Non-shatter Code Along

You are a researcher studying pennycress. From your prior experiments, you've noticed that pennycress reaches physiological maturity (i.e., maximum dry weight) about two weeks prior to harvest maturity when it can actually be harvested. You know that harvest aids have been used to reduce the time between physiological and harvest maturity in canola and rapeseed (closely related brassica species), would that work in pennycress too? You devised an experiment using a naturally-senescing control and three harvest aids: two chemical desiccants and one mechanical method. After two years of experimentation, you want to analyze your data to better understand the use of harvest aids to facilitate earlier pennycress harvest. Use a one-way ANOVA to test the treatment efficacy at reducing seed and biomass moisture, which are the barriers to early harvest.

Load in the libraries

```
library(tidyverse)

library(agricolae)
library(car)

library(readxl)
```

Read in the Non-shatter dataset

```
pc <- read_excel("~/Documents/Non-shatter Data.xlsx")
```

Change categorical treatments to factors.

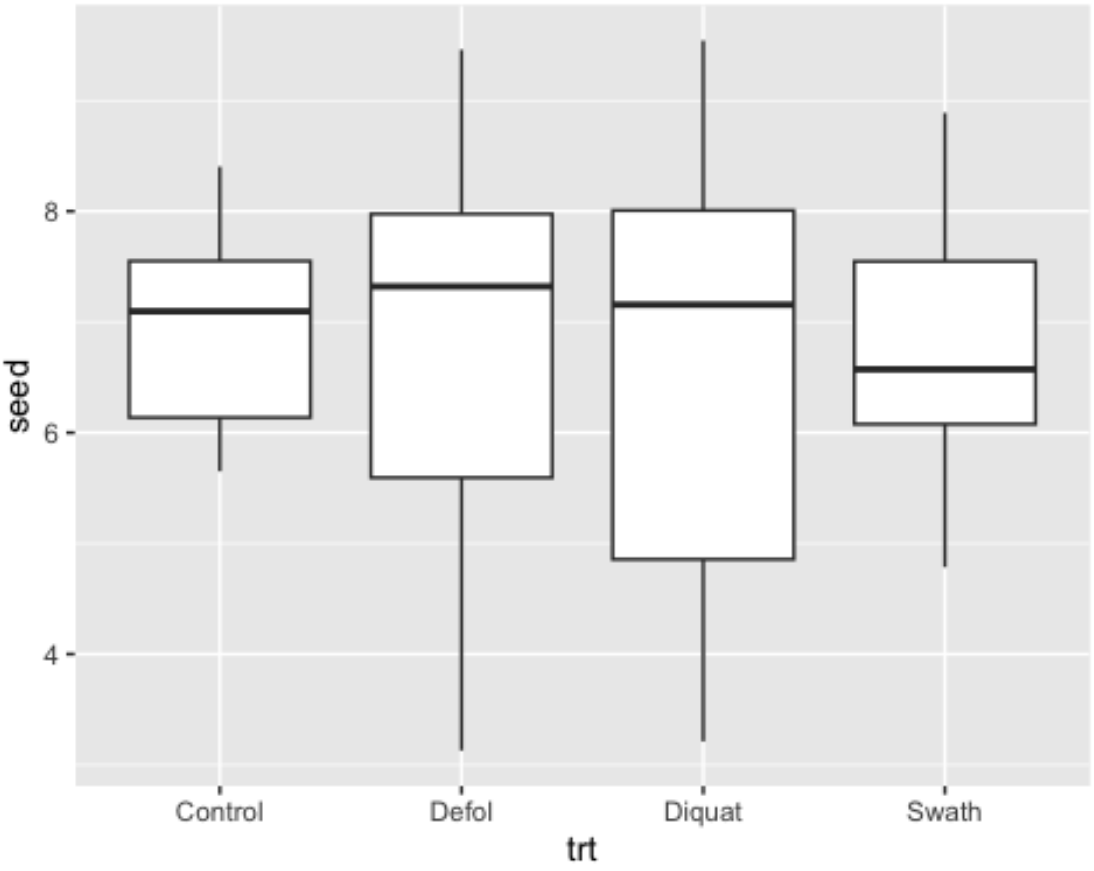
```
pc$rep <- as.factor(pc$rep)
pc$trt <- as.factor(pc$trt)
pc$year <- as.factor(pc$year)

summary(pc)
```

```
##    year    rep         trt        biomass            seed
##  2019:24   1:12   Control: 8   Min.   : 3.817   Min.   :3.129
##  2020:24   2:12   Defol  :16   1st Qu.: 8.549   1st Qu.:5.691
##            3:12   Diquat :16   Median :15.108   Median :7.143
##            4:12   Swath  : 8   Mean   :15.283   Mean   :6.748
##                                 3rd Qu.:19.151   3rd Qu.:7.958
##                                 Max.   :31.328   Max.   :9.542
```

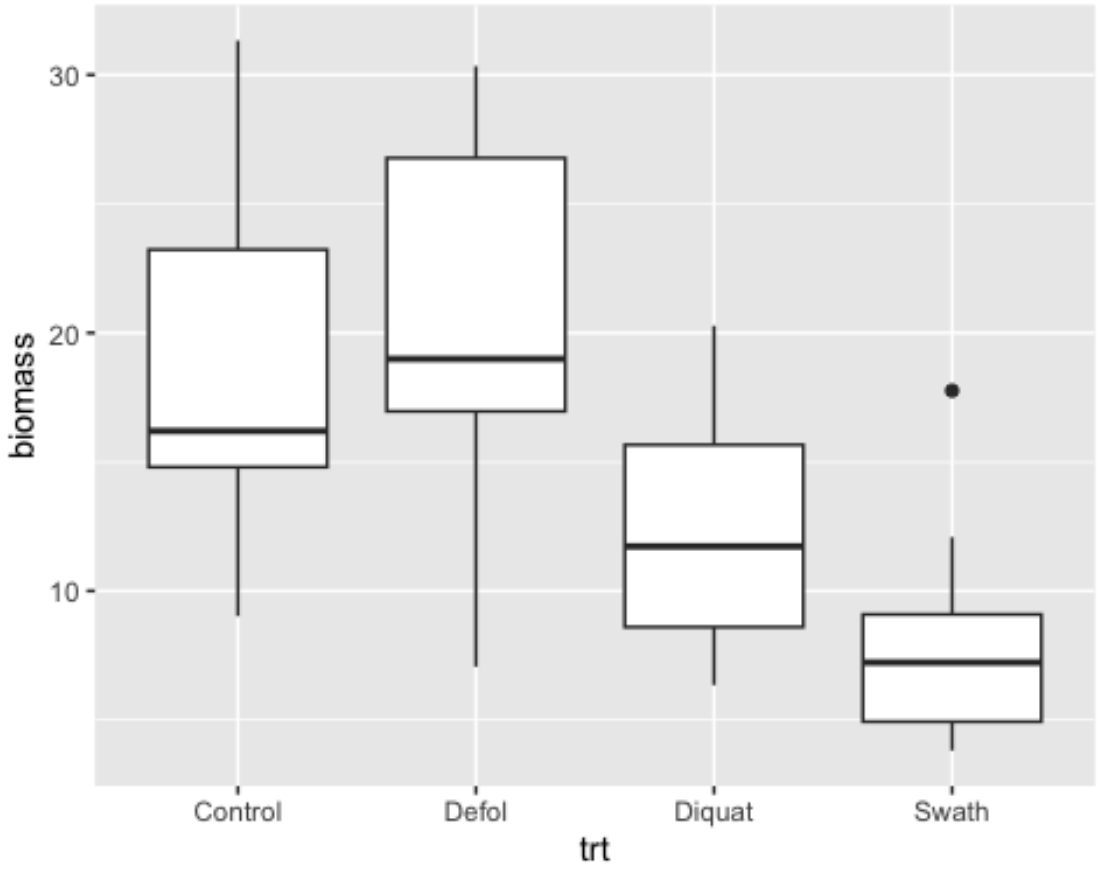
Taking a look at the summary for the non-shatter dataset, you can see the spread of the variables.

Visualize the data

```
ggplot(pc, aes(trt, seed)) + geom_boxplot()
```




```
ggplot(pc, aes(trt, biomass)) + geom_boxplot()
```



Based this figure, it seems as though seed moisture is relatively consistent across treatments whereas biomass moisture is lower in the Diquat and Swath treatments compared to the Control and Defol treatments.

One-way ANOVA

Seed moisture

Write the linear model using the lm() function. The syntax for the model is lm(dependent variable ~ independent variable(s), dataset)

```
seed.lm <- lm(seed ~ trt, pc)
```

Run your linear model through the anova() function. This will run your model through an analysis of variance.

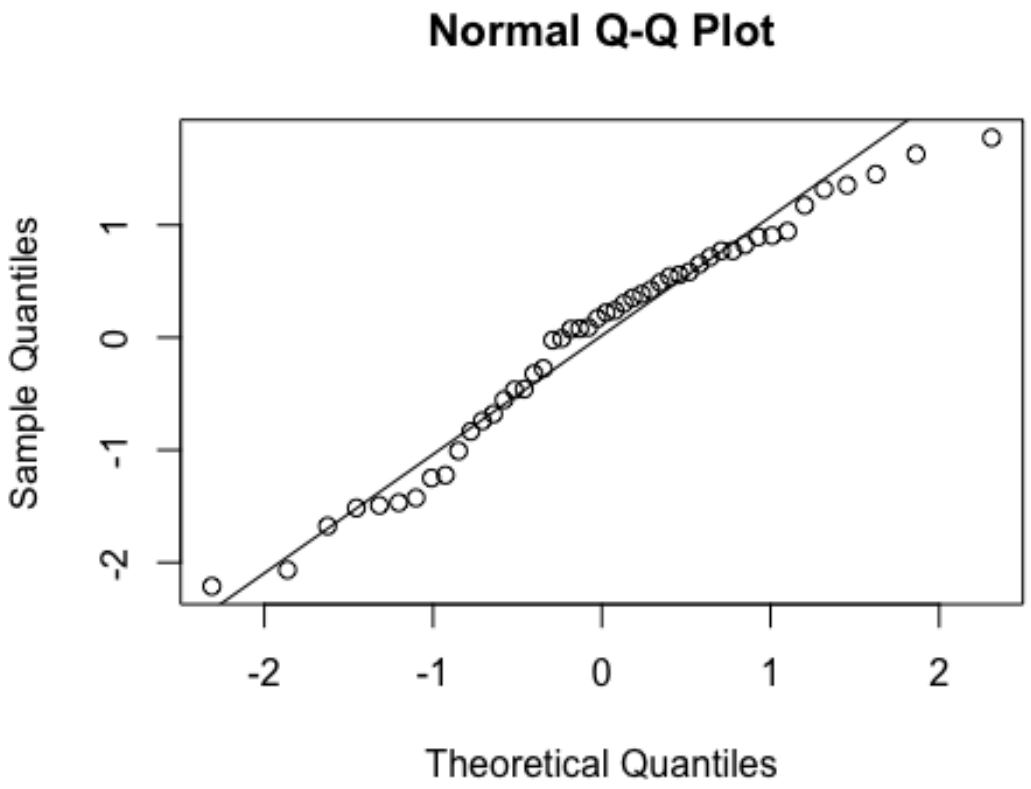
```
anova(seed.lm)
```

```
## Analysis of Variance Table
##
## Response: seed
##           Df  Sum Sq Mean Sq F value Pr(>F)
## trt        3   0.718 0.23937  0.0826 0.9691
## Residuals 44 127.463 2.89689
```

Based on the anova, treatment is **not** significant! This means that there is no analysis to run and that there are no differences between treatments. If you want to present those results, it still good to run the assumptions for the ANOVA to ensure that the model is a good fit to the data, but the most you can present is the average across treatments.

Normally distributed residuals

```
# Visual
qqnorm(rstandard(seed.lm))
qqline(rstandard(seed.lm))
```

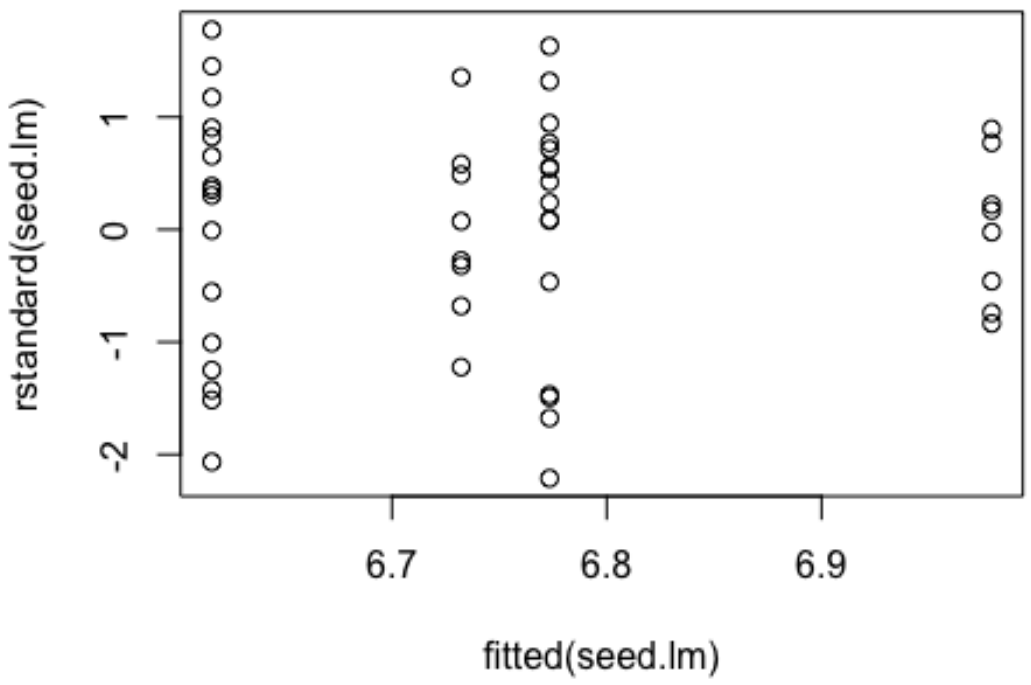



```
# Mathematical
shapiro.test(residuals(seed.lm))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(seed.lm)
## W = 0.96627, p-value = 0.1805
```

The data is normally distributed.

Homogeneity of variances

```
# Visual
plot(fitted(seed.lm), rstandard(seed.lm))
```




```
# Mathematical
leveneTest(seed.lm)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.0561 0.3774
##       44
```

The residuals are homogeneous.

Biomass moisture

```
biomass.lm <- lm(biomass ~ trt, pc)
```

Run your linear model through the anova() function. This will run your model through an analysis of variance.

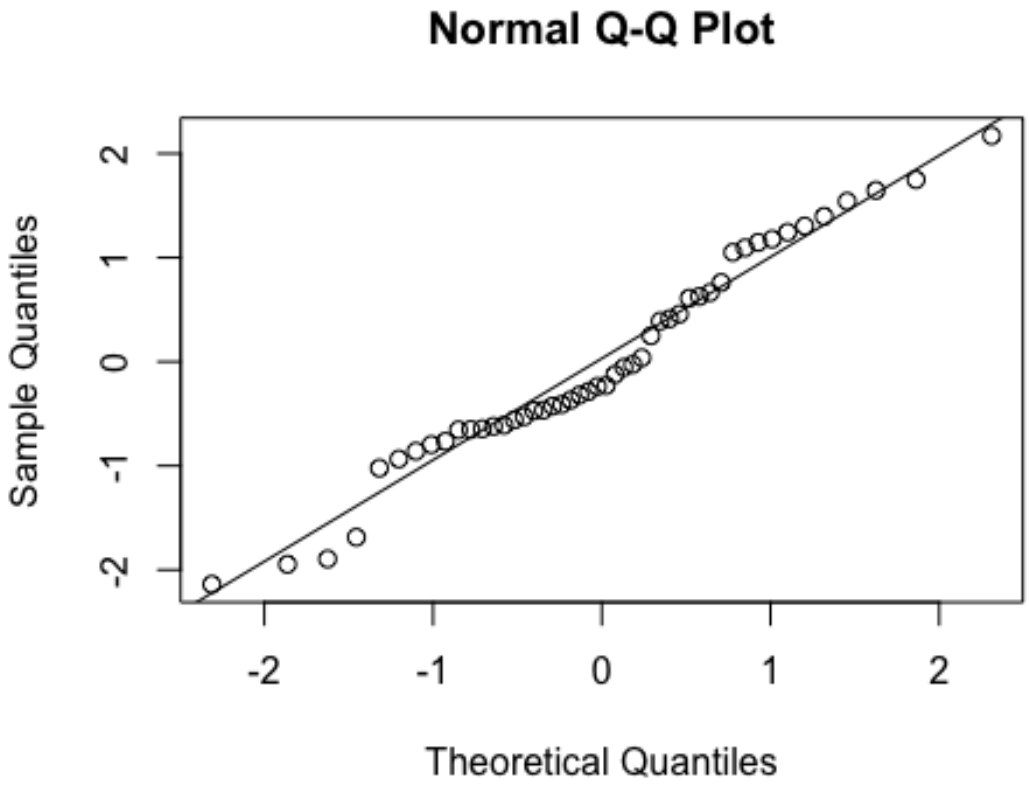
```
anova(biomass.lm)

## Analysis of Variance Table
##
```

```
## Response: biomass
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## trt          3  956.73  318.91  8.3434 0.0001673 ***
## Residuals 44 1681.81   38.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA, treatment is significant to the 0.001 level, which means that there is only a 0.1% chance that the differences between treatments is due to random chance. That's great! Now check the assumptions.

Normally distributed residuals

```
# Visual
qqnorm(rstandard(biomass.lm))
qqline(rstandard(biomass.lm))
```

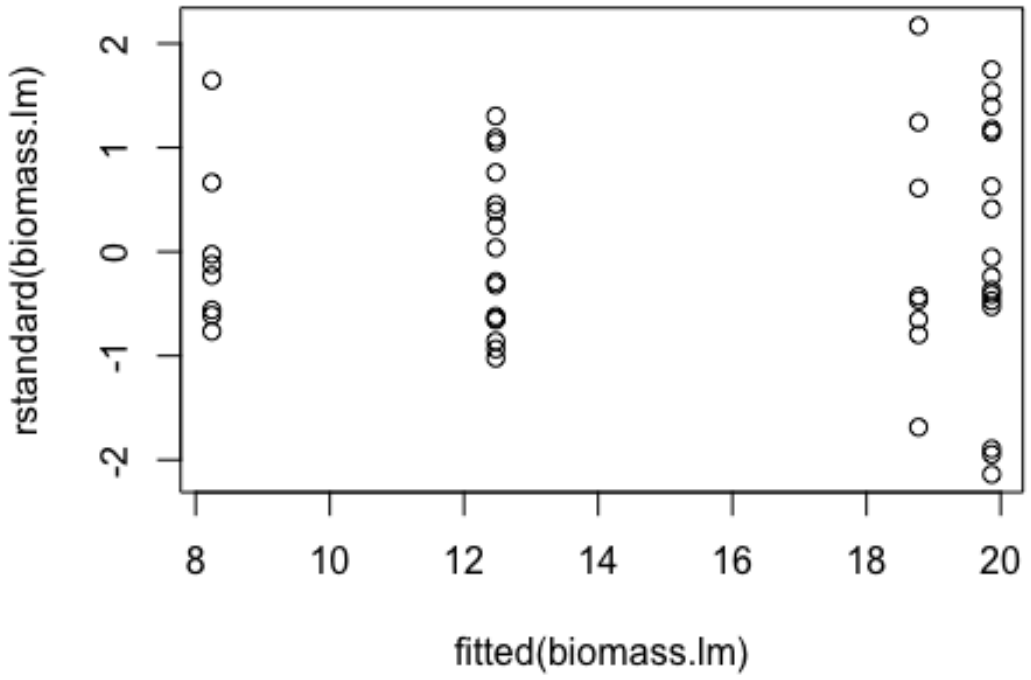



```
# Mathematical
shapiro.test(residuals(biomass.lm))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(biomass.lm)
## W = 0.96994, p-value = 0.252
```

The data is normally distributed.

Homogeneity of variances

```
# Visual
plot(fitted(biomass.lm), rstandard(biomass.lm))
```




```
# Mathemathical
leveneTest(biomass.lm)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.2818 0.2924
##       44
```

The residuals are homogeneous.

Mean separation

Now that we've found the biomass model is significant **AND** has passed is both assumption tests, we can move forward to the mean separation. Let's use Tukey's HSD for this analysis.

```
biomass.HSD <- HSD.test(biomass.lm, "trt")
biomass.HSD$groups
```

```
##           biomass groups
## Defol   19.866349      a
## Control 18.775938     ab
## Diquat  12.473299     bc
## Swath    8.241183      c
```

Final visualization

Create a summary of the data

First, we need to group and summarize our data, so we have means for our bar graph.

```
pc.summary <- pc %>%
  group_by(trt) %>%
  summarize(mean.biomass = mean(biomass))

pc.summary <- data.frame(pc.summary)
pc.summary
```

```
##       trt mean.biomass
## 1 Control    18.775938
## 2   Defol    19.866349
## 3  Diquat    12.473299
## 4   Swath     8.241183
```

Notice the inclusion of the data.frame() function. This converts the summary to a data frame, which we can add our letters of significance to. The summary() function typically outputs a tibble, which you cannot add values to.

Now, we can add the letters of significance to our summary associated with each treatment. Make sure you create your groups vector in the same order as the summary table! The summary table is typically in alphabetical order.

```
biomass.HSD$groups
```

```
##           biomass groups
## Defol   19.866349      a
## Control 18.775938     ab
## Diquat  12.473299     bc
## Swath    8.241183      c
```

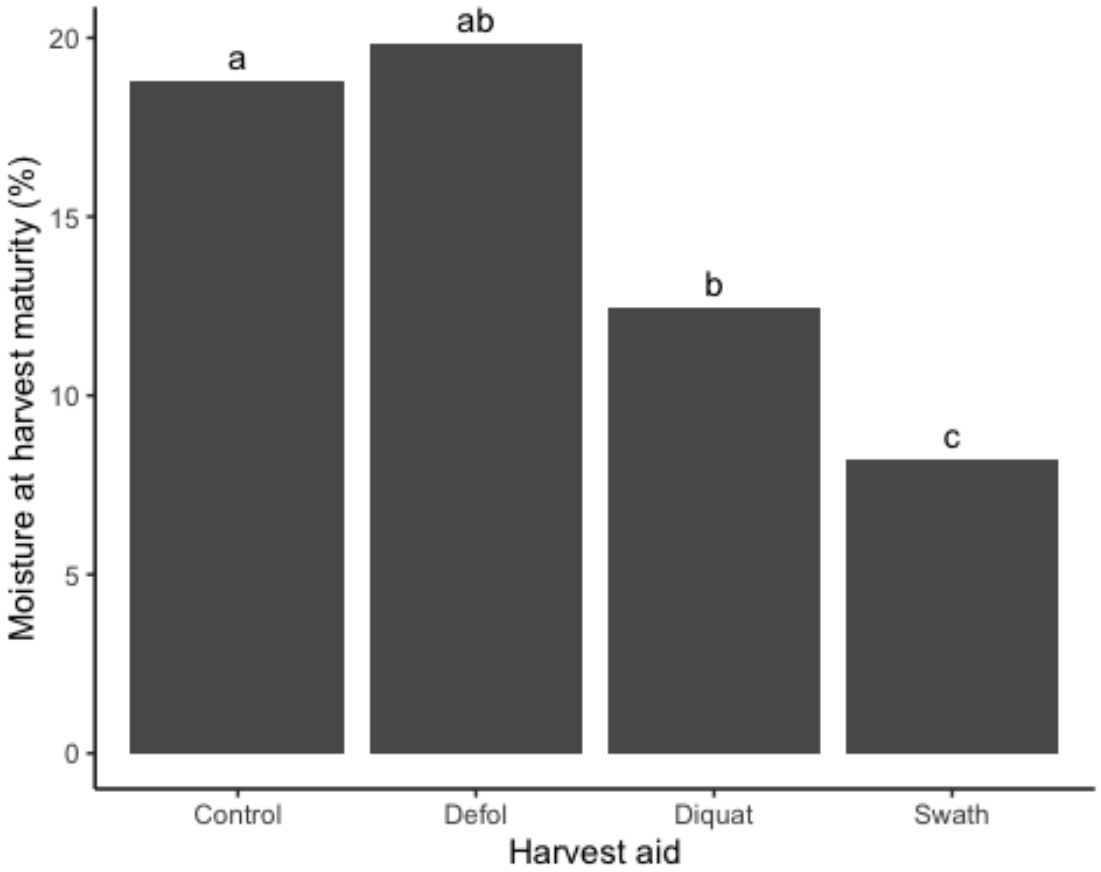
```
# Defol - a
# Control - ab
# Diquat - b
# Swath- c

groups <- c("a", "ab", "b", "c")
pc.final <- cbind(pc.summary, groups)
pc.final
```

```
##       trt mean.biomass groups
## 1 Control    18.775938      a
## 2   Defol    19.866349     ab
## 3  Diquat    12.473299      b
## 4   Swath     8.241183      c
```

Now you have a data set that includes your treatment name, treatment value, and the associated mean separation group.

Make a bar graph

```
ggplot(pc.final, aes(trt, mean.biomass)) +
  geom_bar(stat = "identity") +
  geom_text(label = groups, vjust = -0.5) +
  labs(x = "Harvest aid", y = "Moisture at harvest maturity (%)") +
  theme_classic()
```



Your figure is ready for presentation! The next step is to interpret your results.

The Control and Defol treatments had the highest biomass moisture at harvest maturity while the Swath treatment had the greatest reduction in biomass moisture. It may be interesting to explore why the mean Defol moisture is greater than the Control mean while still similar to the Diquat moisture level.