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"
```

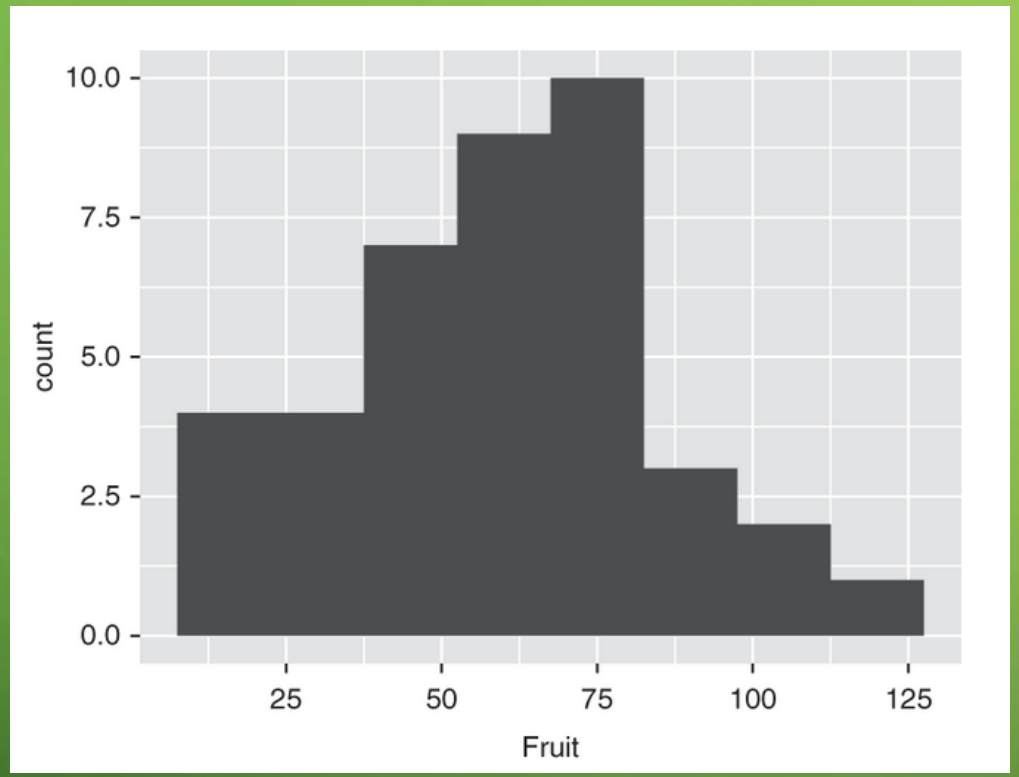
  - Aesthetics layer: refers to the scales onto which we will map our data
  - Geom layer: allows us to choose how the plot will look like.

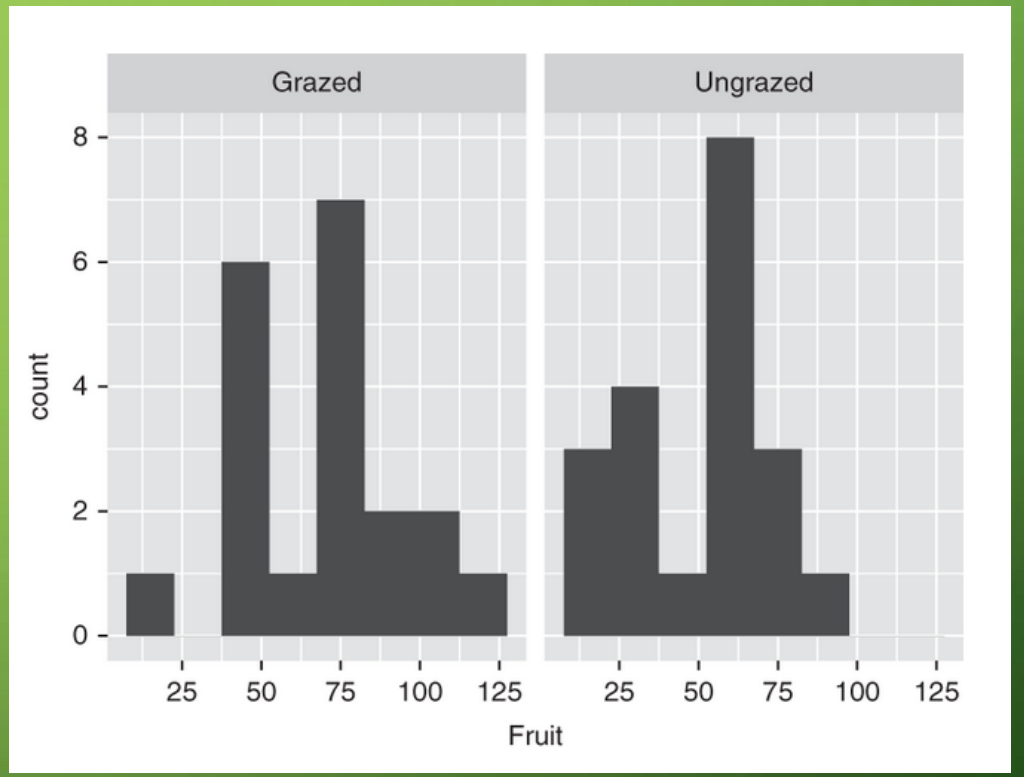
- Optional layers:
  - Theme layer: which controls all the non-data elements of graphics

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)
```

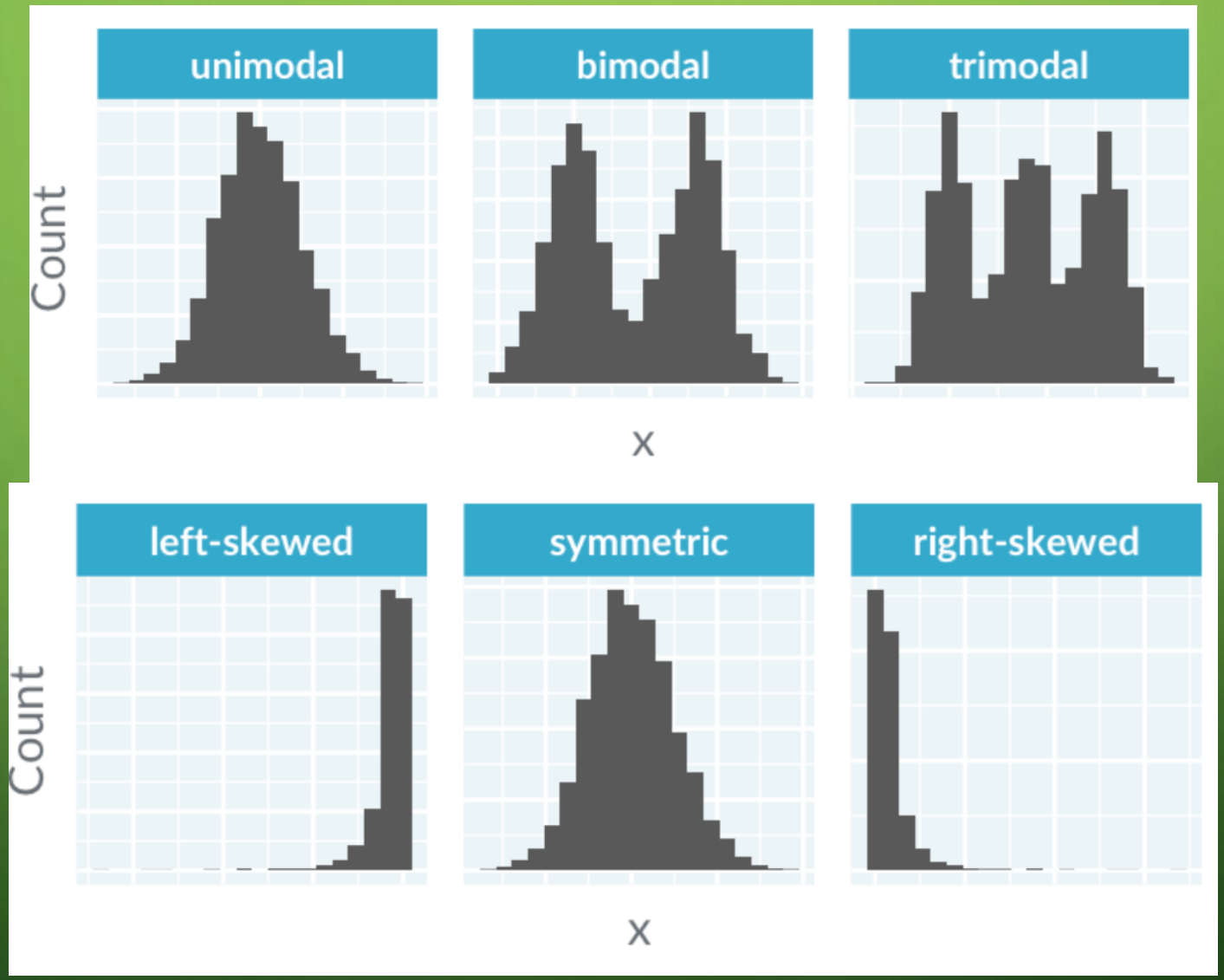
```
ggplot(compensation, aes(x = Fruit)) +
  geom_histogram(binwidth = 15) +
  facet_wrap(~Grazing)
```

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

UNDERSTANDING HISTOGRAMS: MODALITY AND SKEWNESS

Modality

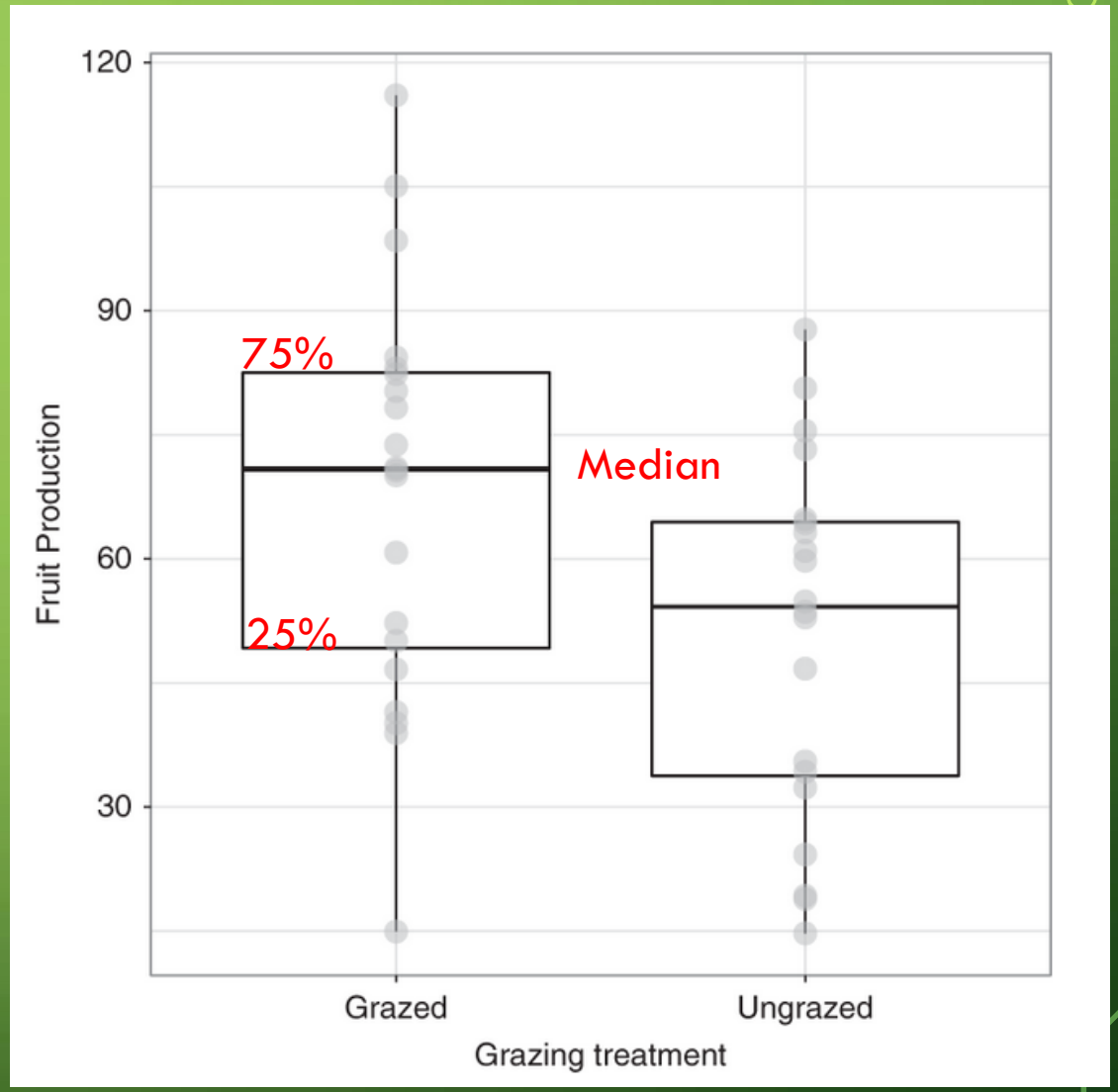
Skewness



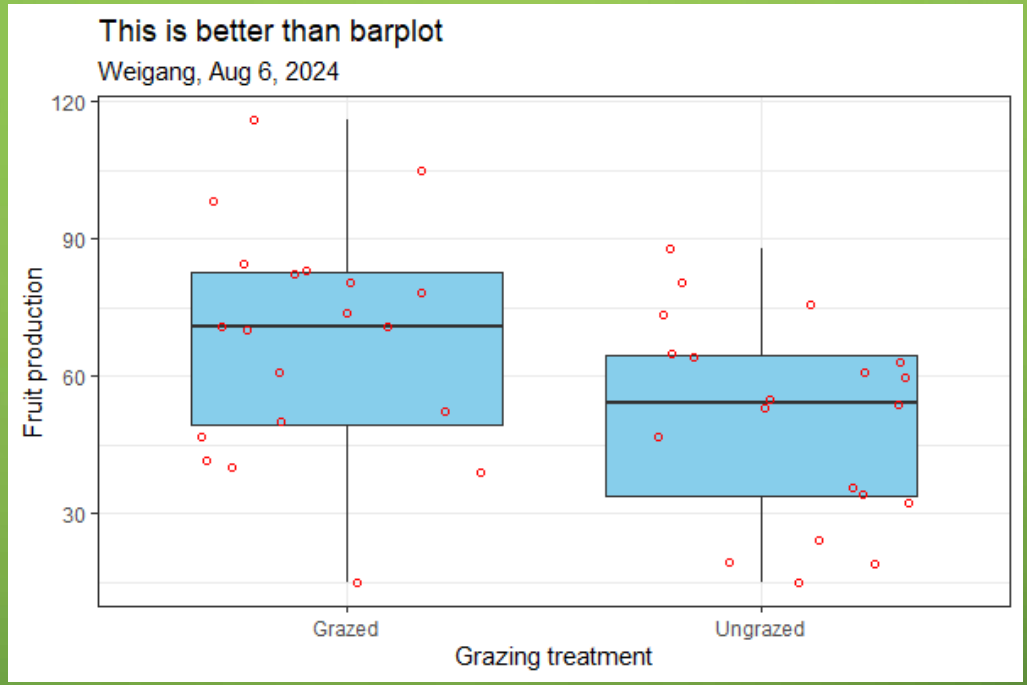
BOX PLOT: NUMERICAL VS CATEGORICAL

```
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()
```

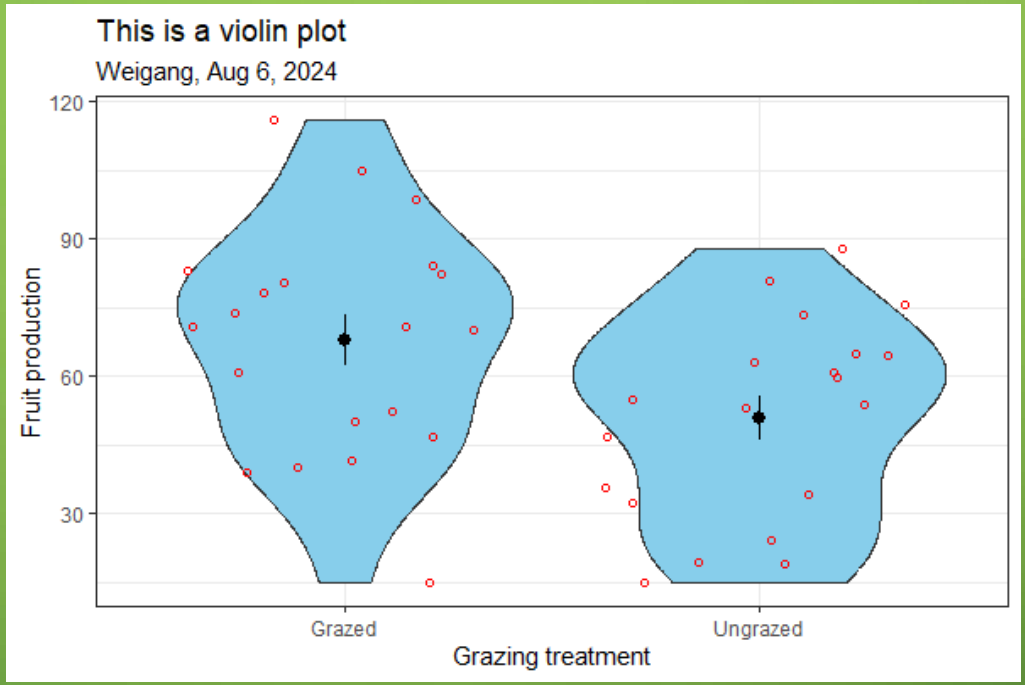


BOX PLOT




```
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")
```

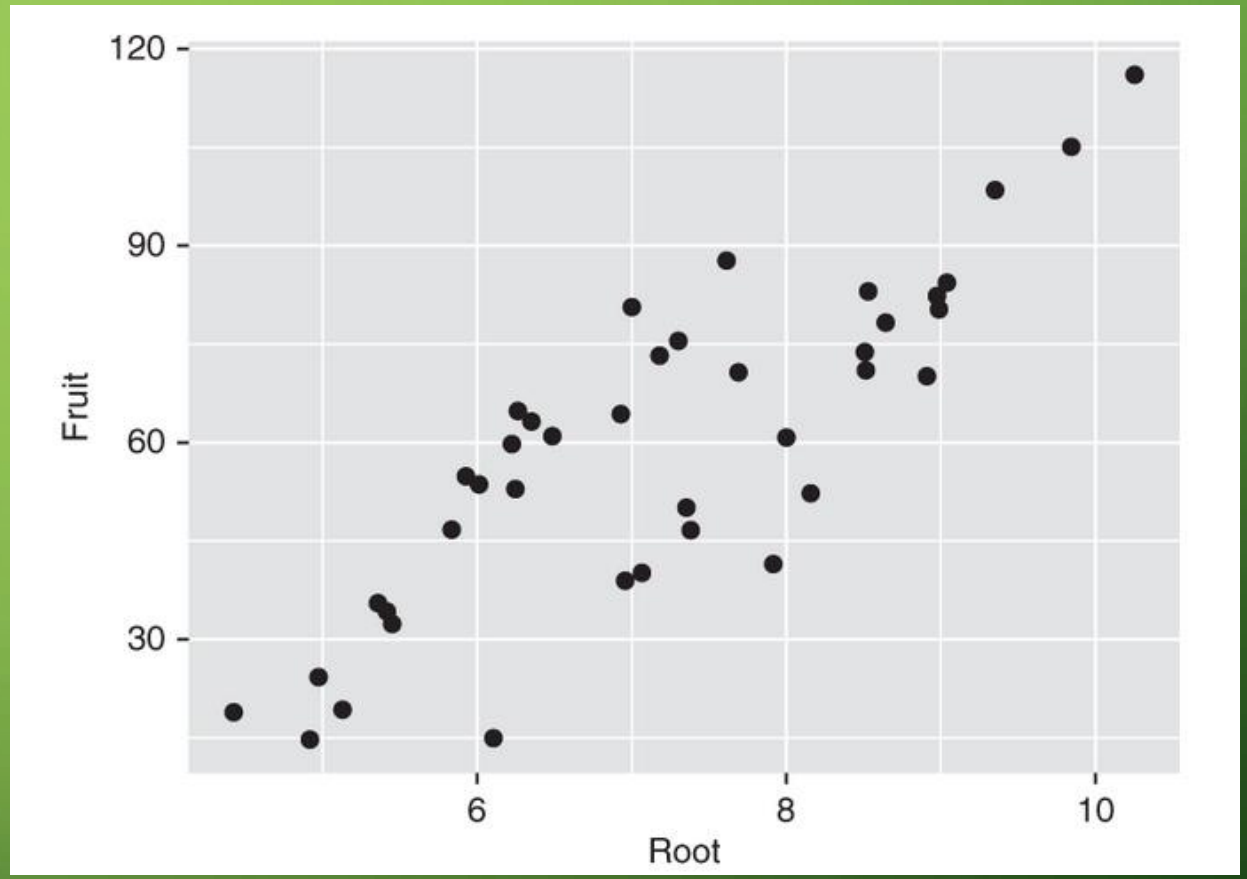
VIOLIN PLOT: NUMERICAL VS CATEGORICAL




```
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")
```

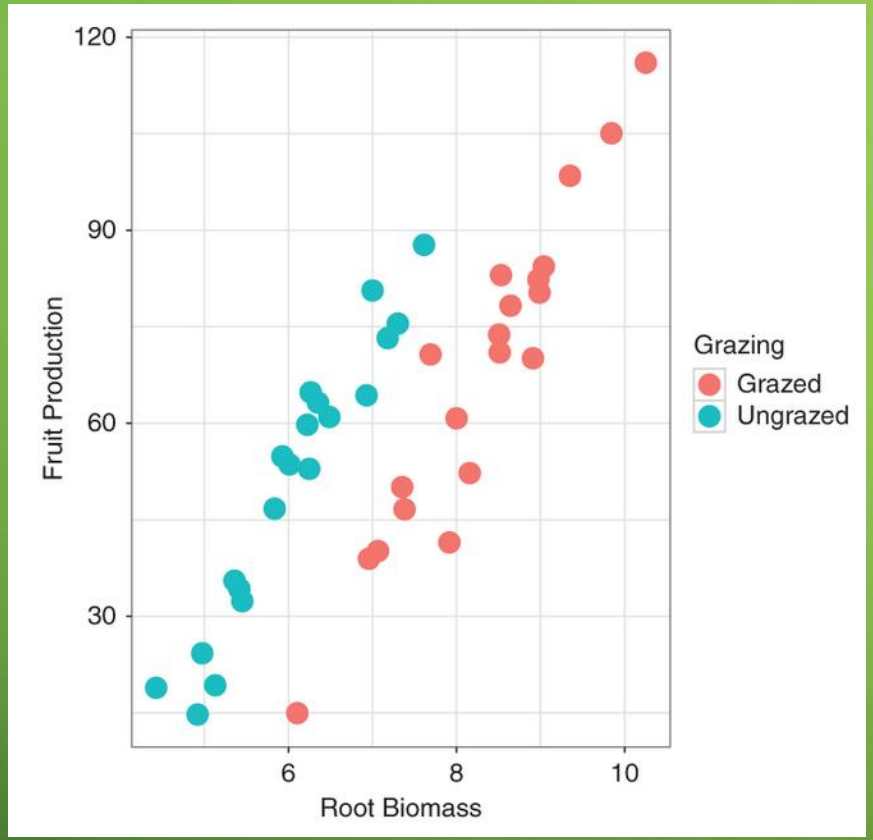
SCATTER PLOT: NUMERICAL VS NUMERICAL

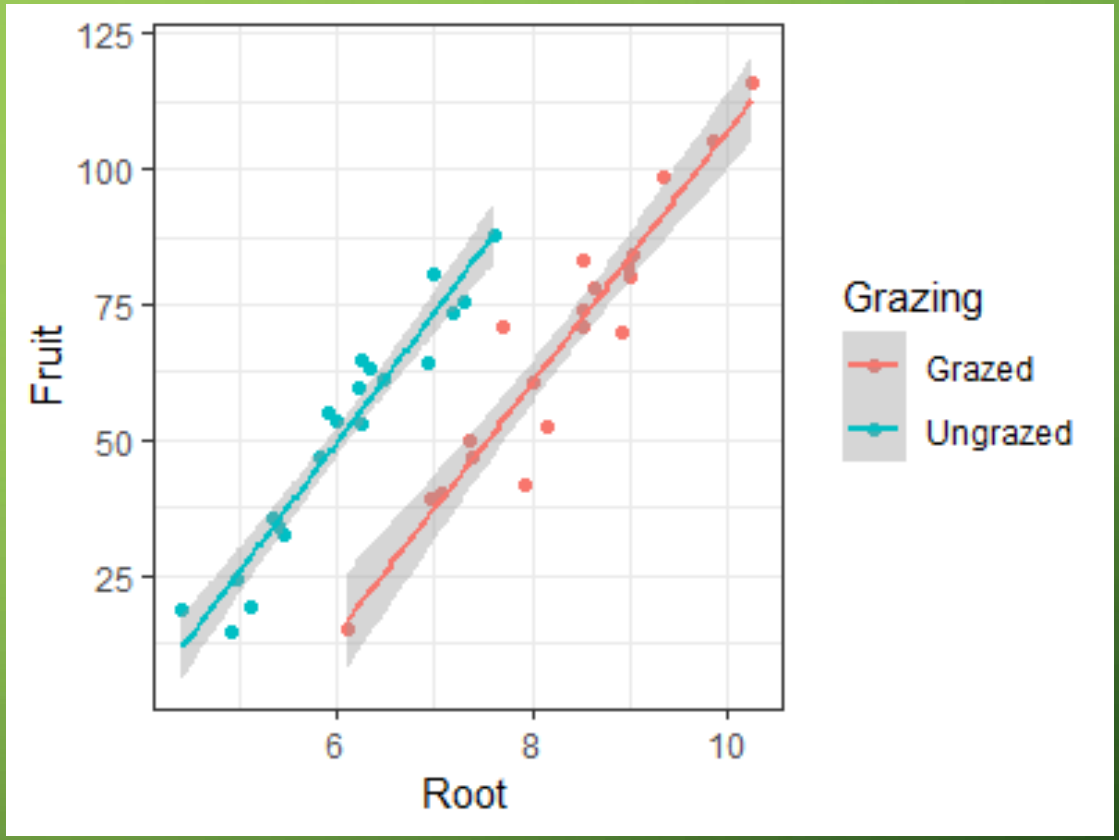
```r
# plotting basics with ggplot
# my tutorial script
# lots and lots of annotation!

# 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”