TUTORIAL 3
Data Visualization with *ggplot2*

@QIU,
HUNTER/CUNY

DATA VISUALIZATION: GRAMMATICAL ELEMENTS OF GRAPHICS

- Three essential grammatical elements (layers) of graphics:
  - Data: the data which we want to plot.

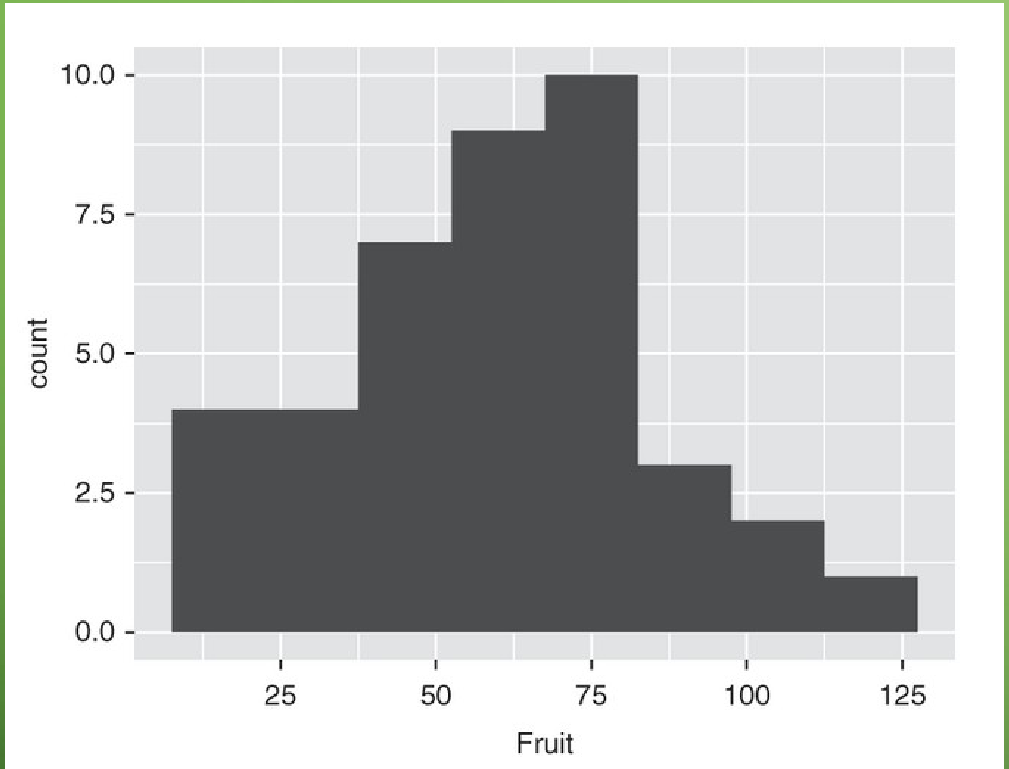
```
> str(iris)
'data.frame':	150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0
 $ Species     : Factor w/ 3 levels "setosa'
```

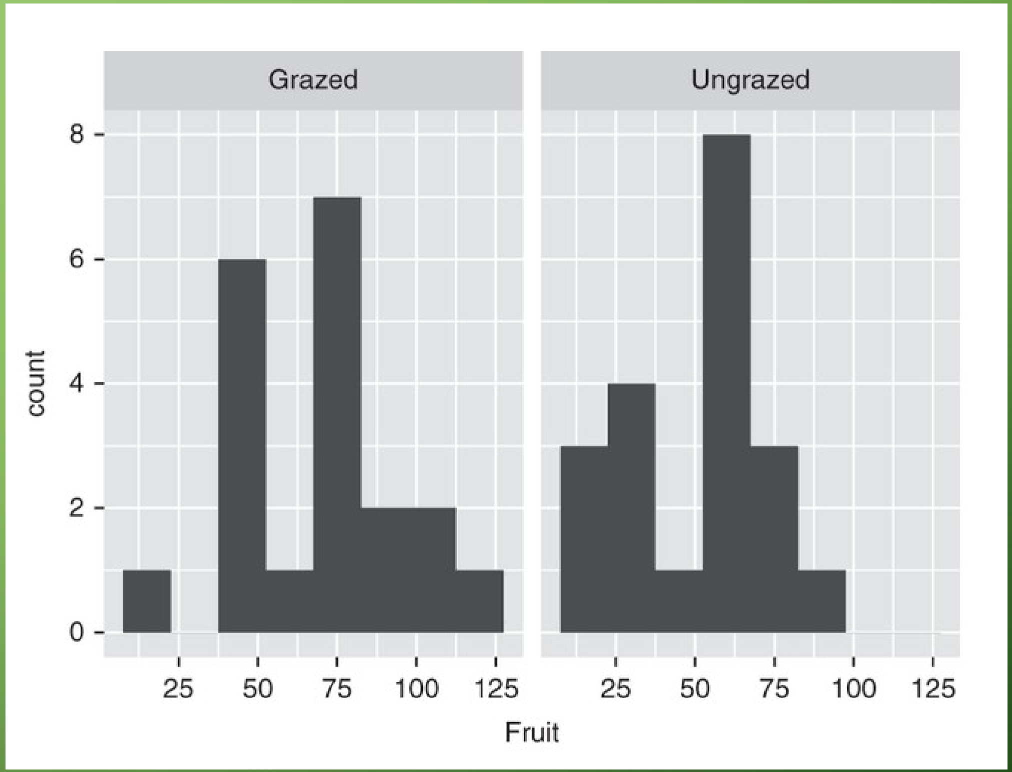
  - Aesthetics layer: refers to the scales onto which we will map our data
  - Geom layer: allows us to choose how the plot will look like.
- Optional layers:
  - Theme layer: which controls all the non-data elements of graphics

ggplot2 geometrics:

- Scatterplot: geom_point(), geom_jitter()
- Line Plot: geom_line()
- Histograms: geom_histogram()
- Box plot: geom_boxplot()
- Bar plot: geom_bar()
- Violin plot: geom_violin()

HISTOGRAM: DISTRIBUTION OF A NUMERICAL VARIABLE






```
ggplot(compensation, aes(x = Fruit)) +
  geom_histogram(bins = 10)
ggplot(compensation, aes(x = Fruit)) +
  geom_histogram(binwidth = 15)
```

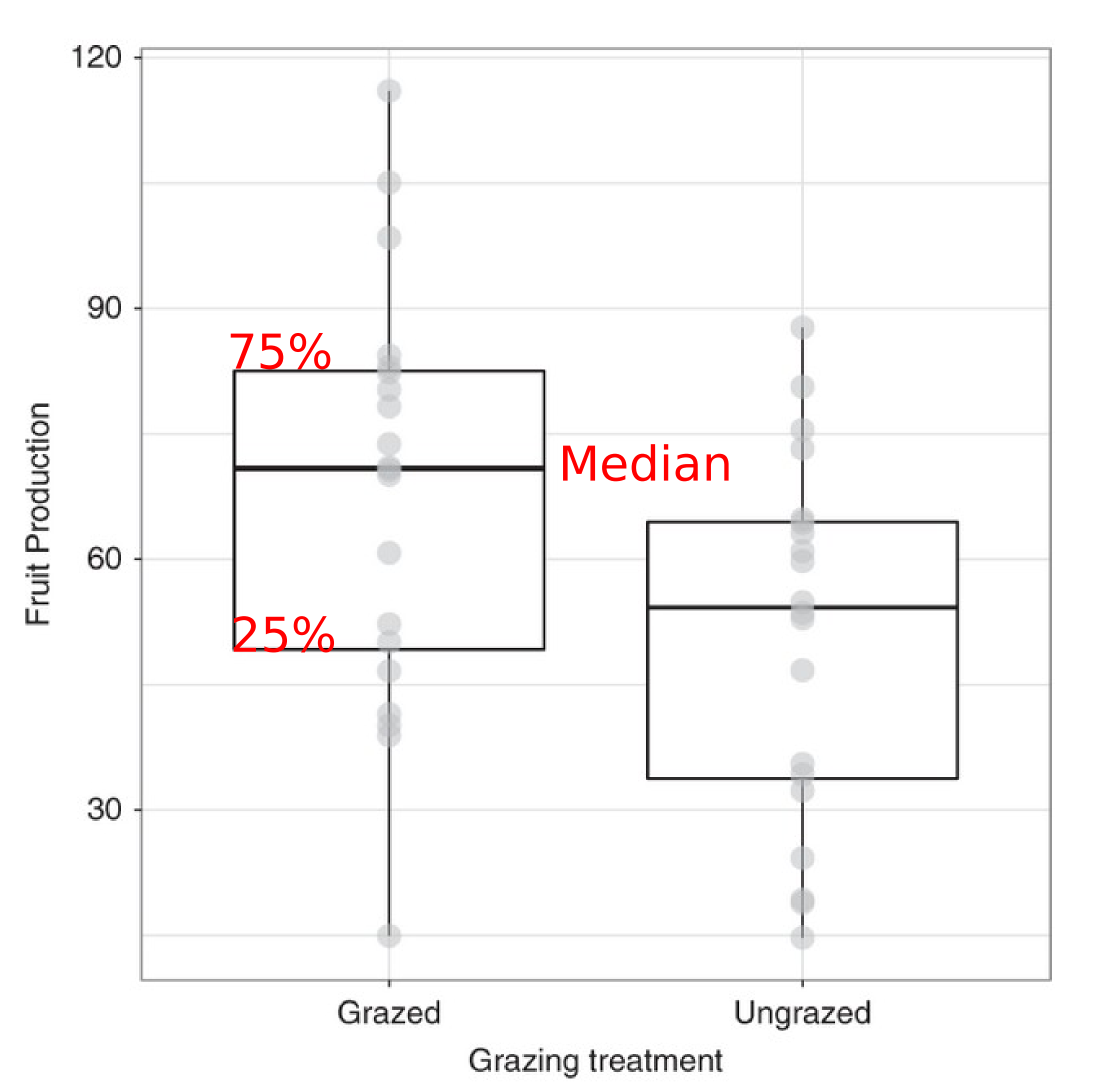
```
ggplot(compensation, aes(x = Fruit)) +
  geom_histogram(binwidth = 15) +
  facet_wrap(~Grazing)
```

Peaks: the most frequent value (not the highest value)

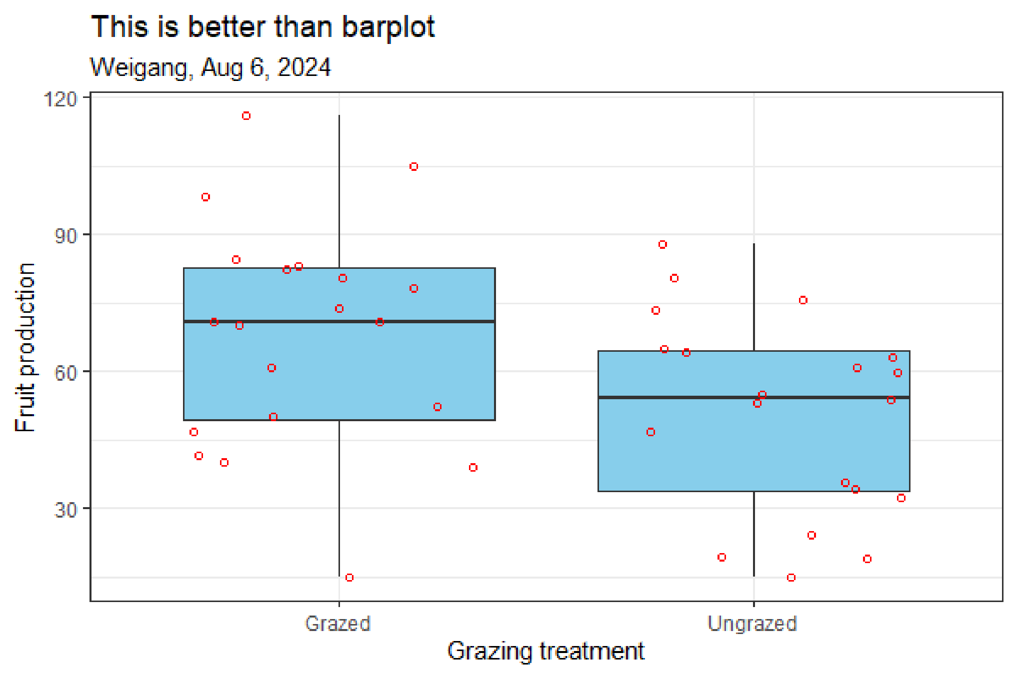
BOX PLOT: NUMERICAL VS CATEGORICAL

```
ggplot(compensation, aes(x = Grazing, y = Fruit)) +
  geom_boxplot() +
  xlab("Grazing treatment") +
  ylab("Fruit Production") +
  theme_bw()
```

```
ggplot(compensation, aes(x = Grazing, y = Fruit)) +
  geom_boxplot() +
  geom_point(size = 4, colour = 'lightgrey', alpha = 0.5) +
  xlab("Grazing treatment") +
  ylab("Fruit Production") +
  theme_bw()
```



BOX PLOT



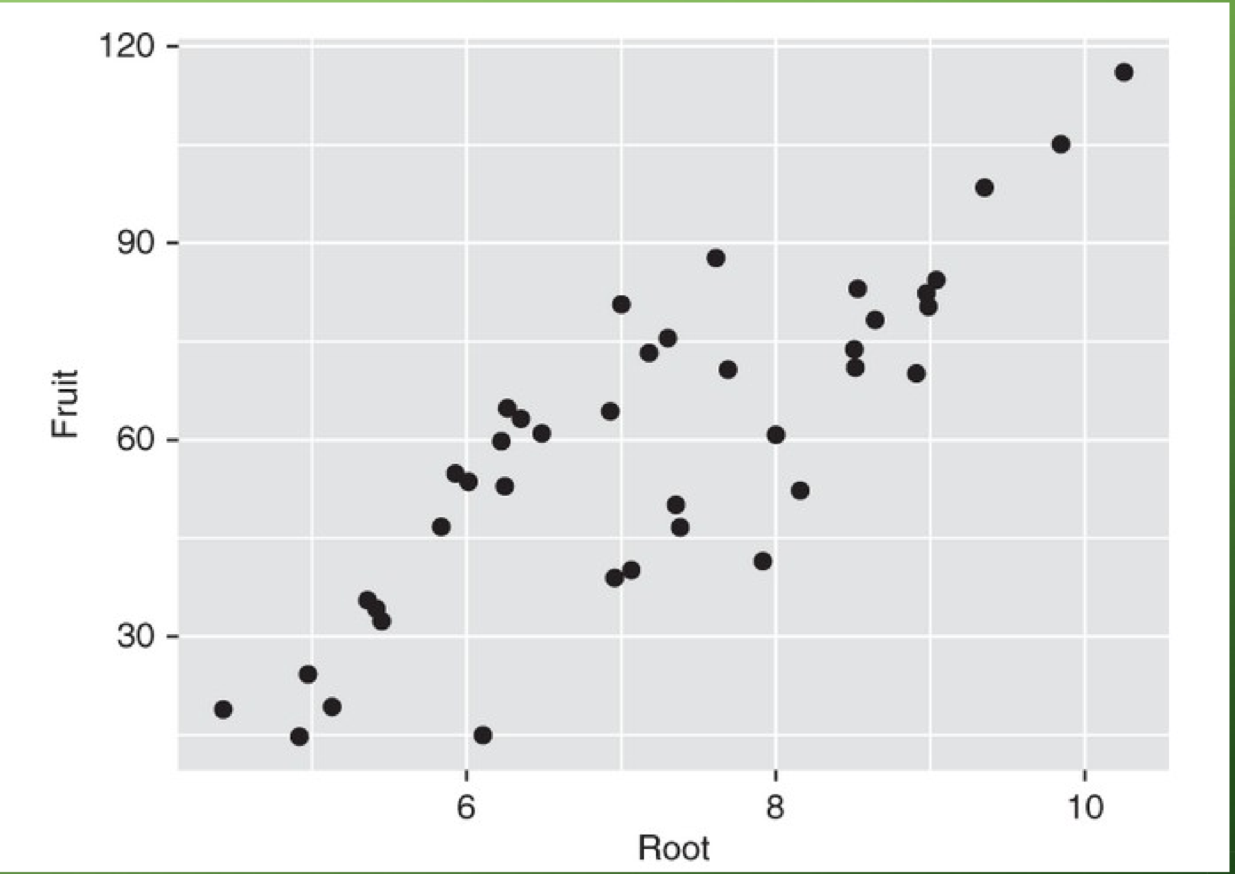

```
compensation %>%
  ggplot(aes(x = Grazing, y = Fruit)) +
  geom_boxplot(fill = "skyblue") +
  geom_jitter(shape = 1, color = "red") # geom_jitter() to show sample sizes!
  theme_bw() +
  xlab("Grazing treatment") +
  ylab("Fruit production") +
  labs(title = "This is better than barplot", subtitle = "Weigang, Aug 6, 2024")
```

VIOLIN PLOT: NUMERICAL VS CATEGORICAL


```
compensation %>%
  ggplot(aes(x = Grazing, y = Fruit)) +
  geom_violin(fill = "skyblue") +
  geom_jitter(shape = 1, color = "red") +
  stat_summary() +
  theme_bw() +
  xlab("Grazing treatment") +
  ylab("Fruit production") +
  labs(title = "This is a violin plot", subtitle = "Weigang, Aug 6, 2024")
```

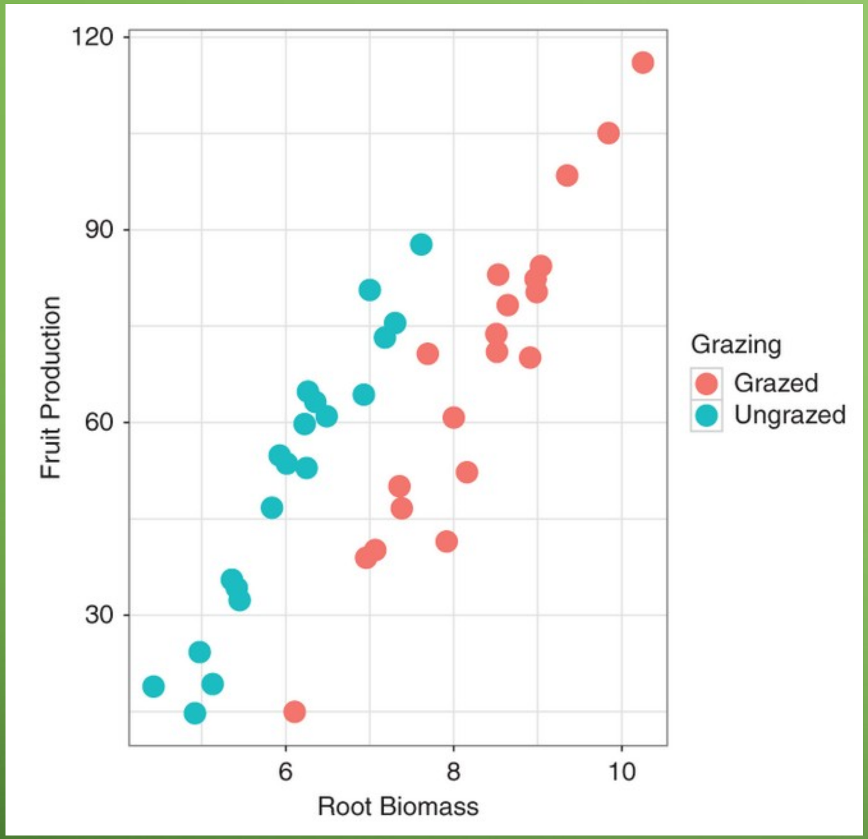
SCATTER PLOT: NUMERICAL VS NUMERICAL

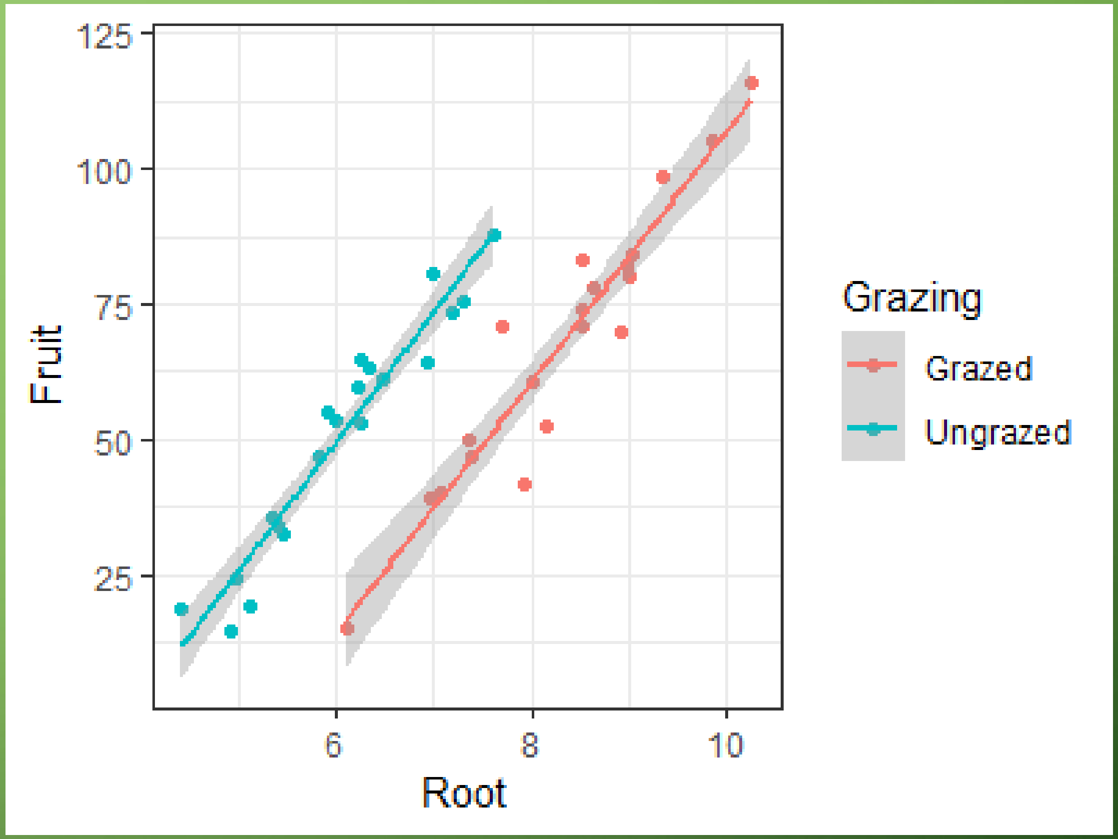
```
# plotting basics with ggplot
# my tutorial script
# lots and lots of annotation!

# libraries I need (no need to install...)
library(dplyr)
library(ggplot2)

# clear the decks
rm(list = ls())

# get the data
compensation <- read.csv('compensation.csv')

# check out the data
glimpse(compensation)

# make my first ggplot picture
ggplot(compensation, aes(x = Root, y = Fruit)) +
  geom_point()
```



- aes(): aesthetic mapping between variables and graph features
- geom_point(): a geometric object

- Map a categorical variable to aes(color = variable)
- Apply geom_smooth(method = “lm”) to show regression line






```
ggplot(compensation, aes(x = Root, y = Fruit, colour = Grazing)) +
  geom_point(size = 5) +
  xlab("Root Biomass") +
  ylab("Fruit Production") +
  theme_bw()
```

```
compensation %>%
  ggplot(aes(x = Root, y = Fruit, color = Grazing)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```

Summary: Data visualization

- Scatterplot – show relations between two numerical variables (e.g., “Fruit” & “Root”)
- Boxplot/Violinplot – show distribution (e.g., median) of a numerical variation (e.g., “Fruit”) with respect to a categorical variable (e.g., “Grazing”)
  - Add “geom_point” or “geom_jitter” to show actual data points
  - A better alternative than barplot
- Histogram/Density – show frequency distribution (e.g., counts in bins) of a numerical variation (e.g., “Fruit”)
- Multidimensional mapping of variables to graphic elements:
  - X-axis
  - Y-axis
  - Color/Fill
  - Panel (“facet_wrap”)

PRACTICE #3

- Show distribution of “Sepal.Length” with a histogram. Show distributions by Species.
- Show distributions of “Sepal.Width” by Species with a boxplot
- Filter the **iris** dataset for species “**versicolor**” and save the result to a variable named “**versicolor**”
- Plot a Petal.Width vs Petal.Length scatter plot using the “versicolor” dataset.
- Let’s check if Petal.Width and Petal.Length for species “versicolor” are correlated.
  - Read the help page of geom_smooth()
    - It will add a linear regression line in the plot that we will use to find the correlation
  - Set “method” argument to “lm” for the geom_smooth layer
- Save all commands to a file “practice-3.R”

TUTORIAL 4
Introductory Statistics with *R*

@QIU, HUNTER/CUNY

Two-sample *t*-test
I. Data & Hypothesis

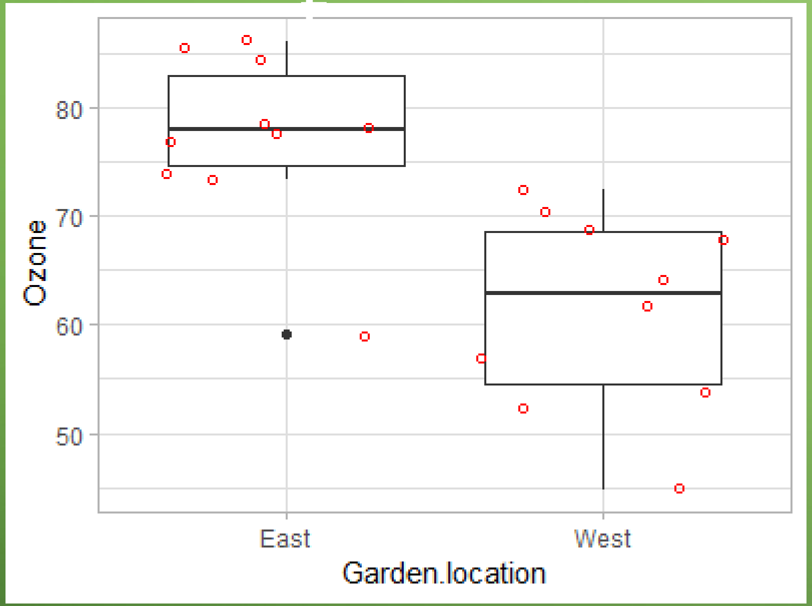
- `ozone ▯ read_csv(“ozone.csv”)`
- Question: Ozone level differs between east/west?
- Null Hypothesis (H_0): No difference

```
glimpse(ozone)

## Observations: 20
## Variables: 3
## $ Ozone           (dbl) 61.7, 64.0, 72.4, 56.8, 52.4, 4...
## $ Garden.location (fctr) West, West, West, West, West, ...
## $ Garden.ID       (fctr) G1, G2, G3, G4, G5, G6, G7, G8...
```

	Ozone	Garden.location	Garden.ID
	<dbl>	<chr>	<chr>
1	61.7	West	G1
2	64	West	G2
3	72.4	West	G3
4	56.8	West	G4
5	52.4	West	G5
6	44.8	West	G6
7	70.4	West	G7
8	67.6	West	G8
9	68.8	West	G9
10	53.7	West	G10
11	59.1	East	G11
12	78.5	East	G12
13	73.9	East	G13
14	86.1	East	G14
15	78	East	G15
16	84.4	East	G16
17	77.7	East	G17
18	76.8	East	G18
19	85.6	East	G19
20	73.3	East	G20

Two-sample *t*-test
II. Data Visualization

Boxplo t




```
ozone %>%
        ggplot(data = ozone, aes(x =
Garden.location, y = Ozone)) +
        geom_boxplot() +
        geom_jitter(shape=1, color="red") +
        theme_bw()
```

Violin plot


```
Ozone %>%
        ggplot(data = ozone, aes(x =
Garden.location, y = Ozone)) +
        geom_violin() +
        geom_jitter(shape=1, color="red") +
        theme_bw()
```

Two-sample *t*-test
III. Run *t*-test

```
# Do a t.test now....
t.test(Ozone ~ Garden.location, data = ozone)
```

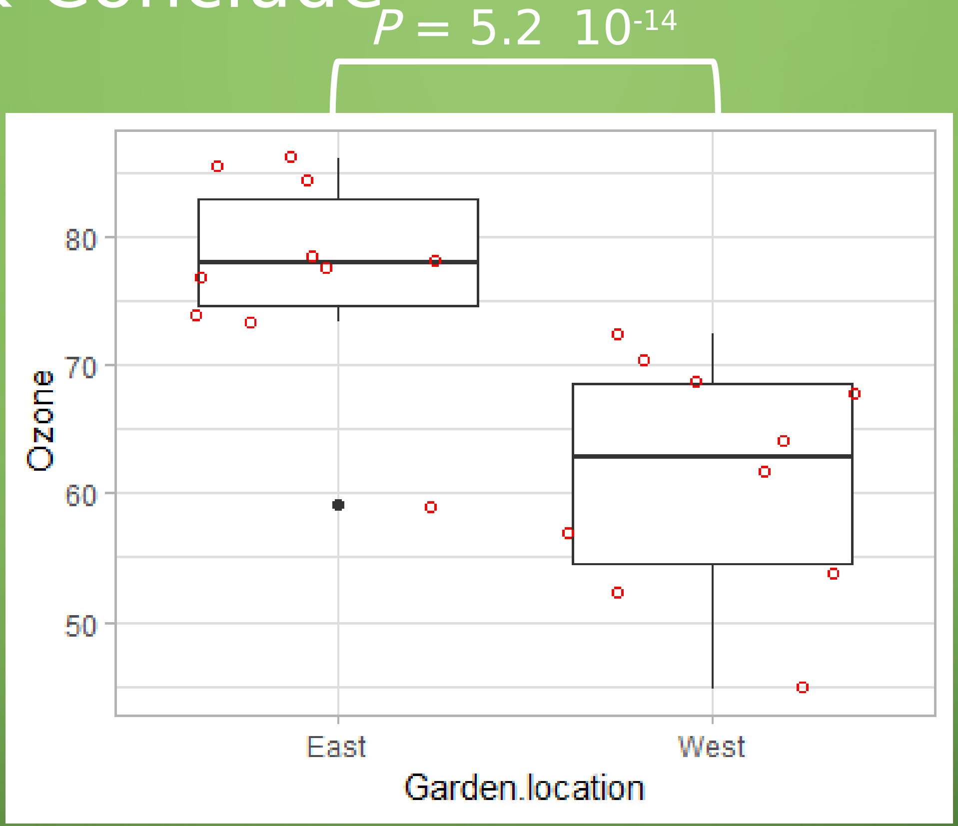
```
##
##  Welch Two Sample t-test
##
## data:  Ozone by Garden.location
## t = 4.2363, df = 17.656, p-value = 0.0005159
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.094171 24.065829
## sample estimates:
## mean in group East mean in group West
##              77.34              61.26
```

P value –

- Probability that observed difference is due to chance
- (more specifically) probability that t >= 4.2363 under null hypothesis (H_0)

T-test

IV. Re-plot & Conclude



Conclusions:

- Statistical conclusion: The null hypothesis (same mean) is rejected at p=5.2 10^{-14}
- Biological conclusion: The ozone level is significantly different between the east & west locations

Linear Regression
I. Data & Hypothesis

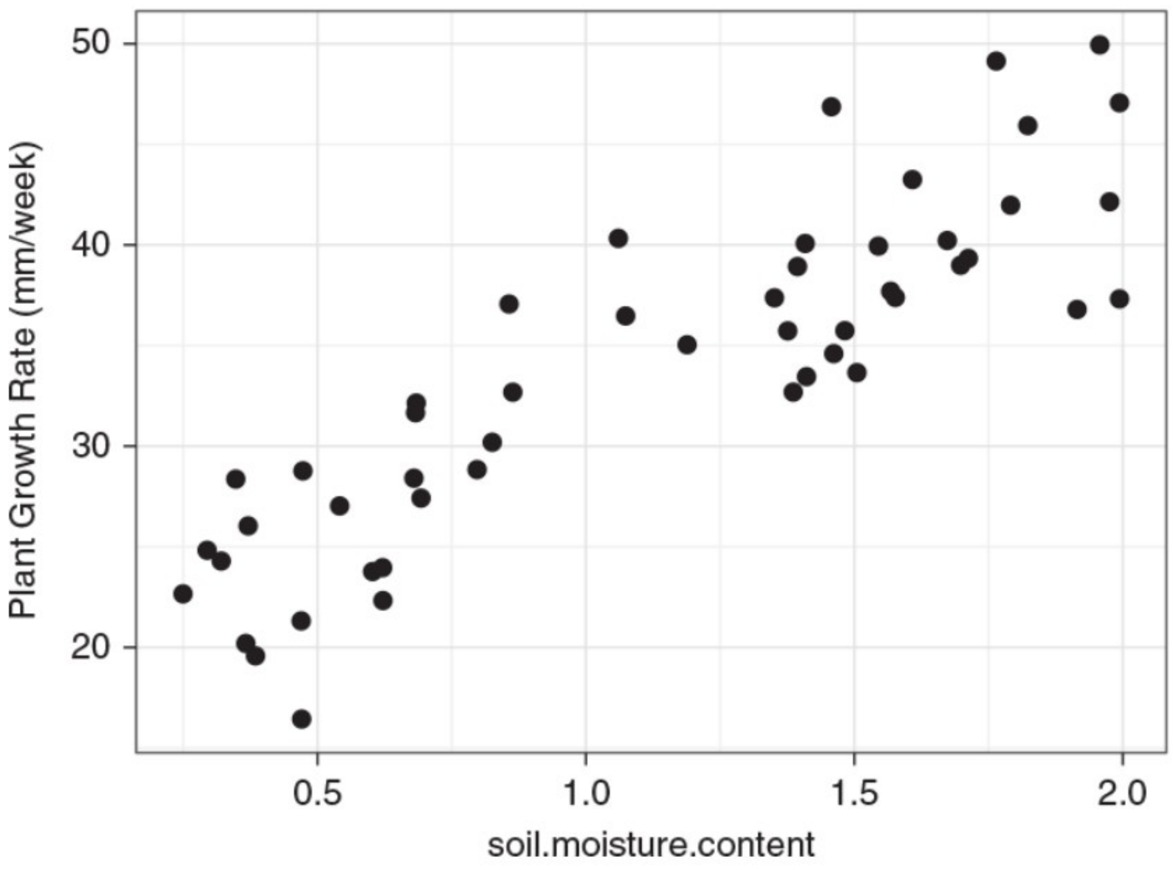
- Biological Question: Does soil moisture affect growth rate?
- Null Hypothesis (H_0): No correlation ($r=0$)

```
glimpse(plant_gr)

## Observations: 50
## Variables: 2
## $ soil.moisture.content  (dbl) 0.4696876, 0.5413106, 1.6...
## $ plant.growth.rate      (dbl) 21.31695, 27.03072, 38.98...
```

```
> plant_gr <- read_csv("plant.growth.rate.csv")
Parsed with column specification:
cols(
  soil.moisture.content = col_double(),
  plant.growth.rate = col_double()
)
> tbl_df(plant_gr)
# A tibble: 50 x 2
   soil.moisture.conte~ plant.growth.ra~
                  <dbl>            <dbl>
 1                0.470             21.3
 2                0.541             27.0
 3                1.70              39.0
 4                0.826             30.2
 5                0.857             37.1
 6                1.61              43.2
 7                0.250             22.7
 8                1.67              40.2
 9                1.46              46.9
10                0.473             28.8
# ... with 40 more rows
```

Linear Regression
II. Visualization

```
ggplot(plant_gr,
       aes(x = soil.moisture.content, y = plant.growth.rate)) +
         geom_point() +
         ylab("Plant Growth Rate (mm/week)") +
         theme_bw()
```



Linear Regression

III. Run linear model

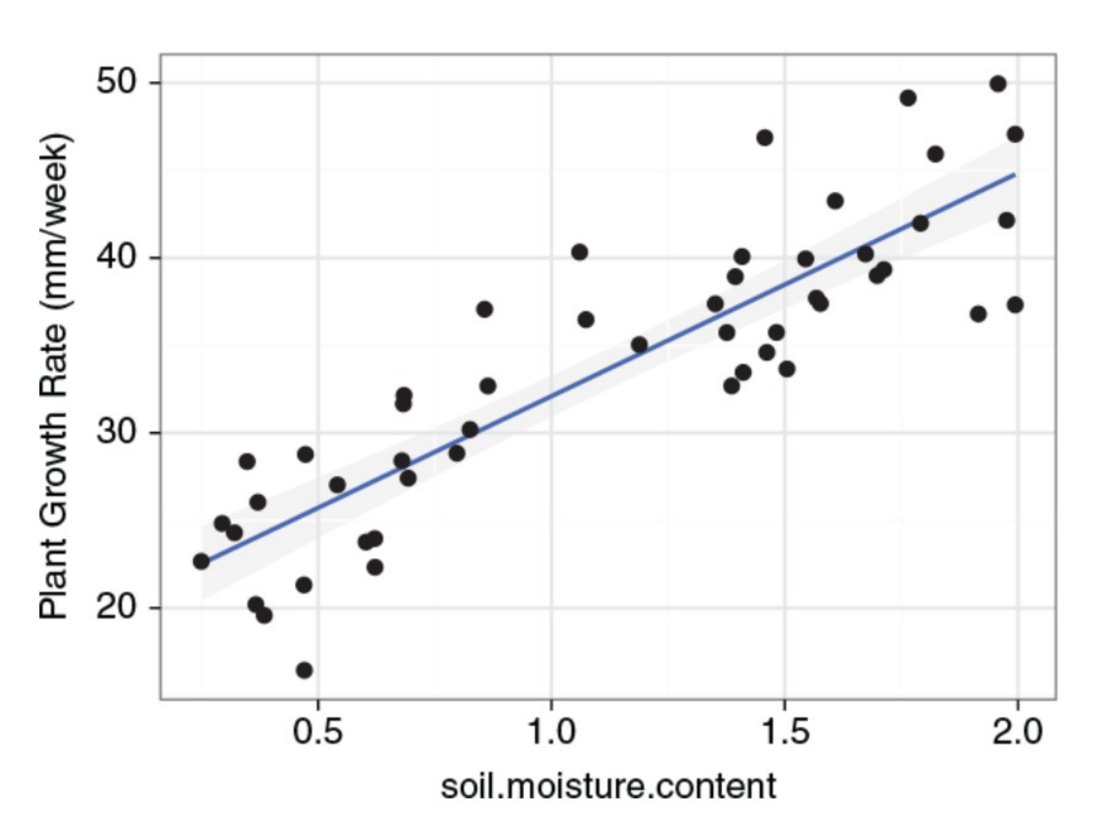
```
model_pgr <- lm(plant.growth.rate ~ soil.moisture.content,
                data = plant_gr)
```

```
summary(model_pgr)

##
## Call:
## lm(formula = plant.growth.rate ~ soil.moisture.content,
##     data = plant_gr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9089 -3.0747  0.2261  2.6567  8.9406
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             19.348      1.283   15.08   <2e-16
## soil.moisture.content   12.750      1.021   12.49   <2e-16
##
## (Intercept)           ***
## soil.moisture.content ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.019 on 48 degrees of freedom
## Multiple R-squared:  0.7648, Adjusted R-squared:  0.7599
## F-statistic: 156.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

Conclusions:

- The null hypothesis (no correlation) is rejected at $p<2.2e\text{-}16$
- The plant growth rate is significantly correlated with soil moisture with $R^2=0.7599$

Linear Regression
IV. Re-plot (add regression line & confidence band)




```r
ggplot(plant_gr, aes(x = soil.moisture.content,
          y = plant.growth.rate)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  ylab("Plant Growth Rate (mm/week)") +
  theme_bw()
```

PRACTICE #4

- Does the "Sepal.Length" differ between the two species "virginica" & "vesicolor"? Perform a *t*-test and include all 4 steps
- How about the "Sepal Width"? Perform a *t*-test and include all 4 steps
- Are the "Sepal.Width" and "Sepal.Length" correlated in the species "setosa"? Show all 4 steps.
- How about in the other two species?
- Batch testing the above correlation in all 3 species at once
- Save all commands to a file "practice-4.R"