Exercise 11 of Applied Biostatistics II

Hector Arribas Julian Niklaus Matthias Rubin

Setup of the experiment

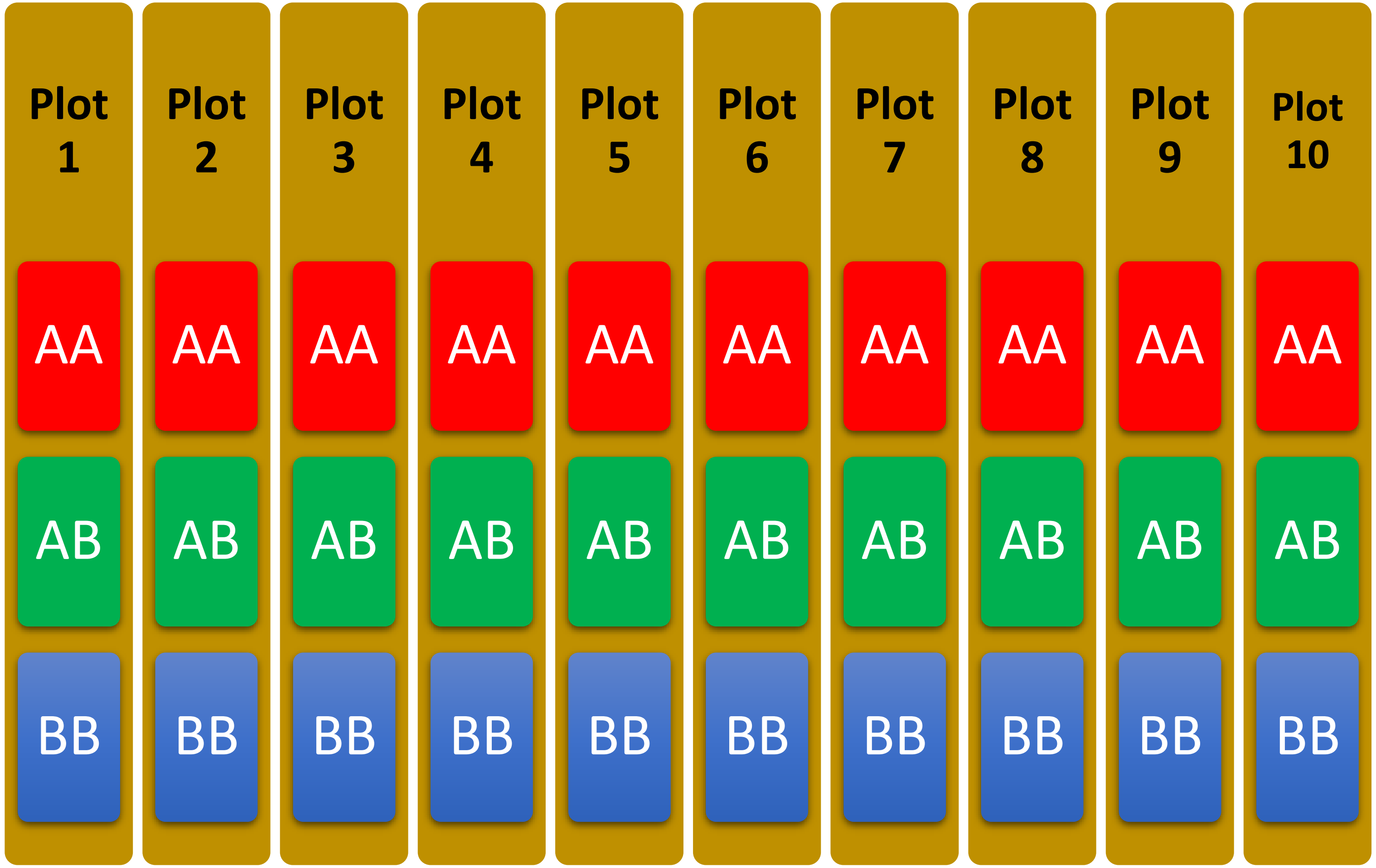
A scientist is interested in how genotype of a strawberry plant affects fruit yield. There are **three levels of genotype** (AA, AB, BB) and **ten plots of land**, **three plants per plot**. Each of the three genotypes is present in each plot.

a) Perform an ANOVA, assuming one-way randomized block design.

b) Repeat the analysis of variance without taking into account land effects.

c) Compare the results in a) and b). Why are the degrees of freedom different? Which result would you use?

Layout of the experiment



Plot 1	Plot 2	Plot 3	Plot 4	Plot 5	Plot 6	Plot 7	Plot 8	Plot 9	Plot 10
AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
AB	AB	AB	AB	AB	AB	AB	AB	AB	AB
BB	BB	BB	BB	BB	BB	BB	BB	BB	BB

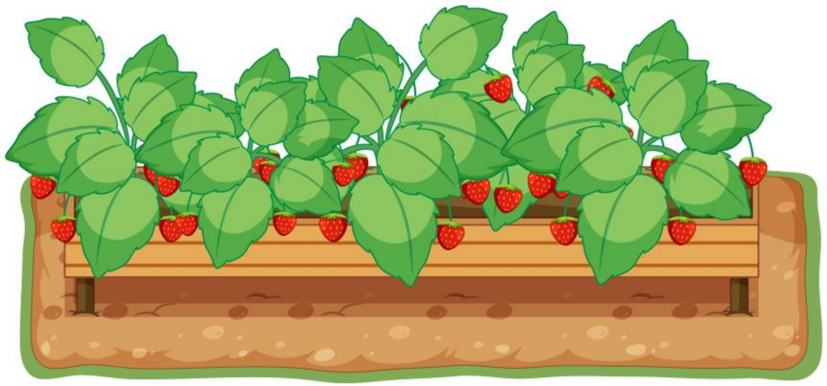
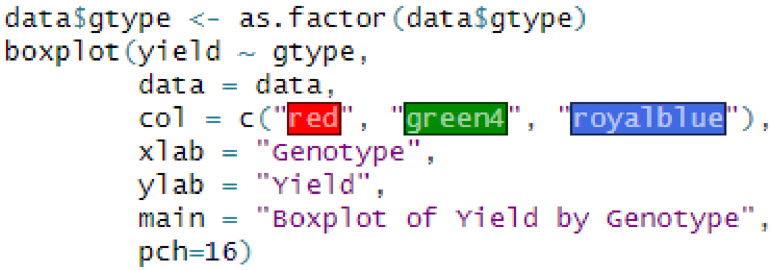
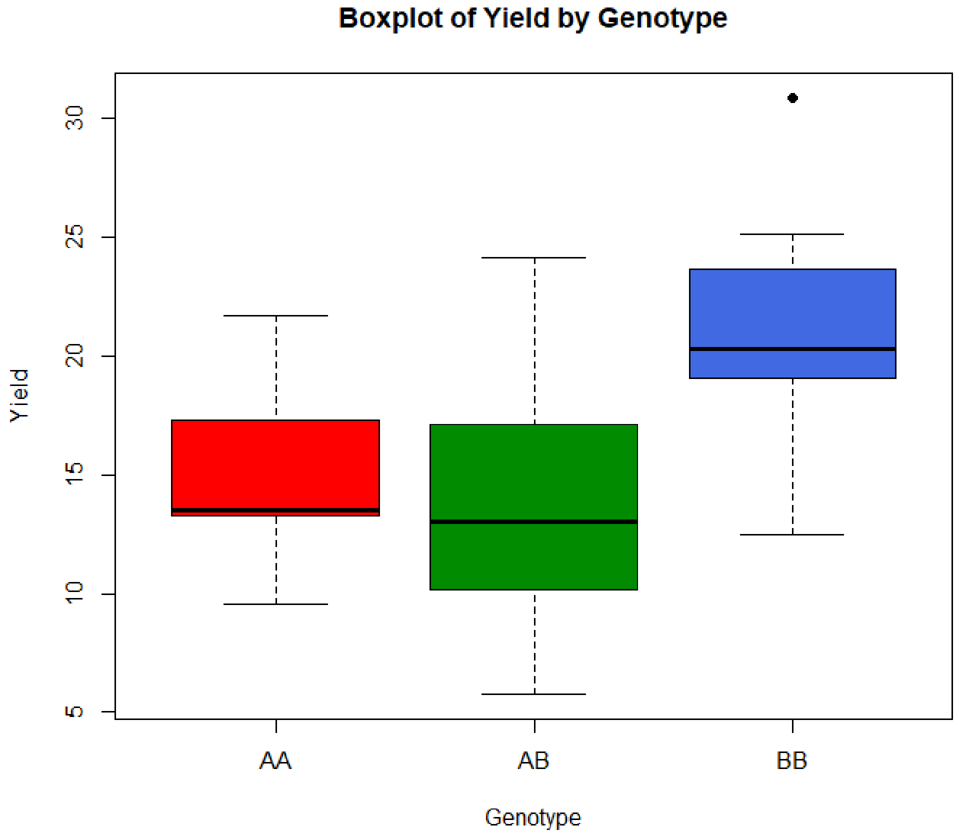


Image by brgfx on Freepick [1]

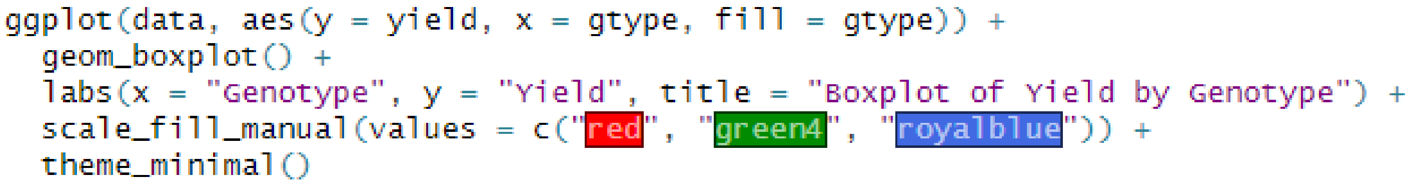
Base R



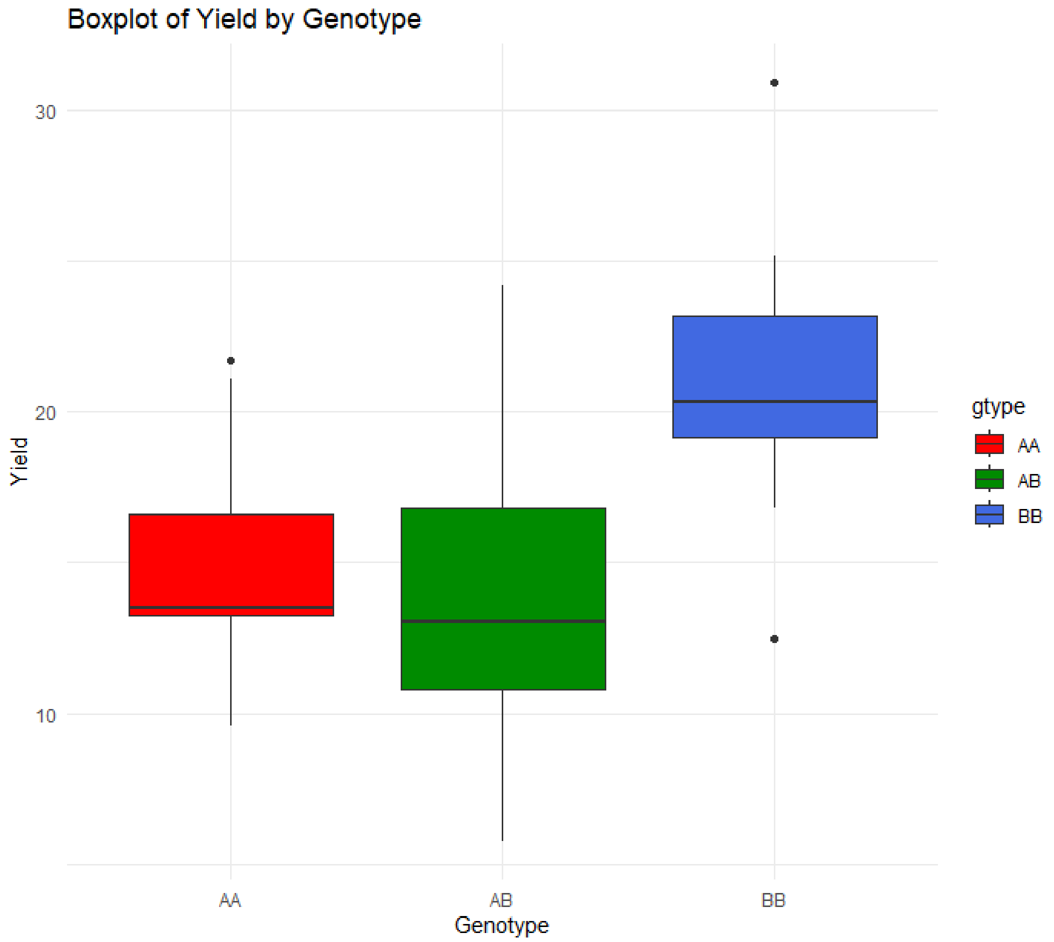
```
data$gtype <- as.factor(data$gtype)
boxplot(yield ~ gtype,
        data = data,
        col = c("red", "green4", "royalblue"),
        xlab = "Genotype",
        ylab = "Yield",
        main = "Boxplot of Yield by Genotype",
        pch=16)
```

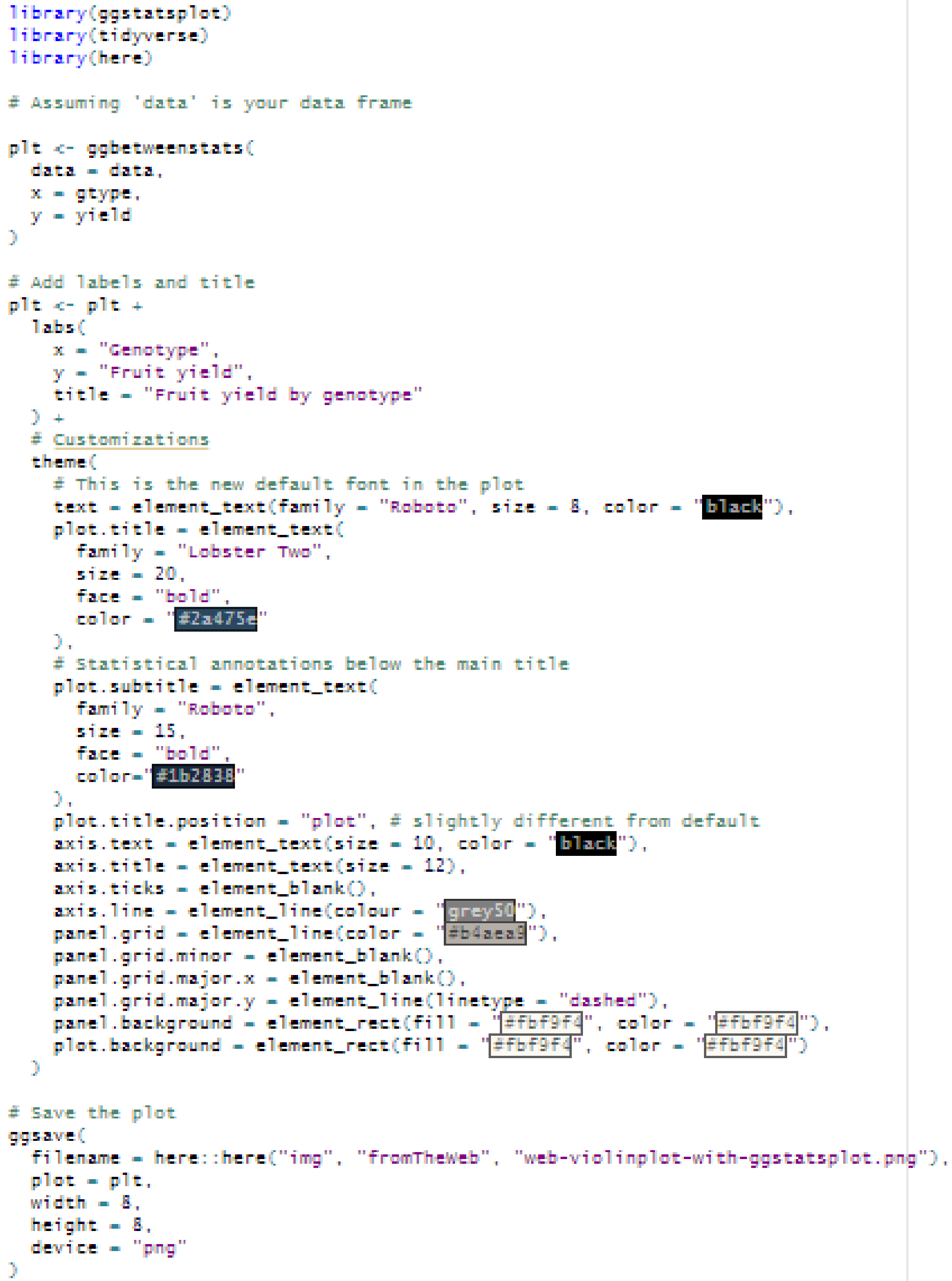


ggplot

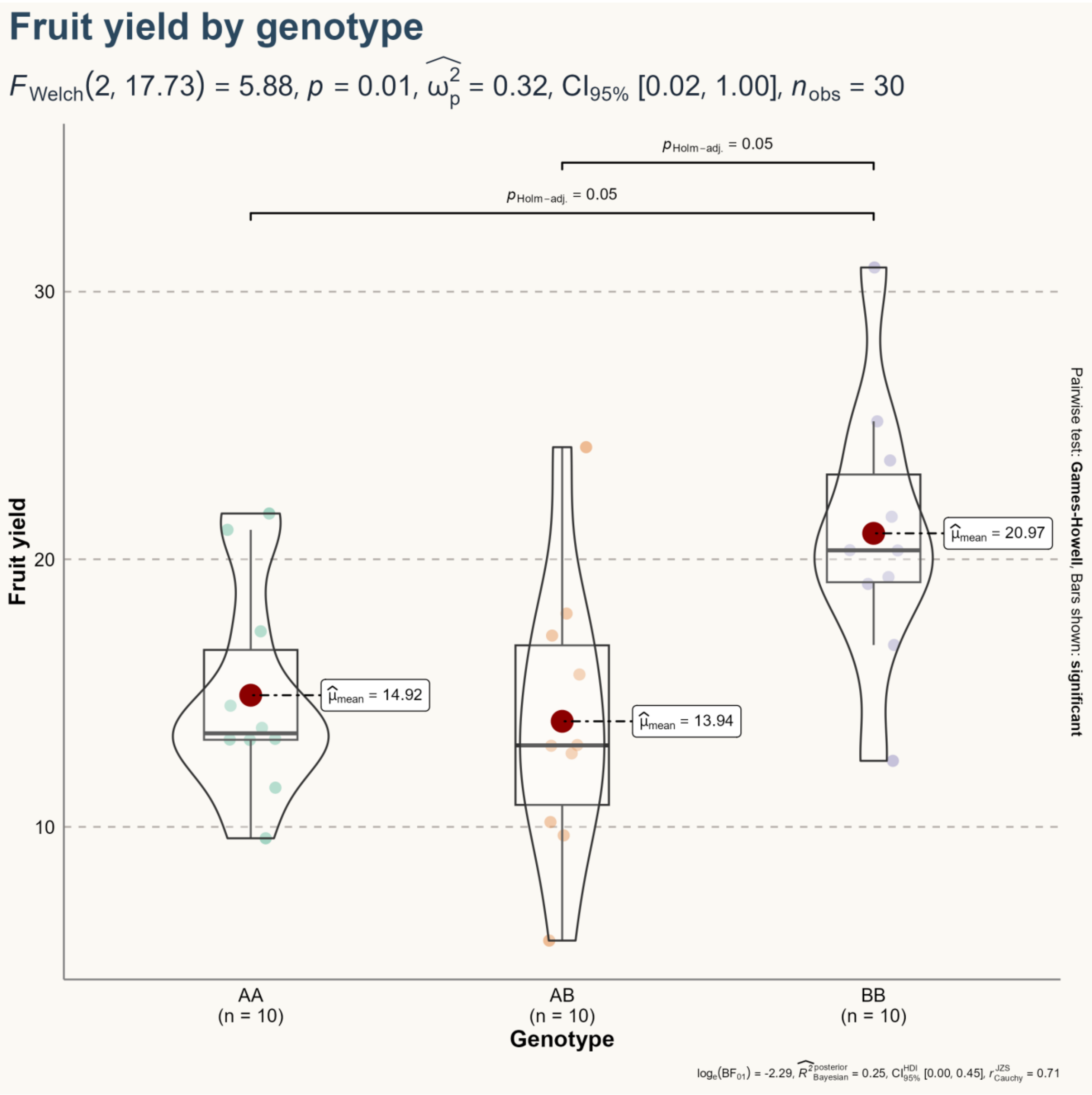


```
ggplot(data, aes(y = yield, x = gtype, fill = gtype)) +
  geom_boxplot() +
  labs(x = "Genotype", y = "Yield", title = "Boxplot of Yield by Genotype") +
  scale_fill_manual(values = c("red", "green4", "royalblue")) +
  theme_minimal()
```

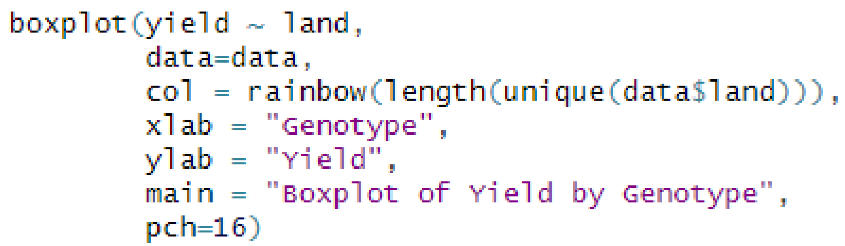




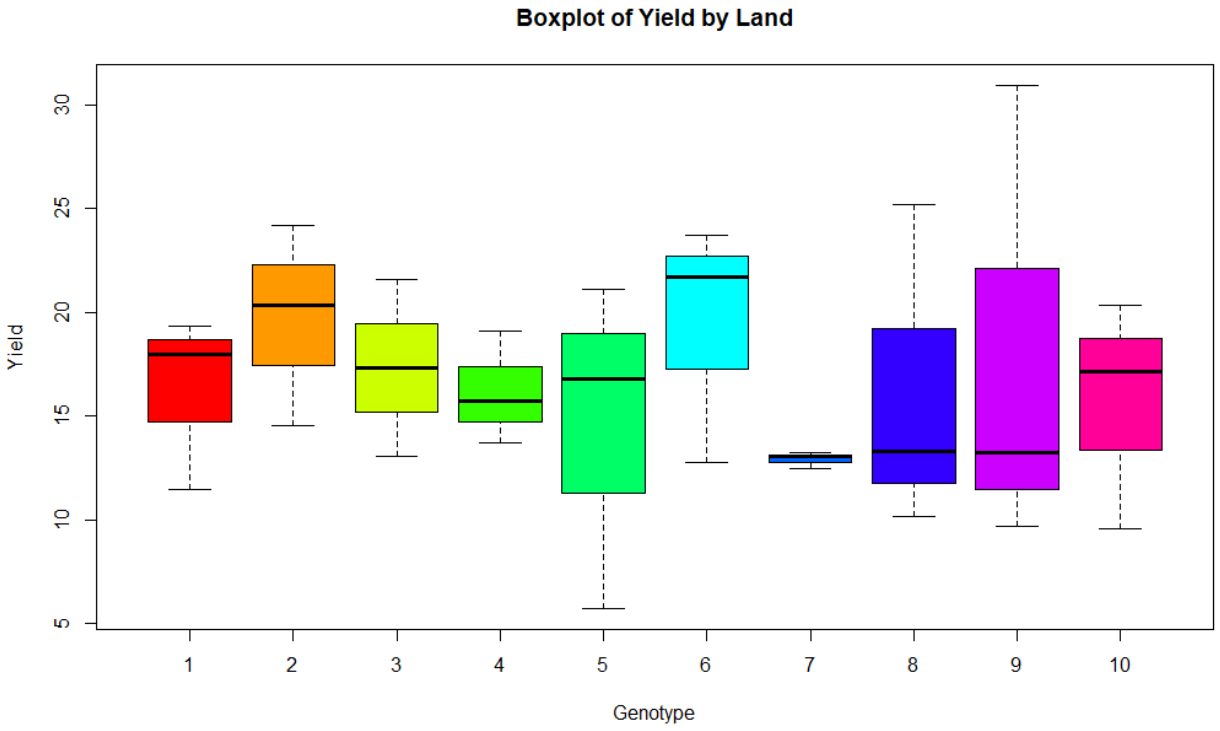
```r
library(ggstatsplot)
library(tidyverse)
library(here)

# Assuming 'data' is your data frame

plt <- ggbetweenstats(
  data = data,
  x = gtype,
  y = yield
)

# Add labels and title
plt <- plt +
  labs(
    x = "Genotype",
    y = "Fruit yield",
    title = "Fruit yield by genotype"
  ) +
  # Customizations
  theme(
    # This is the new default font in the plot
    text = element_text(family = "Roboto", size = 8, color = "black"),
    plot.title = element_text(
      family = "Lobster Two",
      size = 20,
      face = "bold",
      color = "#2a475e"
    ),
    # Statistical annotations below the main title
    plot.subtitle = element_text(
      family = "Roboto",
      size = 15,
      face = "bold",
      color="#1b2838"
    ),
    plot.title.position = "plot", # slightly different from default
    axis.text = element_text(size = 10, color = "black"),
    axis.title = element_text(size = 12),
    axis.ticks = element_blank(),
    axis.line = element_line(colour = "grey50"),
    panel.grid = element_line(color = "#b4aea9"),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank(),
    panel.grid.major.y = element_line(linetype = "dashed"),
    panel.background = element_rect(fill = "#fbf9f4", color = "#fbf9f4"),
    plot.background = element_rect(fill = "#fbf9f4", color = "#fbf9f4")
  )

# Save the plot
ggsave(
  filename = here::here("img", "fromTheWeb", "web-violinplot-with-ggstatsplot.png"),
  plot = plt,
  width = 8,
  height = 8,
  device = "png"
)
```



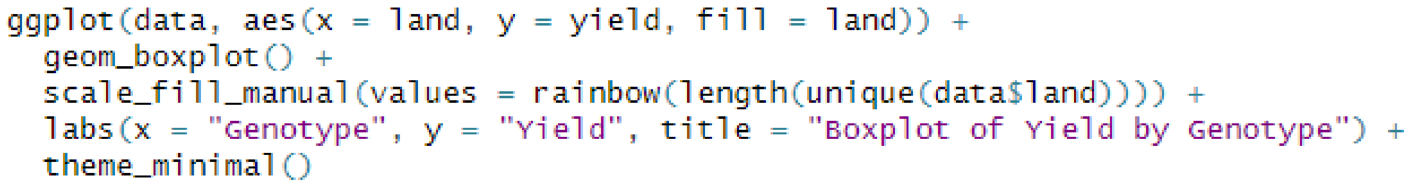
Base R



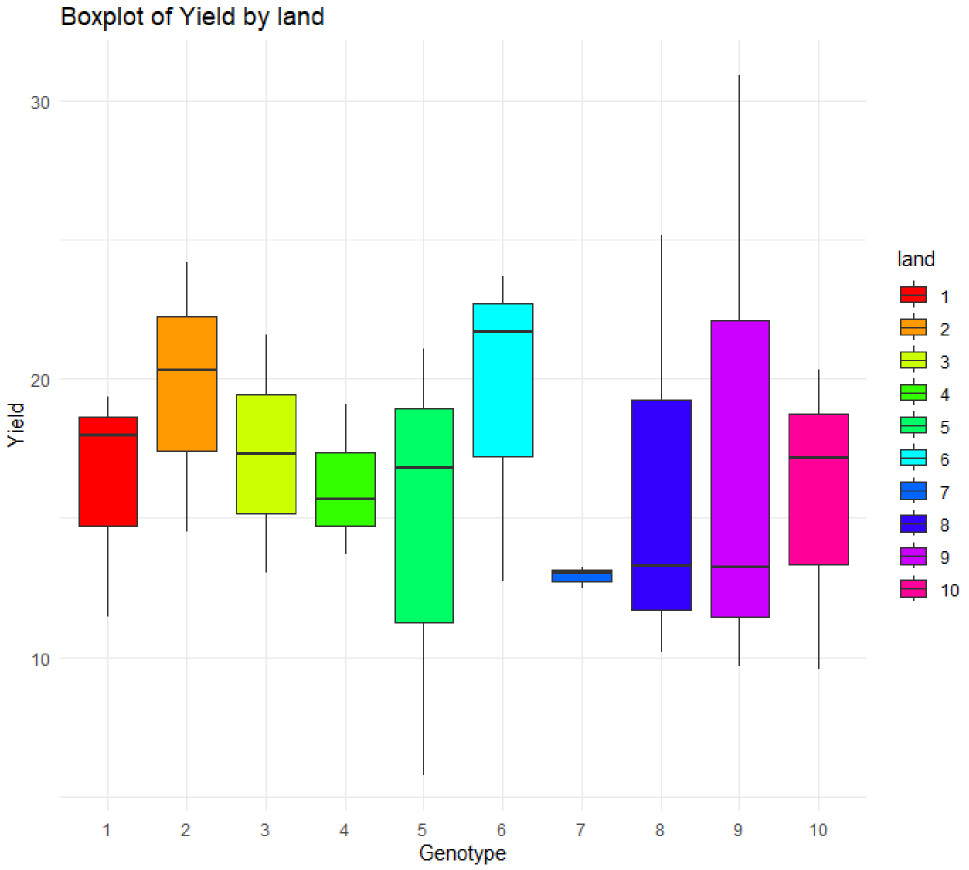
```
boxplot(yield ~ land,
        data=data,
        col = rainbow(length(unique(data$land))),
        xlab = "Genotype",
        ylab = "Yield",
        main = "Boxplot of Yield by Genotype",
        pch=16)
```



ggplot



```
ggplot(data, aes(x = land, y = yield, fill = land)) +
  geom_boxplot() +
  scale_fill_manual(values = rainbow(length(unique(data$land)))) +
  labs(x = "Genotype", y = "Yield", title = "Boxplot of Yield by Genotype") +
  theme_minimal()
```



a) Perform an ANOVA, assuming one-way randomized block design

- ANOVA (ANalysis Of VAriance)
- Basic idea: compare variance within groups to variance between groups
- Basic assumption: variance (“amount of randomness”) is the same in each group
- We test: **H0: $\mu 1 = \mu 2 = \mu 3$** => All genotypes share the same mean fruit yield
- Alternative H1: One or more means are different from the others

a) Perform an ANOVA, assuming one-way randomized block design

A type of experimental design used in statistical analysis:

1. **One-way**: There is only one factor (genotype of strawberries) being studied. We want to determine if it has significant effect on the fruit yield.

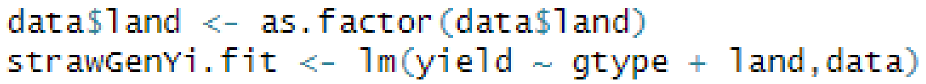
2. **Block**: The land grouped into blocks. Purpose:

- Account for unwanted variability between plots of land (e.g., soil effects)
- Increase the precision of the estimates of genotype effects.

3. **Randomized**: The assignment of genotypes to plots of land is done randomly. This ensures that any observations are not due to systematic biases or other factors related to the plots of land.

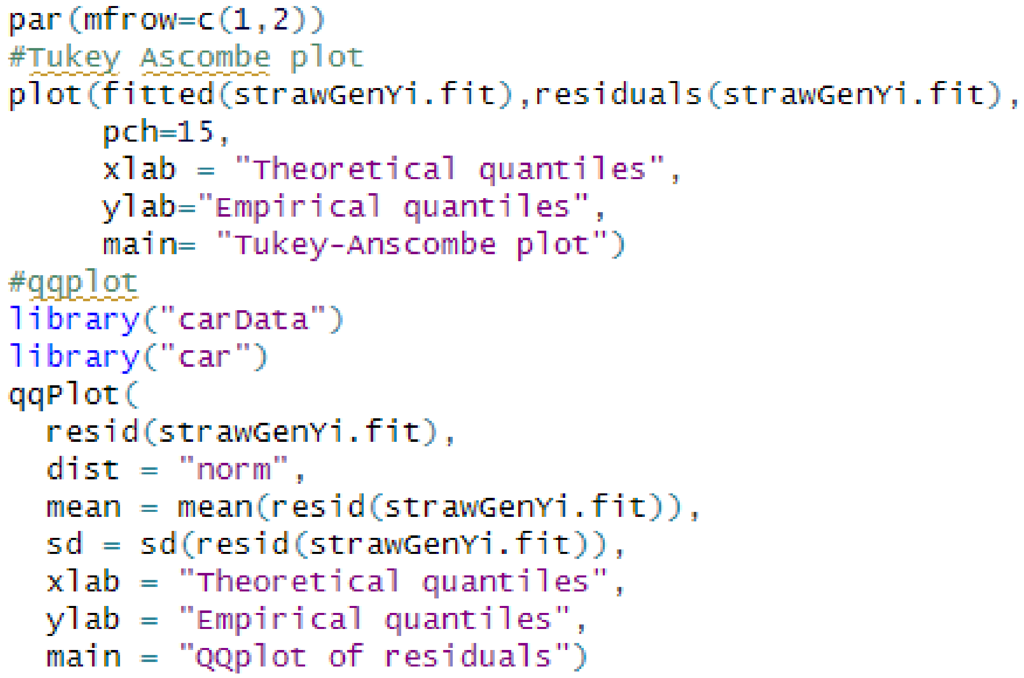
a) Perform an ANOVA, assuming one-way randomized block design

- Model

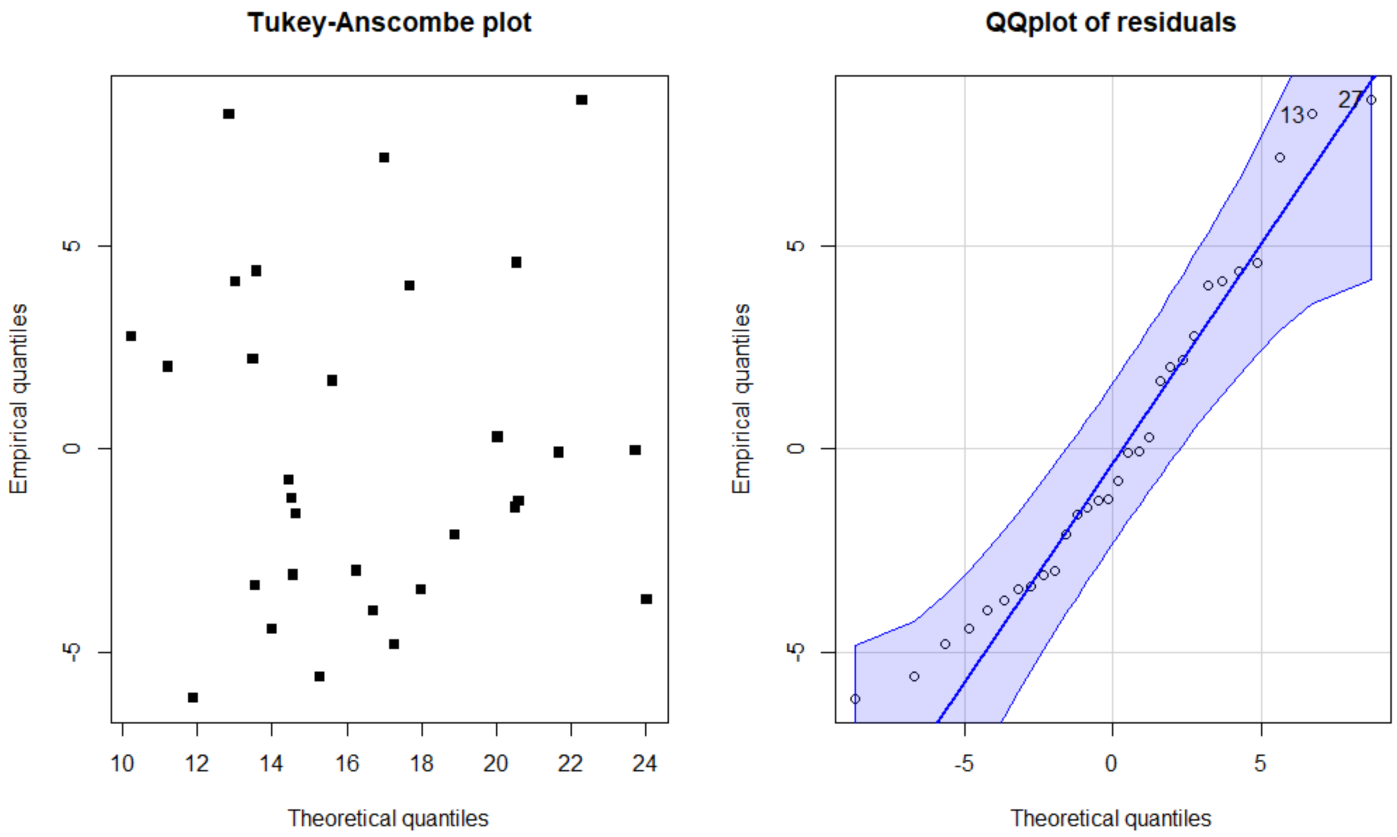


```
data$land <- as.factor(data$land)
strawGenYi.fit <- lm(yield ~ gtype + land,data)
```

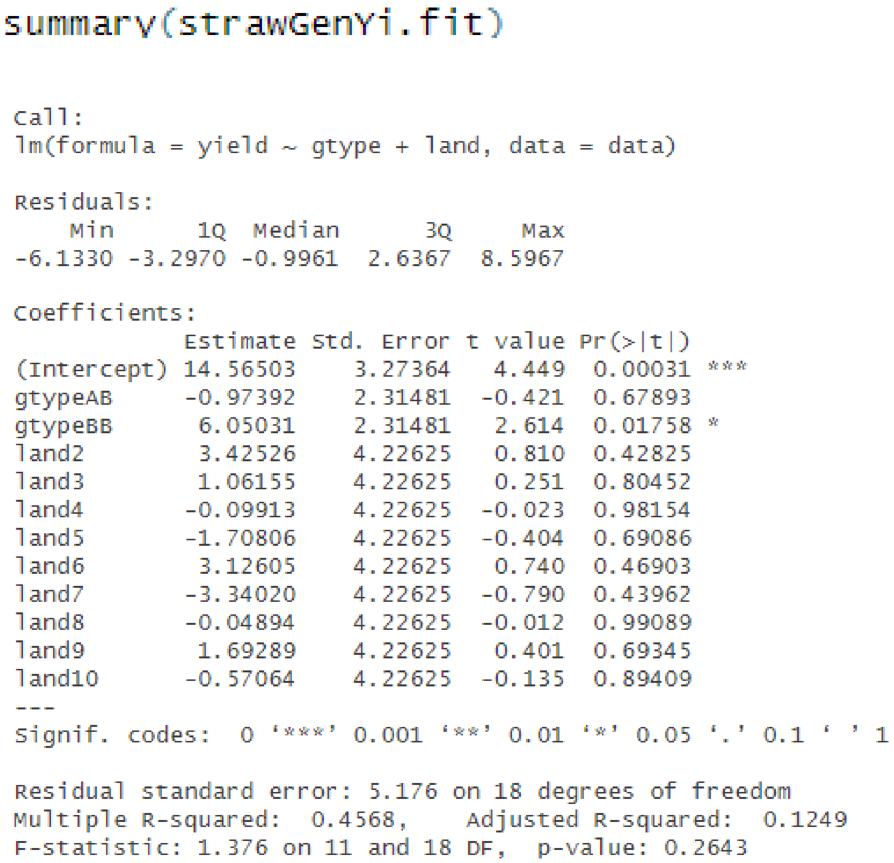
- Check assumptions



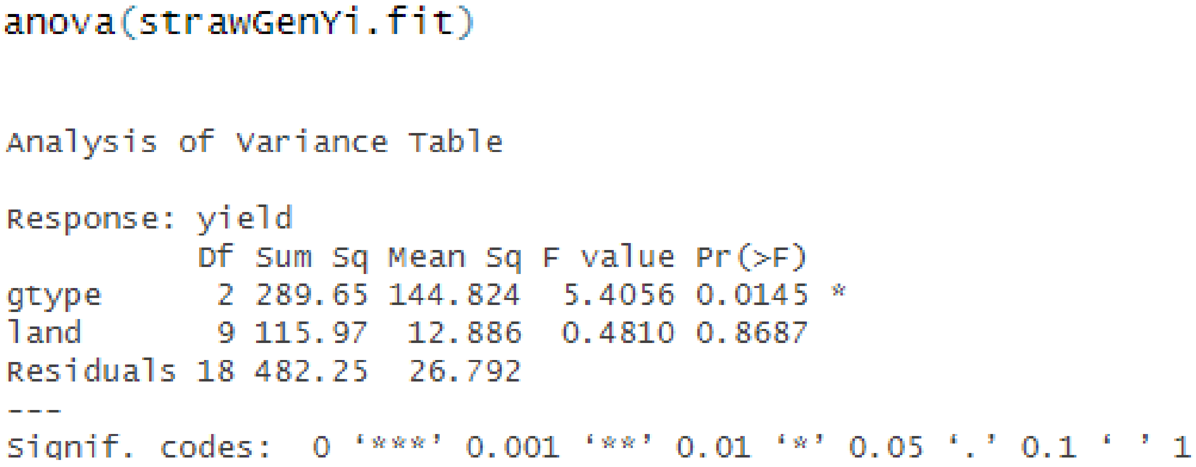
```
par(mfrow=c(1,2))
#Tukey Ascombe plot
plot(fitted(strawGenYi.fit),residuals(strawGenYi.fit),
     pch=15,
     xlab = "Theoretical quantiles",
     ylab="Empirical quantiles",
     main= "Tukey-Anscombe plot")
#qqplot
library("carData")
library("car")
qqPlot(
  resid(strawGenYi.fit),
  dist = "norm",
  mean = mean(resid(strawGenYi.fit)),
  sd = sd(resid(strawGenYi.fit)),
  xlab = "Theoretical quantiles",
  ylab = "Empirical quantiles",
  main = "QQplot of residuals")
```



a) Perform an ANOVA, assuming one-way randomized block design



```
summary(strawGenYi.fit)


Call:
lm(formula = yield ~ gtype + land, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.1330 -3.2970 -0.9961  2.6367  8.5967 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.56503    3.27364   4.449  0.00031 ***
gtypeAB     -0.97392    2.31481  -0.421  0.67893    
gtypeBB      6.05031    2.31481   2.614  0.01758 *  
land2        3.42526    4.22625   0.810  0.42825    
land3        1.06155    4.22625   0.251  0.80452    
land4       -0.09913    4.22625  -0.023  0.98154    
land5       -1.70806    4.22625  -0.404  0.69086    
land6        3.12605    4.22625   0.740  0.46903    
land7       -3.34020    4.22625  -0.790  0.43962    
land8       -0.04894    4.22625  -0.012  0.99089    
land9        1.69289    4.22625   0.401  0.69345    
land10      -0.57064    4.22625  -0.135  0.89409    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.176 on 18 degrees of freedom
Multiple R-squared:  0.4568,	Adjusted R-squared:  0.1249 
F-statistic: 1.376 on 11 and 18 DF,  p-value: 0.2643
```

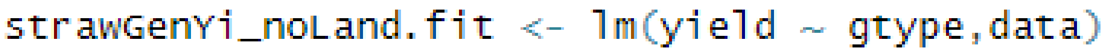


```
anova(strawGenYi.fit)


Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value Pr(>F)  
gtype      2 289.65 144.824  5.4056 0.0145 *
land       9 115.97  12.886  0.4810 0.8687  
Residuals 18 482.25  26.792                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

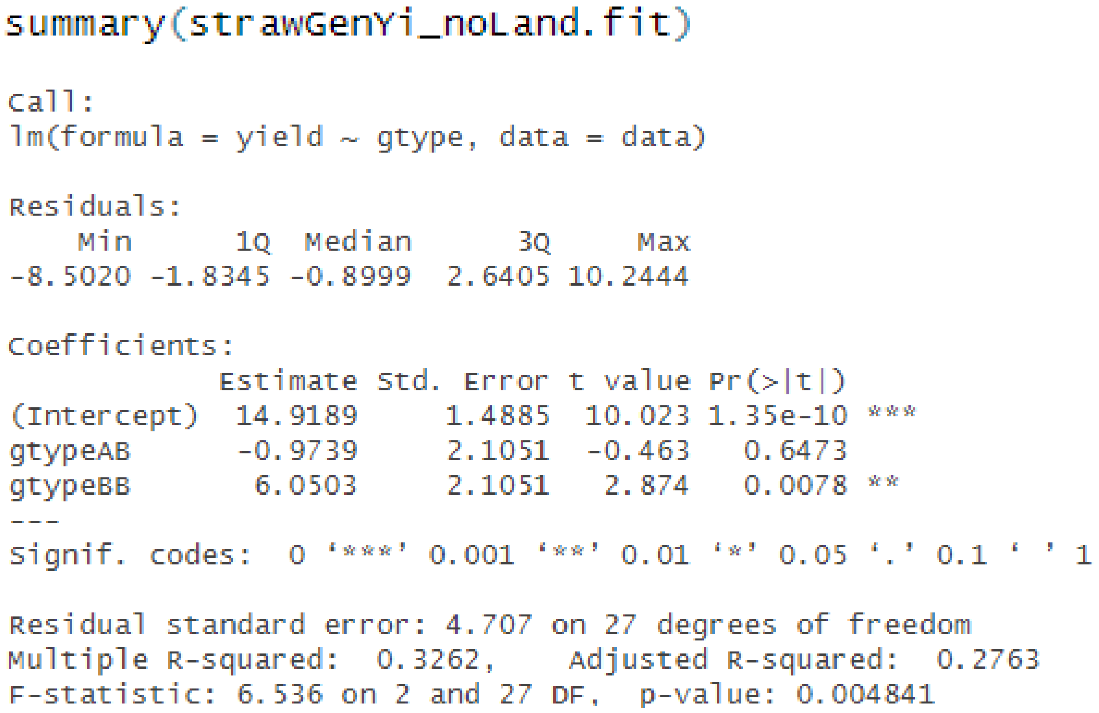
The factor genotype (gtype) is significant on a **5%** level.

The block factor land does not seem to have an influence on the yield.

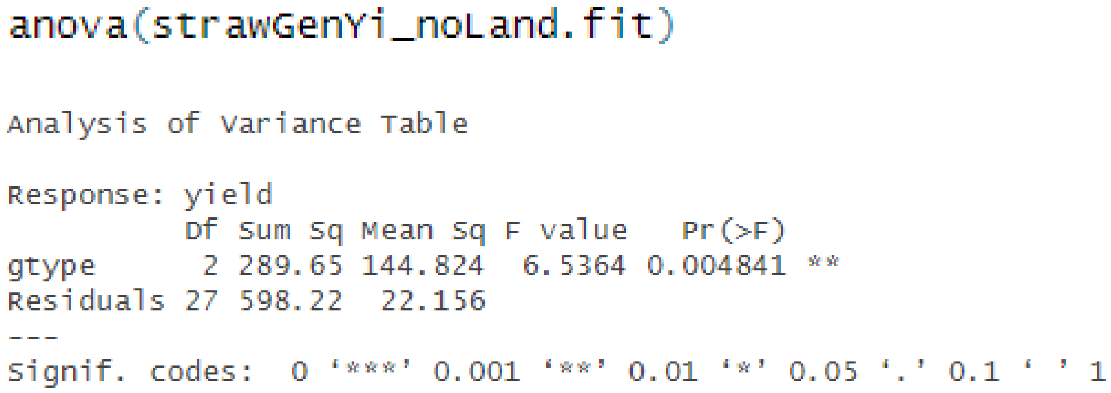
b) Repeat the analysis of variance without taking into account land effects



```
strawGenYi_noLand.fit <- lm(yield ~ gtype,data)
```



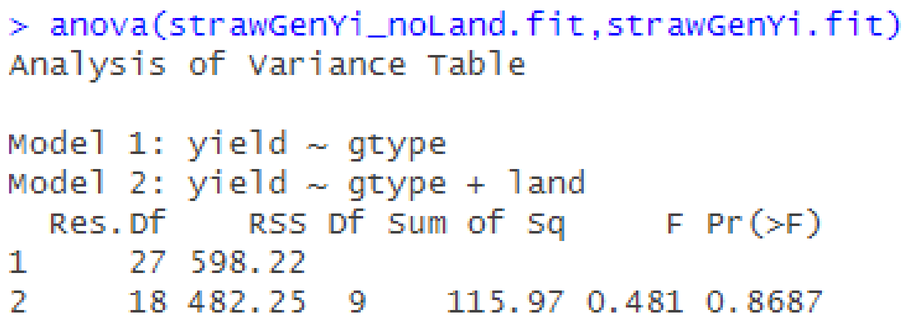
```
summary(strawGenYi_noLand.fit)

Call:
lm(formula = yield ~ gtype, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5020 -1.8345 -0.8999  2.6405 10.2444 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  14.9189     1.4885  10.023 1.35e-10 ***
gtypeAB      -0.9739     2.1051  -0.463   0.6473    
gtypeBB       6.0503     2.1051   2.874   0.0078 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.707 on 27 degrees of freedom
Multiple R-squared:  0.3262,	Adjusted R-squared:  0.2763 
F-statistic: 6.536 on 2 and 27 DF,  p-value: 0.004841
```



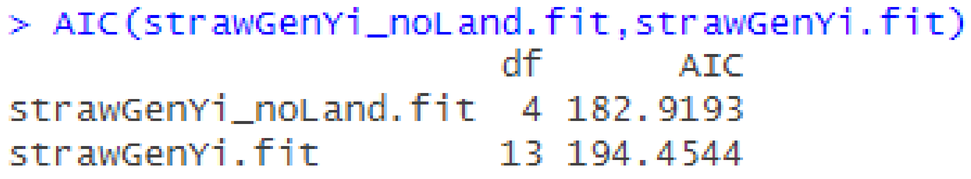
```
anova(strawGenYi_noLand.fit)

Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value   Pr(>F)   
gtype      2 289.65 144.824  6.5364 0.004841 **
Residuals 27 598.22  22.156                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The factor genotype (gtype) is now significant on a **1%** level.

c) Compare the results in a) and b). Why are the degrees of freedom different? Which result would you use?



```
> anova(strawGenYi_noLand.fit,strawGenYi.fit)
Analysis of Variance Table

Model 1: yield ~ gtype
Model 2: yield ~ gtype + land
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     27 598.22
2     18 482.25  9    115.97 0.481 0.8687
```



```
> AIC(strawGenYi_noLand.fit,strawGenYi.fit)
                      df      AIC
strawGenYi_noLand.fit  4 182.9193
strawGenYi.fit        13 194.4544
```

- The factor genotype is **more significant in the simple model**
- F-statistic is **not statistically significant** -> indicates that the full model does **not** provide a significantly better fit
- AIC (Akaike Information Criterion) is **lower for simple model** -> better trade-off between model fit and complexity
- Df are different because the total degrees of freedom in the model remain constant, but a portion of them are allocated to estimate the parameters of the added variables.

References

1: <a href="https://www.freepik.com/free-vector/strawberry-plant-growing-with-soil-cartoon_25672980.htm#query=growing%20strawberries&position=0&from_view=keyword&track=ais&uuid=afdc4e99-9347-4eb4-adcd-0cf323320b31">Image by brgfx</a> on Freepik