Mid Trimester

Nzambuli Daniel

2024-06-22

Question 1

Given the field layout and yields in bushels per acre from an experiment on dusting wheat with sulfur to control stem rust, analyze the data using advanced statistical methods. Following a Latin square design, the treatments applied are:

> A = Dusted before rains C = Dusted once each week E = Control or check
>
> B = Dusted after rains D = Drifting once each week

The yields for each treatment are as follows:

B (4.9)	D (6.4)	E (3.3)	A (9.5)	C (11.8)
C (9.3)	A (4.0)	B (6.2)	E (5.1)	D (5.4)
D (7.6)	C (15.4)	A (6.5)	B (6.0)	E (4.6)
E (6.3)	B (7.6)	C (13.2)	D (8.6)	A (4.9)
A (9.3)	E (6.3)	D (11.8)	C (15.9)	B (7.6)

1. Evaluate the effects of the different treatments on the wheat yield, considering the potential random effects due to environmental variation.

Load libraries

```r
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
```

```r
library(agricolae)
```

Load data

taking:

> Acres – to be the data vertically on the table
>
> Fields – to be the row ie. The yield because of a treatment within the acres

```
q1_data = data.frame(Fields = rep(1:5, each = 5),
                     Acre = rep(1:5, times = 5),
                     Treatment = c('B', 'D', 'E', 'A', 'C',
                                   'C', 'A', 'B', 'E', 'D',
                                   'D', 'C', 'A', 'B', 'E',
                                   'E', 'B', 'C', 'D', 'A',
                                   'A', 'E', 'D', 'C', 'B'),
                     Yield = c(4.9, 6.4, 3.3, 9.5, 11.8,
                               9.3, 4.0, 6.2, 5.1, 5.4,
                               7.6, 15.4, 6.5, 6.0, 4.6,
                               6.3, 7.6, 13.2, 8.6, 4.9,
                               9.3, 6.3, 11.8, 15.9, 7.6))
q1_data
```

```
##    Fields Acre Treatment Yield
## 1       1    1         B   4.9
## 2       1    2         D   6.4
## 3       1    3         E   3.3
## 4       1    4         A   9.5
## 5       1    5         C  11.8
## 6       2    1         C   9.3
## 7       2    2         A   4.0
## 8       2    3         B   6.2
## 9       2    4         E   5.1
## 10      2    5         D   5.4
## 11      3    1         D   7.6
## 12      3    2         C  15.4
## 13      3    3         A   6.5
## 14      3    4         B   6.0
## 15      3    5         E   4.6
## 16      4    1         E   6.3
## 17      4    2         B   7.6
## 18      4    3         C  13.2
## 19      4    4         D   8.6
## 20      4    5         A   4.9
## 21      5    1         A   9.3
## 22      5    2         E   6.3
## 23      5    3         D  11.8
## 24      5    4         C  15.9
## 25      5    5         B   7.6
```

Summary of Yield by treatment

```
treatment_total = q1_data %>%
  group_by(Treatment) %>%
  summarize(Total_yield = sum(Yield))
treatment_total
```

```
## # A tibble: 5 × 2
##   Treatment Total_yield
##   <chr>           <dbl>
## 1 A                34.2
```

```
## 2 B                32.3
## 3 C                65.6
## 4 D                39.8
## 5 E                25.6
```

Correction Factor

$$Correction\ Factor = \frac{Grand\ Total}{N}$$

```
total_yield = sum(treatment_total$Total_yield)
cat("The yields in bushels per acre from an experiment is", total_yield)
```

```
## The yields in bushels per acre from an experiment is 197.5
```

```
cor_fct_q1 = total_yield ^2 / nrow(q1_data)
cat("The correction factor is:\n\t\t", cor_fct_q1)
```

```
## The correction factor is:
##       1560.25
```

Total sum of squares

$$total\ sum\ squares = -Correction\ factor + \sum Yield^2$$

```
tot_ss_q1 = sum(q1_data$Yield^2) - cor_fct_q1
cat("The total sum of squares is:\n\t\t", tot_ss_q1)
```

```
## The total sum of squares is:
##       281.18
```

Row sum of squares

$$rss = \frac{\sum rows^2}{No.\ of\ rows} - CF$$

```
Field_tot_q1 = q1_data %>%
  group_by(Fields) %>%
  summarize(Field_Yield = sum(Yield))
Field_tot_q1
```

```
## # A tibble: 5 × 2
##   Fields Field_Yield
##    <int>       <dbl>
## 1      1        35.9
## 2      2        30
## 3      3        40.1
## 4      4        40.6
## 5      5        50.9
```

```
Rss_q1 = (sum(Field_tot_q1$Field_Yield^2) / nrow(Field_tot_q1)) - cor_fct_q1
```

```
cat("The row sum of squares is:\n\t\t", Rss_q1)
```

```
## The row sum of squares is:
##       46.948
```

Column sum of squares

$$col\ sum\ squares = \frac{\sum column}{No.\ of\ columns} - CF$$

```
Acre_tot_q1 = q1_data %>%
  group_by(Acre) %>%
  summarize(Acre_Yield = sum(Yield))
Acre_tot_q1
```

```
## # A tibble: 5 × 2
##    Acre Acre_Yield
##   <int>      <dbl>
## 1     1       37.4
## 2     2       39.7
## 3     3       41
## 4     4       45.1
## 5     5       34.3
```

```
Css_q1 = (sum(Acre_tot_q1$Acre_Yield^2) / nrow(Acre_tot_q1)) - cor_fct_q1

cat("The column sum of squares is:\n\t\t", Css_q1)
```

```
## The column sum of squares is:
##       13.02
```

Treatment Sum of Squares

$$treatment\ SS = \frac{\sum treatment\ total^2}{No.\ of\ treatments} - CF$$

```
Tss_q1 = sum(treatment_total$Total_yield^2)/ nrow(treatment_total) - cor_fct_q1
cat("The treatment sum of squares is:\n\t\t", Tss_q1)
```

```
## The treatment sum of squares is:
##       190.888
```

Error sum of squares

$$Ess = total\ ss - row\ ss - column\ ss - treatment\ ss$$

```
error_ss_q1 = tot_ss_q1 - Rss_q1 - Css_q1 - Tss_q1
cat("The error sum of squares is:\n\t\t", error_ss_q1)
```

```
## The error sum of squares is:
##       30.324
```

Degree of freedom

```
df_row = nrow(Field_tot_q1) - 1
df_col = nrow(Acre_tot_q1) - 1
df_treat = nrow(treatment_total) - 1
df_total = nrow(q1_data) -1
df_err = df_total - df_row - df_col - df_treat
```

Mean Squares

```
ms_row = Rss_q1 / df_row
ms_col = Css_q1/ df_col
ms_treat = Tss_q1/ df_treat
ms_err = error_ss_q1/df_err
```

F-values

```
f_treat = ms_treat/ ms_err
cat("The f-statistic is:\n\t\t", f_treat)
```

```
## The f-statistic is:
##       18.88484
```

P- value

```
p_treat = pf(f_treat, df_treat, df_total, lower.tail = F)
cat("The p-value for treatment is:\n\t\t",p_treat)
```

```
## The p-value for treatment is:
##       3.901254e-07
```

```
p_acre = pf(ms_col/ms_err, df_col, df_total, lower.tail = F)
cat("The p-value for Acre is:\n\t\t",p_acre)
```

```
## The p-value for Acre is:
##       0.3025007
```

```
p_field = pf(ms_row/ms_err, df_row, df_total, lower.tail = F)
cat("The p-value for Field is:\n\t\t",p_field)
```

```
## The p-value for Field is:
##       0.006414349
```

Interpretation

> Taking a significance level of `0.05` for H_0 there is no statistically significant difference in the mean bushel yield and H_1 there is at least one mean yield that is statistically and significantly different from the rest of the means

> - **Acres –** Fail to reject H_0 since the p-value is larger than `0.05` . Conclude that the mean yields given different acres has no statistically significant difference.
> - **Field –** Reject H_0 since the p-value is lower than `0.05`. Conclude that one of the mean yields given different Fields is statistically significant from the other mean yields
> - **Treatment –** Reject H_0 since the p-value is lower than `0.05`. Conclude that one of the mean yields given Treatment is statistically significant from the other mean yield

Using AOV in R

```
q1_data$Treatment = as.factor(q1_data$Treatment)
q1_aov = aov(Yield~Fields + Acre + Treatment, data = q1_data)
summary(q1_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Fields       1  32.97   32.97  10.354  0.00477 ** 
## Acre         1   0.01    0.01   0.004  0.95014    
## Treatment    4 190.89   47.72  14.988 1.48e-05 ***
## Residuals   18  57.31    3.18                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

> At a significance level of `0.05` for H_0 there is no statistically significant difference in the mean bushel yield and H_1 there is at least one mean yield that is statistically and significantly different from the rest of the means
>
> - **Acres –** Fail to reject H_0 since the p-value is larger than `0.05` . Conclude that the mean yields given different acres has no statistically significant difference.
> - **Field –** Reject H_0 since the p-value is lower than `0.05`. Conclude that one of the mean yields given different Fields is statistically significant from the other mean yields
> - **Treatment –** Reject H_0 since the p-value is lower than `0.05`. Conclude that one of the mean yields given Treatment is statistically significant from the other mean yield
>
> This mirrors the calculated conclusion.

2. if statistically necessary, determine which specific treatments significantly differ from each other in terms of yield. Include a visual output to compare the treatments.

Treatment has at least one mean that is statistically and significantly different from the other mean Yield

This prompts the need for a post-hoc analysis to identify which treatment had the different yield

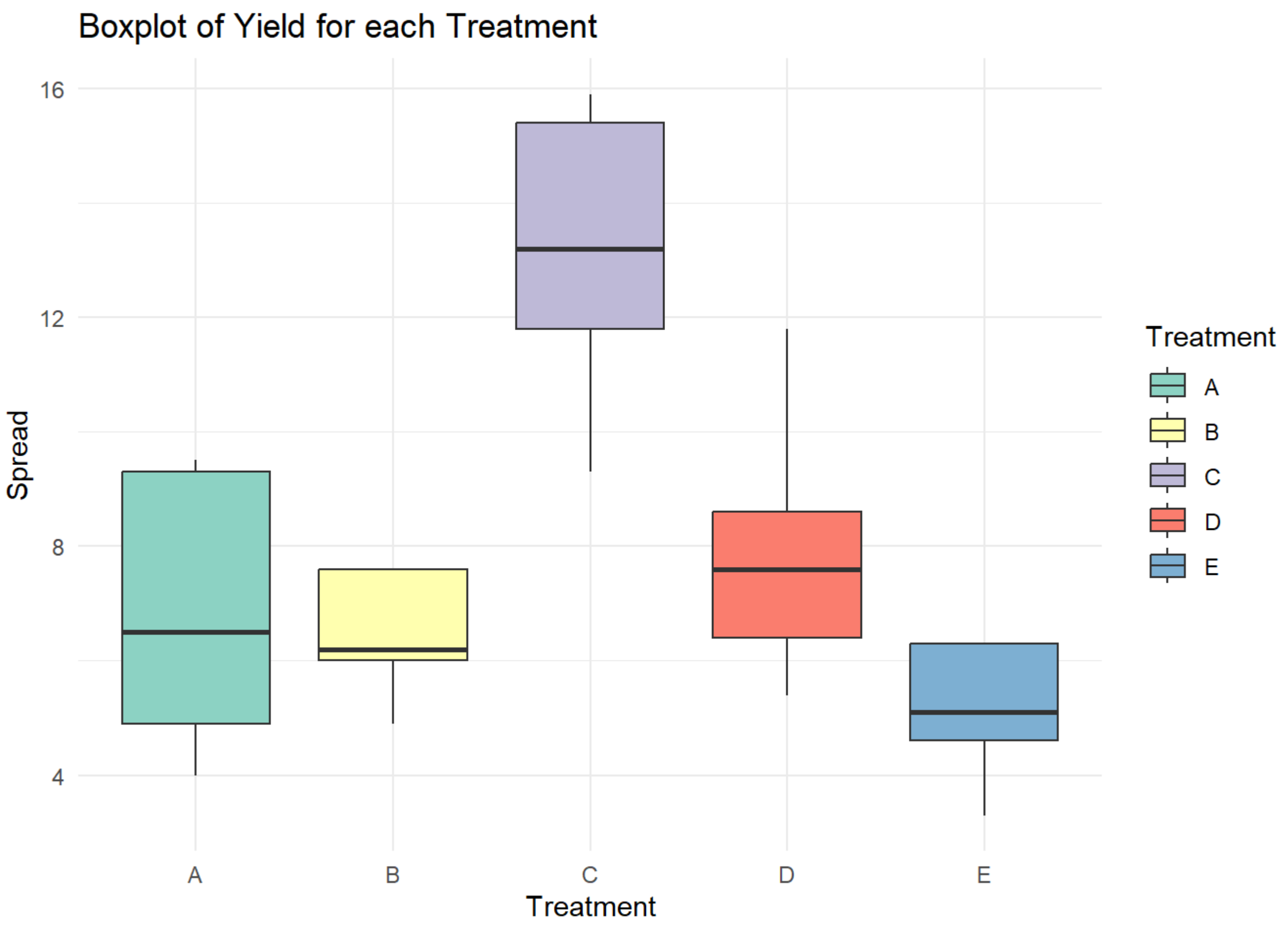
```
lsd_q1 = LSD.test(aov(Yield~Treatment, data = q1_data), "Treatment", p.adj = "none")
lsd_q1$groups
```

```
##   Yield groups
## C 13.12      a
## D  7.96      b
```

```
## A  6.84     bc
## B  6.46     bc
## E  5.12      c
```

```
library(ggplot2)

ggplot(q1_data, aes(x = Treatment, y = Yield, fill = Treatment))+
  geom_boxplot()+
  labs(
    title = "Boxplot of Yield for each Treatment",
    x = "Treatment",
    y = "Spread"
  )+
  theme_minimal()+
  scale_fill_brewer(palette = "Set3")
```



Explanation

> There are 3 distinguishable groups in the data based on Yield for a given treatment
>
> 1. **A** has the largest inter quartile range IQR indicating large variability in Yield.
> 2. **E and B** have the lowest IQR meaning the yield is more consistent for these two treatments
> 3. **C** appears to have a higher overall Yield. It may be the best performing treatment. However it also has high variation in its Yield
>
> - **C** – has a mean yield of `13.12 bushels per acre` and is in its own group which is highlighted in the box-plot with its median yield being above `12`. Its lower and upper quartiles are also above the the quartiles of the other treatments

- **D, A, B** can be grouped together as they have a mean Yield that is close to each other. Investigating further

```
sd(c(7.96, 6.84, 6.46))
```

```
## [1] 0.779829
```

```
median( q1_data[q1_data$Treatment %in% c("A", "B","D"), "Yield"])
```

```
## [1] 6.5
```

> the standard deviation of the means of **A, D, B** is less than `1` this indicates that the mean values have minimal fluctuation between them.
>
> Checking the box-plot the values the IQR for A overlaps with the IQR of B and D. The median value of the yield from the three treatments is also within the range $(6, 8)$.
>
> - **A, B, E** have similar groupings with mean yields that fall in the range $(5, 7)$

```
sd(c(6.84, 6.46, 5.12))
```

```
## [1] 0.9035486
```

```
median( q1_data[q1_data$Treatment %in% c("A", "B","D"), "Yield"])
```

```
## [1] 6.5
```

> The IQR spread of the Yield from A also blankets the yield from both B and E . They have a median Yield $\approx$ less than `6.5`

3. Interpret the results in the context of agricultural practices for controlling stem rust, considering both statistical significance and practical implications.

From the analysis:

1. **Dusting once each week (C)** – has the highest yield suggesting regular application of sulfur is the most effective method to control stem rust and maximize wheat yields.
2. The events **Dusting before rains (A)** and **Dusting after rains (B)** have lower yields. This indicates that timing sulfur application around a rain event is less effective in improving plant yields
3. A and B have similar mean yields similar to **The control (E)**. This enforces the claim that dusting around rain events may not improve the overall yield of wheat by controlling stem rust
4. **Drifting once each week (D)** – This has higher yields than dusting after rain. However, the improvement is not high to justify this as a viable option.

Advice

- For optimal control of stem rust and to maximize wheat yield, adopting a practice of dusting once each week (C) is recommended. It can be augmented by improving the plant resistance to stem rust by genetic modification GM. This is because the
- Practices based solely on dusting around rain events (either before or after) is less effective and should be avoided if the goal is to maximize yield

Question 2

In a digestion trial carried out with 6 shorthorn steers, each animal received one of 6 rations in 6 successive periods following a Latin square design. The coefficients of digestibility of nitrogen for each steer and period are as follows:

steer	period 1	period 2	period 3	period 4	period 5	period 6
1	61.1 (B)	69.3 (D)	67.6 (C)	61.9 (F)	58.8 (A)	65.2 (E)
2	56.9 (A)	59.1 (F)	64.0 (D)	61.0 (C)	65.7 (E)	56.6 (B)
3	66.5 (C)	62.2 (A)	61.1 (B)	66.2 (E)	62.0 (F)	62.2 (D)
4	66.7 (E)	67.4 (B)	65.1 (F)	65.1 (D)	69.6 (C)	52.7 (A)
5	67.8 (D)	64.7 (C)	63.6 (E)	53.2 (A)	61.7 (B)	62.0 (F)
6	71.4 (F)	67.5 (E)	55.8 (A)	63.2 (B)	68.0 (D)	62.9 (C)

The rations consist of: - A: Hay alone - B: Various mixtures of hay and barley (Rations B, C, D, E, F)

1. Assess the effects of different rations on the digestibility of nitrogen. Include an analysis of variance (ANOVA) to test the significance of the ration effects, period effects, and steer effects.

Load the data

```
q2_data = data.frame(Steer = rep(1:6, each = 6),
                     Period = rep(1:6, times = 6),
                     Treatment = c('B', 'D', 'C', 'F', 'A', 'E',
                                   'A', 'F', 'D', 'C', 'E', 'B',
                                   'C', 'A', 'B', 'E', 'F', 'D',
                                   'E', 'B', 'F', 'D', 'C', 'A',
                                   'D', 'C', 'E', 'A', 'B', 'F',
                                   'F', 'E', 'A', 'B', 'D', 'C'),
                     Digestion = c(61.1, 69.3, 67.6, 61.6, 58.8,65.2,
                                   56.9, 59.1, 64.0, 61.0, 65.7, 56.6,
                                   66.5, 62.2, 61.1, 66.2, 62.0, 62.2,
                                   66.7, 67.4, 63.6, 53.2, 61.7, 62.0,
                                   67.8, 64.7, 63.6, 53.2, 61.7, 62.0,
                                   71.4, 67.5, 55.8, 63.2, 68.0, 62.9))
q2_data
```

```
##    Steer Period Treatment Digestion
## 1      1      1         B      61.1
## 2      1      2         D      69.3
## 3      1      3         C      67.6
## 4      1      4         F      61.6
```

```
## 5      1      5          A      58.8
## 6      1      6          E      65.2
## 7      2      1          A      56.9
## 8      2      2          F      59.1
## 9      2      3          D      64.0
## 10     2      4          C      61.0
## 11     2      5          E      65.7
## 12     2      6          B      56.6
## 13     3      1          C      66.5
## 14     3      2          A      62.2
## 15     3      3          B      61.1
## 16     3      4          E      66.2
## 17     3      5          F      62.0
## 18     3      6          D      62.2
## 19     4      1          E      66.7
## 20     4      2          B      67.4
## 21     4      3          F      63.6
## 22     4      4          D      53.2
## 23     4      5          C      61.7
## 24     4      6          A      62.0
## 25     5      1          D      67.8
## 26     5      2          C      64.7
## 27     5      3          E      63.6
## 28     5      4          A      53.2
## 29     5      5          B      61.7
## 30     5      6          F      62.0
## 31     6      1          F      71.4
## 32     6      2          E      67.5
## 33     6      3          A      55.8
## 34     6      4          B      63.2
## 35     6      5          D      68.0
## 36     6      6          C      62.9
```

Anova

```
q2_data$Treatment = as.factor(q2_data$Treatment)
str(q2_data)
```

```
## 'data.frame':    36 obs. of  4 variables:
##  $ Steer    : int  1 1 1 1 1 1 2 2 2 2 ...
##  $ Period   : int  1 2 3 4 5 6 1 2 3 4 ...
##  $ Treatment: Factor w/ 6 levels "A","B","C","D",..: 2 4 3 6 1 5 1 6 4 3 ...
##  $ Digestion: num  61.1 69.3 67.6 61.6 58.8 65.2 56.9 59.1 64 61 ...
```

```
q2_aov = aov(Digestion~Steer + Period + Treatment, data = q2_data)
summary(q2_aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)  
## Steer        1    5.8    5.83   0.434 0.5153  
## Period       1   54.8   54.79   4.079 0.0531 .
## Treatment    5  210.5   42.09   3.133 0.0227 *
## Residuals   28  376.2   13.43                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis

H_0 there is no statistically significant difference in the digestion coefficient of nitrigen for short horn steers.

H_1 there is at-least one digestion coefficient of nitrogen that is statistically and significant different in the short horn steers

Findings

based on a significance level of `0.1`

- **Treatment** – Given that the p-value is less than the significant value. We reject H_0 and conclude that at-least one mean digestion coefficient of nitrogen is different for the 6 rations.
- **Period** – Given that the p-value is less than the significant value. We reject H_0 and conclude that at-least one mean digestion coefficient of nitrogen is different for the 6 periods.
- **Steers** – Given that the p-value is greater than the significant value. We fail to reject H_0 and conclude that the mean digestion coefficient of nitrogen is not statistically or significantly different for all the 6 steers.

2. Where necessary, identify which specific rations differ significantly from each other in terms of digestibility. Include a visual output to compare the treatments.

Investigation for what features cause the significant difference in the mean digestion of nitrogen for

- rations
- period

is needed

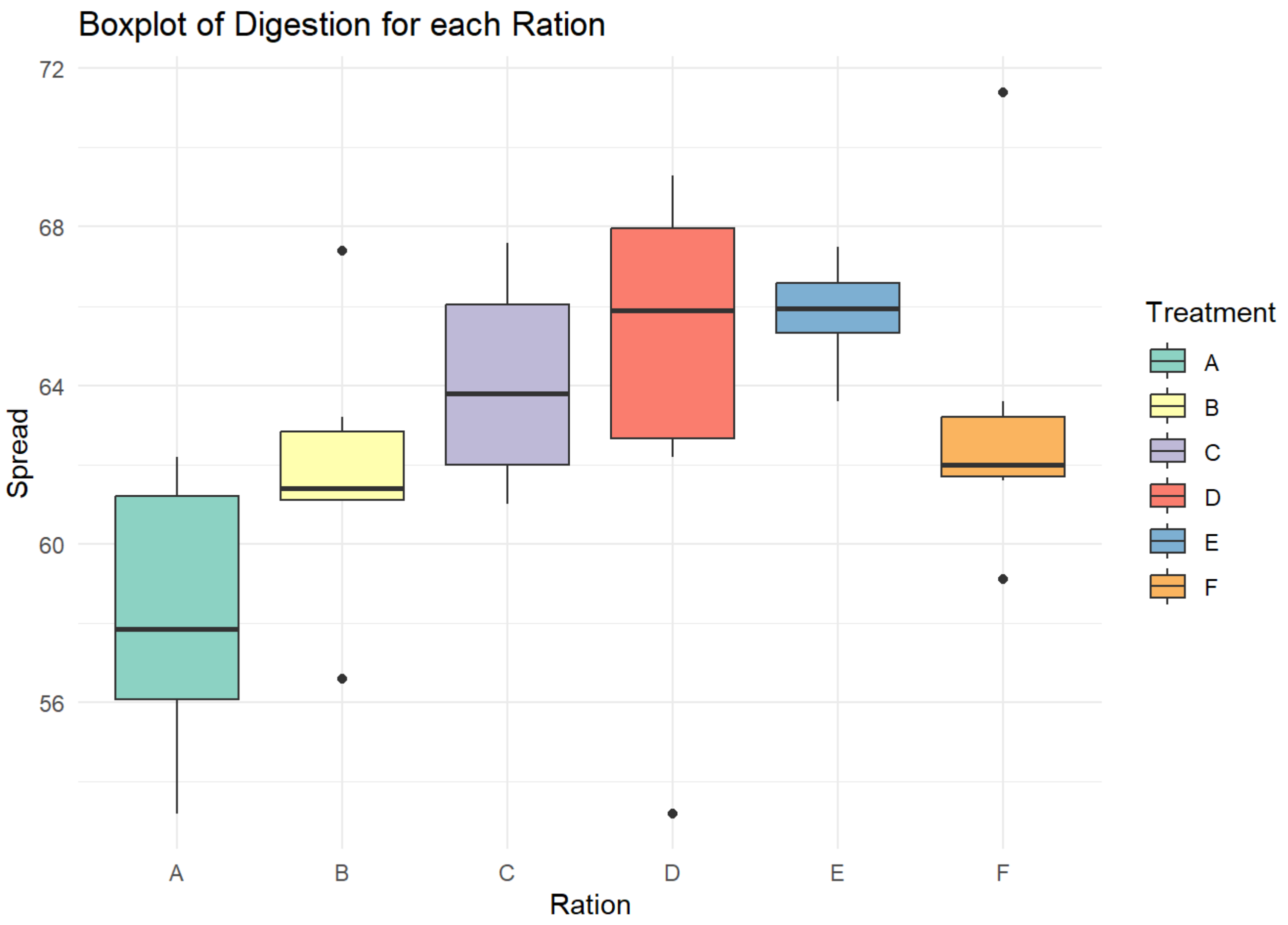
Pos-hoc

```
lsd_q2 = LSD.test(aov(Digestion~Treatment, data = q2_data), "Treatment", p.adj = "none")
lsd_q2$groups
```

```
##   Digestion groups
## E  65.81667      a
## D  64.08333      a
## C  64.06667      a
## F  63.28333      a
## B  61.85000     ab
## A  58.15000      b
```

Plot the relationship

```
ggplot(q2_data, aes(x = Treatment, y = Digestion, fill = Treatment))+
  geom_boxplot()+
  labs(
    title = "Boxplot of Digestion for each Ration",
    x = "Ration",
    y = "Spread"
  )+
  theme_minimal()+
  scale_fill_brewer(palette = "Set3")
```



Observation

1. There are two groups one with **E, D, C, F** and **B** and the other containing **A** and **B only**
2. There are extreme values in **B** , **D** and **F**
3. **A** has the lowest values. In spite of A and B being grouped together, most of the values in A fall in between the **lower extreme** and the **first quartile** of B
4. A appears to be the most significantly different value from the other rations based on digestion of nitrogen
5. E has the **highest median with the narrowest IQR** indicating that it has the most consistent digestion coefficient of nitrogen

3. Discuss the implications of your findings for nutritional strategies in cattle farming, highlighting any potential benefits or drawbacks of specific rations based on the digestibility results.

to achieve the most consistent results in steer digestion

- Rations **C** and **E** should be given. This ensures consistent digestion of nitrogen while also getting relatively high proportions of nitrogen absorption in the digestive tract of the steers compared to the nitrogen intake

when the rations **C** and **E** are not accessible **B, D and F** can be used as substitutes. Although they have some extreme values they do produce consistently high median nitrogen absorption for nitrogen intake.

Avoid

- Ration A should be avoided because it produces the lowest overall nitrogen digestion coefficients.
- However, if it is an essential ration it should be supplemented with one of the higher digestible rations to try and offset its low coefficient of digestible nitrogen