Problem Set 1, Winter 2021

Adam Ten Hoeve

1/15/2021

The data used for this problem set comes from Snedecor & Cochran (p. 218), which was derived from an experiment conducted by Lowe (1935). Lowe wanted to learn more about how much fat doughnuts absorb when cooked in different kinds of fat. He tested four kinds of fats (fat_type). He cooked six identical batches of doughnuts using each type of fat. Each batch contained 24 doughnuts, and the total fat (in grams) absorbed by the doughnuts in each batch was recorded (total_fat)

Loading the data:

```
doughnuts <- read.csv("doughnuts.csv",header=TRUE,sep=",") # Loads the CSV file into memory. You may ne
names(doughnuts) <- c("fat_type", "total_fat")
doughnuts$fat_type_factor <- as.factor(doughnuts$fat_type) # Creates a new variable and tells R that th
```

Question 1 - 10 points

Compute the mean and standard deviation (note: you have sample data, not population data) for each group, and then create a bar plot to visually display the means.

```
# Compute group means
doughnuts.summary <- doughnuts %>%
                     group_by(fat_type_factor) %>%
                     summarize(mean.tf=mean(total_fat), sd.tf=sd(total_fat), n=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

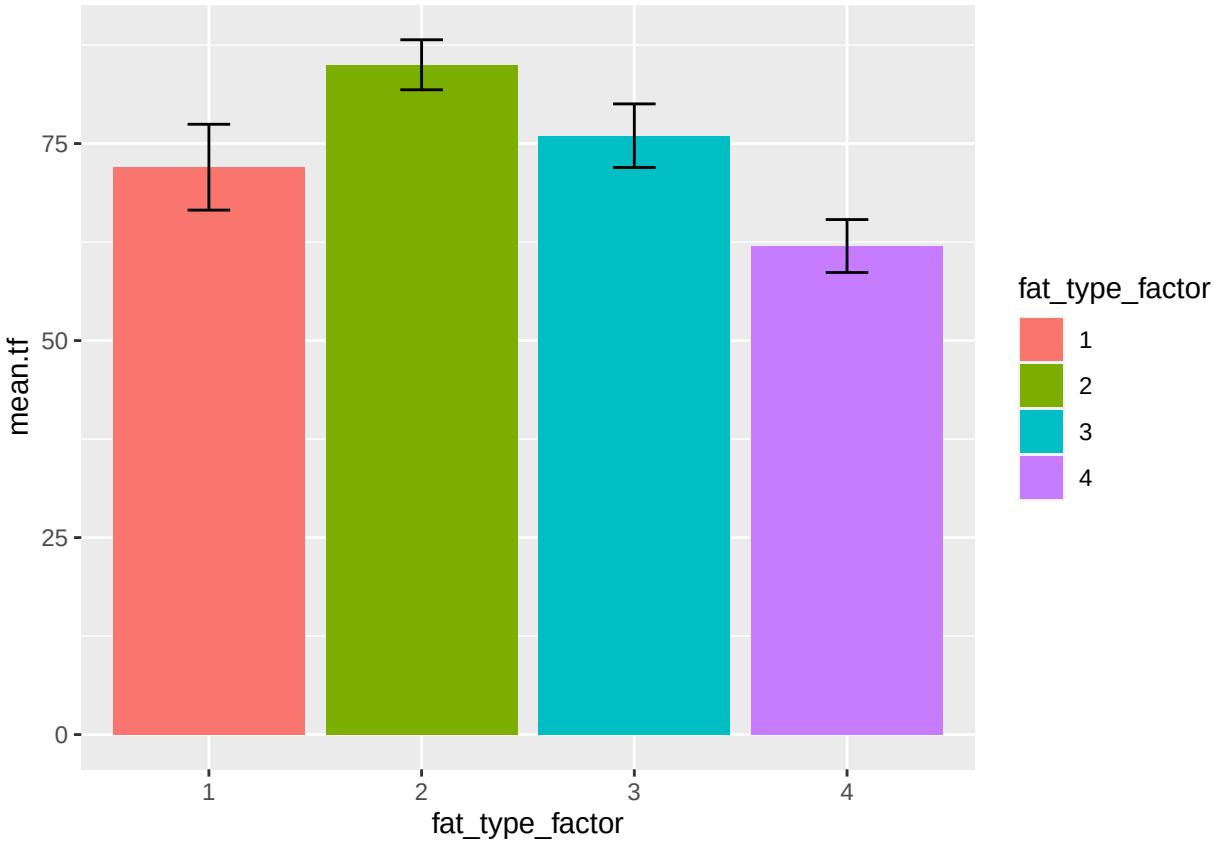
```
doughnuts.summary
```

```
## # A tibble: 4 x 4
##   fat_type_factor mean.tf sd.tf     n
##   <fct>             <dbl> <dbl> <int>
## 1 1                    72 13.3      6
## 2 2                    85  7.77     6
## 3 3                    76  9.88     6
## 4 4                    62  8.22     6
```

```
# Compute standard deviations for each group
# Above code calculates both mean and sd per group.


# Code for visualization
```

```
g<-ggplot(data=doughnuts.summary, aes(x=fat_type_factor, y=mean.tf, fill=fat_type_factor)) +
    geom_col()
g<- g + geom_errorbar(aes(ymin=mean.tf-sd.tf/sqrt(n), ymax=mean.tf+sd.tf/sqrt(n)), width=.2)
g
```



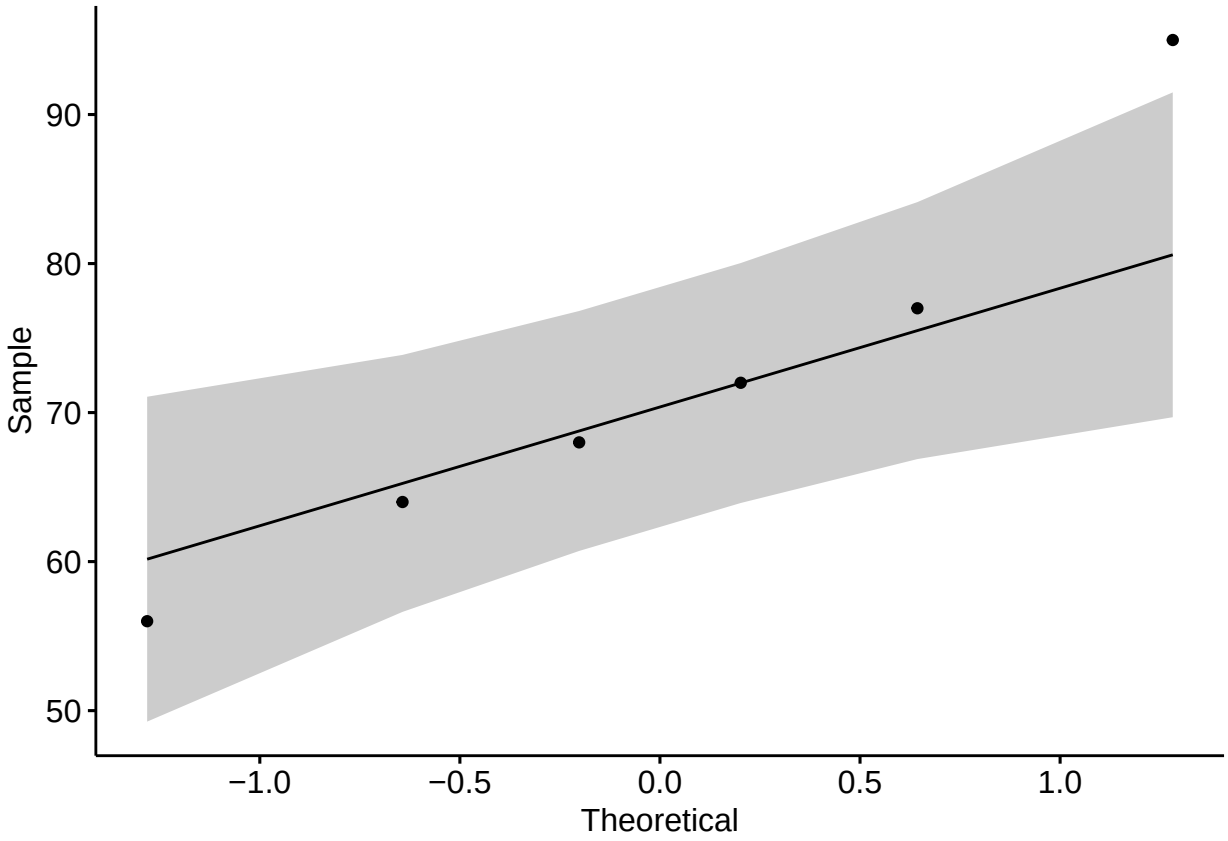
Question 2 - 10 points

Assess the assumption of normality visually and quantitatively and comment on how well the data met this assumption.

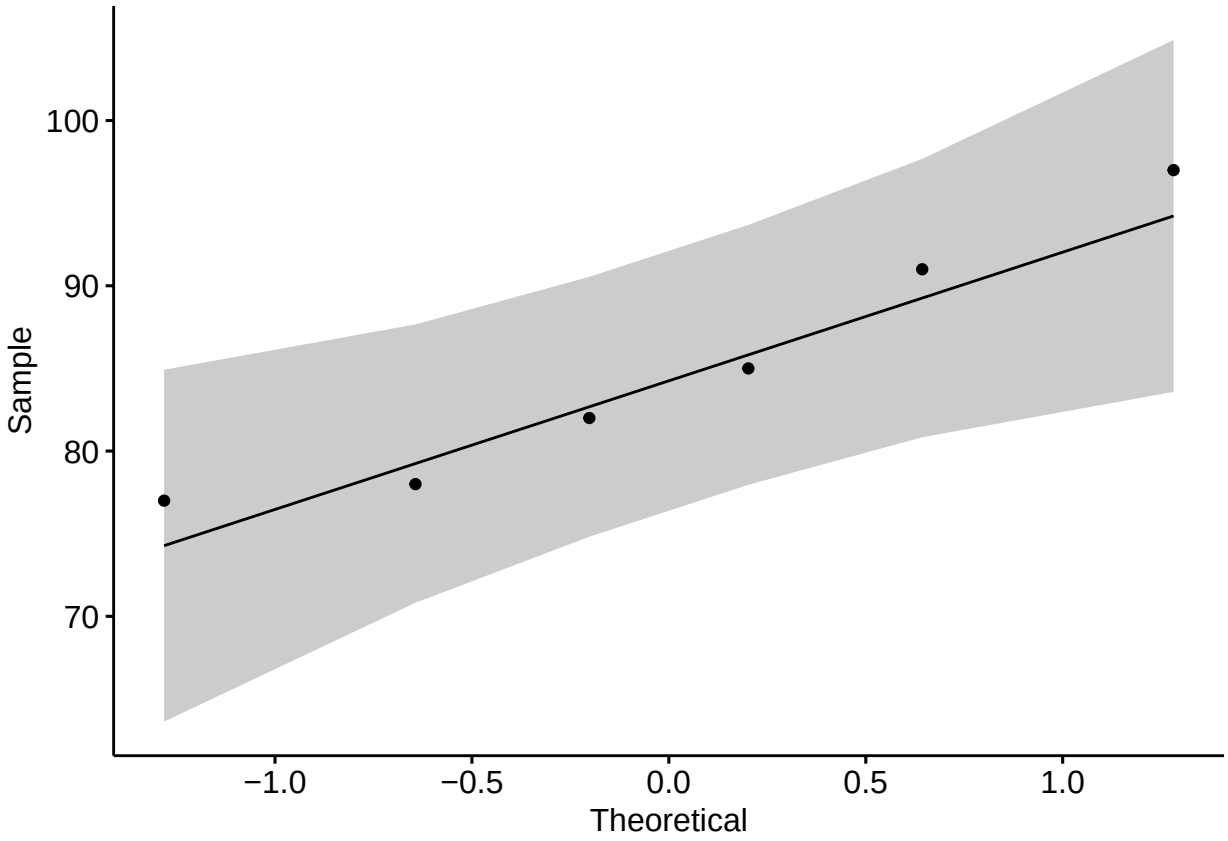
```
doughnuts[doughnuts$fat_type == 1, ]
```

```
##   fat_type total_fat fat_type_factor
## 1        1        64               1
## 2        1        72               1
## 3        1        68               1
## 4        1        77               1
## 5        1        56               1
## 6        1        95               1
```

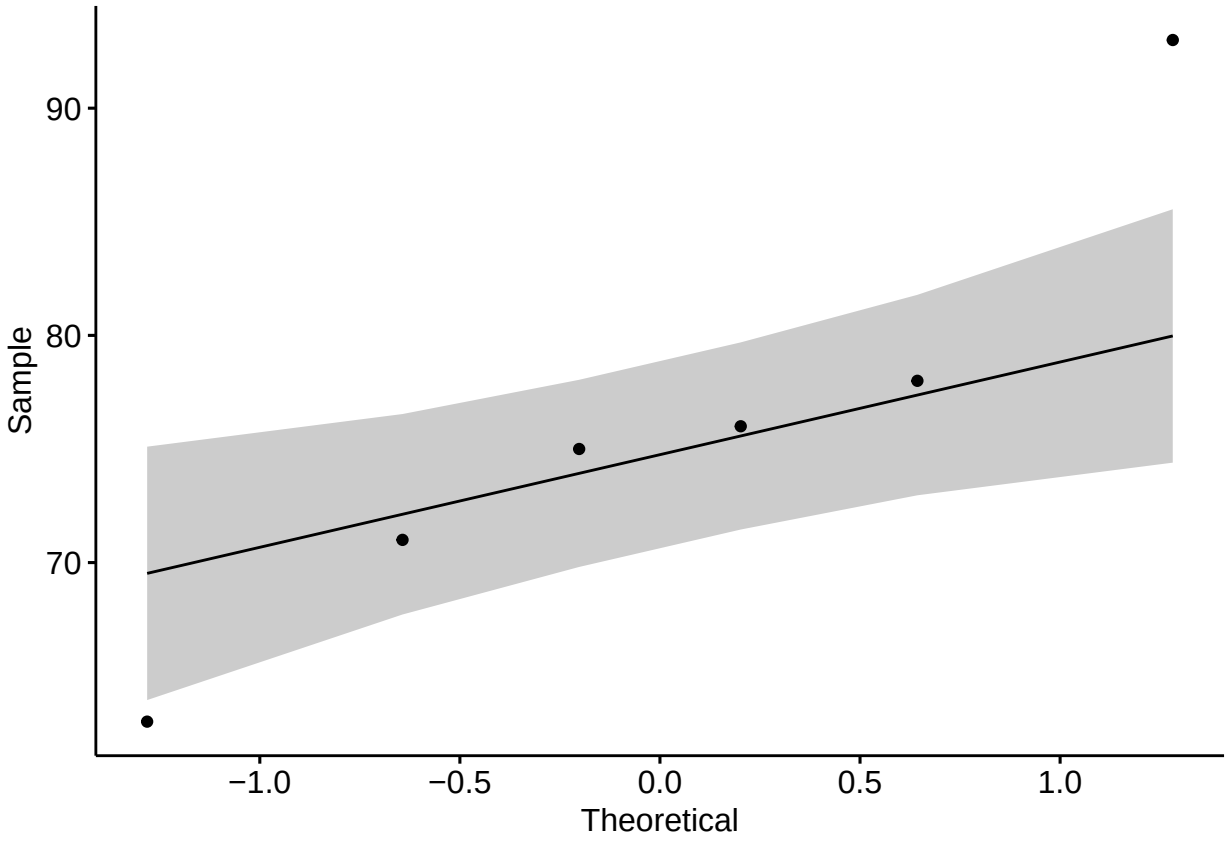
```
# Code for visual assessment
# Use a QQplot to visually assess the data.
ggqqplot(doughnuts[doughnuts$fat_type == 1, ]$total_fat)
```



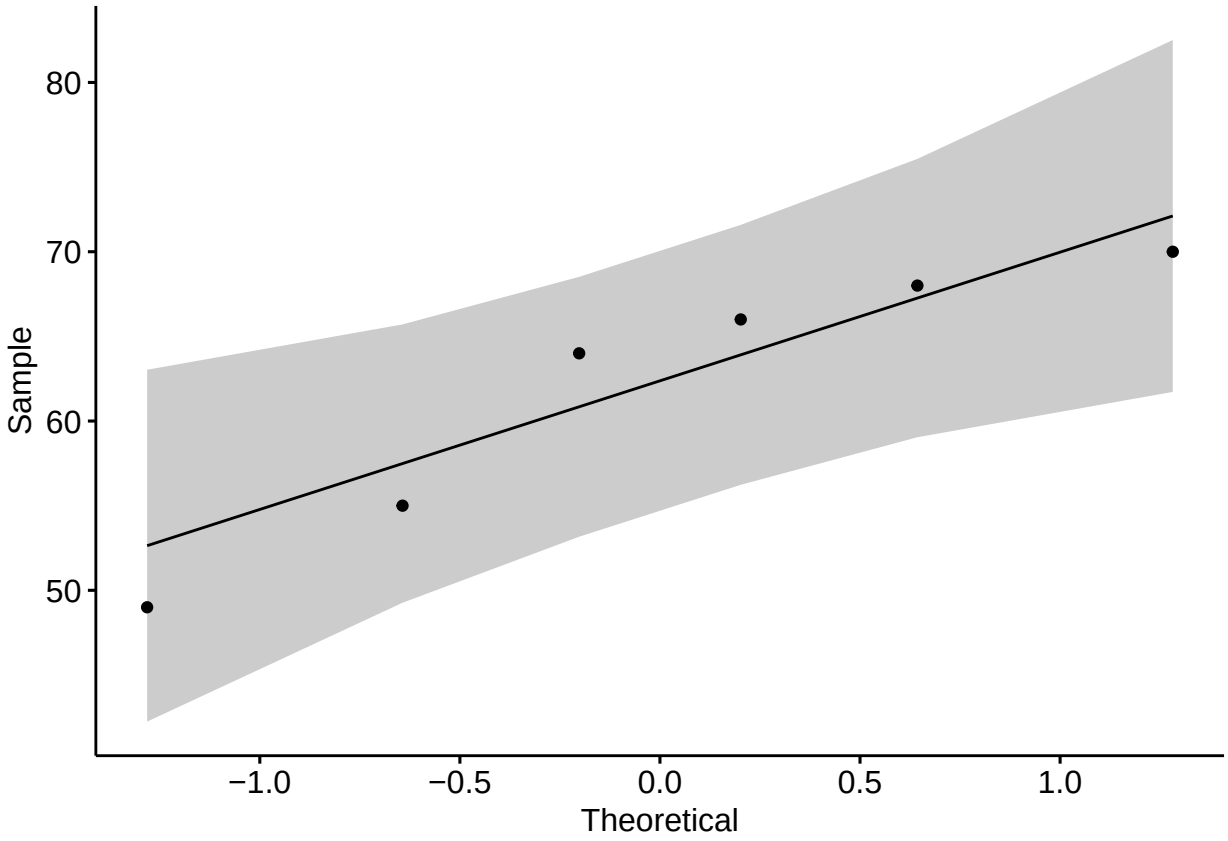

```
ggqqplot(doughnuts[doughnuts$fat_type == 2, ]$total_fat)
```




```
ggqqplot(doughnuts[doughnuts$fat_type == 3, ]$total_fat)
```




```
ggqqplot(doughnuts[doughnuts$fat_type == 4, ]$total_fat)
```




```
# Code for quantitative assessment
# We can use a Shapiro-Wilks Test to test the normality of our data.
# We test all of the groups for normality.
# All tests returns a large p-value, so all groups fail to reject the null.
shapiro.test(doughnuts[doughnuts$fat_type == 1, ]$total_fat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  doughnuts[doughnuts$fat_type == 1, ]$total_fat
## W = 0.95004, p-value = 0.7406
```

```
shapiro.test(doughnuts[doughnuts$fat_type == 2, ]$total_fat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  doughnuts[doughnuts$fat_type == 2, ]$total_fat
## W = 0.93162, p-value = 0.5926
```

```
shapiro.test(doughnuts[doughnuts$fat_type == 3, ]$total_fat)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  doughnuts[doughnuts$fat_type == 3, ]$total_fat
## W = 0.9334, p-value = 0.6066
```

```
shapiro.test(doughnuts[doughnuts$fat_type == 4, ]$total_fat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  doughnuts[doughnuts$fat_type == 4, ]$total_fat
## W = 0.88836, p-value = 0.3097
```

Write your comments about how well the data meet the normality assumption below:

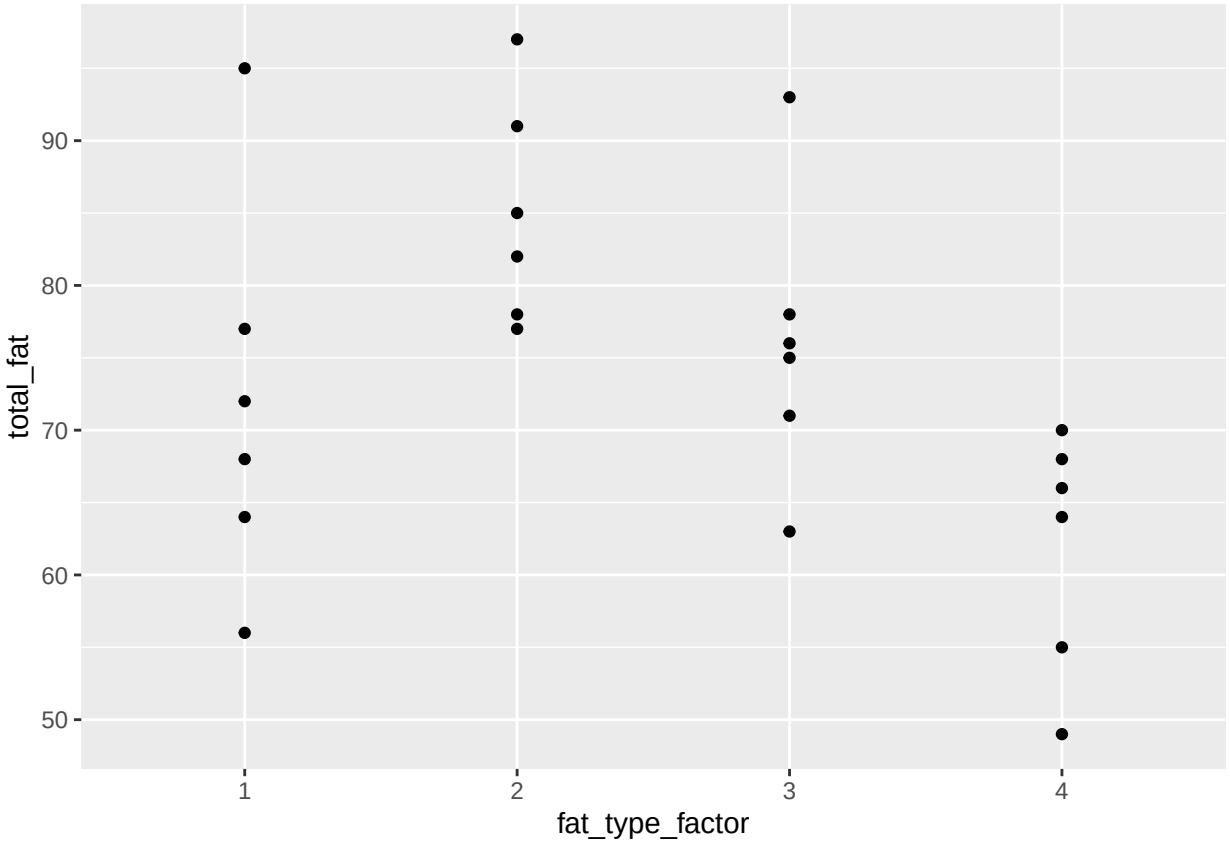
To visually assess the data, we can use QQplots. If the data follows the theoretical distribution, then we can say that the data is approximately normal. For the most part, our data follows this pattern for all levels of `fat_type`. The only potential exception is when `fat_type==3`, which has a single outlier. As a whole, the QQplots show that the levels are all approximately normally distributed.

For a quantitative assessment, we can use the Shapiro-Wilk Test on each of the levels of `fat_type`. From this test, we see that all 4 levels have a large p-value. Therefor, all fat types fail to reject the null, and we can assume that they are all normally distributed.

Question 3 - 10 points

Assess the assumption of equality of variances visually and quantitatively and comment on how well the data met this assumption.

```
# Code for visual assessment
# Plot the points per group to see how spread out they are.
ggplot(doughnuts, aes(x=fat_type_factor, y=total_fat)) + geom_point()
```




```
# Code for quantitative assessment
# Use the Brown-Forsythe test for equal variances
levene.test(doughnuts$total_fat, doughnuts$fat_type_factor) # Defaults to Brown-Forsythe
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  doughnuts$total_fat
## Test Statistic = 0.3434, p-value = 0.7942
```

Write your comments about how well the data meet the equality of variances assumption below:

From the figure, we can see that there is some difference in the spread between the groups. The first and third groups have a larger variance than the second and fourth groups. But to determine if this is enough to be statistically different, we should conduct some actual tests.

A test we can use to compare the equality of variances among many groups is the Brown-Forsythe test, which is a special kind of Lavene's test. From the test, we get a p-value of 0.7942 which is large, so we fail to reject the null. Therefor, we can conclude that there is not a significant difference between the groups' variances, and can move forward with the equality of variances assumption.

Question 4 - 10 points

Regardless of assumptions, Conduct a one-way ANOVA and interpret the result in the context of the research question. Please be sure to display your ANOVA results using the summary() function.

```
# Code for the ANOVA
aov.model <- aov(total_fat~fat_type_factor, doughnuts)
summary(aov.model)
```

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## fat_type_factor  3   1636   545.5   5.406 0.00688 **
## Residuals       20   2018   100.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Write your interpretation of the results below:

From the ANOVA results, we get a p-value of 0.00688 which is lower than the standard significance level of $\alpha = 0.05$. Therefor, we reject the null and assert that there is a significant difference between the means of the groups. This means that one of the fats was absorbed more by the doughnuts than another fat, on average. We can be confident in these results because of the confirmation to the normality and equal variances assumptions that we did earlier.

Question 5 - 10 points

When the null hypothesis in ANOVA is rejected, you conclude that at least one group mean is different than the others. You may then wonder which of the means is different. There are numerous tests that have been developed to answer this question. These are sometimes referred to as "post hoc" tests because they are usually done after an ANOVA has returned a significant result. For this question, you will do two things:

1) Do some reading on your own to find a post hoc test that is appropriate for use after an ANOVA. There are numerous tests available, but you only need to pick one.

Write your choice of ANOVA post hoc test below:

We will use Tukey's Method. This test performs the pair-wise mean comparisons while keeping the group error rate below our significance level, in our case 0.05.

2) Conduct this post hoc test on the doughtnuts data and determine which mean/s are different.

```
# Code for the post hoc test of your choice
TukeyHSD(aov.model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = total_fat ~ fat_type_factor, data = doughnuts)
##
```

```
## $fat_type_factor
##     diff        lwr       upr     p adj
## 2-1   13  -3.232221 29.232221 0.1461929
## 3-1    4 -12.232221 20.232221 0.8998057
## 4-1  -10 -26.232221  6.232221 0.3378150
## 3-2   -9 -25.232221  7.232221 0.4270717
## 4-2  -23 -39.232221 -6.767779 0.0039064
## 4-3  -14 -30.232221  2.232221 0.1065573
```

Write your determination about which mean/s are different below:

The Tukey method provides pairwise comparisons for mean equality. Any pair that has an adjusted p-value less than our significance level means that the pair's means are statistically different. From the results, the only pair of groups that has a smaller p-value than 0.05 are group 4 and group 2, which has an adjusted p-value of ~ 0.0039. Therefor, the second fat was absorbed by the doughnuts more than the fourth fat. This conclusion is supported by our initial bar plot in Question 1, where the error bars of the second and fourth groups do not overlap.