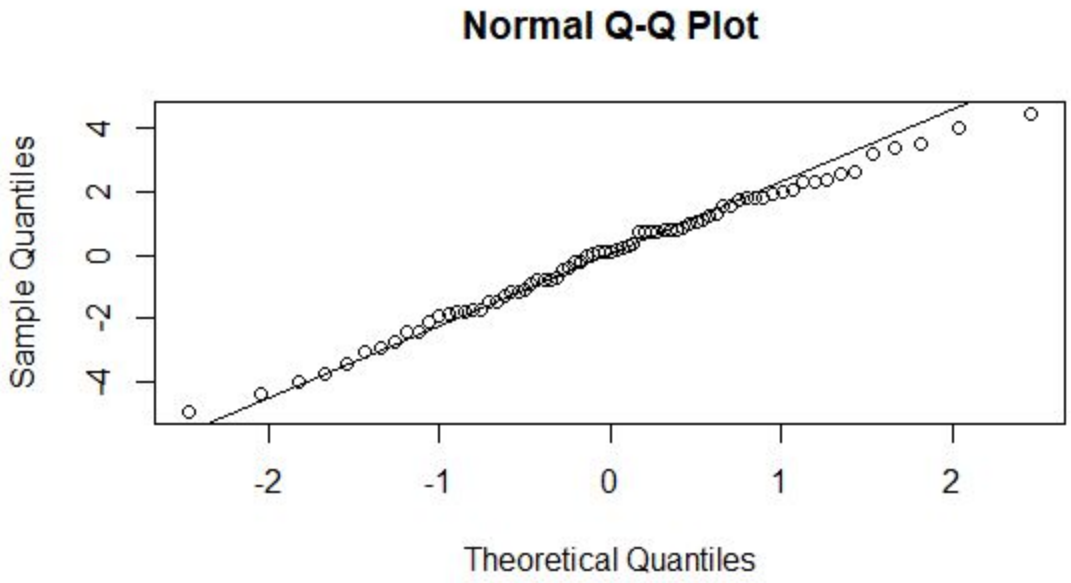
	A	B	C
Means	**3.3000**	**3.2680**	**5.2333**
Std. Dev	**2.2401**	**2.4645**	**2.2477**
Sample Size	**24.0000**	**25.0000**	**27.0000**

For the standard deviations, Diet A is 2.2401, for B it is 2.4645, and for C it is 2.2477. Despite B having a slightly higher standard deviation, one could argue that all of the Diets' standard deviations are similar enough. Plus, we can also observe the sample sizes which are relatively the same.

III. Diagnostics

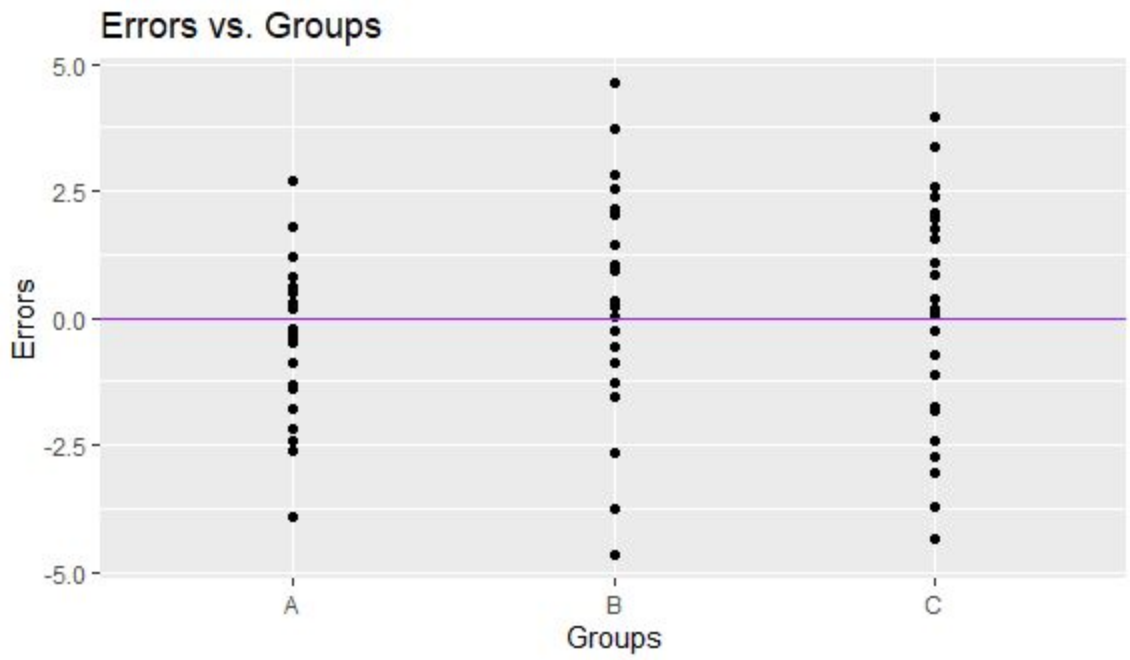
Making assumptions requires data visualization, so we used the summary as an evidence for our diagnostics. According to the box plot, there are a total of 3 outliers in the data set. There are 2 outliers in Diet A, which is 8.5 and 9, and 1 outlier in Diet B, which is -2.1. Looking at both the semi-studentized and studentized residuals yields us no additional outliers. We can remove those outliers for this data in order to perform a more accurate analysis.

Here is the QQ plot that represents the data:



In order to perform Single Factor ANOVA, several assumptions would have to be made. The first major assumption of ANOVA is that each sample has to be drawn independently from one another. Though this isn't 100% certain, we can conclude based on the nature of the study that each sample was independently picked. The second assumption is that each sample is taken from a normally distributed population. A method of finding this out is via the Shapiro-Wilkes Test. Using the Shapiro Wilkes test yields us a p-value of 0.909, and because it is significantly higher than any realistic significance level, we fail to reject the null hypothesis, and can conclude that each of the samples of the diet data has been taken from a normally distributed population.

The final assumption we have to make is that the variance of data per data group must be equal. We can find this out by using an error dot plot.



As shown in the graphs, there is an equal spread of the data, and therefore we can assume that the data has equal variance.

IV. Analysis

The null hypothesis of this study is that all the averages of all the diets are equal with one another. The alternate hypothesis is that at least one of the diets differs from the rest.

Using the F test, we get our test statistic of 9.2133. The corresponding p-value for this test statistic is 0.0002806. Using the power of the test at a 95% confidence interval, we get our power calculation to be approximately 0.9719.

The different confidence interval multipliers we get with a 95% confidence interval, we get a Tukey multiplier of 2.394561, a Scheffe multiplier of 2.50107, and a Bonferroni multiplier of 2.452872. We use the smallest multiplier, which is the Tukey multiplier. The Confidence Interval of Diet C - Diet A is (1.010697, 3.846878), the confidence interval between Diet C - Diet B is (0.3565169, 3.1268165), and the confidence interval between Diet B - Diet A is (-0.77702196, 2.1444620).

V. Interpretation

In the context of this data set, the null hypothesis represents that all the sample averages are the same, which says that these Diets are equally effective with one another. The alternate hypothesis suggests that since at least one of the means differs from the rest, which suggests that at least one of the diets is more effective than the others. We can find out which is more effective through looking at confidence intervals.

Since the p-value is significantly less than any reasonable significance level, we reject the null hypothesis and can conclude that there is enough statistical evidence that at least one of the means differ, and in other words, at least one of the diets was more or less effective than the others. Because of a high power test, we are very likely to be correct in our conclusion to reject the null hypothesis.

Looking at the confidence intervals, more specifically between the C-A difference and the C-B difference, we are 95% confident to see a difference that is going to be positive. That means that out of the three Diets, Diet C is the most effective in losing weight. Also, when we look at the confidence interval for the difference between B and A, most of the time Diet B would be more effective than Diet A, but it could also be possible that both diets share the same effectiveness.

VI. Conclusion

If we go back to the initiation of the project, the question was about based on the data given, what's the best diet plan to choose in order to lose more weight. From a first look in the summary of the values, it was looking like Diet C was the most effective diet to lose weight, but we had to ensure this using statistics. Based on our assumptions taken after removing certain outliers and performing diagnostics, we were able to conduct a test using Single Factor ANOVA. By conducting calculations in our analysis and also testing the null and alternative hypotheses, we're able to answer this question now. Plus we were able to use confidence intervals of these

pairs to determine approximately the level of difference each diet has with one another. In fact, the Single Factor ANOVA approach suggested that we have to use diet C as the best diet plan to lose weight, as the both Diets A and B are equal, but as a whole they are less effective than C.

VII. Appendix

```
#Importing the dataset


loseit = read.csv("~/Desktop/loseit.csv")
loseit = data.frame(Loss,Diet)


#SUMMARY


#Calculating the means of each type of diet
group.means = by(loseit$Loss,loseit$Diet,mean)
Group.means


##Scatter plot for showing the amount of weight loss in different
groups
plot(group.means,xaxt = "n",pch = 19,col = "red",xlab = "Three
different diets",ylab = "Average of the weight loss",main = "Losing
weight by group",type = "b")



#Visualizing the data by creating a boxplot
boxplot(loseit$Loss ~ loseit$Diet, main = "Weight loss by group",ylab
= "Weight Loss")
```

```
#Summarizing the data
group.means =  by(loseit$Loss,loseit$Diet,mean)
group.sds = by(loseit$Loss,loseit$Diet,sd)
group.nis = by(loseit$Loss,loseit$Diet,length)
the.summary = rbind(group.means,group.sds,group.nis)
the.summary = round(the.summary,digits = 4)
colnames(the.summary) = names(group.means)
rownames(the.summary) = c("Means","Std. Dev","Sample Size")
the.summary


#Calculating the five-number summary
fivenum(loseit$Loss)


#DIAGNOSTICS

#Identifying values for a cut off
CO3 = which((loseit$Diet=="A" & loseit$Loss > 7) | (loseit$Diet=="B"
& loseit$Loss < -2))
CO3

the.model = lm(Loss ~ Diet, data = loseit)
```

```
#semi-studentized model
loseit$ei = the.model$residuals
nt = nrow(loseit) #Calculates the total sample size
a = length(unique(loseit$Diet)) #Calculates the value of a
SSE = sum(loseit$ei^2) #Sums and squares the errors (finds SSE)
MSE = SSE/(nt-a) #Finds MSE
eij.star = the.model$residuals/sqrt(MSE)
alpha = 0.05
t.cutoff= qt(1-alpha/(2*nt), nt-a)
CO.eij = which(abs(eij.star) > t.cutoff)
CO.eij #No outliers via this method, so we disregard it

rij = rstandard(the.model) #studentized method
CO.rij = which(abs(rij) > t.cutoff)
CO.rij #No outliers via this method

outliers = CO3
outliers
new.data = loseit[-outliers,]
new.model = lm(Loss ~ Diet,data = new.data)
new.model #new data and model with the removed outliers

new.data

qqnorm(new.model$residuals) #QQ Plots
qqline(new.model$residuals)
```

```
ei = new.model$residuals
the.SWtest = shapiro.test(ei) #Tests for Normality
the.SWtest #Since our p-value was relatively large, we fail to reject
the null, and support that our data is normally distributed at any
reasonable significance level (1%, 5%, 10%).

library(ggplot2) #Tests for equal variance
qplot(Diet, ei, data = new.data) + ggtitle("Errors vs. Groups") +
xlab("Groups") + ylab("Errors") + geom_hline(yintercept = 0,col =
"purple") #Since our vertical spread is relatively equal with each
other, we can say that there is equal variance between them.

#ANALYSIS OF DATA
group.means = by(new.data$Loss,new.data$Diet,mean)
group.sds = by(new.data$Loss,new.data$Diet,sd)
group.nis = by(new.data$Loss,new.data$Diet,length)
the.summary = rbind(group.means,group.sds,group.nis)
the.summary = round(the.summary,digits = 4)
colnames(the.summary) = names(group.means)
rownames(the.summary) = c("Means","Std. Dev","Sample Size")
the.summary

anova.table = anova(new.model)
Anova.table #ANOVA Table
```

```
F.stat = anova.table[1,4]
F.stat

p.val = anova.table[1,5]
p.val #Because it is very less than any reasonable significance
level, we reject the null hypothesis

#Finding a power of the test
give.me.power = function(ybar,ni,MSE,alpha){
  a = length(ybar) # Finds a
  nt = sum(ni) #Finds the overall sample size
  overall.mean = sum(ni*ybar)/nt # Finds the overall mean
  phi = (1/sqrt(MSE))*sqrt( sum(ni*(ybar - overall.mean)^2)/a) #Finds
the books value of phi
  phi.star = a *phi^2 #Finds the value of phi we will use for R
  Fc = qf(1-alpha,a-1,nt-a) #The critical value of F, use in R's
function
  power = 1 - pf(Fc, a-1, nt-a, phi.star)# The power, calculated
using a non-central F
  return(power)
}

MSE = anova.table[2,3]

the.power = give.me.power(group.means,group.nis,MSE,0.05)
the.power
```

```
give.me.CI = function(ybar,ni,ci,MSE,multiplier){
  if(sum(ci) != 0 & sum(ci !=0 ) != 1){
    return("Error - you did not input a valid contrast")
  } else if(length(ci) != length(ni)){
    return("Error - not enough contrasts given")
  }
  else{
    estimate = sum(ybar*ci)
    SE = sqrt(MSE*sum(ci^2/ni))
    CI = estimate + c(-1,1)*multiplier*SE
    result = c(estimate,CI)
    names(result) = c("Estimate","Lower Bound","Upper Bound")
    return(result)
  }
}

nt = sum(group.nis)
a = length(group.means)
alpha = 0.05

Tuk = qtukey(1-alpha,a,nt-a)/sqrt(2)
Tuk #Since this is the smallest, we use this
S = sqrt((a-1)*qf(1-alpha, a-1, nt-a))
S
```

```
g=3
B = qt(1-alpha/(2*g),nt-a)
B

#Interval for C-A
c1 = c(-1,0,1)
give.me.CI(group.means,group.nis,c1,MSE,Tuk)

#Interval for C-B
c2 = c(0,-1,1)
give.me.CI(group.means,group.nis,c2,MSE,Tuk)

#Interval for B-A
c3 = c(-1,1,0)
give.me.CI(group.means,group.nis,c3,MSE,Tuk)
```