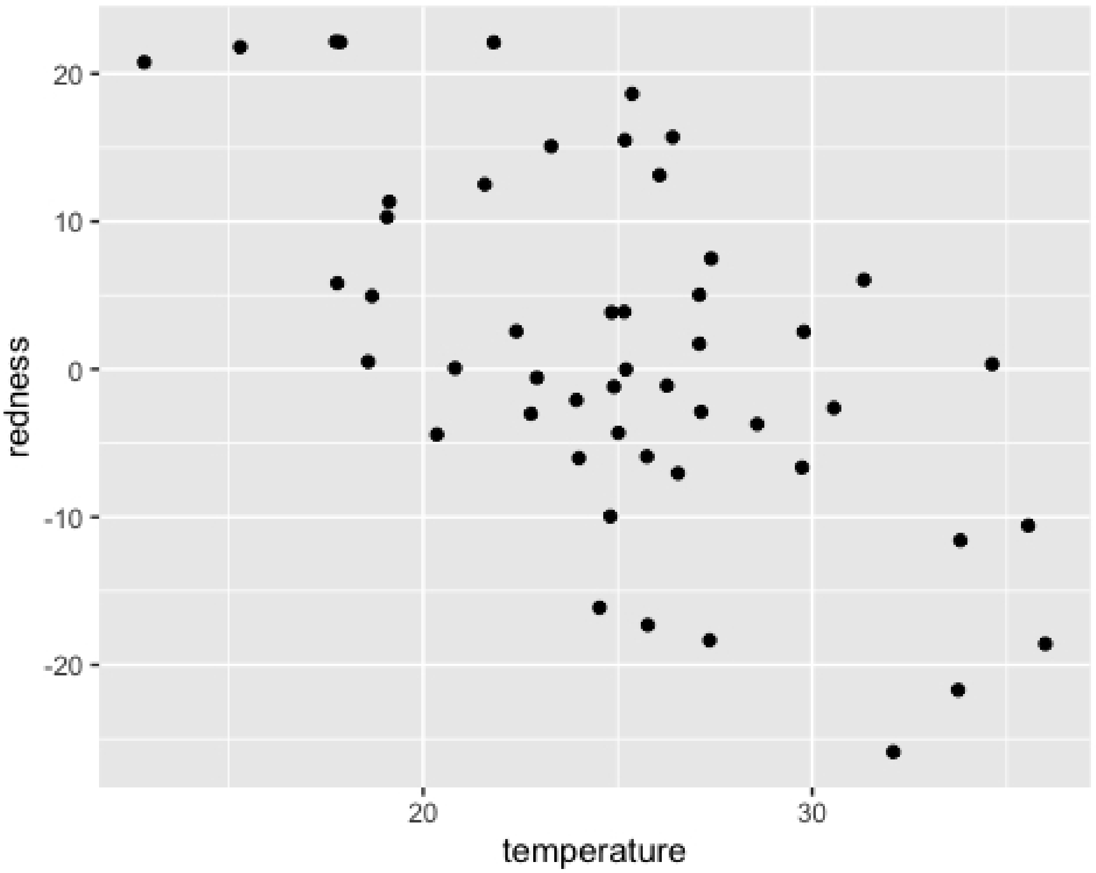
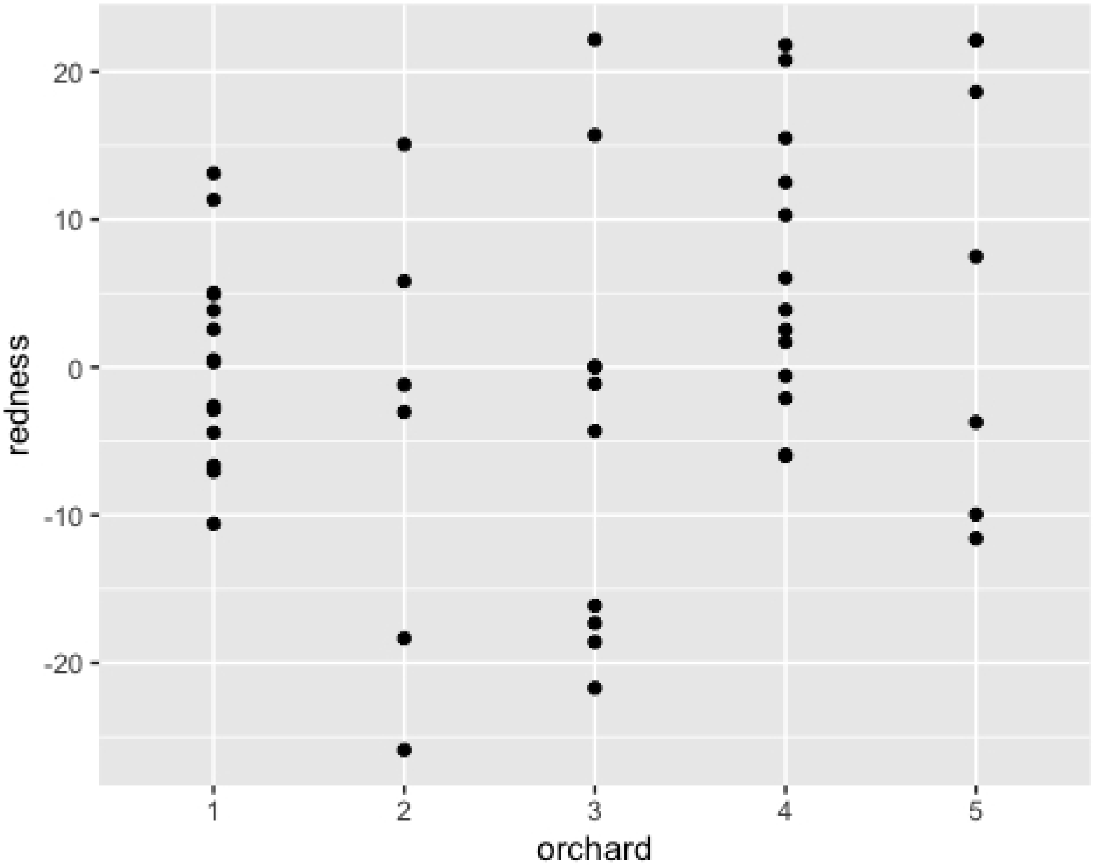
Question 1

a. Load the data into R and call it redness_df. Create the following: • a plot showing temperature and redness • a scatterplot and boxplots of redness between orchards • a summary of the data, such as the mean and standard deviation of redness by orchard

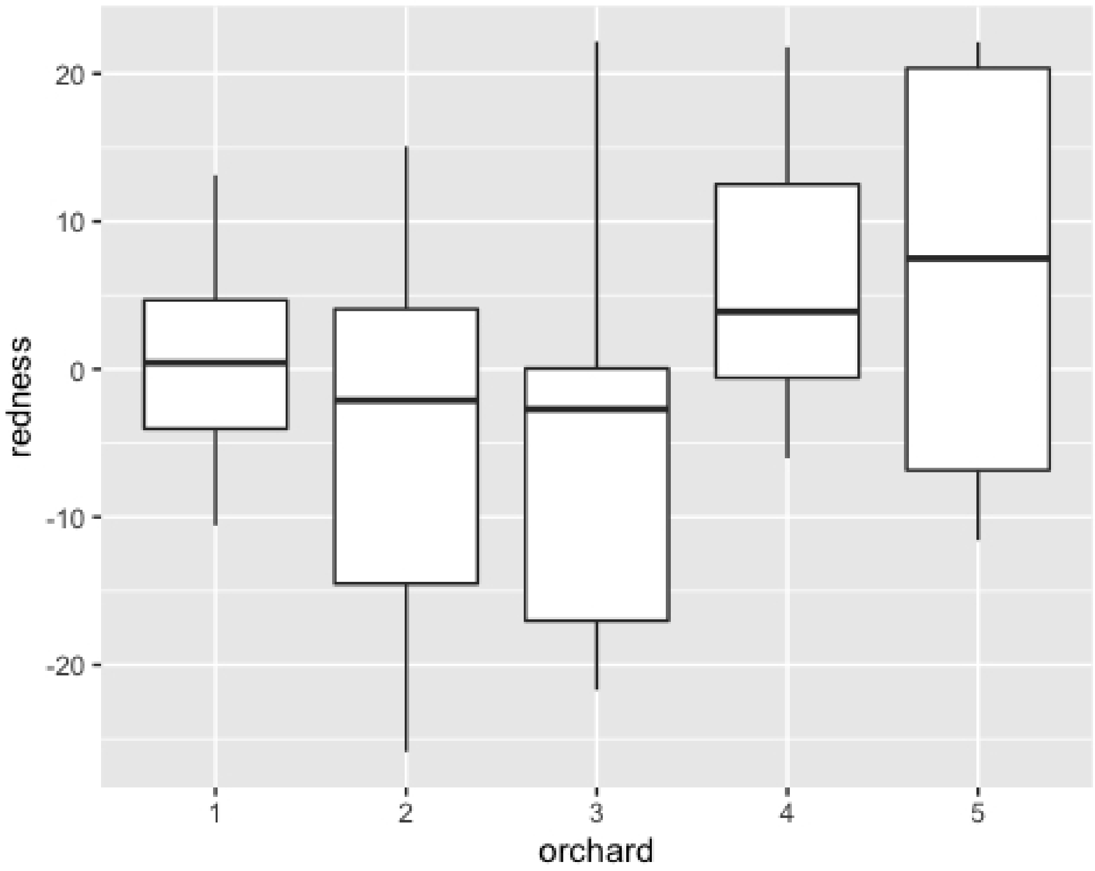
```
redness_df = read.csv("Redness.csv", stringsAsFactors = TRUE)

library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

redness_df <- redness_df %>%
  mutate(orchard = factor(orchard)) %>% as.data.frame()

str(redness_df)

## 'data.frame':    50 obs. of  3 variables:
##  $ orchard    : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ temperature: num  18.6 18.7 26.1 29.7 19.1 ...
##  $ redness    : num  0.516 4.948 13.121 -6.639 11.335 ...

ggplot(redness_df, aes(x=temperature, y= redness)) + geom_point()
```



```
ggplot(redness_df, aes(x=orchard, y= redness)) + geom_point()
```



```
ggplot(redness_df, aes(x=orchard, y= redness)) + geom_boxplot()
```



```
redness_df %>% group_by(orchard) %>%
  summarise(mean = mean(redness),
            sd = sd(redness),
            n = length(redness))
```

```
## # A tibble: 5 x 4
##   orchard   mean    sd     n
##   <fct>    <dbl> <dbl> <int>
## 1 1        0.539  6.84    14
## 2 2       -4.59  15.2      6
## 3 3       -4.12  14.8     10
## 4 4        6.19   9.35    13
## 5 5        6.45  14.9      7
```

b. Based on the plots and output you created in Question 1. a describe what features/relationships you see.

The plot between temperature and redness shows that as the temperature increases, the redness of the apple decreases. This indicates a negative relationship.

The scatterplot and the boxplot between orchard and redness shows that they overlap. This indicates that there is no difference between orchards.

The summary shows that orchard 2 has the lowest mean whereas orchard 5 has the highest mean. Orchard 2 also has the highest standard deviation, but it has the lowest sample size. In contrast, orchard 1 has the lowest standard deviation, but it has the highest sample size. Perhaps, sample sizes affect how the standard deviation turned out for each orchards (in terms of redness).

c. Based on the plots and output you created in Question 1. a explain why analysing the data using an ANOVA may not be appropriate.

Using ANOVA analysis for this data may not be appropriate because of these reasons:

- measurements are made on clusters(orchards) of related statistical units
- the sample size between orchards are not the same and ANOVA analysis is not better suited for dealing with missing values
- temperature and redness is observed to have a relationship

d. Using the LMM checklist we introduced in the lectures, indentify the response, fixed effect (explanatory component), and the structural component (random effect). Give reasons for your choice for the fixed and random part.

The response is redness. The fixed effect (explanatory component) is temperature. The random effect (structural component) is orchard.

The fixed effect is temperature because it is said that temperature affects the redness of apples (condition of interest). The random effect is orchard because it is the physical structure of the study (acts as blocking).

e. Using the lmer package, fit an intercept only model and a model with the fixed effect included. Call the models m0 and m1 respectively. Use the code below and alter the arguments appriopriately.

```
library(lme4)

## Loading required package: Matrix

m0 <- lmer(redness ~ (1|orchard), data=redness_df)
m1 <- lmer(redness ~ temperature + (1|orchard), data=redness_df)
```

f. Test the signficance of the fixed effect using anova(m0,m1). Is there evidence to suggest significance of the fixed effect? Interpret what the coefficient for the fixed effect implies for this analysis.

```
anova(m0,m1)

## refitting model(s) with ML (instead of REML)
## Data: redness_df
## Models:
## m0: redness ~ (1 | orchard)
## m1: redness ~ temperature + (1 | orchard)
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0    3 395.49 401.23 -194.75   389.49
## m1    4 373.49 381.14 -182.75   365.49 23.999  1  9.638e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: redness ~ temperature + (1 | orchard)
##    Data: redness_df
##
## REML criterion at convergence: 363.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.73680 -0.66823 -0.04499  0.66214  2.01396
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  orchard  (Intercept) 10.46    3.234
##  Residual             84.47    9.191
## Number of obs: 50, groups:  orchard, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  36.2959     6.7332   5.391
## temperature  -1.4016     0.2563  -5.469
##
## Correlation of Fixed Effects:
##             (Intr)
## temperature -0.956
```

There is an evidence that suggests that the fixed effect model (m1: temperature) is significant because the p-value is lower than 0.05 so we reject the null hypothesis that both models are the same. In addition, the AIC, BIC and logLik of the fixed effect is lower than the model without the fixed effect (m0).

The coefficient for the fixed effect implies that as the temperature increases, the redness of apples decreases by 1.4.

f. Test the signficance of the random effect using using the code below, with the appropriate arguments changed. Is there evidence to suggest that the random effect is necessary in the model?

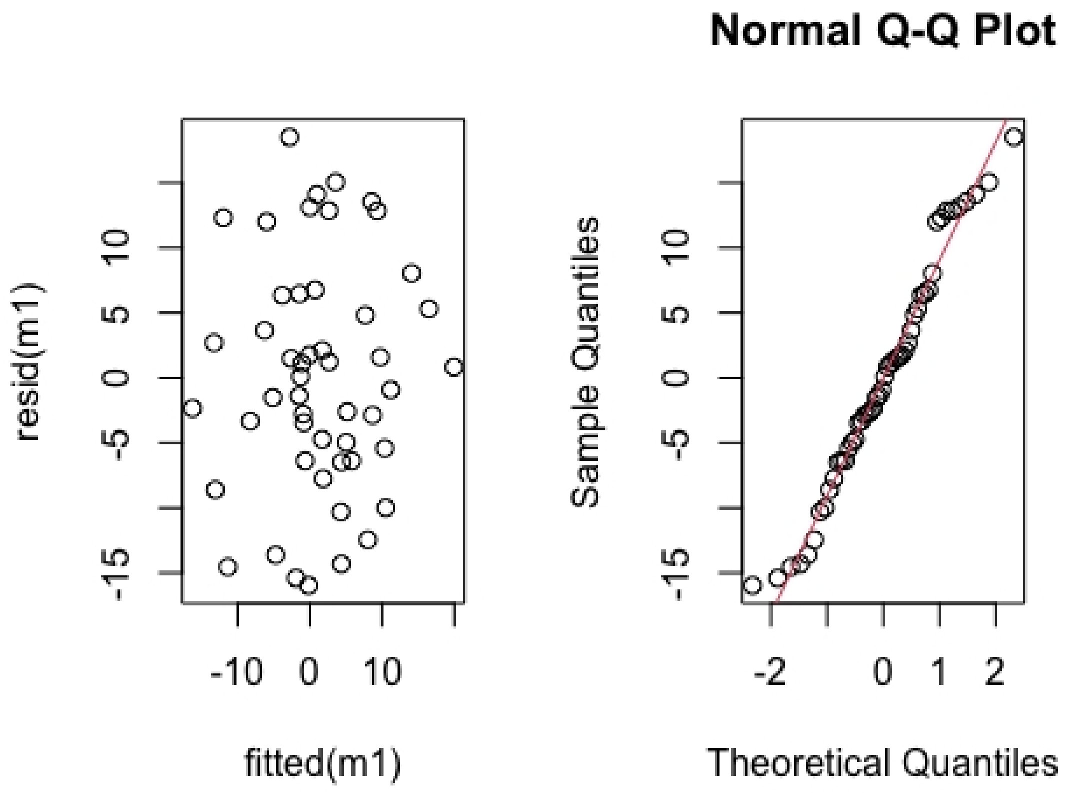
```
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step

ranova(m1)

## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## redness ~ temperature + (1 | orchard)
##                npar  logLik    AIC     LRT Df Pr(>Chisq)
## <none>            4 -181.68 371.36
## (1 | orchard)     3 -182.15 370.31 0.95168  1     0.3293
```

The variability introduced by the random effect (orchard) is explaining less variability in the response. This is because the p-value is greater than 0.05. In addition, both the AIC and LRT suggests that the random effect is not necessary in the model.

g. Check the model assumptions of the linear mixed effects models using the code below, with the appropriate arguments changed. Have they been met? Explain why or why not.

```
par(mfrow = c(1,2))
plot(resid(m1)~fitted(m1))
qqnorm(resid(m1)); qqline(resid(m1), col=2)
```



Assumptions:

- homogeneity of variance
- normality of the residuals

The residual plot shows that the residuals are evenly and symmetrically distributed meeting the assumption of homogeneity of variance.

The qqplot shows that the points are lying close to the line resembling a straight line, except for the tails meeting the assumption of normality of the residuals.

h. Suppose that the experiment was run again, this time taking several apples from 10 randomly selected trees, in 5 different orchards. Give the random effects for this study and and explain if they are nested or crossed.

The random effects are trees and orchards.

It is a crossed random effect because in 5 different orchards, several apples are taken from 10 randomly selected apple trees. So for each tree, the redness of an apple in orchard 1 is similar to that of in orchard 2.

Furthermore, it didn’t say that for each orchards, several apples are taken from 10 randomly selected apple trees; this would’ve indicate a hierarchical relationship which would’ve been a nested random effect.