Data Science

Statistical Data Modelling

Task 3

Case Study: Assignment: One-Way and Two-Way ANOVA in Crop Yield Study

Introduction:

This assignment focuses on analyzing the effects of fertilizer type and planting density on crop yield. The study includes observations from an imaginary experiment conducted to determine the optimal combination of fertilizer type and planting density to achieve maximum crop yield.

Objective:

The objective of this assignment is to perform a one-way ANOVA and a two-way ANOVA to determine the effects of fertilizer type and planting density on crop yield.

Data Description:

The dataset "Crop.data.anova.zip"includes 96 observations, with continuous and categorical variables. The dataset includes the following variables:

Fertilizer : A categorical variable indicating the type of fertilizer used, with three possible values: 1, 2, and 3.
Density: A categorical variable indicating the planting density, with two possible values: Low and High converted into dummy 1,2.
Crop Yield: A continuous variable indicating the crop yield in bushels per acre.
block: different block of cultivation.
Methodology:
To analyze the effects of fertilizer type and planting density on crop yield, the following approach will be adopted:

Data Cleaning: The dataset should be cleaned for missing values and outliers.

Data Exploration: An exploratory data analysis should be conducted to identify trends, patterns, and relationships in the data.

Post-Hoc Analysis: If significant differences are found in the ANOVA tests, a post-hoc analysis should be conducted to determine which groups are significantly different from each other.

Report should include the following:

- One way and two way ANOVA hypothesis testing process and assumptions
- Interpretation of the results.
- Validity of the model and summary of the findings.

The statistical technique known as Analysis of Variance, or ANOVA, is used to examine how the group means in a sample differ from one another. It is a t-test modification that is used to compare two groups' means. ANOVA makes it possible to compare means across several groups at once. ANOVA compares the alternative hypothesis—that at least one group mean differs—with the null hypothesis, which states that all group means are equal. The null hypothesis is rejected and there is evidence of a significant difference in at least one group mean if the test's p-value is less than a predefined significance level (usually 0.05).

Data source: crop.data.csv (given by the institute)

Installing and loading packages, reading the data source file.

```
install.packages(c("ggplot2", "ggpubr", "tidyverse", "broom", "AICcmodavg"))


library(ggplot2)
library(ggpubr)
library(tidyverse)
library(broom)
library(AICcmodavg)
library(psych)

# It is common for factors to be read as quantitative variables when importing a dataset into R
# to avoid that, it is better to define data types at the time of reading data
cropdata <- read.csv("D://crop.data.csv", header = TRUE, colClasses = c("factor", "factor", "factor", "numeric"))
```

Inspecting data, using describe function to view descriptive statistics.

```
# DATA INSPECTION ################################
# describes function from psych library
# descriptive statistics
describe(cropdata)
```

```
           vars  n   mean   sd median trimmed  mad    min    max range skew kurtosis   se
density*      1 96   1.50 0.50   1.50    1.50 0.74   1.00   2.00   1.0 0.00    -2.02 0.05
block*        2 96   2.50 1.12   2.50    2.50 1.48   1.00   4.00   3.0 0.00    -1.39 0.11
fertilizer*   3 96   2.00 0.82   2.00    2.00 1.48   1.00   3.00   2.0 0.00    -1.53 0.08
yield         4 96 177.02 0.66 177.06  177.01 0.68 175.36 179.06   3.7 0.11     0.01 0.07
> |
```

The values of skewness and kurtosis are significantly low (density, block & fertilizer variables) meaning data normally distributed.

Viewing data types of each column.

```
# data types of each column
str(cropdata)
```

```
'data.frame':   96 obs. of  4 variables:
 $ density   : Factor w/ 2 levels "1","2": 1 2 1 2 1 2 1 2 1 2 ...
 $ block     : Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...
 $ fertilizer: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ yield     : num  177 178 176 178 177 ...
> |
```

Displaying dimensions.

```
# no of rows and columns
dim(cropdata)
```

```
[1] 96  4
> |
```

Viewing summary.

```
# summary of the dataset
summary(cropdata)
```

```
 density block  fertilizer     yield
 1:48    1:24   1:32       Min.   :175.4
 2:48    2:24   2:32       1st Qu.:176.5
         3:24   3:32       Median :177.1
         4:24              Mean   :177.0
                           3rd Qu.:177.4
                           Max.   :179.1
>
```

The density variable has two classes, the block variable has four classes, and the fertilizer variable has three classes, as can be seen above. The variables yield are all quantitative (can compute minimum, median, mean, maximum).

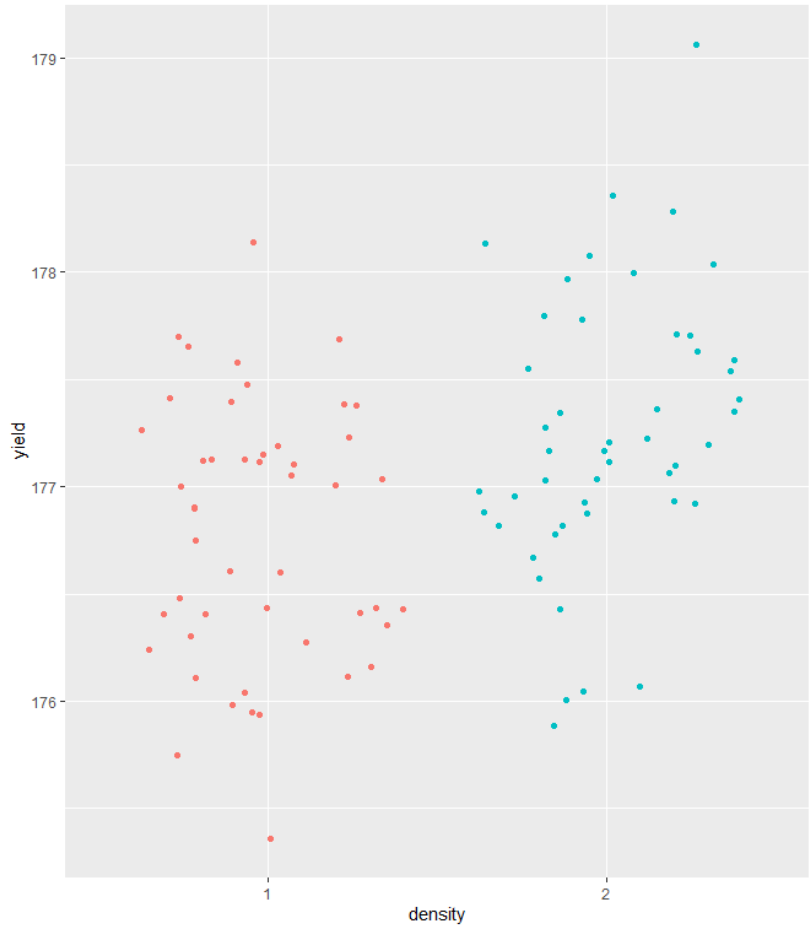
Checking NA values.

```
# check NA values in all columns
colSums(is.na(cropdata))
```

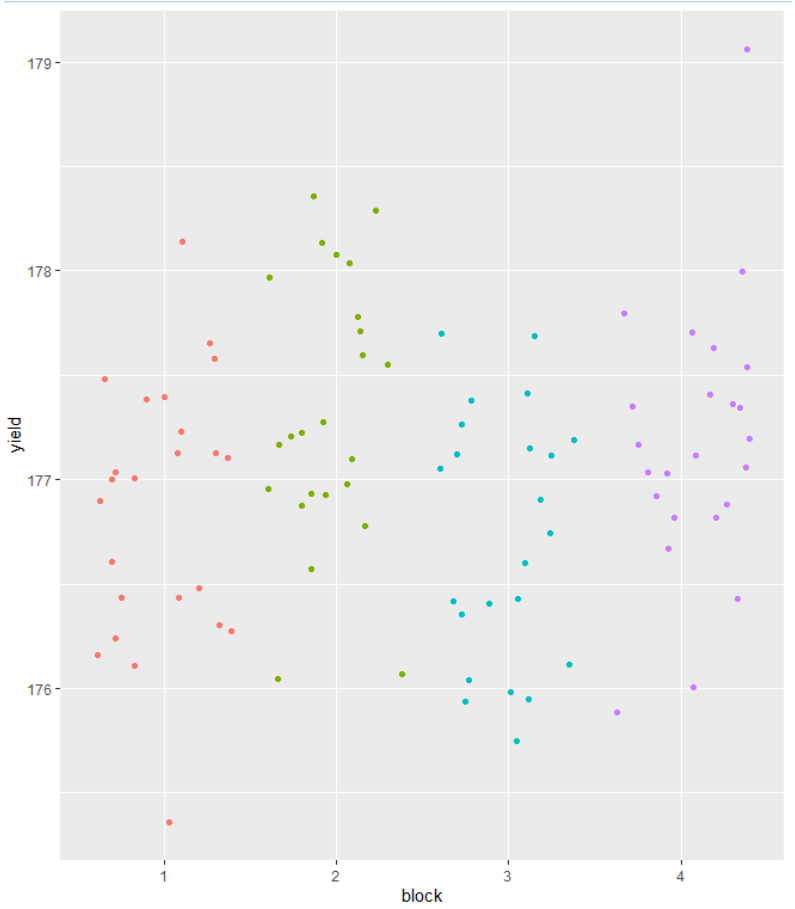
```
density      block fertilizer      yield
      0          0          0          0
```

Viewing distributions against each parameter.

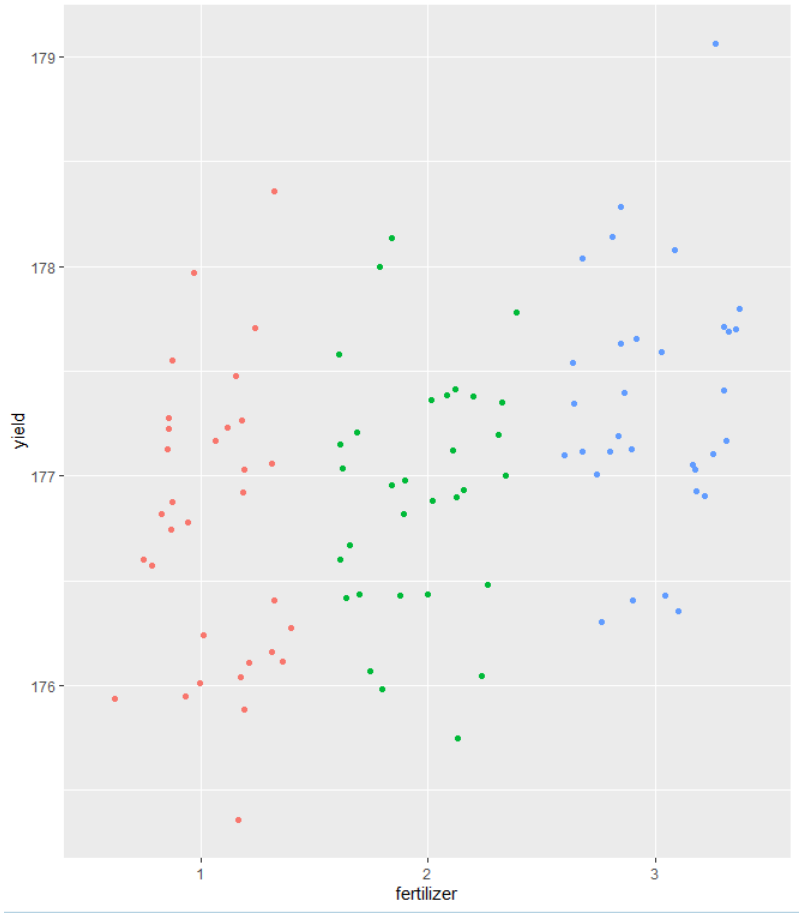
```
# check distributions

library(ggplot2)

# scatter plot
ggplot(cropdata) + aes(x = density, y = yield , color = density) + geom_jitter() + theme(legend.position = "none")
```




```
ggplot(cropdata) + aes(x = block , y = yield , color = block ) + geom_jitter() + theme(legend.position = "none")
```




```
ggplot(cropdata) + aes(x = fertilizer , y = yield , color = fertilizer ) + geom_jitter() + theme(legend.position = "none")
```



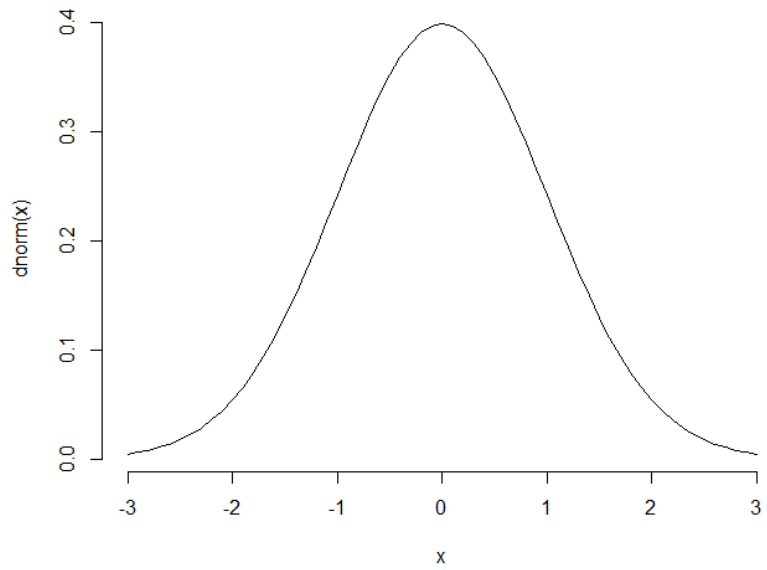
Creating a density variable bell curve to visualize the distribution (determining whether or not the data are normally distributed).

dnorm is the density function for the normal distribution.

```
# bell curve
x <- cropdata$density
summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -3.0    -1.5     0.0     0.0     1.5     3.0
> 
```

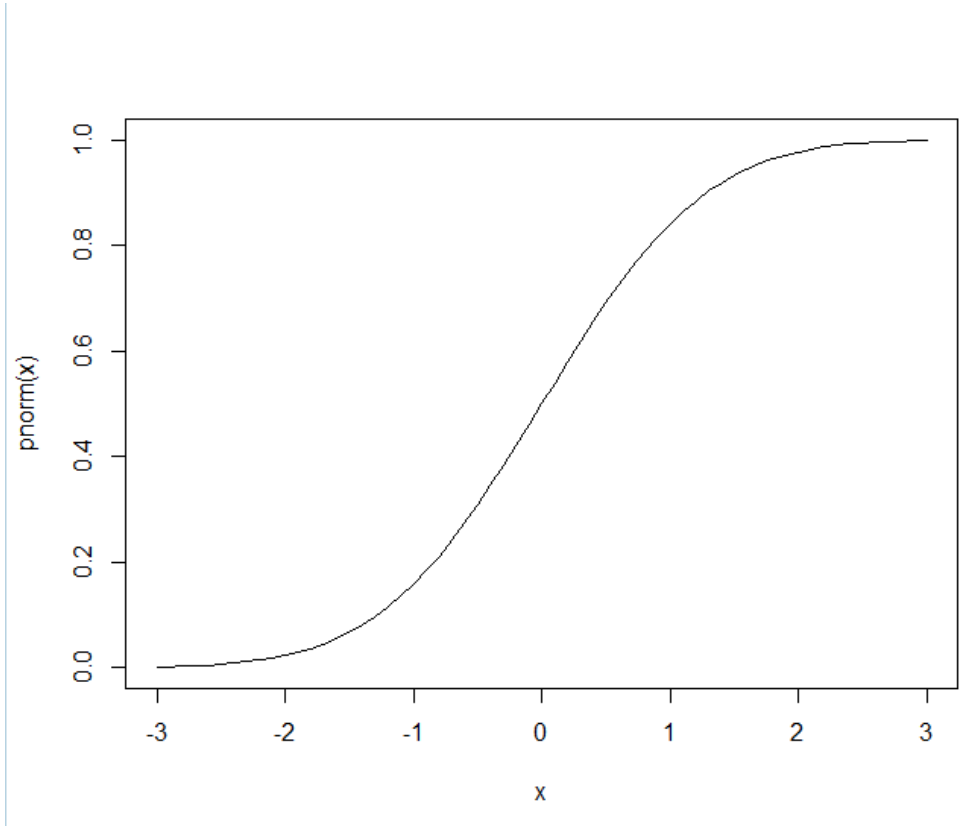
```
x = seq(-3, 3, 0.1)
plot(x = x, y = dnorm(x), type="l", bty="n")
```



Generating the normal distribution probability curve.

pnorm is the probability function (the integral of the density function)

```
plot(x, pnorm(x), type="l")
```



Generating the normal probability plot.

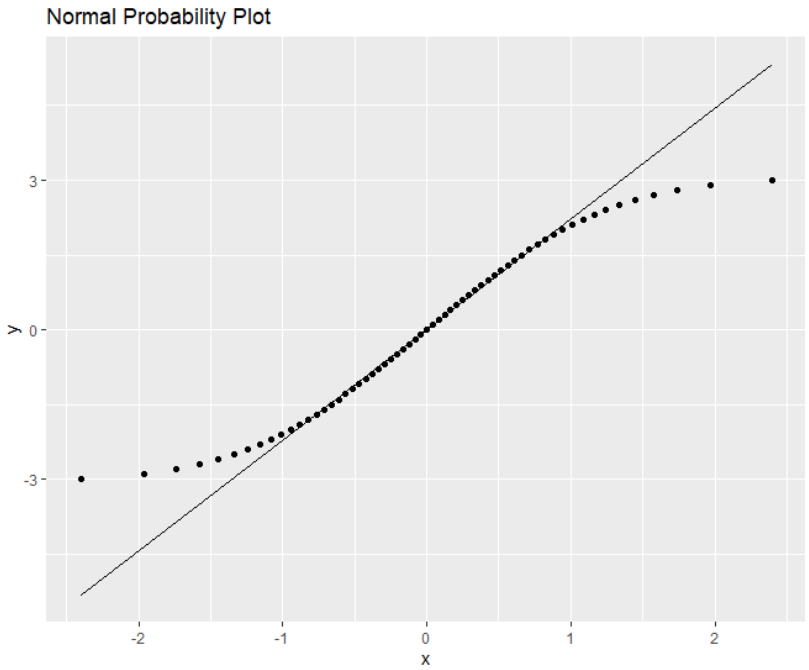
```
# Normal Probability Plot
ggplot(data.frame(x), aes(sample = x)) + stat_qq() + stat_qq_line() + labs(title = "Normal Probability Plot")
```

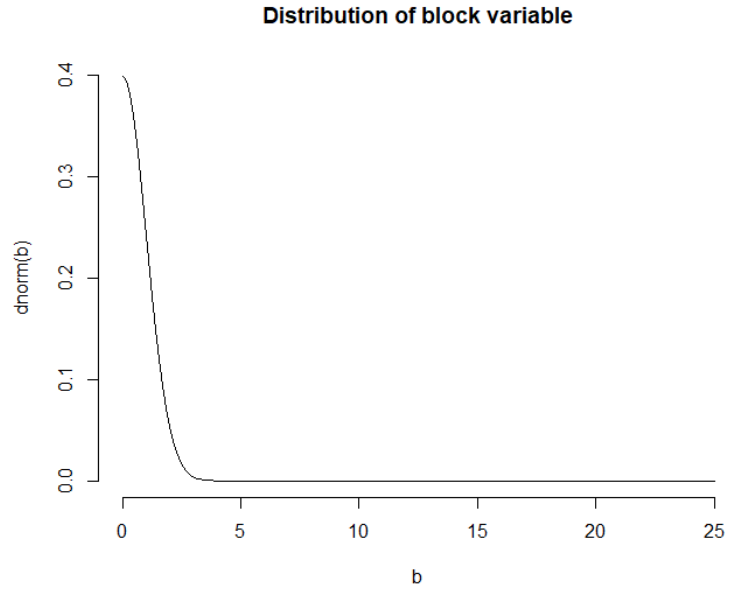
Generating a plot for the block variable.

```
# block variable
b <- cropdata$block
summary(b)
```

```
 1  2  3  4
24 24 24 24
```

```
b = seq(0, 25, 0.1)
plot(x = b, y = dnorm(b), type="l", bty="n", main="Distribution of block variable")
```

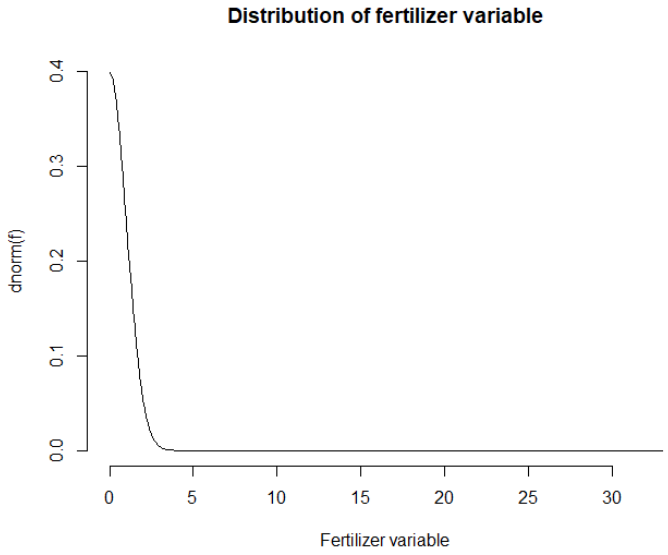

Normal Probability Plot
y
x
3
0
-3
-2
-1
0
1
2


Distribution of block variable
dnorm(b)
b
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25

Generating a plot for the fertilizer variable.

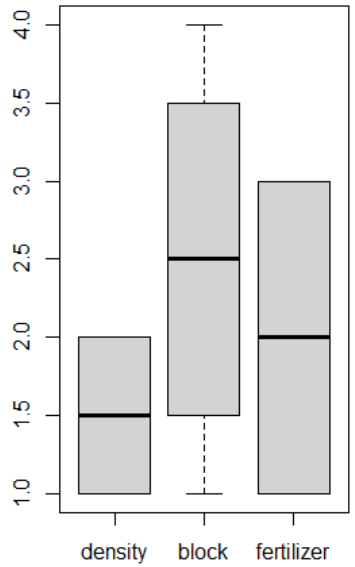
```
# fertilizer variable
f <- cropdata$fertilizer
summary(f)
```

```
f = seq(0, 33, 0.1)
plot(x = f, y = dnorm(f), type="l", bty="n", main="Distribution of fertilizer variable", xlab="Fertilizer variable")
```



Finding outliers.

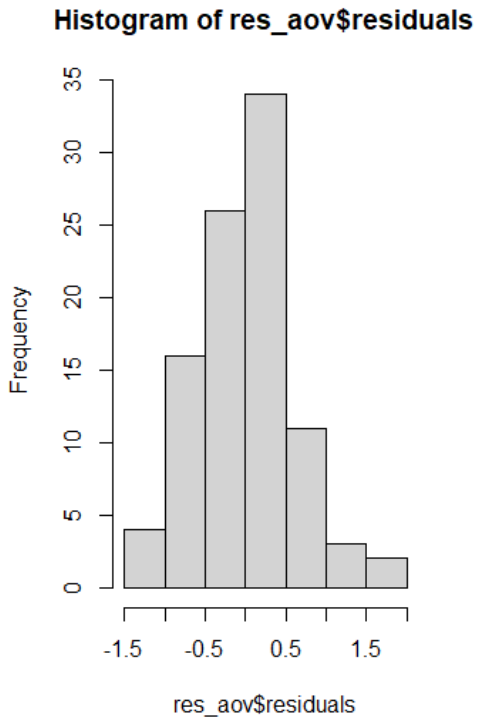
```r
# Handling outliers ##########################

# Create box plots for the numerical columns
boxplot(cropdata[c("density", "block", "fertilizer")])
```



From the above, no outliers can be found.

The parametric test ANOVA is based on certain presumptions. The data is assumed to be regularly distributed in an ANOVA. Additionally, homogeneity of variance—the idea that group variance should be about equal—is assumed by the ANOVA. Additionally, an ANOVA is predicated on the observations' independence from one another.

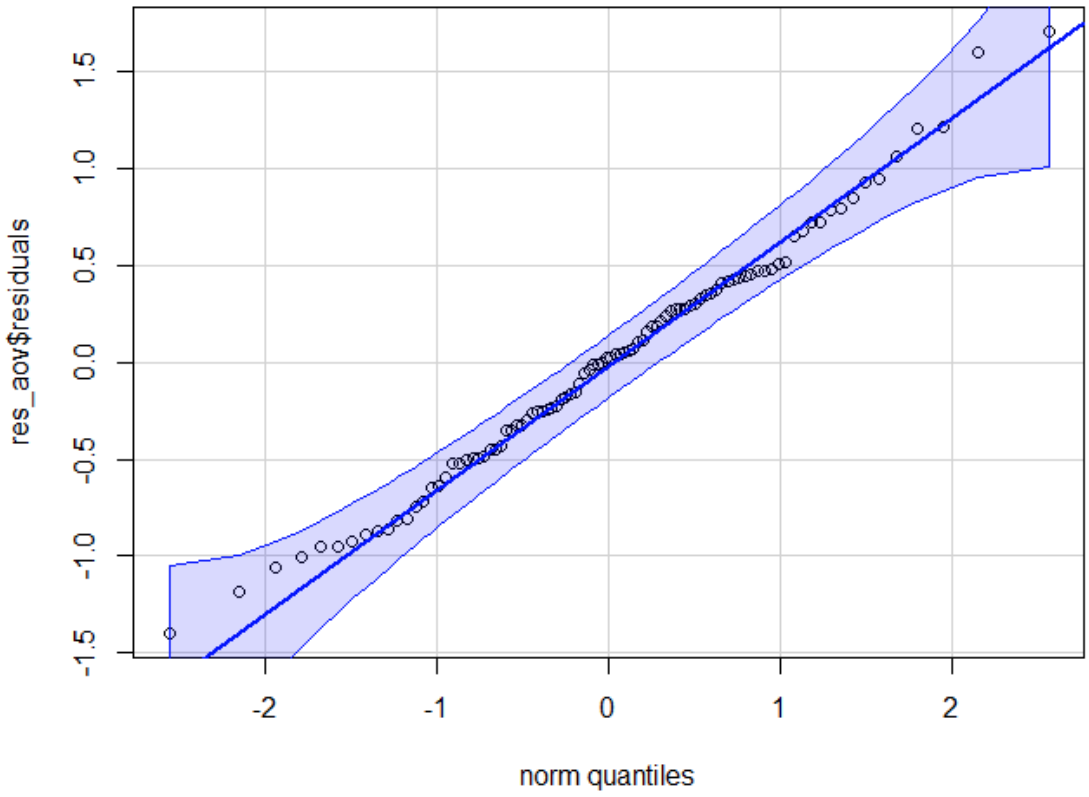
```r
# Perform the ANOVA test

# Normality
res_aov <- aov(yield ~ fertilizer, data = cropdata)
# check normality visually
par(mfrow = c(1, 2)) # combine plots

# histogram
hist(res_aov$residuals)
```



It is observable that the distribution is normal.

Generating a QQ Plot (quantile-quantile plot).

```
# QQ-plot - Normality checking
library(car)

# id = FALSE to remove point identification
qqPlot(res_aov$residuals, id = FALSE)
```



The correlation between the provided sample and the normal distribution is depicted in the above QQ plot.

Perform one way ANOVA: This has one independent variable only.

```
# fertilizer = independent variable
one.way <- aov(yield ~ fertilizer, data = cropdata)

summary(one.way)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
fertilizer   2   6.07  3.0340   7.863  7e-04 ***
Residuals   93  35.89  0.3859
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

- **Residuals** = Residual variance is the total variation that cannot be addressed by the independent variables.
- **Df** = degrees of freedom for the residuals (the total number of observations minus one and minus the number of levels in the independent variables) and degrees of freedom for the independent variable (the number of levels in the variable minus 1).
- **Sum Sq** = The total variation between the group means and the overall mean is known as the sum of squares.
- **Mean Sq** = the sum of squares divided by the degrees of freedom for each parameter to determine the mean of the sum of squares.
- **F value** = test statistic obtained using the F test. This is calculated by dividing the mean square of the residuals by the mean square of each independent variable. The greater the F value, the more probable it is that the independent variable-caused variation is real.
- **Pr(>F)** = the F statistic's p value. This is the likelihood that, had the null hypothesis—that there is no difference between the group means.

The type of fertilizer which has been used appears to have a significant effect on the crop yield, as seen by the low fertilizer's p value variable. ($p < 0.001$).

Perform Two-Way ANOVA test: also called factorial ANOVA, uses two independent variables.

```r
# Two-way ANOVA
two.way <- aov(yield ~ fertilizer + density, data = cropdata)

summary(two.way)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
fertilizer   2  6.068   3.034   9.073 0.000253 ***
density      1  5.122   5.122  15.316 0.000174 ***
Residuals   92 30.765   0.334
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Observations:

- the residual variance has been reduced.
- Due to p-values < 0.001, it can be seen that there is a significant with respect to statistics between planting density and fertilizer.

Adding interactions between variables:

Occasionally, there is an interaction effect rather than an additive impact between two of the independent variables. An asterisk rather than a plus sign can be used in the model to assess if two variables have an collaboration impact in an ANOVA.

```r
# Adding interactions between variables
interaction <- aov(yield ~ fertilizer*density, data = cropdata)

summary(interaction)
```

```
                   Df Sum Sq Mean Sq F value   Pr(>F)
fertilizer          2  6.068   3.034   9.001 0.000273 ***
density             1  5.122   5.122  15.195 0.000186 ***
fertilizer:density  2  0.428   0.214   0.635 0.532500
Residuals          90 30.337   0.337
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

Observations:

- Sum Sq = it is observable that there is a sum-of-squared value (0.428) on the variable 'fertilizer.density'
- p value = p-value is high / 0.532500
- Few variations may be explained by the relationship between planting density and fertilizer.

Including a block variable:

It is advised to add that parameter as a block variable in the model if the data has in any way grouped the experimental treatments or if the data contains confounding variables that could alter the connection.

```
# Adding a blocking variable

blocking <- aov(yield ~ fertilizer + density + block, data = cropdata)

summary(blocking)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
fertilizer   2  6.068   3.034   9.018 0.000269 ***
density      1  5.122   5.122  15.224 0.000184 ***
block        2  0.486   0.243   0.723 0.488329
Residuals   90 30.278   0.336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Observations:

- The "block" variable is most likely not contributing much information to the model because of its less sum of squares value and high p-value.

- The “block” variable has no effect on the sum-of-squares for either of the 2 independent variables, indicating that it has no effect on the amount of deviation in the dependent variable that they are able to explain.

Finding the best fit model: the model that explains for the dependent variable's variation the best.

For evaluating model fit, the AIC (Akaike information criterion) is a useful tool. By balancing the difference explained against the number of elements (parameters) utilized, AIC determines the statistic value of each model. The lowest Akaike information criterion score is preferable since it indicates more information.

```r
# Find the best-fit model = the model that best explains the variation in the dependent variable.

library(AICcmodavg)

model.set <- list(one.way, two.way, interaction, blocking)
model.names <- c("one.way", "two.way", "interaction", "blocking")

aictab(model.set, modnames = model.names)
```

```
Model selection based on AICc:

            K   AICc Delta_AICc AICcWt Cum.Wt     LL
two.way     5 173.86       0.00   0.71   0.71 -81.59
blocking    7 176.93       3.08   0.15   0.86 -80.83
interaction 7 177.12       3.26   0.14   1.00 -80.92
one.way     4 186.41      12.56   0.00   1.00 -88.99

>
```

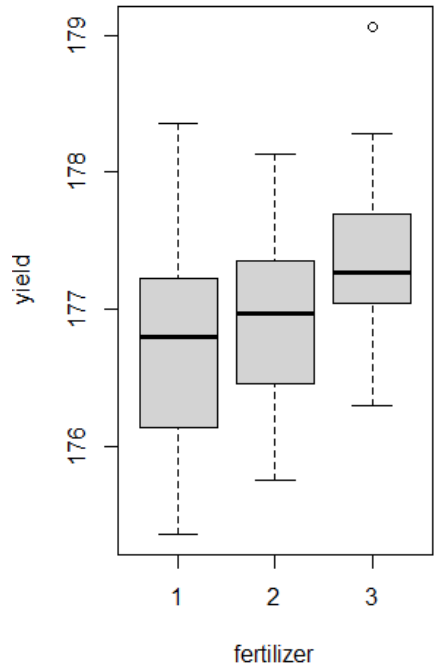
Observations:

- The best fit for the data has a lowest Akaike information criterion (AIC) score which is showed in the first row in the table.
- With the lowest Akaike information criterion value and 71% of the AIC weight, the two-way model explains 71% of the total difference in the dependent variable that the entire set of models is responsible for.
- Although the blocking term model contributes 15% more to the AIC weight than the best model, it is likely not good to include in the results.

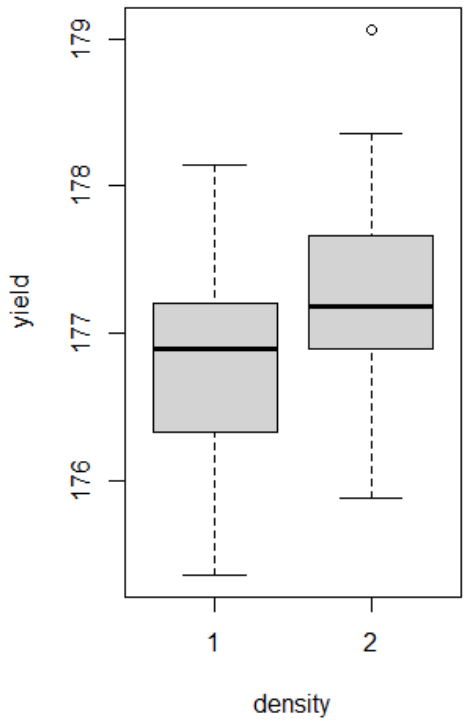
Checking Homoscedasticity (homogeneity of variances).

The basic idea of homoscedasticity states that the variances of the several groups under comparison are equal or comparable. This is a crucial assumption of parametric statistical tests since the latter are sensitive to even sample variances, which can lead to biased and skewed test findings.

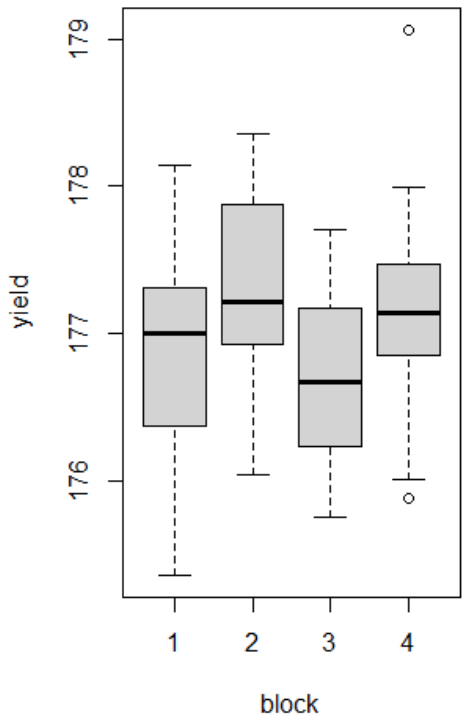
```
# Check for homoscedasticity ###############################################

# Equality of variances - homogeneity
# Boxplot - fertilizer
boxplot(yield ~ fertilizer, data = cropdata)
```



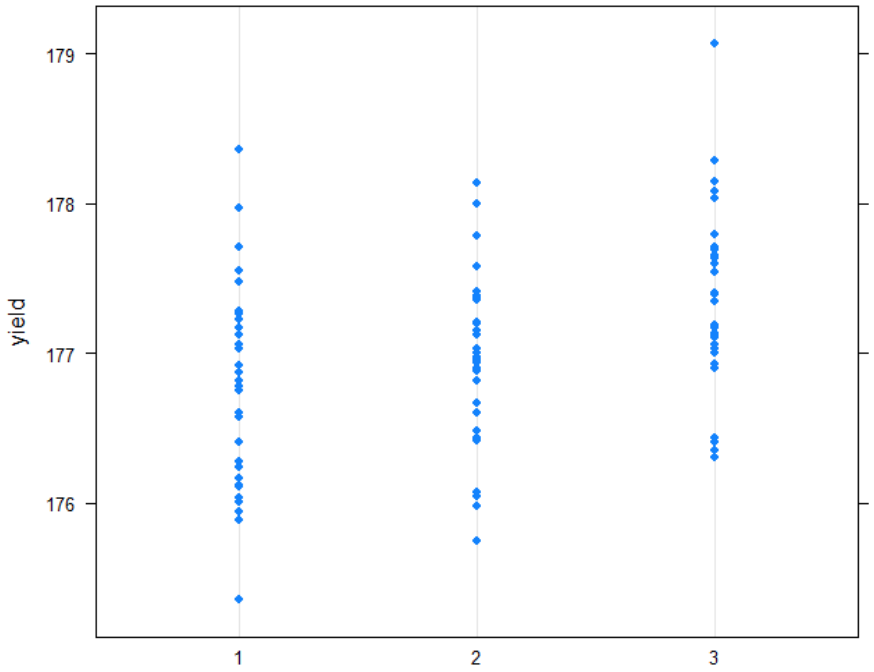
```
boxplot(yield ~ density, data = cropdata)
```




```
boxplot(yield ~ block, data = cropdata)
```




```
# Dotplot
library("lattice")

dotplot(yield ~ fertilizer, data = cropdata)
```



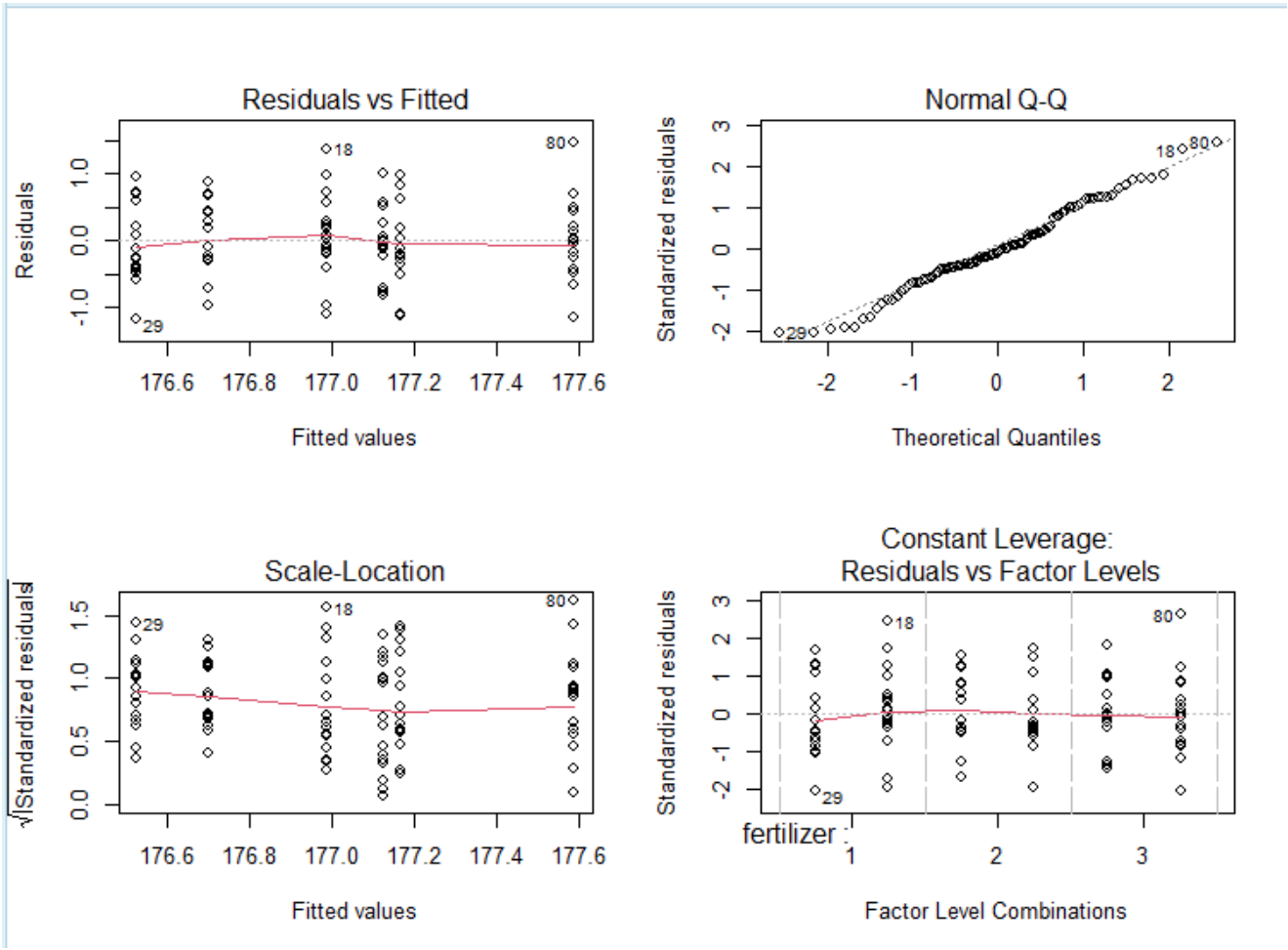
Diagnostic plots.

The residuals, or unexplained variance, are displayed in the diagnostic plots over the whole range of the observed data. There should not be any significant outliers in the model that could lead to research bias, as indicated by the red line indicating the mean of the residuals being horizontal and centered on zero (or one, in the scale-location plot).

Q-Q plot plots.

The normal Q-Q plot displays a regression between your model's actual residuals and the theoretical residuals of a perfectly-homoscedastic model; the closer the residuals are to a slope of 1, the better.

```
par(mfrow=c(2,2))
plot(two.way)
par(mfrow=c(1,1))
```



It is evident from the above that the model satisfies the homoscedasticity assumption.

Post-hoc test.

A Tukey's Honestly Significant Difference (Tukey's HSD) post-hoc test can be used for pairwise comparisons to determine whether groups are statistically different from each other. An ANOVA only indicates whether there are variations between group means, not the specific differences.

```
# post-hoc test

tukey.two.way<-TukeyHSD(two.way)

tukey.two.way
```

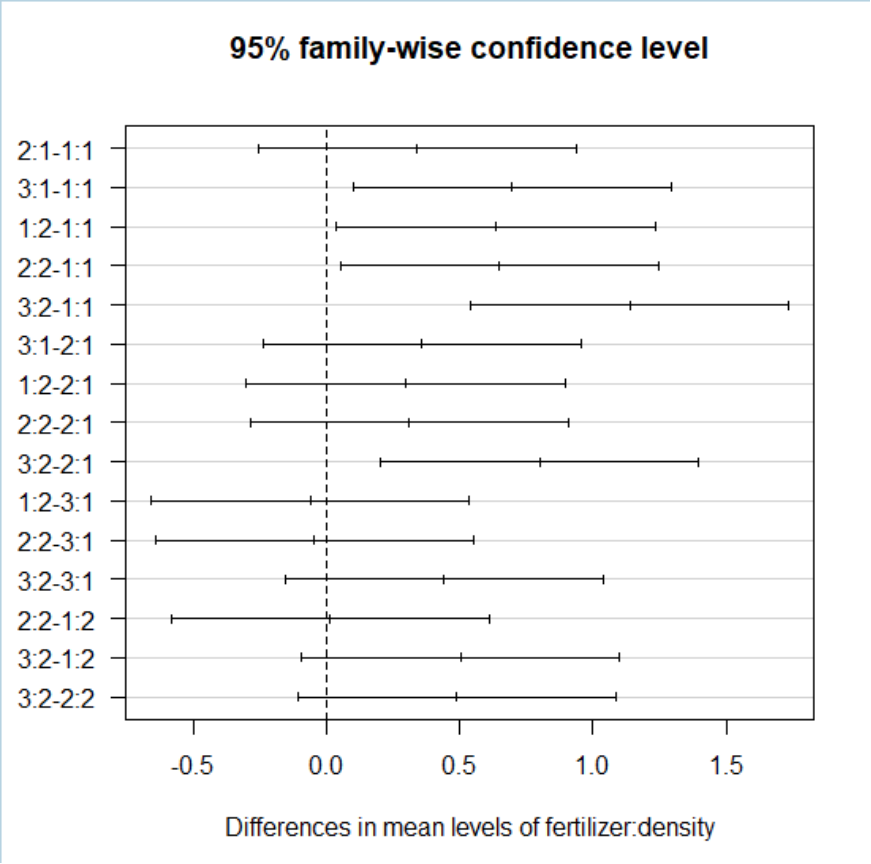
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ fertilizer + density, data = cropdata)

$fertilizer
         diff         lwr       upr     p adj
2-1 0.1761687 -0.16822506 0.5205625 0.4452958
3-1 0.5991256  0.25473179 0.9435194 0.0002219
3-2 0.4229569  0.07856306 0.7673506 0.0119381

$density
        diff       lwr       upr     p adj
2-1 0.461956 0.2275204 0.6963916 0.0001741

> |
```

It is evident from the aforementioned post-hoc test results that there are statistically significant differences ($p < 0.05$) between fertilizer types 3 and 2 as well as between fertilizer groups 3 and 1, but not between fertilizer groups 2 and 1. The two distinct planting density levels also differ significantly from one another.

- A result is unlikely to be explained by chance or random causes alone if it is statistically significant.

- The probability value, or p value, indicates a finding's statistical importance. A p value of 0.05 or less is typically regarded as statistically significant in analysis.

```
# find which group means are statistically different from one another / another ANOVA test

tukey.plot.aov<-aov(yield ~ fertilizer:density, data=cropdata)
tukey.plot.test<-TukeyHSD(tukey.plot.aov)
plot(tukey.plot.test, las = 1)
```



Wherever zero is excluded from the 95% confidence interval, there are substantial groupwise differences. Put another way, this indicates that the pairwise differences' p value is less than 0.05.

- 3:1-1:1, 1:2-1:1, 2:2-1:1, 3:2-1:1, and 3:2-2:1 are the significantly different combinations of fertilizer and planting density.

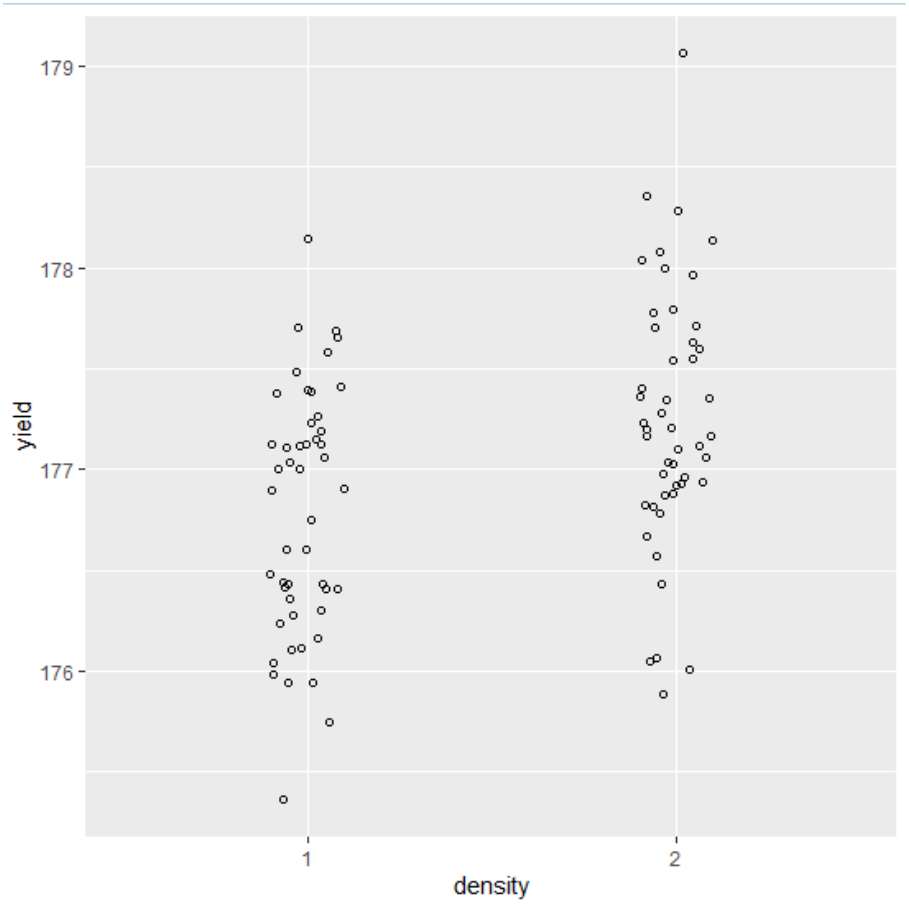
Creating a data frame including labels of groups.

```
# Make a data frame with the group labels
# We can make three labels for our graph
# A = 1:1
# B = all intermediate combinations
# C = 3:2

mean.yield.data <- cropdata %>% group_by(fertilizer, density) %>% summarise(yield = mean(yield))

mean.yield.data$group <- c("a","b","b","b","b","c")

mean.yield.data
```

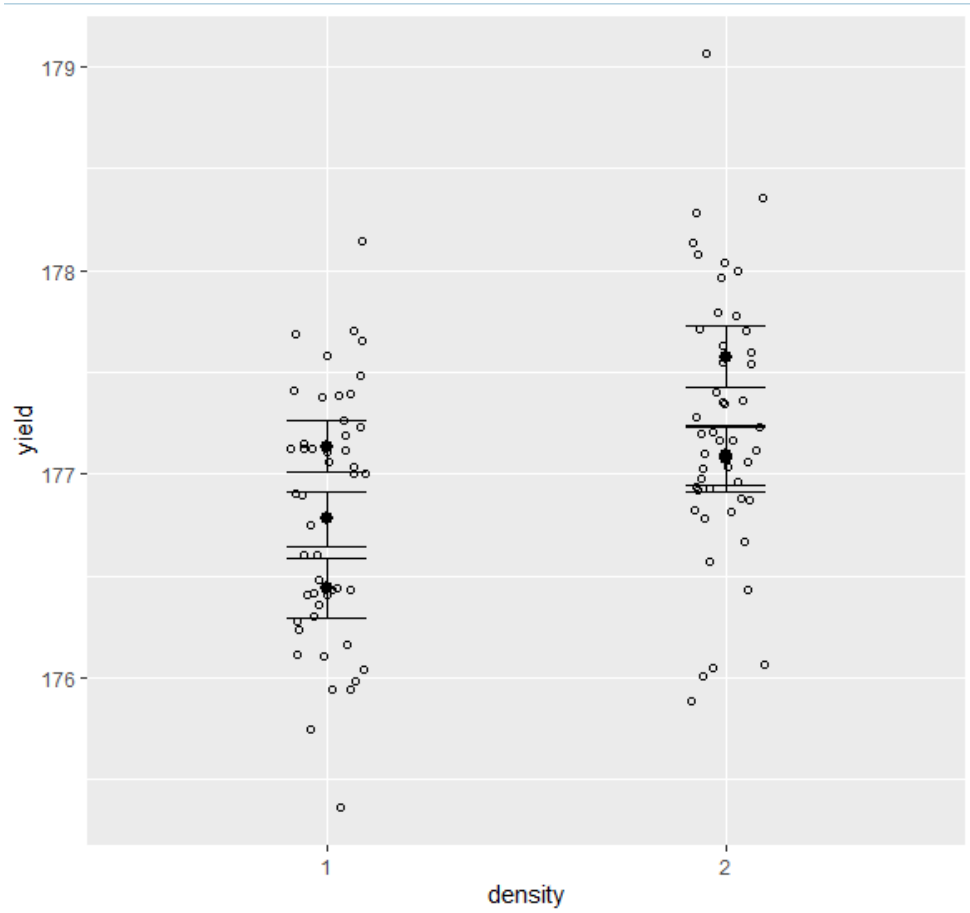
```
# A tibble: 6 x 4
# Groups:   fertilizer [3]
  fertilizer density yield group
  <fct>      <fct>   <dbl> <chr>
1 1          1        176. a
2 1          2        177. b
3 2          1        177. b
4 2          2        177. b
5 3          1        177. b
6 3          2        178. c
> |
```

Plot row data.

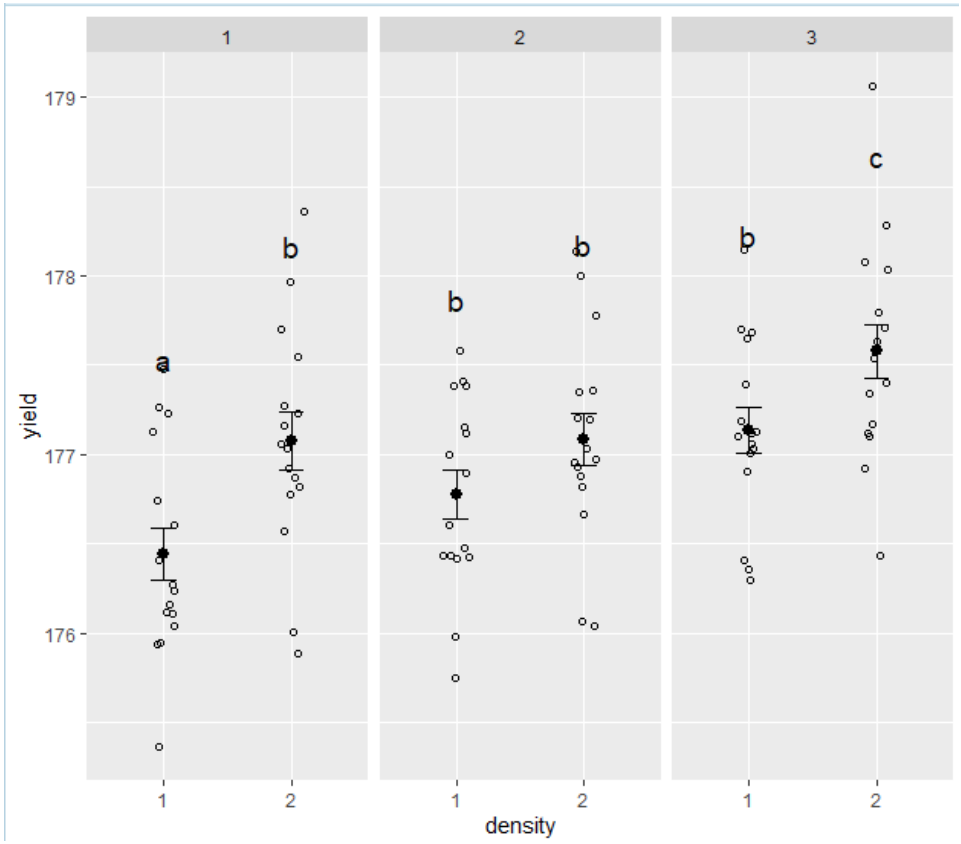
```r
#Plot the raw data

two.way.plot <- ggplot(cropdata, aes(x = density, y = yield, group=fertilizer)) +
  geom_point(cex = 1.5, pch = 1.0,position = position_jitter(w = 0.1, h = 0))

two.way.plot
```



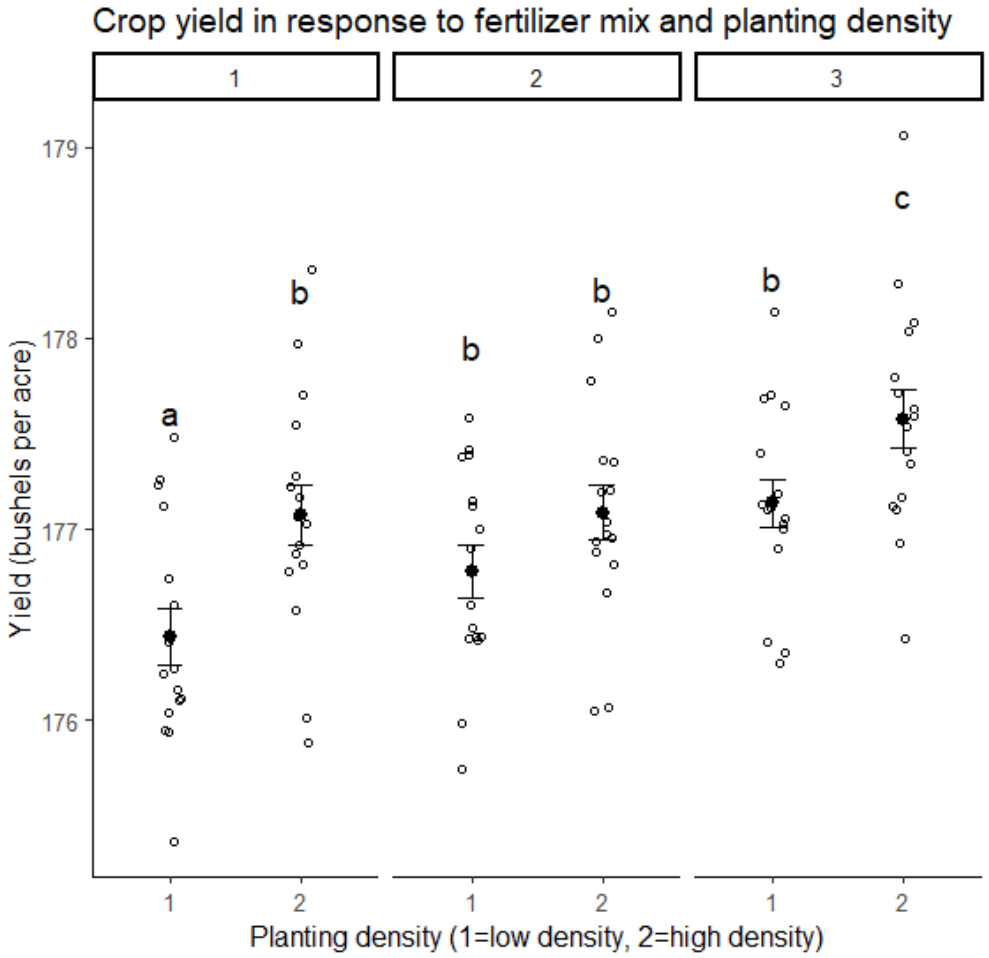
```r
# Add the means and standard errors to the graph

two.way.plot <- two.way.plot + stat_summary(fun.data = 'mean_se', geom = 'errorbar', width = 0.2) +
  stat_summary(fun.data = 'mean_se', geom = 'pointrange') + geom_point(data=mean.yield.data, aes(x=density, y=yield))

two.way.plot
```



```
# hard to read so Split up the data
# showing which groups are different from one another / use facet_wrap function

two.way.plot <- two.way.plot + geom_text(data=mean.yield.data, label=mean.yield.data$group, vjust = -8, size = 5) + facet_wrap(~ fertilizer)

two.way.plot
```



```
two.way.plot <- two.way.plot + theme_classic2() +
  labs(title = "Crop yield in response to fertilizer mix and planting density", x = "Planting density (1=low density, 2=high density)", y = "Yield (bushels per acre)")

two.way.plot
```



Observations:

- It is observed that by fertilizer type and planting density, there is a statistically significant variation in average crop yield
  (F (1) = 15.316, p < 0.001)
  (F (2) = 9.018, p < 0.001)

- Fertilizer mix 3 produced an average yield that was higher than that of mix 1 fertilizer type (0.59 bushels per acre) and mix 2 fertilizer type (0.42 bushels/acre), according to a Tukey post-hoc test. Another important factor was planting density, whereby planting density 2 produced an average yield that was 0.46 bushels per acre higher than planting density 1.

- In planting density 2, the highest yield increases were observed and mix 3 fertilizer type in a subsequent groupwise comparison, indicating that this combination of treatments was most helpful for crop growth in this test model setup.

References

Alex. (2019, Jun 1). *Linear Regression Summary(lm): Interpretting in R*. Retrieved from boostedml.com: https://boostedml.com/2019/06/linear-regression-in-r-interpreting-summarylm.html

amazon.com. (n.d.). *What's the Difference Between Linear Regression and Logistic Regression?* Retrieved from amazon.com: https://aws.amazon.com/compare/the-difference-between-linear-regression-and-logistic-regression/

Arvind Shukla. (2023, Aug 7). *Neural Networks are Decision Trees*. Retrieved from www.linkedin.com: https://www.linkedin.com/pulse/neural-networks-decision-trees-arvind-shukla/

developers.google.com. (n.d.). *Classification: ROC Curve and AUC*. Retrieved from developers.google.com: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Dr. Osman Dag. (2022, Mar 4). *How to Remove Outliers from Data in R* . Retrieved from universeofdatascience.com: https://universeofdatascience.com/how-to-remove-outliers-from-data-in-r/

Gustav Willig. (2023, Jan 17). *Decision Tree vs Logistic Regression*. Retrieved from gustavwillig.medium.com: https://gustavwillig.medium.com/decision-tree-vs-logistic-regression-1a40c58307d0

John . (2020, Jan 19). *How to Remove Outliers in R*. Retrieved from www.r-bloggers.com: https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r/

makemeanalyst.com. (n.d.). *Normal Probability Plot in R using ggplot2*. Retrieved from makemeanalyst.com: https://makemeanalyst.com/statistics-with-r/normal-probability-plot-in-r-using-ggplot2/
methodenlehre.github.io. (n.d.). *Graphics with ggplot2*. Retrieved from methodenlehre.github.io: https://methodenlehre.github.io/SGSCLM-R-course/graphics-with-ggplot2.html
Niam Zaki Zamani. (2021, Jan 18). *Linear Regression on Student Grade Prediction*. Retrieved from rpubs.com: https://rpubs.com/niamzaki/student_grade_prediction
Peter Bruce, Andrew Bruce. (n.d.). *Chapter 4. Regression and Prediction*. Retrieved from www.oreilly.com: https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch04.html
Rebecca C. Steorts. (n.d.). *Comparison of Linear Regression with K-Nearest*. Retrieved from www2.stat.duke.edu: https://www2.stat.duke.edu/~rcs46/lectures_2017/03-lr/03-knn.pdf
Rohit Kundu. (2022, Sep 13). *Confusion Matrix: How To Use It & Interpret Results [Examples]*. Retrieved from www.v7labs.com: https://www.v7labs.com/blog/confusion-matrix-guide
Safa Mulani. (2022, Aug 3). *Outlier Analysis in R - Detect and Remove Outliers*. Retrieved from www.digitalocean.com: https://www.digitalocean.com/community/tutorials/outlier-analysis-in-r
stackoverflow.com. (2023, May 3). *Difference between Logistic Regression and Decision Trees*. Retrieved from stackoverflow.com: https://stackoverflow.com/questions/76161673/difference-between-logistic-regression-and-decision-trees
typeset.io. (n.d.). *What is the difference between KNN regression and linear regression?* . Retrieved from typeset.io: https://typeset.io/questions/what-is-the-difference-between-knn-regression-and-linear-1pad331c0a
www.ibm.com. (n.d.). *What is logistic regression?* . Retrieved from www.ibm.com: https://www.ibm.com/topics/logistic-regression
www.rdocumentation.org. (n.d.). *plot: Generic X-Y Plotting*. Retrieved from www.rdocumentation.org: https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/plot
www.statisticssolutions.com. (n.d.). *Correlation (Pearson, Kendall, Spearman)*. Retrieved from www.statisticssolutions.com: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/
www.sthda.com. (n.d.). *QQ-plots: Quantile-Quantile plots - R Base Graphs* . Retrieved from www.sthda.com: http://www.sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs