

## **Week 2: Data/Variable Types and ggplot and statistics (as Opposed to “Upper Case” Staistics)**

Scott Schwartz

September 19, 2021

## Class Check Round 1

<https://pollev.com/sta> (6 questions)

## Loading Data: Review

```
#install.packages("tidyverse")           <- not required since this is
library(tidyverse)                      # <- preinstalled on jupyterhub

# http://database.coffeeinstitute.org/
# https://github.com/rfordatascience/tidytuesday/blob/master/
#           data/2020/2020-07-07
coffee_ratings <- read_csv("coffee_ratings.csv")
# the printout below is normal even though RStudio colors it red

## Rows: 1338 Columns: 36
## -- Column specification -----
## Delimiter: ","
## chr (18): species, owner, country_of_origin, farm_name, mill, company, altit...
## dbl (18): total_cup_points, aroma, flavor, aftertaste, acidity, body, balanc...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

## Viewing Data: Review

```
coffee_ratings %>% glimpse() # useful printout isn't actually a tibble
```

```
## Rows: 1,338
## Columns: 36
## $ total_cup_points      <dbl> 90.58, 89.92, 89.75, 89.00, 88.83, 88.83, 88.75, ~
## $ species                <chr> "Arabica", "Arabica", "Arabica", "Arabica", "Arab~
## $ owner                  <chr> "metad plc", "metad plc", "grounds for health adm~
## $ country_of_origin       <chr> "Ethiopia", "Ethiopia", "Guatemala", "Ethiopia", ~
## $ farm_name               <chr> "metad plc", "metad plc", "san marcos barrancas \~
## $ mill                    <chr> "metad plc", "metad plc", NA, "wolensu", "metad p~
## $ company                 <chr> "metad agricultural developmet plc", "metad agric~
## $ altitude                <chr> "1950-2200", "1950-2200", "1600 - 1800 m", "1800--~
## $ region                  <chr> "guji-hambela", "guji-hambela", NA, "oromia", "gu~
## $ producer                <chr> "METAD PLC", "METAD PLC", NA, "Yidnekachew Dabess~
## $ in_country_partner      <chr> "METAD Agricultural Development plc", "METAD Agri~
## $ harvest_year             <chr> "2014", "2014", NA, "2014", "2014", "2013", "2012~
## $ grading_date             <chr> "April 4th, 2015", "April 4th, 2015", "May 31st, ~
## $ variety                  <chr> NA, "Other", "Bourbon", NA, "Other", NA, "Other", ~
## $ processing_method        <chr> "Washed / Wet", "Washed / Wet", NA, "Natural / Dr~
## $ aroma                    <dbl> 8.67, 8.75, 8.42, 8.17, 8.25, 8.58, 8.42, 8.25, 8~
```

## RStudio + knitter: An aside

```
coffee_ratings # %>% head() # %>% knitr::kable()
```

Since we use pdf output in STA130 we won't get into the weeds on this, but...

Table printouts in RStudio are excellent; not so much for pdf output; but for html there's

- <https://bookdown.org/yihui/rmarkdown-cookbook/kable.html>

knitr::kable()

you don't actually have to use library(knitr) if you use knitr::

- <https://bookdown.org/yihui/rmarkdown-cookbook/kableextra.html>

library(kableExtra) # for even more control

# Types of Data/Variables

All artwork thanks to @allison\_horst!

## Numerical/Quanitative

### CONTINUOUS

measured data, can have  $\infty$  values within possible range.



I AM 3.1" TALL  
I WEIGH 34.16 grams

### DISCRETE

OBSERVATIONS CAN ONLY EXIST AT LIMITED VALUES, OFTEN COUNTS.



I HAVE 8 LEGS  
and  
4 SPOTS!

@allison\_horst

## Categorical/Qualitative

### NOMINAL

UNORDERED DESCRIPTIONS



i'm a  
snail!!



-i'm a  
butterfly!

### ORDINAL

ORDERED DESCRIPTIONS



-i am  
unhappy.  
-i am  
OK.  
-i am  
Awesome!!!

### BINARY

ONLY 2 MUTUALLY EXCLUSIVE OUTCOMES



@allison\_horst

What should these apply to?

Data or Variables? What's the difference?

# Binary Data/Variables



- **Binary** just means **two-level** (either/or) **categorical** **Binary** can always be represented as **logical (*boolean*)** by just asking "Is it the first category? TRUE or FALSE?"
- Q: "Is this animal extinct?" A: TRUE/FALSE

So **binary/two-level** (either/or) **categorical** variables are equivalent to **logical** TRUE/FALSE (***boolean***) variables

# Boolean Data/Variables

A logical type has “&” and “|” rules

---

“&” (and) operator

TRUE & TRUE = TRUE

TRUE & FALSE = FALSE

“|” (or) operator

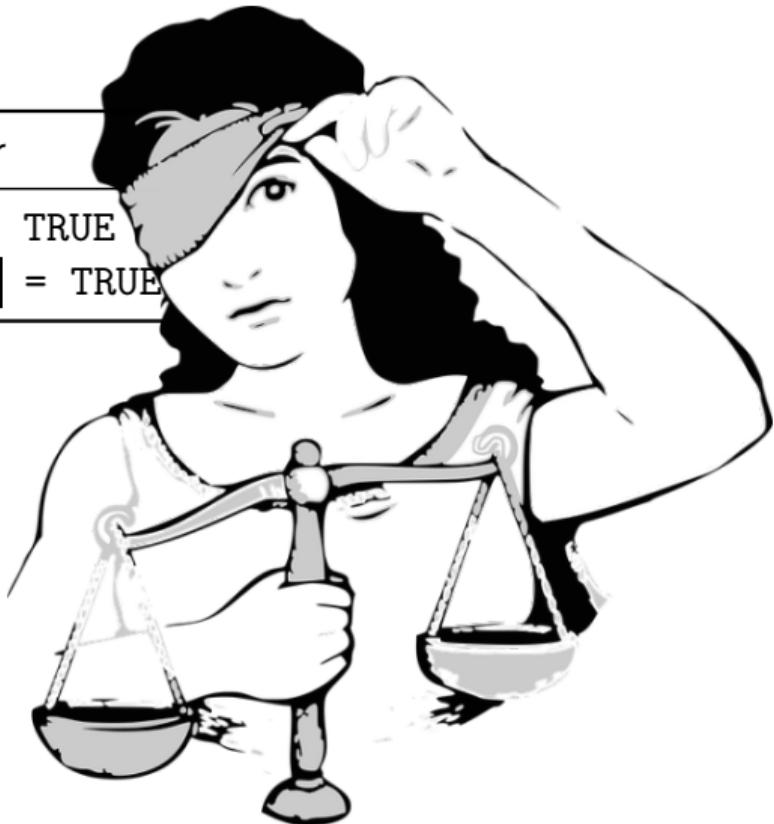
TRUE | TRUE = TRUE

TRUE | FALSE = TRUE

---

**Comparisons** create logical types

- < or <=                    123 < 1.23
- > or >=                    123 >= 1.23
- == or !=                    123 != 1.23
- !(123 == 1.23)
- !!TRUE = !FALSE = TRUE

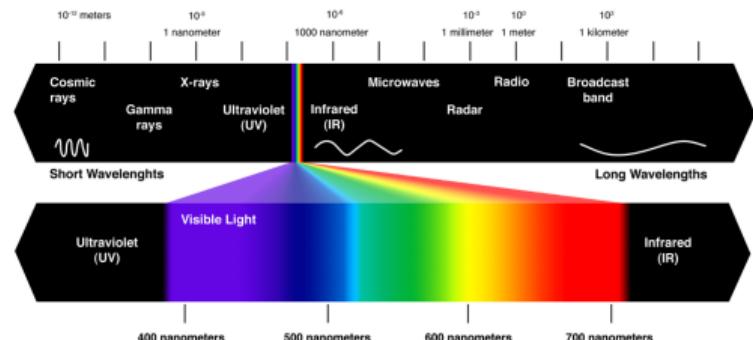


# R Data/Variables

ALL R types ARE “numbers”; but, these exist on a Quantitative-Qualitative spectrum:

	1.23	123	TRUE/FALSE	as.factor("char str")
	<b>Continuous</b>	<b>Discrete/Ordinal</b>	<b>Binary</b>	<b>Nominal</b>
class	numeric	numeric	logical	factor
typeof	double	double	logical	integer
is.integer	FALSE	FALSE	FALSE	FALSE
	"float"	"integer"	"boolean"	"factor"

- 123 and 1.23 aren't different for R
- 123  $\equiv$  **Ordinal**; factor  $\equiv$  **Nominal**
- R has logical **Binary** variables
- **Careful:** class, typeof, integer have unexpected meanings in R...



## Class Check Round 2

<https://pollev.com/sta> (6 questions)

## Changing R Data/Variables

“R alchemy” to change data types across the “R data type spectrum” is possible when sensible

---

as.character(1.23)	as.logical(as.numeric("0"))	as.numeric(FALSE)
as.numeric("1")	as.logical(as.numeric("1"))	as.numeric(TRUE)

---

which is actually exactly what  
as.factor("char str") does!

R does this automatically, sometimes.

When it does it's called ***coercion***:

---

TRUE + 1.23	paste(TRUE + 1.23, " = 2.23")
FALSE == 1	paste(TRUE, " = 1")

---



You can guess/figure out what the paste function does, right?

## Class Check Round 3

<https://pollev.com/sta> (3 questions)

# Visualizing Different Data Types with ggplot2

`library(ggplot)`



**Why care about data types?** So we use the appropriate visualization for each type.

ggplot2 is part of the tidyverse. Standard usage is based on a grammar of graphics, and there are many learning resources available (including the R4DS textbook, the official cheatsheet, and the DoSS Toolkit)! A big recommendation and hint though (as usual – but this time from ggplot2 itself) is to search and find answers online!

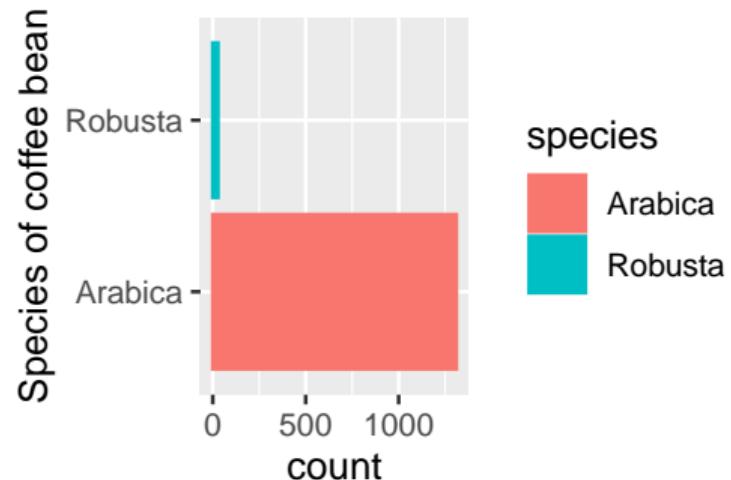
# Data Types and their ggplots

```
# Code chunk figure/text ratio options: fig.height=2, fig.width=3  
ggplot(data=coffee_ratings, # difference between `color` & `fill`?  
        aes(x=species, color=species, fill=species)) +  
  geom_bar() + labs(x="Species of coffee bean") + coord_flip()
```

① What kind of variable is species?

- A. Nominal Categorical
- B. Ordinal Categorical
- C. Continuous
- D. Boolean

② A **barplot** (geom\_bar) is NOT an appropriate visualization for what variable type?



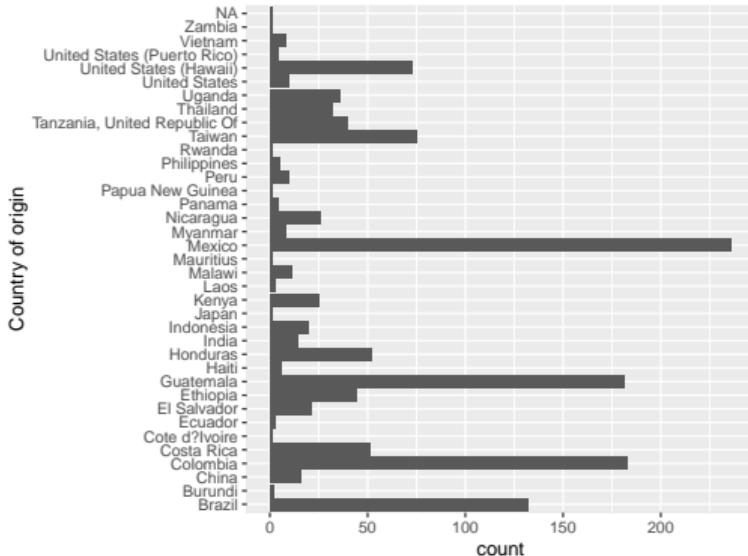
# Data Types and their ggplots

```
# Code chunk figure/text ratio options: fig.height=4.5, fig.width=6
coffee_ratings %>% ggplot(aes(x=country_of_origin)) +
  geom_bar() + labs(x="Country of origin") + coord_flip()
# Can this visualization be improved? Hint: google "order geom_bar"
```

① What kind of variable is country\_of\_origin?

- A. Nominal Categorical
- B. Ordinal Categorical
- C. Continuous
- D. Boolean

② What are interesting features of this **barplot** (`geom_bar`) compared to the last?



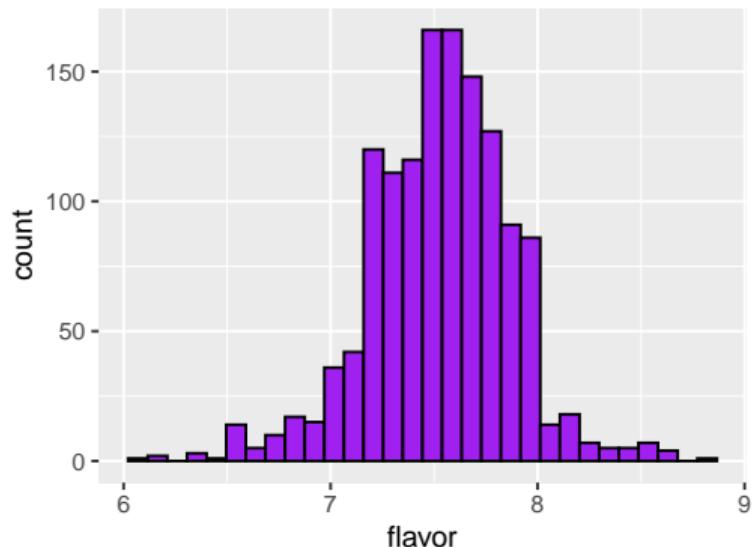
# Data Types and their ggplots

```
# Code chunk figure/text ratio options: fig.height=3, fig.width=4
coffee_ratings %>% ggplot(aes(x=flavor)) +
  geom_histogram(bins=30, color="black", fill="blue")
# What's the `bins` parameter do? What's the right choice for `bins`?
```

① What kind of variable is flavor?

- A. Nominal Categorical
- B. Ordinal Categorical
- C. Continuous
- D. Boolean

② A **histogram** (`geom_histogram`) is an appropriate visualization for what variable type?



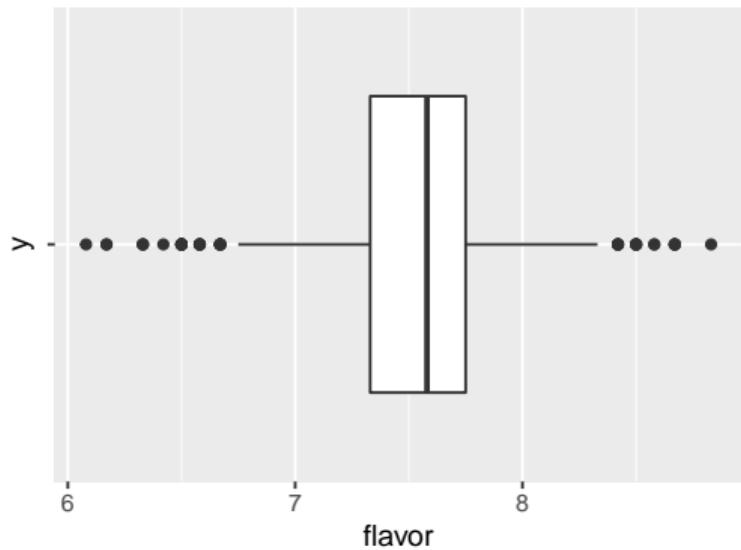
# Data Types and their ggplots

```
# Code chunk figure/text ratio options: fig.height=3, fig.width=4  
coffee_ratings %>% ggplot(aes(x=flavor, y="")) + geom_boxplot()  
# How does this code differ from the code for `geom_histogram()`?
```

- ① A **boxplot** (`geom_boxplot`) is an appropriate visualization for what variable type?

- A. Nominal Categorical
- B. Ordinal Categorical
- C. Continuous
- D. Boolean

- ② Why is the middle line of the box plot located more towards the right?

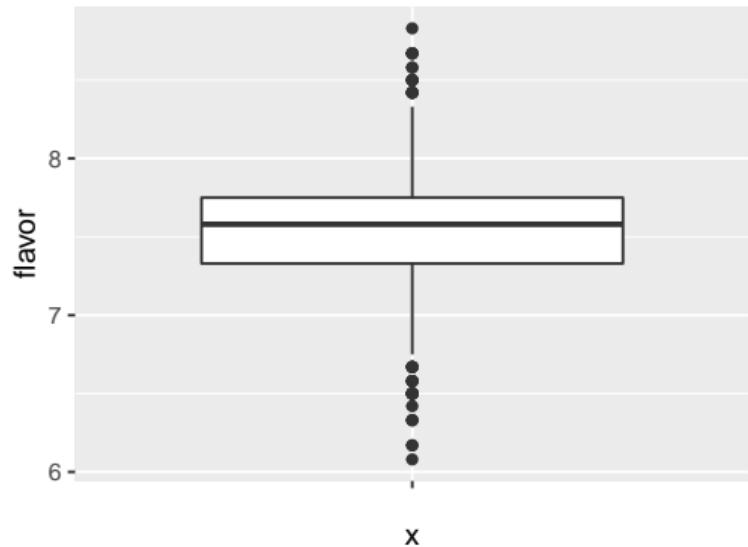


# Data Types and their ggplots

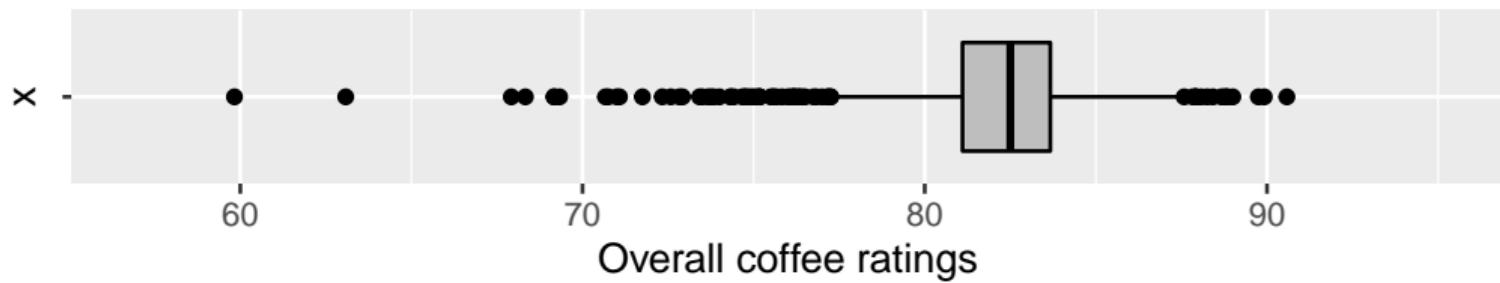
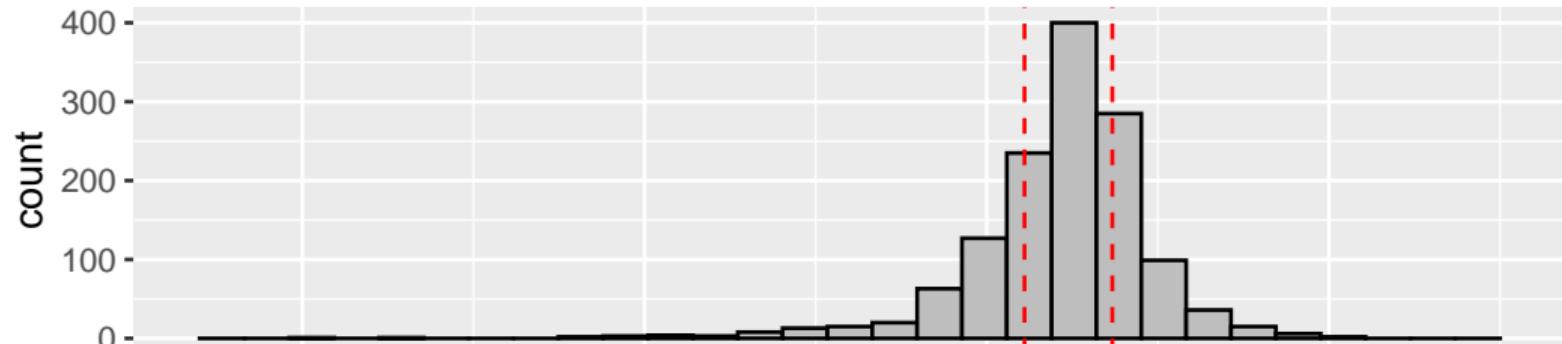
```
# Code chunk figure/text ratio options: fig.height=3, fig.width=4  
coffee_ratings %>% ggplot(aes(x="", x=flavor)) + geom_boxplot()  
# Something changed? What? And do you like it better? Why or why not?
```

## Boxplot (geom\_boxplot) components

- ① Median (50th data percentile)
- ② Interquartile range (IQR) box  
(25th to 75th data percentiles)
- ③ Whiskers cover farthest “outliers”  
 $\leq 1.5 \times \text{IQR}$  from the “IQR box”
- ④ “Outliers” are the data points  
more extreme than the above rule

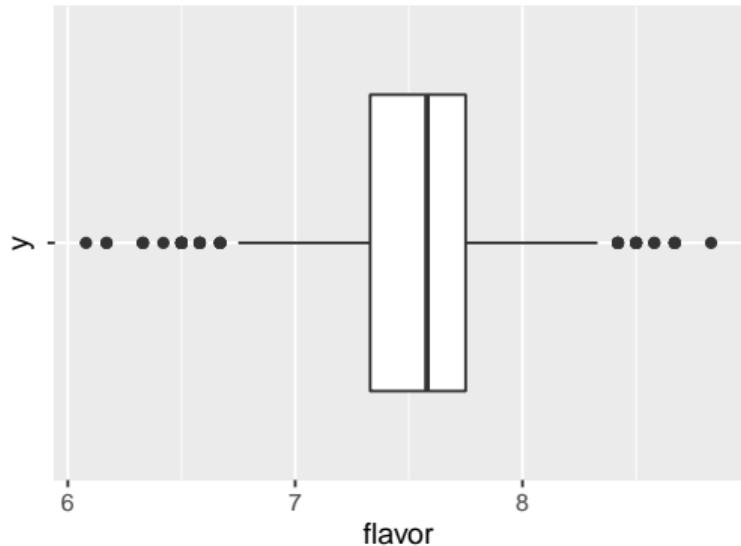
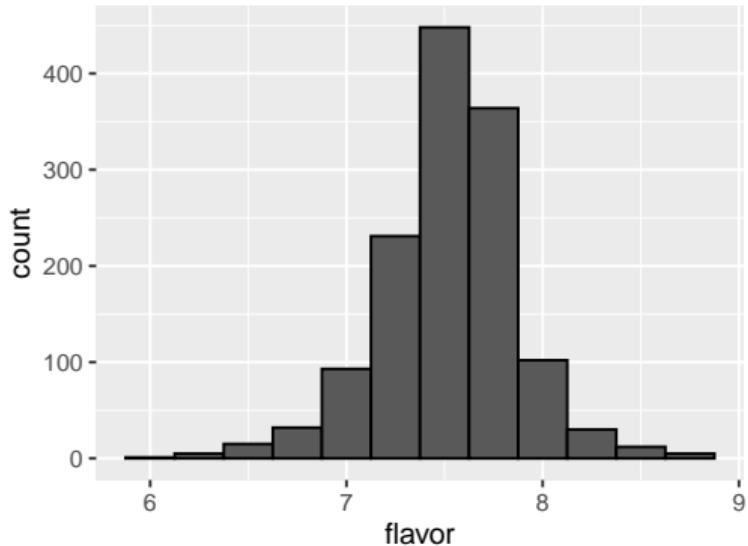


## Histograms Versus Boxplots



# Histograms Versus Boxplots

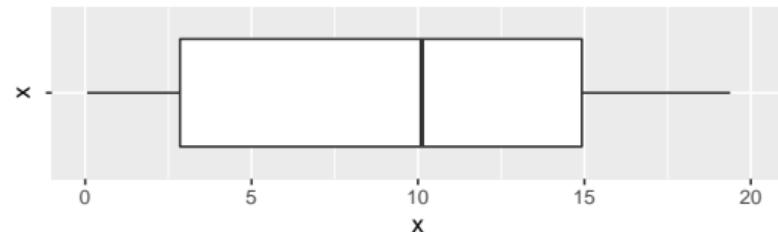
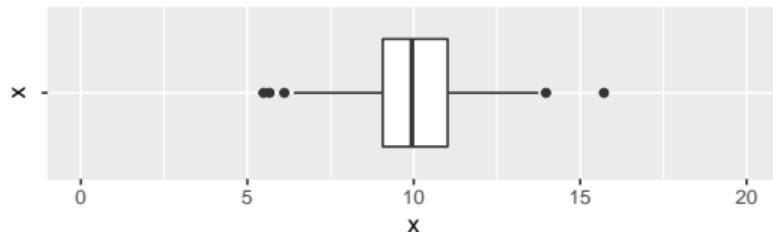
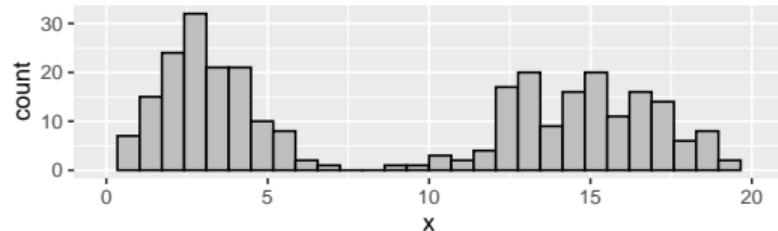
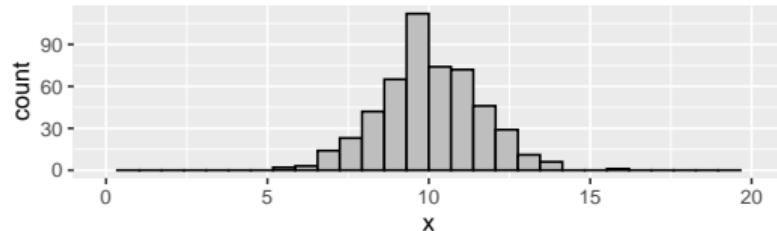
```
# Code chunk figure/text ratio options: fig.height=3, fig.width=4  
geom_histogram(bins=12, color="black") VERSUS geom_boxplot()  
# What reasons are there to prefer boxplots over histograms, if any?
```



## Which do you like better?

# Code chunk figure/text ratio options: fig.height=3.2

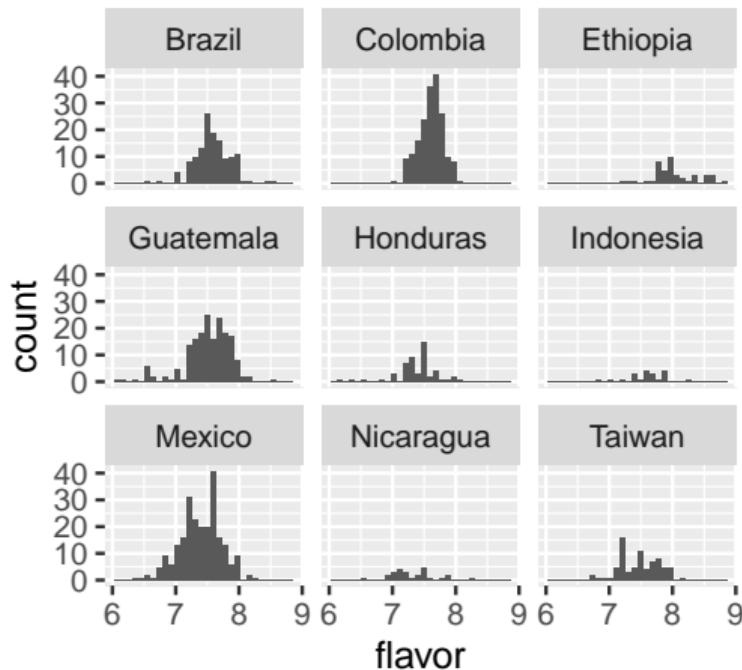
geom\_histogram(bins=30, color="black", fill="gray") VERSUS geom\_boxplot()



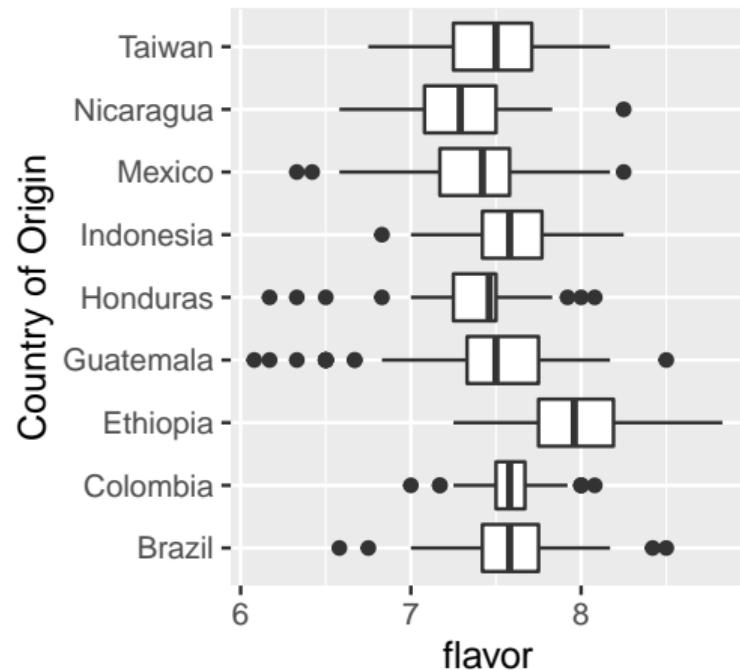
library(gridExtra) # gives functionality to arrange the plots as above  
library(egg) # also gives functionality to arrange the plots as above

# Which do you like better?

(Google "free yaxis facet\_wrap")



(But what is this missing?)



Why would this be important?

## (First Order) Distributional Characteristics: Center/Location statistics

**Median:** 50<sup>th</sup> percentile of the data

- Half of the data is less than or equal to the median
- Half of the data is greater than or equal to the median

**Mean:** the average value in the data

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

**Mode:** the most frequent data value

① What happens to the **Median/Mean** as the largest data point increases? *Is the **Median** or **Mean** more “robust” relative to the largest data point?*

② For which type of variable is the **Mode** the least meaningful? Why?

- A. Nominal Categorical
- B. Ordinal Categorical
- C. Continuous
- D. Boolean

## (Second Order) Distributional Characteristics: Spread/Scale statistics

IQR (from the **boxplot**):

$$75^{\text{th}} \text{ percentile} - 25^{\text{th}} \text{ percentile}$$

---

**Range:**

$$\max_{i=1,\dots,n} x_i - \min_{i=1,\dots,n} x_i$$

---

**Variance:** (almost)  
average squared distance from mean

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

**Standard Deviation:**  
Square root of the **Variance**

$$s = \sqrt{s^2}$$

**Squared Units... don't mean much**

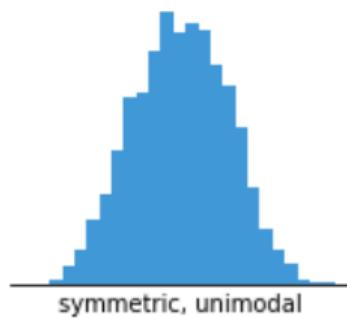
"Average 'squared distance'?"

**Original Units... are interpretable**

& range is often  $\sim 5$  standard deviations

## (Higher Order) Distributional Characteristics: Skewness + Modality

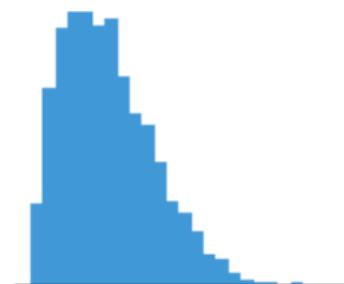
---



symmetric, unimodal



skew left



skew right



uniform



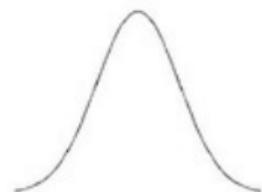
bimodal



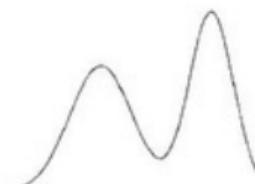
multimodal

## (Higher Order) Distributional Characteristics: Skewness + Modality

---



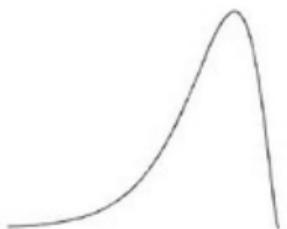
Unimodal



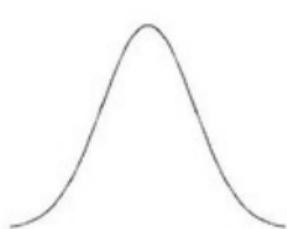
Bimodal



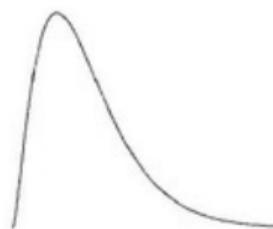
Multimodal



Left-Skewed



Symmetric



Right-Skewed

## Review Quiz

1 What kind of variable is **sweetness**?

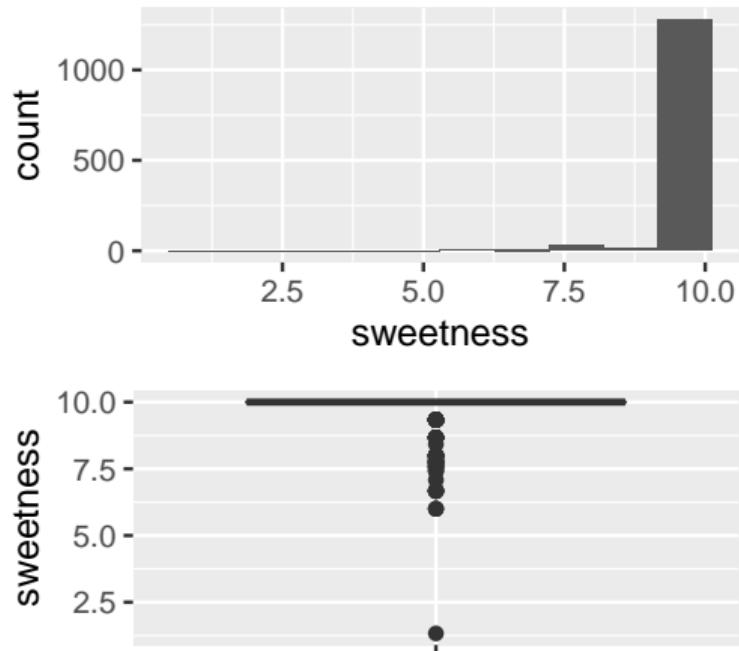
- A. Nominal Categorical
- B. Ordinal Categorical
- C. Continuous
- D. Boolean

Is it

2 **symmetric, left, or right-skewed?**

3 **unimodal, bimodal, or multimodal?**

4 What's the approximate distributional **Median, Mean, and Mode?**



## Self-Quiz

### Ask Yourself . . .

- ① What's the difference between continuous numerical / discrete numerical / nominal categorical / ordinal categorical / binary / logical (boolean) Data types?
- ② What's the difference between data and a variable?
- ③ What's the difference between data and a distribution?
- ④ What is the difference between geom\_ 's bar, histogram, and boxplot?
- ⑤ Do you know how to make them in R, and place (and size them) in RStudio?
- ⑥ Do you know how to put multiple boxplots in the same ggplot2 figure?
- ⑦ Do you know the mean, median, range, IQR, var, and sd statistics R functions?
- ⑧ Can you visually describe data distributions based on their data types, visualizations, the above statistics (or their rough approximations), and characterizations of their ***modality***, ***skewness*** and, ***outliers***?

# Describing Numerical/Categorical Distributions

## Focus of this weeks Problem Set and TUT

- What is the shape of the data and where is it located? Where is its center?
- For categorical data, which category occurred the most and least frequently?
- Is the observed data concentrated near a particular value or category?
- What are the mean, median, mode of data or a distribution, and what makes them different?
- Is it symmetric, with the data centered in the middle?
- What is the standard deviation and interquartile range of data or a distribution?
- Is it left-skewed (with a long left tail and the data mostly to the right of this)?
- Is it right-skewed (with a long right tail and the data mostly to the left of this?)
- How much spread is there in the data (and relative to what)?
- Are the tails of the distribution heavy-tailed, potentially producing lots of outliers?
- Are there any outliers, i.e., extreme values in a data set?
- Are the tails of the distribution thin-tailed, so the data doesn't really produce outliers?
- Is the data or distribution unimodal, bimodal, multimodal, or uniform?
- If there are modes, how many are there and where are they?

## Rstudio Demo

- ① Click this [jupyterhub](#) repo launcher link