R Notebook

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.3
```

```
fertilizer = c(
  rep("Blend X", 5),
  rep("Blend Y", 5),
  rep("Blend Z", 5)
)

df = data.frame(
  Fertilizer = fertilizer,
  Wheat = c(
    123, 156, 112, 100, 168,
    135, 130, 176, 120, 155,
    156, 180, 147, 146, 193
    ),
  Corn = c(
    128, 150, 174, 116, 109,
    175, 132, 120, 187, 184,
    186, 138, 178, 176, 190
    ),
  Soy = c(
    166, 178, 187, 153, 195,
    140, 145, 159, 131, 126,
    185, 206, 188, 165, 188
    ),
  Rice = c(
    151, 125, 117, 155, 158,
    167, 183, 142, 167, 168,
    175, 173, 154, 191, 169
    )
)

df_graph <- df |>
  pivot_longer(
    cols = Wheat:Rice,
    names_to = "Crop",
    values_to = "Yield"
  )

df_graph
```

```
## # A tibble: 60 x 3
##    Fertilizer Crop  Yield
```

```
##    <chr>      <chr> <dbl>
##  1 Blend X    Wheat   123
##  2 Blend X    Corn    128
##  3 Blend X    Soy     166
##  4 Blend X    Rice    151
##  5 Blend X    Wheat   156
##  6 Blend X    Corn    150
##  7 Blend X    Soy     178
##  8 Blend X    Rice    125
##  9 Blend X    Wheat   112
## 10 Blend X    Corn    174
## # i 50 more rows
```

Assumptions

Assumption #1: You have one dependent variable that is measured at the continuous level.

Assumption #2: You have one independent variable that consists of three or more categorical, independent groups.

Assumption #3: You should have independence of observations, which means that there is no relationship between the observations in each group of the independent variable or among the groups themselves.

Assumption #4: There should be no significant outliers in the three or more groups of your independent variable in terms of the dependent variable.

Assumption #5: Your dependent variable should be approximately normally distributed for each group of the independent variable. Remark: Given the P-values for all treatment types are $p > 0.05$, then the weights are approximately normally distributed for all groups.

Assumption #6. You have homogeneity of variances (i.e., the variance of the dependent variable is equal in each group of your independent variable).

Hypothesis

Null hypothesis: There is no significant interaction effect on yield between fertilizer and crop.

Alternative hypothesis: There is significant interaction effect on yield between fertilizer and crop.

Checking of Assumptions

Assumption #1: You have one dependent variable that is measured at the continuous level.

Remark: The dependent variable (Yield) is continuous

Assumption #2: You have two independent variables that consist of two or more categorical, independent groups.

Remark: The two independent variables are Fertilizers and Crops and are independent groups.

Assumption #3: You should have independence of observations, which means that there is no relationship between the observations in each group of the independent variable or among the groups themselves.

Remark: There exists independence of observations between all groups of fertilizers, as well as all groups of crops.

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'tibble' was built under R version 4.4.3

## Warning: package 'readr' was built under R version 4.4.3

## Warning: package 'purrr' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.4
## v forcats   1.0.1     v readr     2.1.5
## v ggplot2   4.0.1     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tidyquant)
```

```
## Warning: package 'tidyquant' was built under R version 4.4.3

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Warning: package 'xts' was built under R version 4.4.3

## Warning: package 'zoo' was built under R version 4.4.3

## Warning: package 'quantmod' was built under R version 4.4.3

## Warning: package 'TTR' was built under R version 4.4.3

## Warning: package 'PerformanceAnalytics' was built under R version 4.4.3
```

```
## -- Attaching core tidyquant packages ----------------------- tidyquant 1.0.11 --
## v PerformanceAnalytics 2.0.8      v TTR                0.24.4
## v quantmod             0.4.28     v xts                0.14.1
## -- Conflicts ------------------------------------------ tidyquant_conflicts() --
## x zoo::as.Date()                 masks base::as.Date()
## x zoo::as.Date.numeric()         masks base::as.Date.numeric()
## x dplyr::filter()                masks stats::filter()
## x xts::first()                   masks dplyr::first()
## x dplyr::lag()                   masks stats::lag()
## x xts::last()                    masks dplyr::last()
## x PerformanceAnalytics::legend() masks graphics::legend()
## x quantmod::summary()            masks base::summary()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggdist)
```

```
## Warning: package 'ggdist' was built under R version 4.4.3
```

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.4.3
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.4.3

##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(effectsize)
```

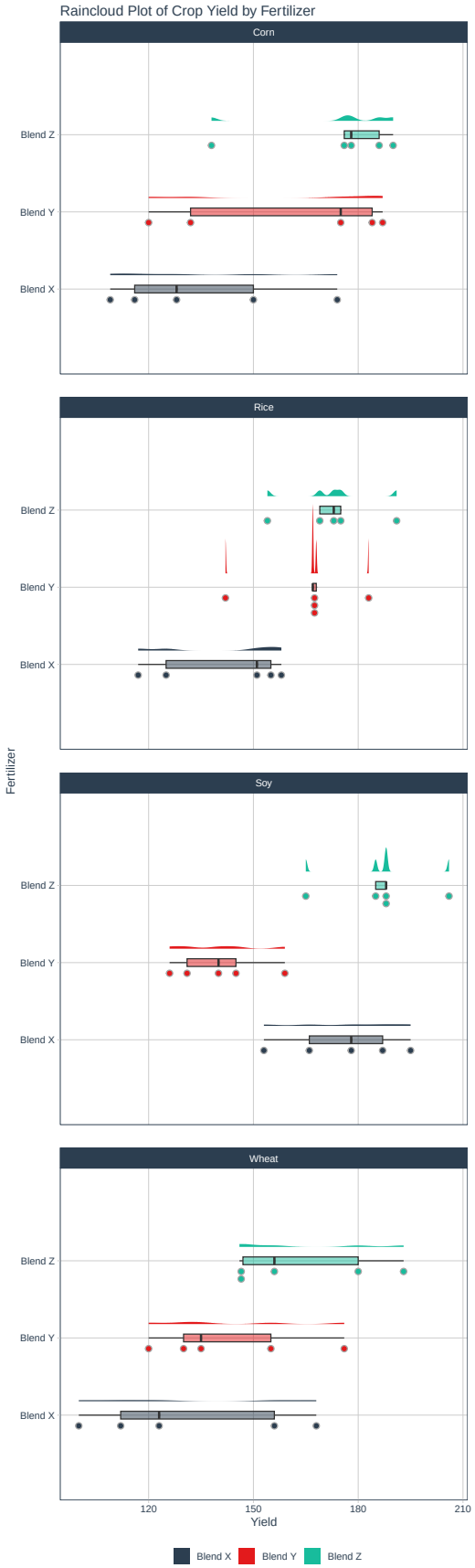
```
## Warning: package 'effectsize' was built under R version 4.4.3
```

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.4.3
```

```r
df_graph %>%
  ggplot(aes(x=factor(Fertilizer), y=Yield, fill=Fertilizer)) +

  stat_halfeye(
    adjust = 0.3,
    justification = -0.2,
    .width = 0,
    point_colour = NA,
    width = 1
  ) +
  geom_boxplot(
    width = 0.1,
    outlier.color = NA,
    alpha = 0.5,
    show.legend = FALSE
  ) +
  stat_dots(
    side = "left",
    justification = 1.1,
    binwidth = 2,
    show.legend = FALSE
  ) +
  facet_wrap(~ Crop, ncol=1, scales="free_y") +
  scale_fill_tq() +
  theme_tq() +
  labs(
    title = "Raincloud Plot of Crop Yield by Fertilizer",
    x = "Fertilizer",
    y = "Yield",
    fill = ""
  ) +
  coord_flip()
```


Raincloud Plot of Crop Yield by Fertilizer
Corn
Rice
Soy
Wheat
Blend Z
Blend Y
Blend X
Fertilizer
120
150
180
210
Yield
Blend X
Blend Y
Blend Z

Assumption #4: There should be no significant outliers in the three or more groups of your independent variables in terms of the dependent variable.

Remark: By visual inspection of the raincloud plot of crop yields, there are no significant outliers for any of the fertilizer groups in any crop groups in terms of the dependent variable.

```
get_stats = function(x){
  valid = sum((!is.na(x)))
  missing = sum((is.na(x)))
  mean_x = (mean(x, na.rm=TRUE))
  sd_x = (sd(x, na.rm=TRUE))
  skew_x = moments::skewness(x, na.rm = TRUE)
  se_skew = sqrt(6/valid)
  kurt_x = moments::skewness(x, na.rm = TRUE)
  se_kurt = sqrt(24/valid)
  shap_wilk_w = shapiro.test(x)$statistic
  shap_wilk_p = shapiro.test(x)$p.value

  return(data.frame(
    Valid = valid,
    Missing = missing,
    Mean = mean_x,
    SD = sd_x,
    Skewness = skew_x,
    SE_Skewness = se_skew,
    Kurtosis = kurt_x,
    SE_Kurtosis = se_kurt,
    Shapiro_W = shap_wilk_w,
    Shapiro_P = shap_wilk_p
  ))
}

descriptive_table = df_graph %>%
  group_by(Fertilizer, Crop) %>%
  summarise(get_stats(Yield), .groups='drop')

descriptive_table
```

```
## # A tibble: 12 x 12
##    Fertilizer Crop  Valid Missing  Mean    SD Skewness SE_Skewness Kurtosis
##    <chr>      <chr> <int>   <int> <dbl> <dbl>    <dbl>       <dbl>    <dbl>
##  1 Blend X    Corn      5       0  135.  26.6   0.512         1.10   0.512
##  2 Blend X    Rice      5       0  141.  18.8  -0.424         1.10  -0.424
##  3 Blend X    Soy       5       0  176.  16.7  -0.258         1.10  -0.258
##  4 Blend X    Wheat     5       0  132.  29.1   0.239         1.10   0.239
##  5 Blend Y    Corn      5       0  160.  31.3  -0.405         1.10  -0.405
##  6 Blend Y    Rice      5       0  165.  14.7  -0.640         1.10  -0.640
##  7 Blend Y    Soy       5       0  140.  12.9   0.408         1.10   0.408
##  8 Blend Y    Wheat     5       0  143.  22.3   0.542         1.10   0.542
##  9 Blend Z    Corn      5       0  174.  20.7  -1.22          1.10  -1.22
## 10 Blend Z    Rice      5       0  172.  13.3   0.0220        1.10   0.0220
## 11 Blend Z    Soy       5       0  186.  14.6  -0.205         1.10  -0.205
## 12 Blend Z    Wheat     5       0  164.  21.1   0.452         1.10   0.452
## # i 3 more variables: SE_Kurtosis <dbl>, Shapiro_W <dbl>, Shapiro_P <dbl>
```

Assumption #5: Your dependent variable should be approximately normally distributed for each group of the independent variable.

Remark: Given the P-values for all fertilizer types and crop groups are $p > 0.05$, then the yields are approximately normally distributed for all groups.

```
df_graph %>%
  group_by(Crop) %>%
  group_map(~ leveneTest(Yield ~ Fertilizer, data=.x))
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## [[1]]
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.4506 0.6476
##       12
##
## [[2]]
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2   0.371 0.6977
##       12
##
## [[3]]
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.2405 0.7899
##       12
##
## [[4]]
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.2546 0.7793
##       12
```

Assumption #6. You have homogeneity of variances (i.e., the variance of the dependent variable is equal in each group of your independent variable).

Remark: With $p > 0.05$ from all the Levene Tests, we can say that there is equal variance of the dependent variable across all fertilizer type groups.

Computation

```r
df_anova = aov(Yield ~ Fertilizer * Crop, data = df_graph)
summary(df_anova)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## Fertilizer       2   8783    4391   9.933 0.000245 ***
## Crop             3   3412    1137   2.572 0.064944 .
## Fertilizer:Crop  6   6226    1038   2.347 0.045555 *
## Residuals       48  21220     442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n\n\n")
```

```r
eta_partial_squared = eta_squared(df_anova, partial=TRUE)
eta_partial_squared
```

```
## # Effect Size for ANOVA (Type I)
##
## Parameter       | Eta2 (partial) |       95% CI
## -------------------------------------------------
## Fertilizer      |           0.29 | [0.11, 1.00]
## Crop            |           0.14 | [0.00, 1.00]
## Fertilizer:Crop |           0.23 | [0.00, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

```r
df_descriptive_stats = data.frame(Fertilizer = c("Blend X",
                                                 "",
                                                 "",
                                                 "",
                                                 "Blend Y",
                                                 "",
                                                 "",
                                                 "",
                                                 "Blend Z",
                                                 "",
                                                 "",
                                                 ""),
                                  Crop = c("Corn",
                                           "Rice",
                                           "Soy",
                                           "Wheat",
                                           "Corn",
                                           "Rice",
                                           "Soy",
                                           "Wheat",
                                           "Corn",
                                           "Rice",
                                           "Soy",
                                           "Wheat"),
                                  Mean = c(descriptive_table$Mean),
```

```
                    SD = c(descriptive_table$SD),
                    Valid = c(descriptive_table$Valid)
                  )
df_descriptive_stats
```

```
##    Fertilizer  Crop  Mean       SD Valid
## 1     Blend X  Corn 135.4 26.60451     5
## 2              Rice 141.2 18.82020     5
## 3               Soy 175.8 16.69431     5
## 4             Wheat 131.8 29.05512     5
## 5     Blend Y  Corn 159.6 31.27779     5
## 6              Rice 165.4 14.74110     5
## 7               Soy 140.2 12.87245     5
## 8             Wheat 143.2 22.33159     5
## 9     Blend Z  Corn 173.6 20.70749     5
## 10             Rice 172.4 13.25896     5
## 11              Soy 186.4 14.57052     5
## 12            Wheat 164.4 21.05469     5
```

```
tukey_results <- df_graph %>%
  group_by(Crop) %>%
  do(tidy(TukeyHSD(aov(Yield ~ Fertilizer, data = .)))) %>%
  ungroup()

tukey_results <- tukey_results %>%
  separate(contrast, into = c("Group1","Group2"), sep = "-")

tukey_results
```

```
## # A tibble: 12 x 9
##    Crop  term   Group1 Group2 null.value estimate conf.low conf.high adj.p.value
##    <chr> <chr>  <chr>  <chr>       <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
##  1 Corn  Ferti~ Blend~ Blend~          0    24.2    -20.6       69.0    0.352
##  2 Corn  Ferti~ Blend~ Blend~          0    38.2     -6.60      83.0    0.0984
##  3 Corn  Ferti~ Blend~ Blend~          0    14      -30.8       58.8    0.690
##  4 Rice  Ferti~ Blend~ Blend~          0    24.2     -2.43      50.8    0.0764
##  5 Rice  Ferti~ Blend~ Blend~          0    31.2      4.57      57.8    0.0221
##  6 Rice  Ferti~ Blend~ Blend~          0     7.00   -19.6       33.6    0.767
##  7 Soy   Ferti~ Blend~ Blend~          0   -35.6    -60.6      -10.6    0.00655
##  8 Soy   Ferti~ Blend~ Blend~          0    10.6    -14.4       35.6    0.513
##  9 Soy   Ferti~ Blend~ Blend~          0    46.2     21.2       71.2    0.000926
## 10 Wheat Ferti~ Blend~ Blend~          0    11.4    -29.8       52.6    0.746
## 11 Wheat Ferti~ Blend~ Blend~          0    32.6     -8.57      73.8    0.129
## 12 Wheat Ferti~ Blend~ Blend~          0    21.2    -20.0       62.4    0.384
```

Report

A two-way ANOVA was conducted to determine if the yield of crops were significantly affected by crop type and fertilizer type. All assumptions were determined to be achieved. Normality was observed through Shaphiro-Wilk test ($p > 0.05$) for all groups. Homogeneity of variances was also observed through Levene Test ($p > 0.05$) for all crop types. Then, the two-way ANOVA was conducted and we determined that fertilizers had significant effect on the yield ($p = 0.000245 < 0.05$). Crop type may have had some effect

($p = 0.064944 \sim 0.05$) but was not strictly significant. Then, the combination between fertilizers and crop types suggests significant interaction between them ($p = 0.045555 < 0.05$). Next, partial eta squared values are as follows: Fertilizer (0.29, 95% CI [0.11, 1.00]) explains 29 percent of variance, large effect; Crop (0.14, 95% CI [0.00, 1.00] explains 14 percent of variance, medium effect; Fertilizer:Crop (0.23, 95% CI [0.00, 1.00] explains 23 percent of variance, moderate effect.

Finally, the Tukey HSD results indicate the following:

Corn No pairwise differences were statistically significant (all adjusted p-values > 0.05).

For example, the mean difference between Blend Z and Blend X was 38.2, 95% CI [-6.60, 83.0], p = 0.098.

Interpretation: Corn yields did not differ significantly among the three fertilizers.

Rice Blend Z produced significantly higher yields than Blend X (mean difference = 31.2, 95% CI [4.57, 57.83], p = 0.022).

The other comparisons (Blend Y vs Blend X, Blend Z vs Blend Y) were not significant (p > 0.05).

Interpretation: For Rice, Blend Z performed better than Blend X; the other fertilizers were similar.

Soy Blend Y yielded significantly less than Blend X (mean difference = -35.6, 95% CI [-60.56, -10.64], p = 0.007).

Blend Z was not significantly different from Blend X (p = 0.51), but was significantly higher than Blend Y (mean difference = 46.2, 95% CI [21.24, 71.16], p = 0.001).

Interpretation: Blend Y underperformed compared to the other blends; Blend Z matched Blend X.

Wheat No significant differences were observed between any fertilizer pair (all p > 0.05).

Interpretation: Wheat yields were similar across all three fertilizer blends.

Finally, given the significant effects Fertilizer and Fertilizer:Crop had on the variance of the data, we reject the **Null hypothesis:** There is no significant interaction effect on yield between fertilizer and crop.

Github Link: https://github.com/SylTana/APM1111-QUIJANO-JULIAN_PHILIP-Statistical-Theory-/tree/main/FA9