PSY 503: Foundations of Statistical Methods in Psychological Science

Data Generating Processes (DGPs)

Descriptives/Visualization, ggplot primer

Suyog Chandramouli

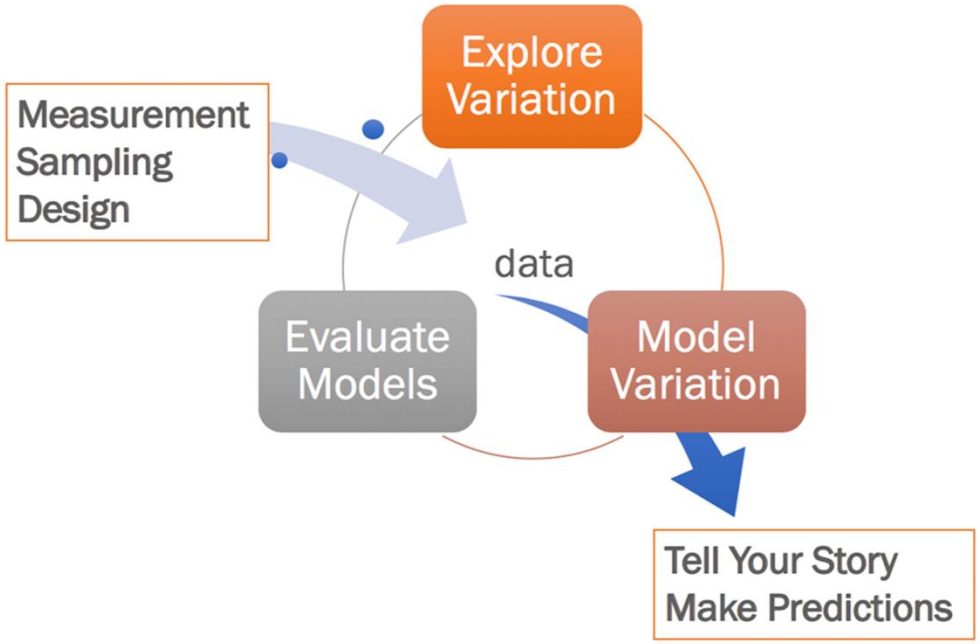
Zoom & 311 PSH (Princeton University)

15th September, 2025

Why Statistics?

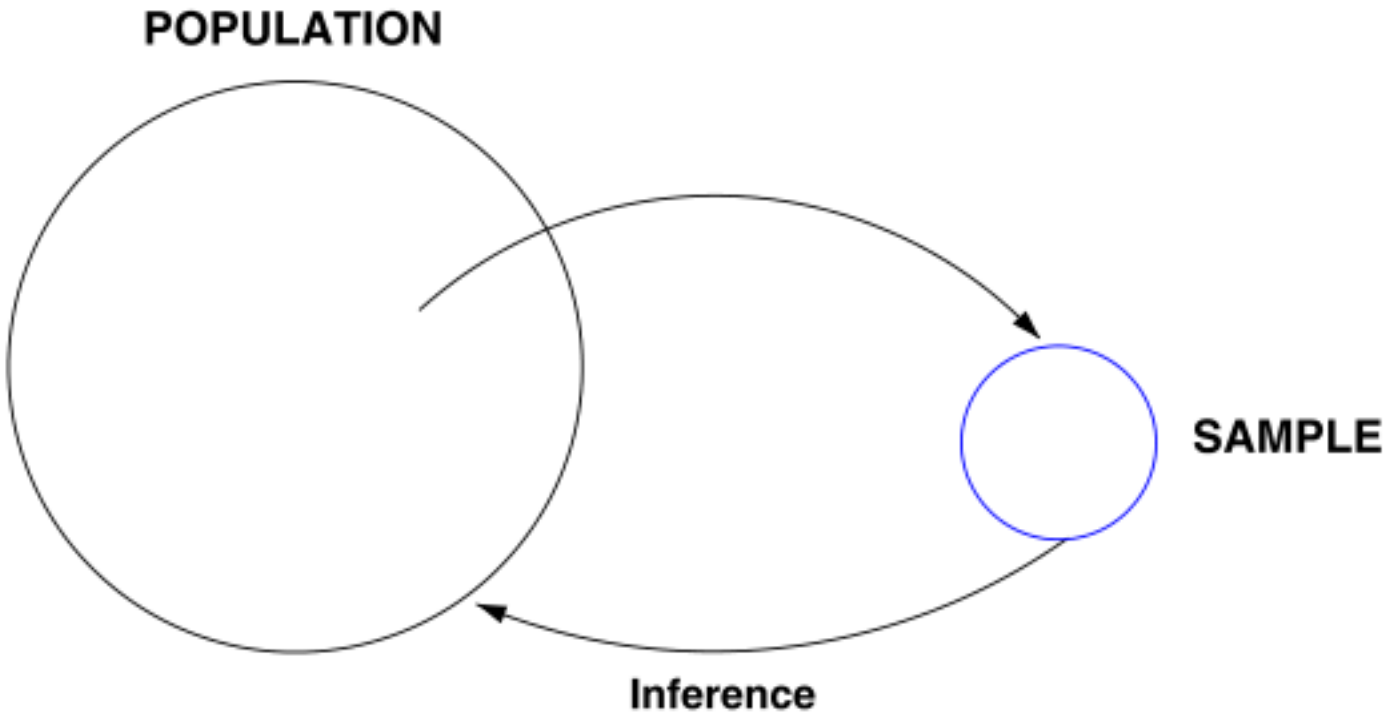
- Statistics helps us to:
  - make sense of variation
  - make general claims from limited data

- Statistics requires careful thought (Simpson's paradox)
- Merely summary statistics can be misleading.
- Examining the data, and EDA help with understanding variation.



Why bother with understanding variation in data?

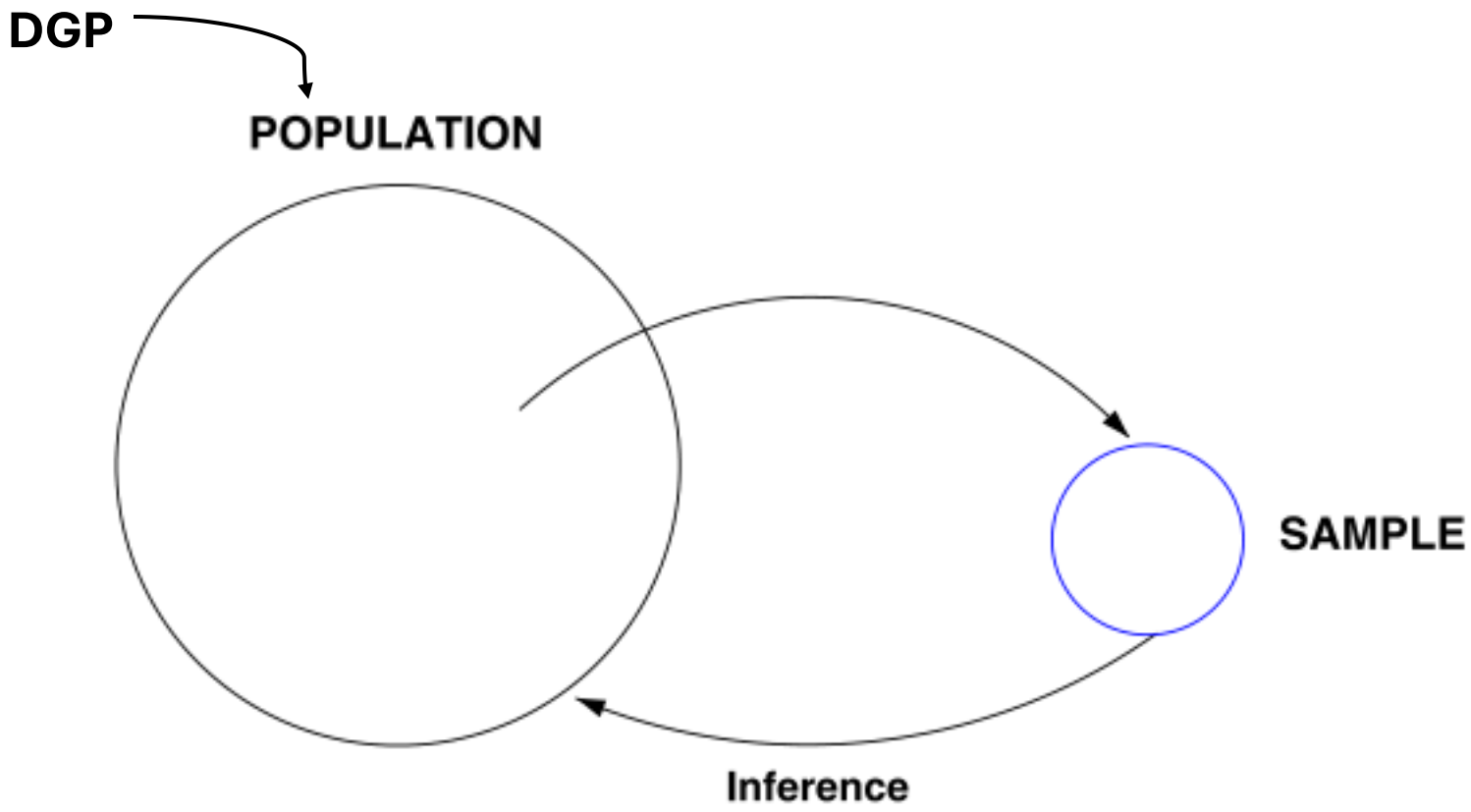
We want to infer what may be going on in the population level



Why bother with understanding variation in data?

We want to infer what may be going on in the population level

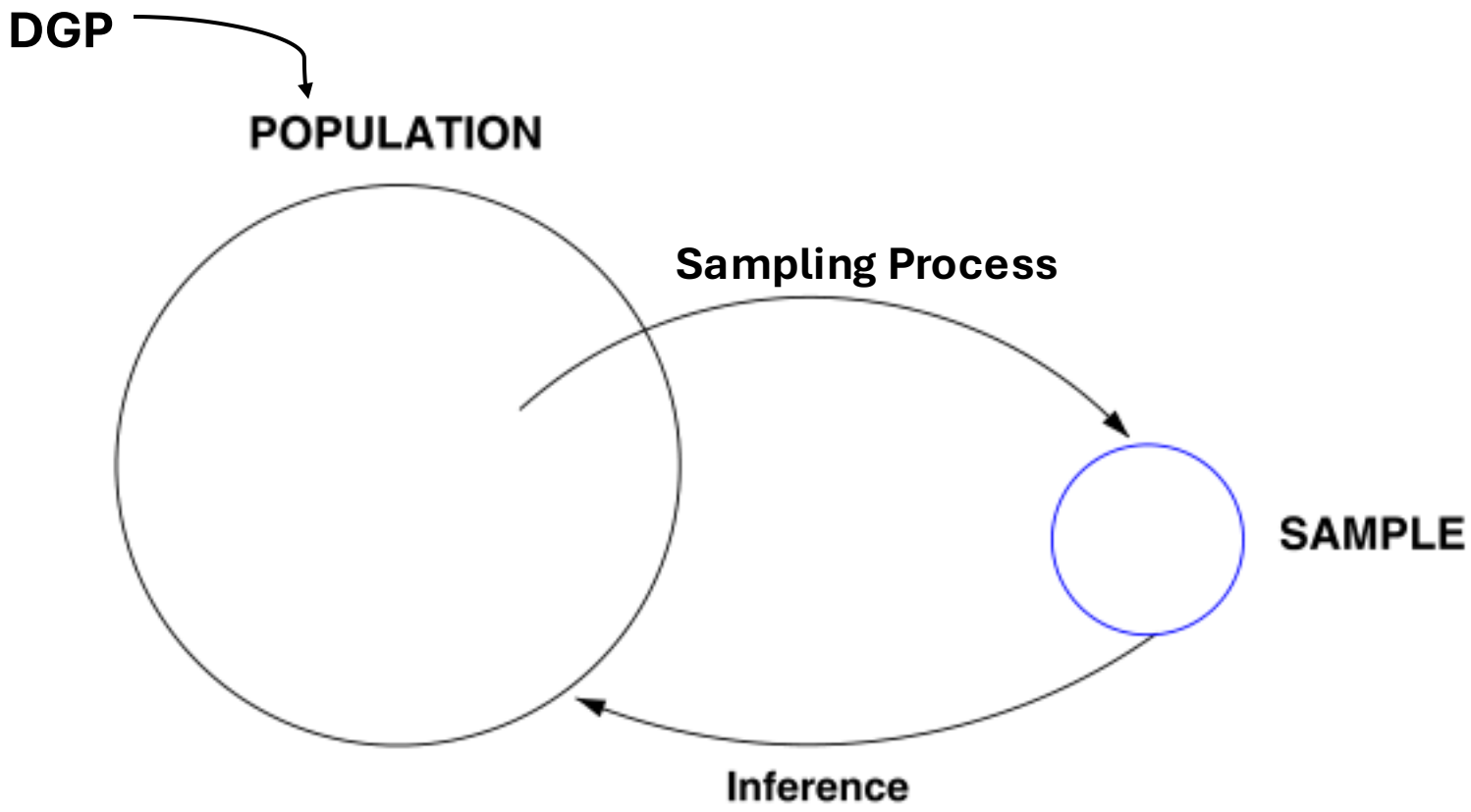
An unknown true "Data Generating Process" (DGP) generates the population.



Why bother with understanding variation in data?

We want to infer what may be going on in the population level

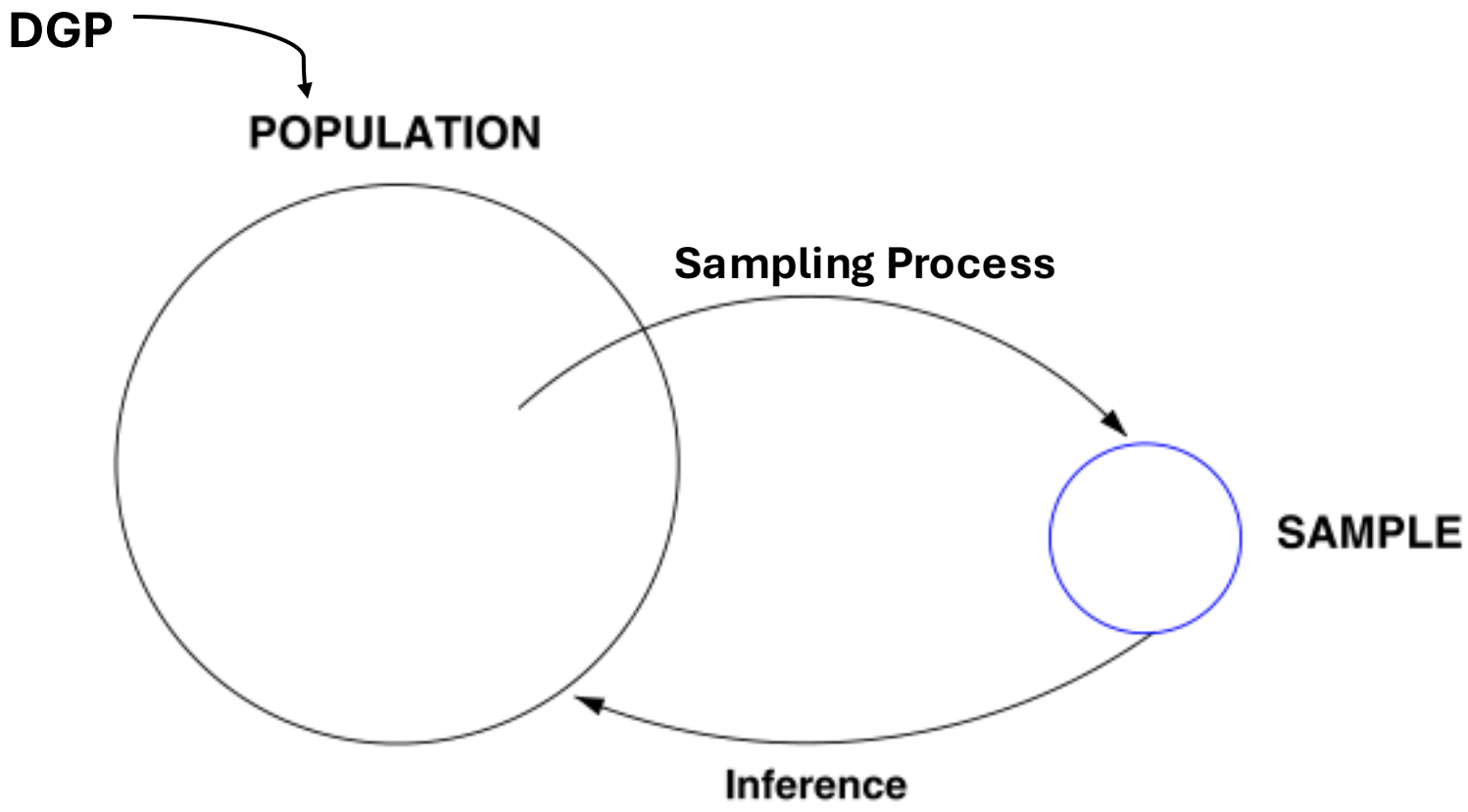
An unknown true "Data Generating Process" (DGP) generates the population.



Why bother with understanding variation in data?

We want to infer what may be going on in the population level

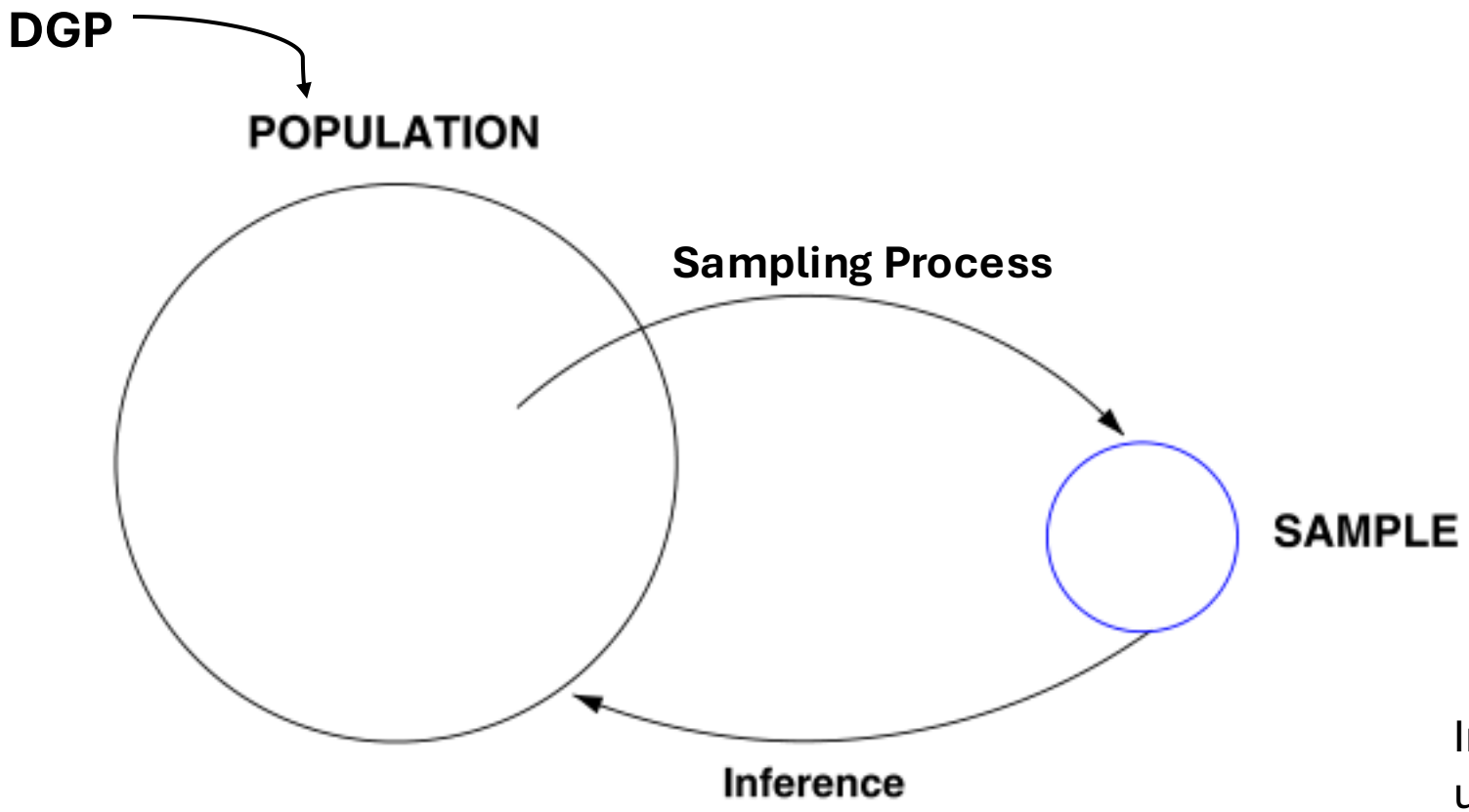
An unknown true "Data Generating Process" (DGP) generates the population.



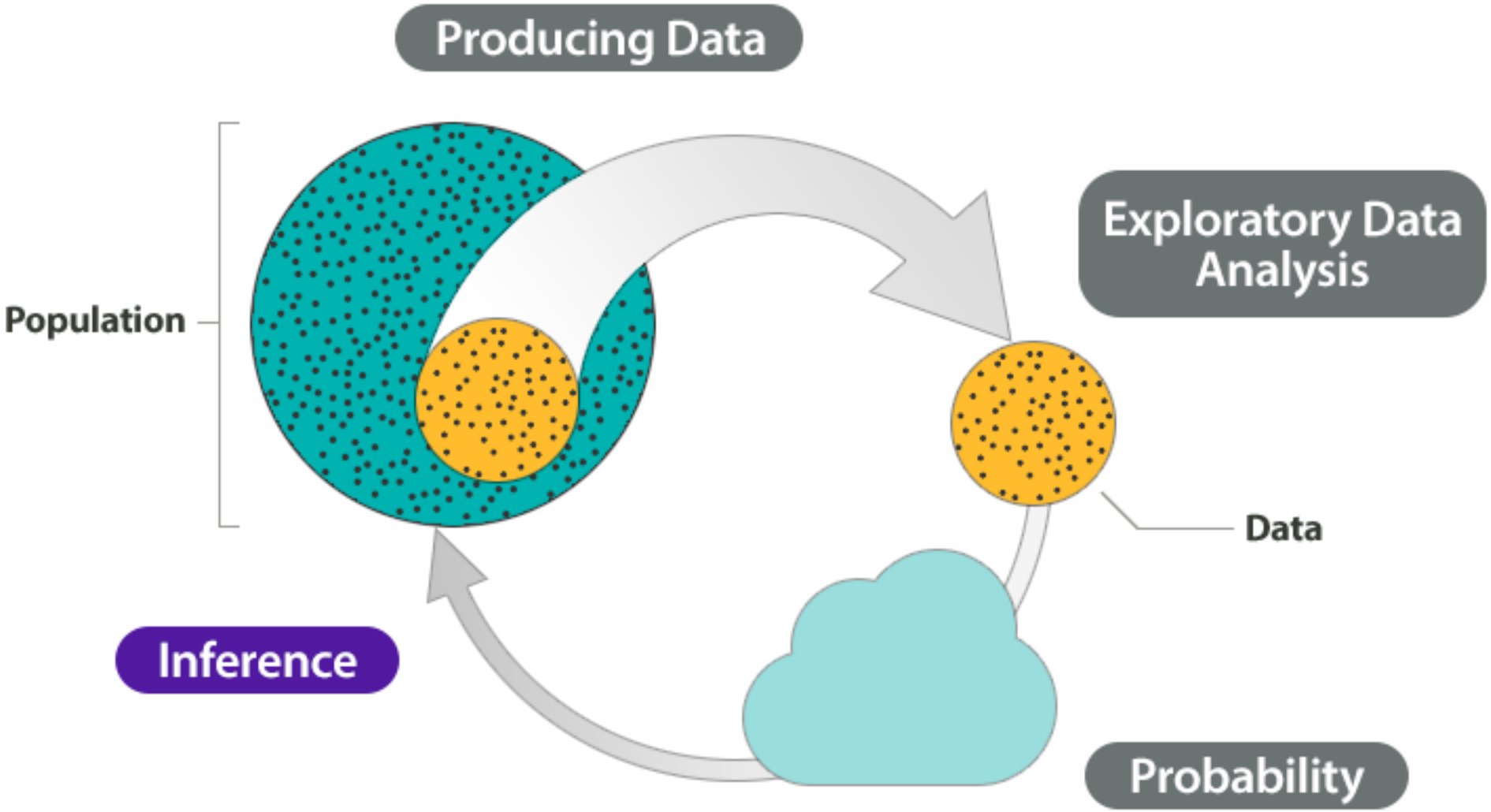
Why bother with understanding variation in data?

We want to infer what may be going on in the population level

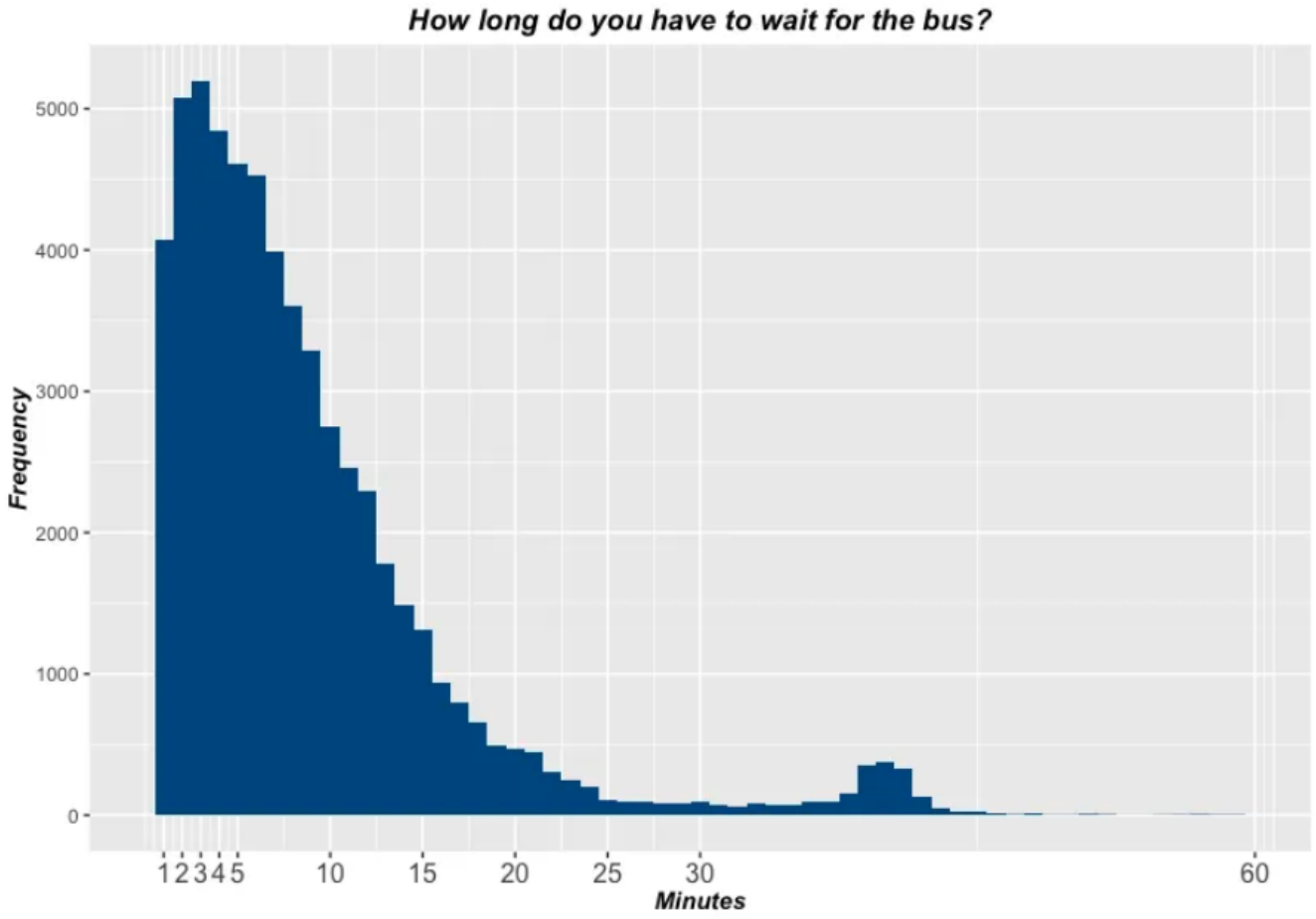
An unknown true "Data Generating Process" (DGP) generates the population.



Inference under uncertainty due to "sampling error"


Producing Data
Population
Exploratory Data Analysis
Data
Inference
Probability

EDA for hypothesis generation



Discuss: Why does the distribution look the way it does?

Approaches to infer DGP

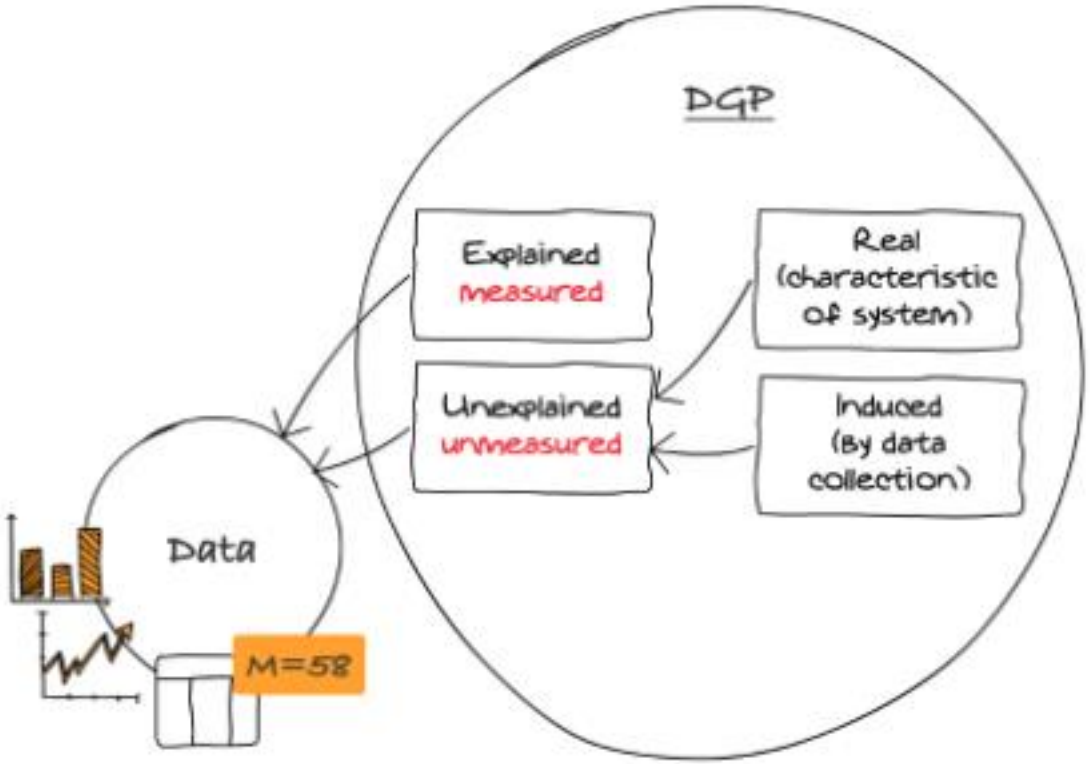
Bottom-up:

- This approach begins with the observed data.
- By examining the data distribution, one might make educated guesses or inferences about the underlying processes that produced it

Approaches to infer DGP

Bottom-up:

- This approach begins with the observed data.
- By examining the data distribution, one might make educated guesses or inferences about the underlying processes that produced it

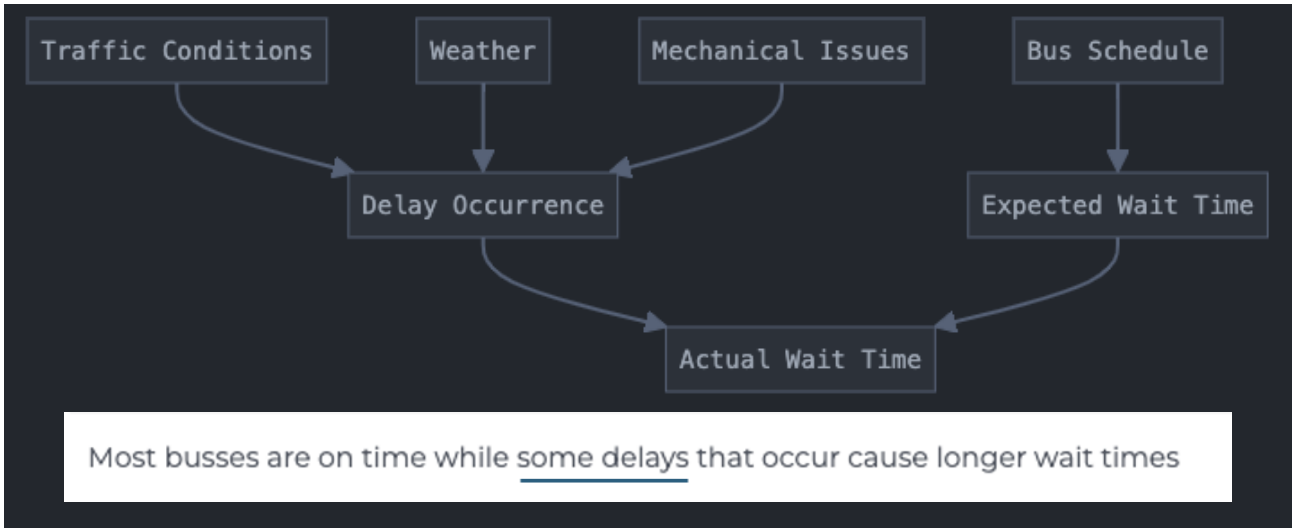


Approaches to infer DGP

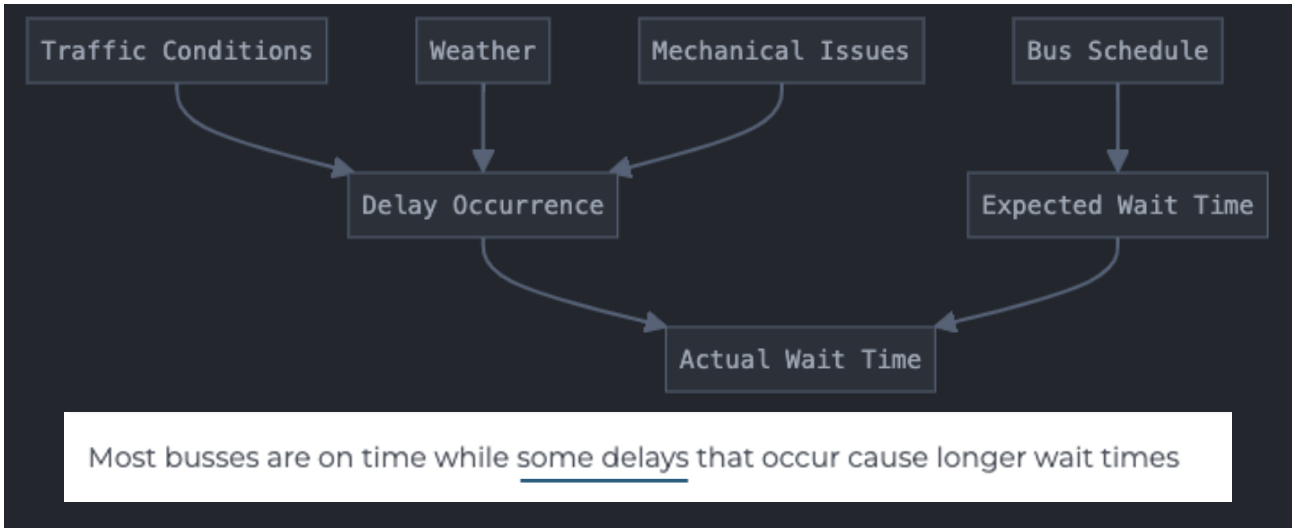
Bottom-up:

- This approach begins with the observed data.
- By examining the data distribution, one might make educated guesses or inferences about the underlying processes that produced it

Possible guesses about DGP

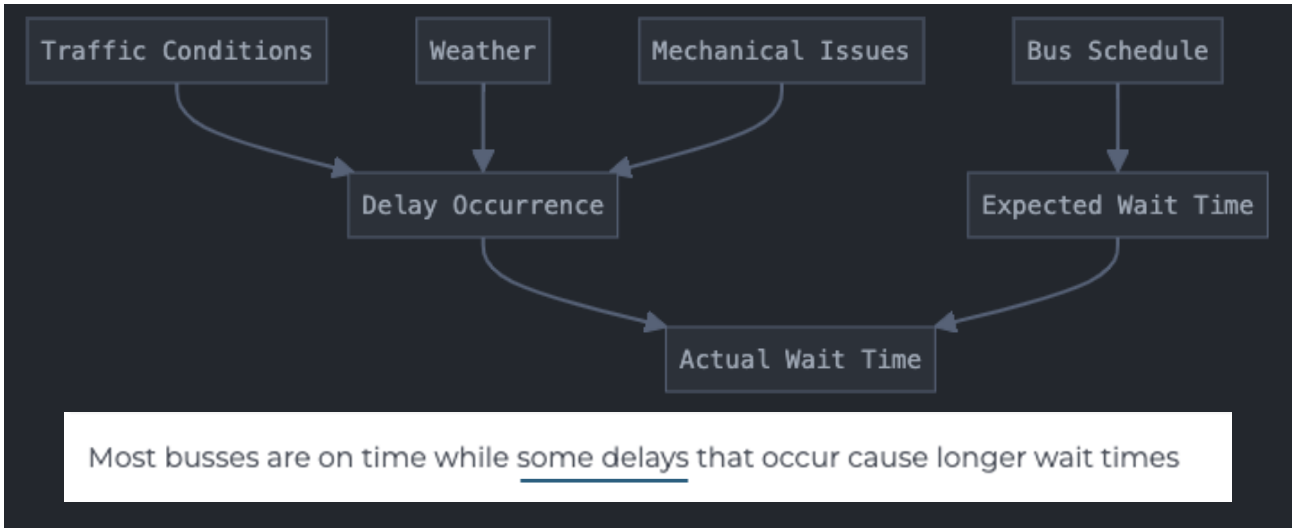


Possible guesses about DGP

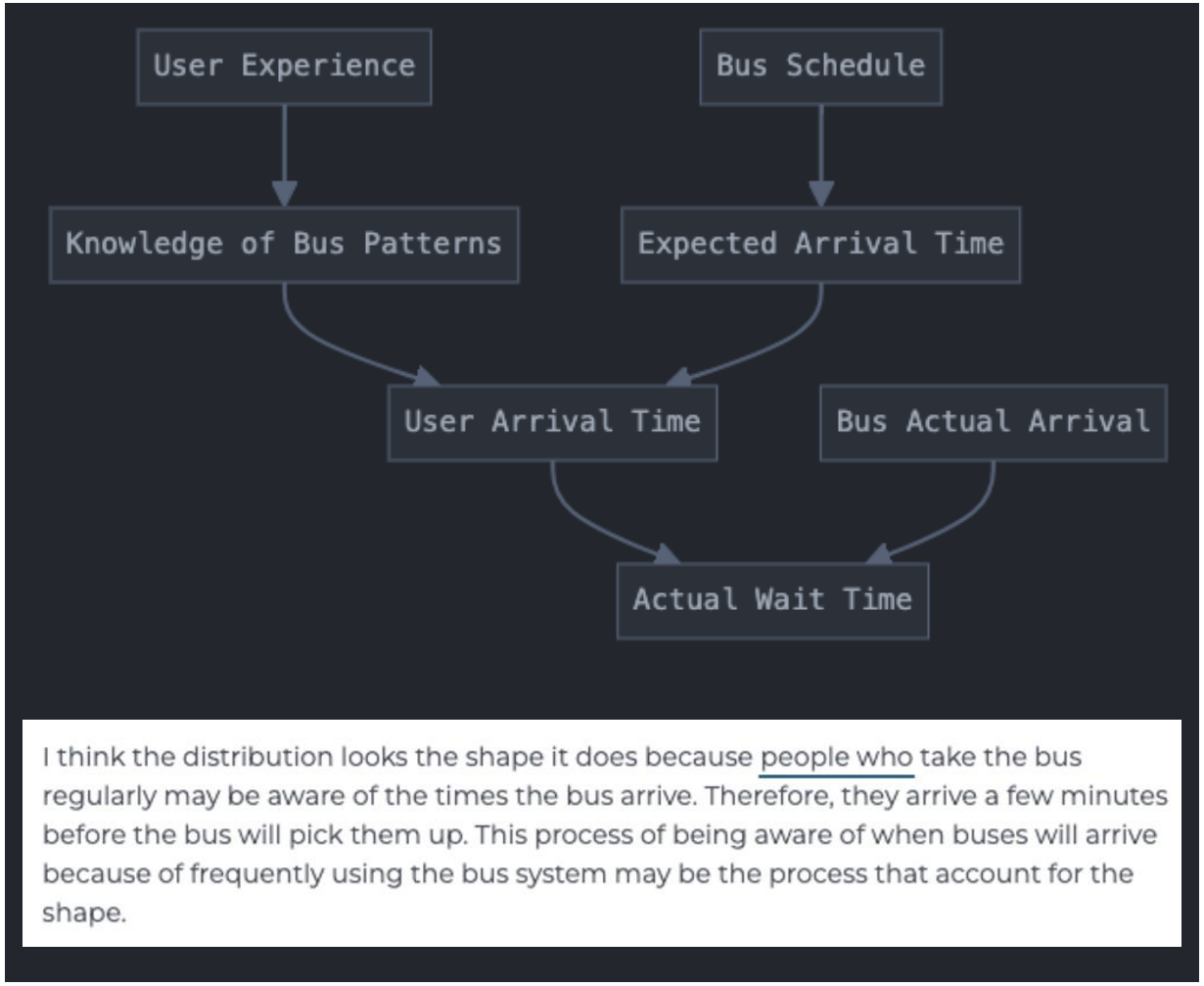


Explantory variables try to explain the outcome

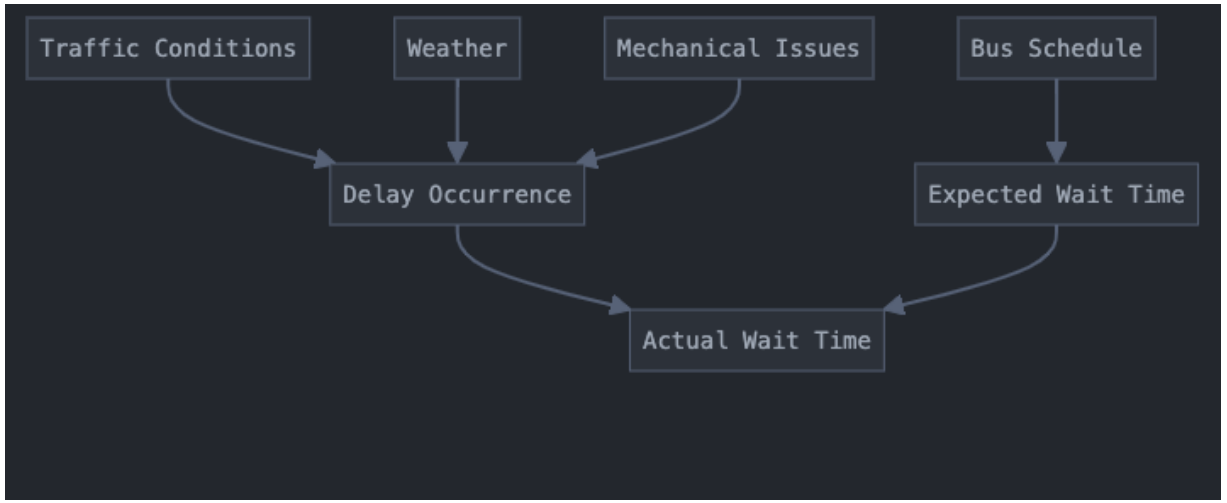
Possible guesses about DGP



Explantory variables try to explain the outcome

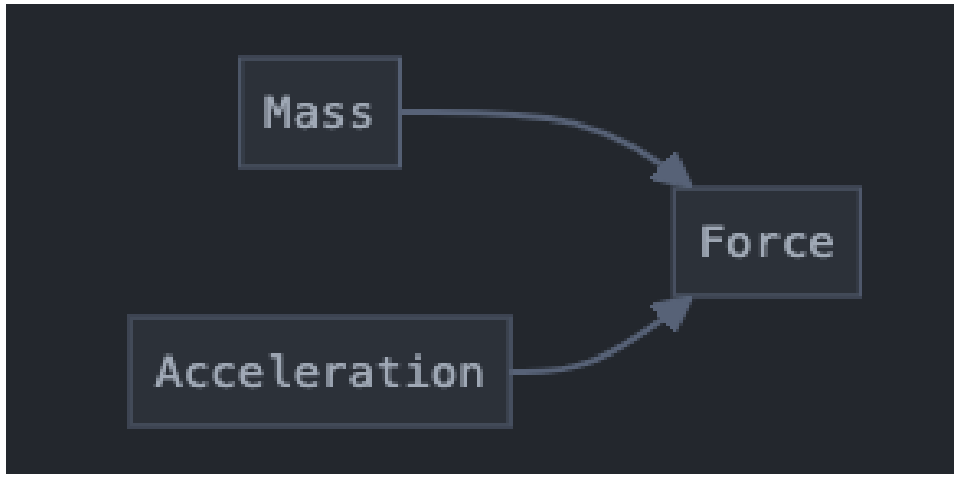


Notion of a model



- Mathematically represents these relationships
- Sets up initial conditions

Notion of a model



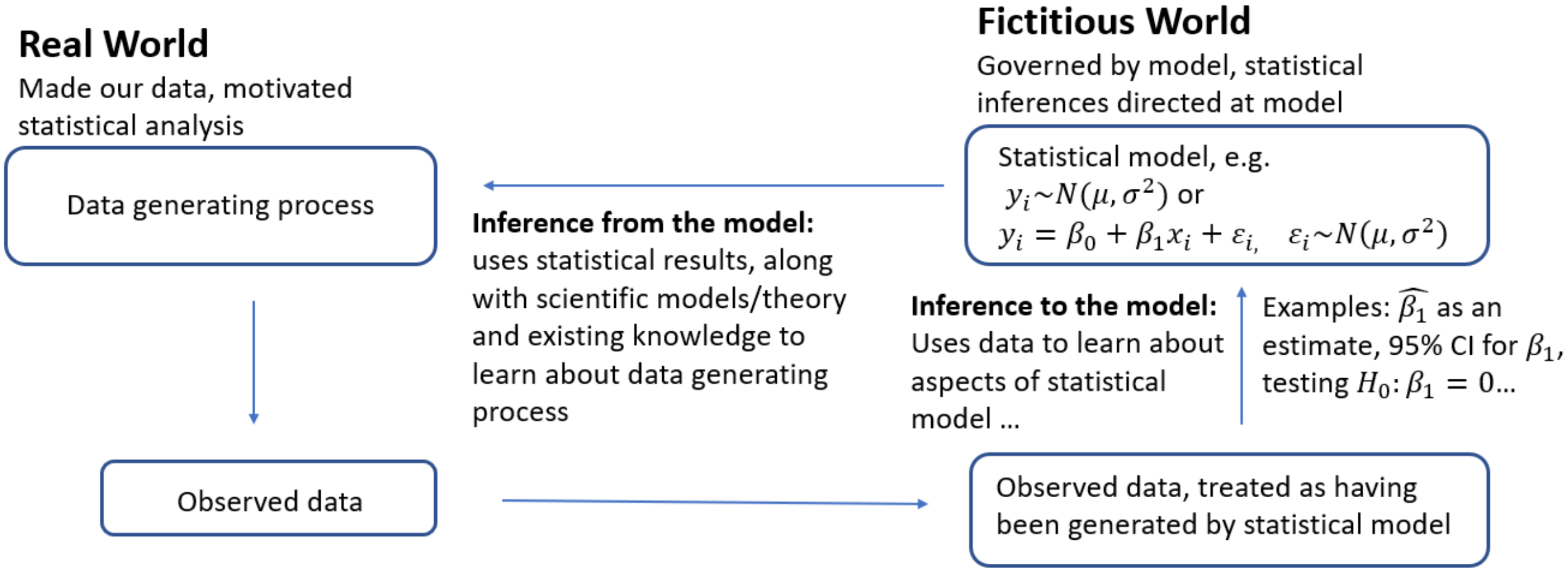
- Mathematically represents these relationships
- Sets up initial conditions

Law:

$$F = m.a$$

Model:

$$F = m.a + \in$$


Real World
Made our data, motivated statistical analysis
Data generating process
Observed data
Fictitious World
Governed by model, statistical inferences directed at model
Statistical model, e.g.
$y_i \sim N(\mu, \sigma^2)$ or
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(\mu, \sigma^2)$
Inference from the model:
uses statistical results, along with scientific models/theory and existing knowledge to learn about data generating process
Inference to the model:
Uses data to learn about aspects of statistical model …
Examples: $\widehat{\beta_1}$ as an estimate, 95% CI for β_1, testing $H_0: \beta_1 = 0$…
Observed data, treated as having been generated by statistical model

Discussion

- Let’s assume that the waiting time follows a normal distribution:

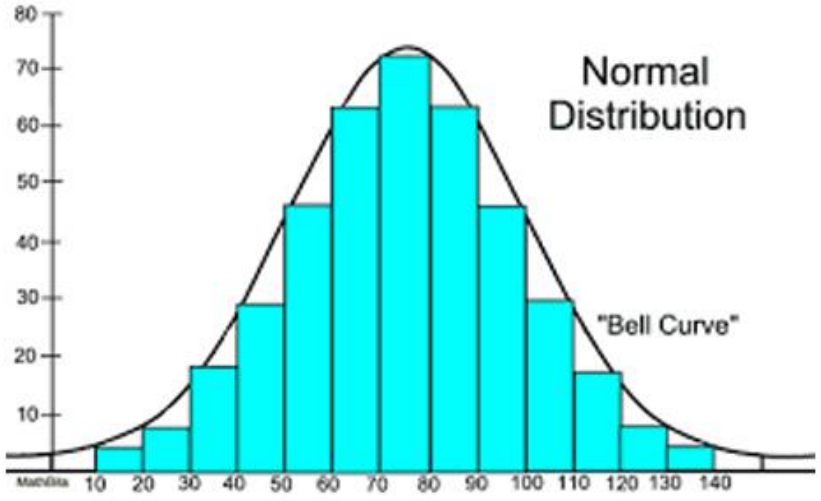


Figure 3: Normal Distribution

How do your hypotheses change in case wait times are normally distributed?

Approaches to infer DGP

Bottom-up:

- This approach begins with the observed data.
- By examining the data distribution, one might make educated guesses or inferences about the underlying processes that produced it

Top-down:

- This approach relies on pre-existing knowledge or theories about the system or phenomenon in question to inform our understanding of the DGP

Approaches to infer DGP

Bottom-up:

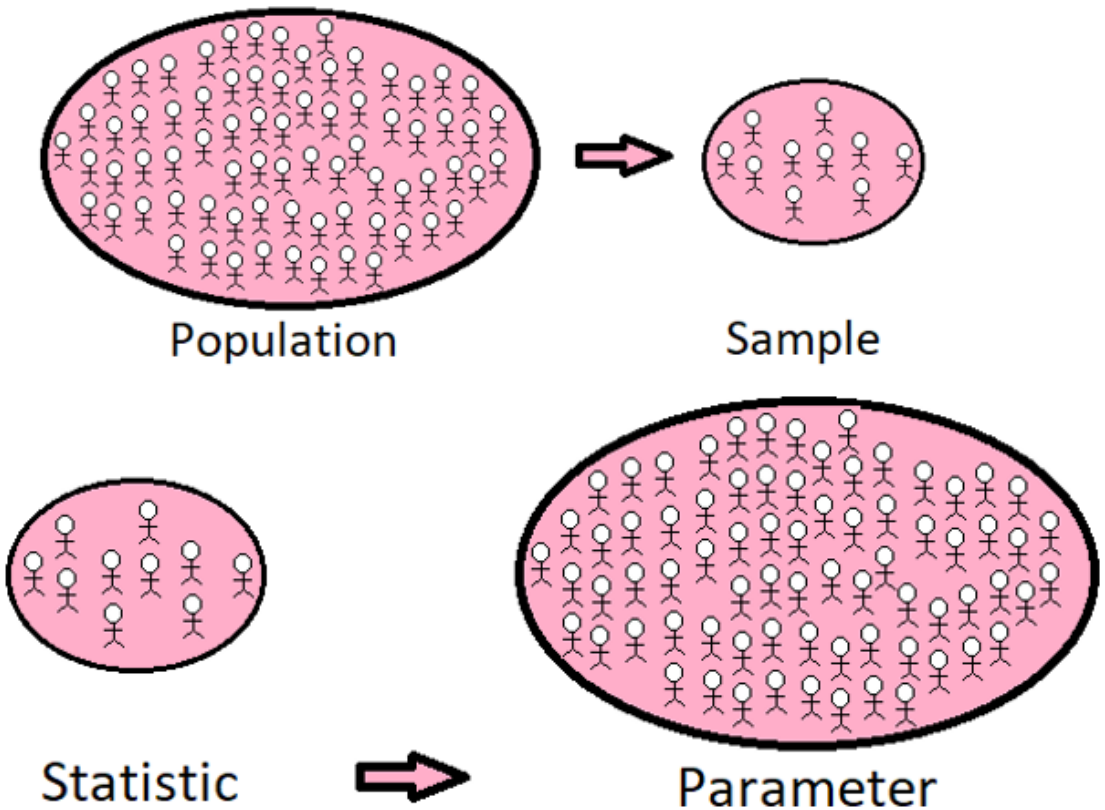
- This approach begins with the observed data.
- By examining the data distribution, one might make educated guesses or inferences about the underlying processes that produced it

Top-down:

- This approach relies on pre-existing knowledge or theories about the system or phenomenon in question to inform our understanding of the DGP

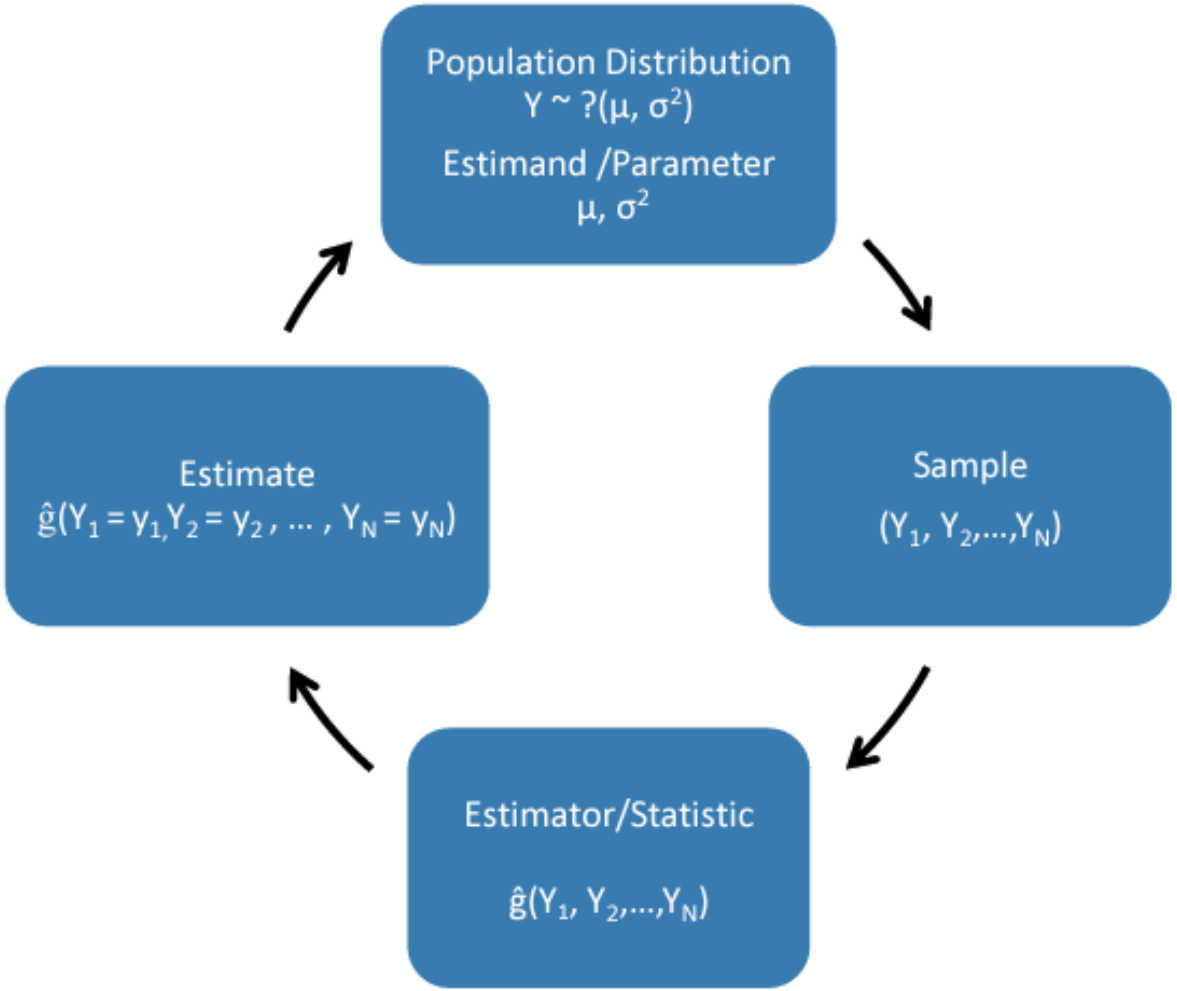
Discuss:
Top-down approaches to inferring DGPs

Summary statistics / Population parameters



- Take a sample from a larger population
- Calculate a statistic from the sample to estimate a parameter for the larger population

Inference loop in terms of statistics/ parmeters



Descriptive Statistics / Summary statistics

- Can be informative when used appropriately.

Descriptive Statistics / Summary statistics

- To go from lots of numbers to a few numbers

84	48	-74	-71	-57	27	-16	12	-33	-48	25	-30	6	83	45	29	-72	57	23	-35
-16	61	32	86	27	96	83	-41	27	-57	-76	60	-13	83	38	-75	-58	-30	-58	16
-94	-100	-80	-22	-63	-84	15	52	75	-58	88	-51	54	-73	-50	-80	-19	-9	36	26
77	86	84	-99	-47	-94	-4	-60	-69	74	32	-7	98	0	3	8	31	38	38	-46
88	70	89	33	-53	13	44	22	98	32	70	92	90	-41	52	3	62	-65	-93	-26
29	-89	-94	-83	5	97	17	-59	-88	-61	25	-84	72	-22	-51	99	41	-37	21	65
73	51	-22	21	71	18	-68	92	-82	-20	33	95	95	26	33	60	39	-29	-62	18
6	43	17	70	-68	-12	51	3	73	-29	1	-66	82	-72	26	-52	-98	59	48	34
5	79	66	-29	-35	-95	39	-80	-39	61	96	-56	74	-95	-30	75	-12	98	58	9
-46	-6	21	93	47	47	24	-46	39	11	-92	22	50	-88	-35	52	64	7	36	-40

Descriptive Statistics / Summary statistics

- To go from lots of numbers to a few numbers

84	48	-74	-71	-57	27	-16	12	-33	-48	25	-30	6	83	45	29	-72	57	23	-35
-16	61	32	86	27	96	83	-41	27	-57	-76	60	-13	83	38	-75	-58	-30	-58	16
-94	-100	-80	-22	-63	-84	15	52	75	-58	88	-51	54	-73	-50	-80	-19	-9	36	26
77	86	84	-99	-47	-94	-4	-60	-69	74	32	-7	98	0	3	8	31	38	38	-46
88	70	89	33	-53	13	44	22	98	32	70	92	90	-41	52	3	62	-65	-93	-26
29	-89	-94	-83	5	97	17	-59	-88	-61	25	-84	72	-22	-51	99	41	-37	21	65
73	51	-22	21	71	18	-68	92	-82	-20	33	95	95	26	33	60	39	-29	-62	18
6	43	17	70	-68	-12	51	3	73	-29	1	-66	82	-72	26	-52	-98	59	48	34
5	79	66	-29	-35	-95	39	-80	-39	61	96	-56	74	-95	-30	75	-12	98	58	9
-46	-6	21	93	47	47	24	-46	39	11	-92	22	50	-88	-35	52	64	7	36	-40

Types of statistics. Those that measure

a) Sameness / Central Tendency

- Mean, Median, Mode

Descriptive Statistics / Summary statistics

- Can be informative when used appropriately.
- Discuss:
  - Example where mean is misleading.

Descriptive Statistics / Summary statistics

- To go from lots of numbers to a few numbers

84	48	-74	-71	-57	27	-16	12	-33	-48	25	-30	6	83	45	29	-72	57	23	-35
-16	61	32	86	27	96	83	-41	27	-57	-76	60	-13	83	38	-75	-58	-30	-58	16
-94	-100	-80	-22	-63	-84	15	52	75	-58	88	-51	54	-73	-50	-80	-19	-9	36	26
77	86	84	-99	-47	-94	-4	-60	-69	74	32	-7	98	0	3	8	31	38	38	-46
88	70	89	33	-53	13	44	22	98	32	70	92	90	-41	52	3	62	-65	-93	-26
29	-89	-94	-83	5	97	17	-59	-88	-61	25	-84	72	-22	-51	99	41	-37	21	65
73	51	-22	21	71	18	-68	92	-82	-20	33	95	95	26	33	60	39	-29	-62	18
6	43	17	70	-68	-12	51	3	73	-29	1	-66	82	-72	26	-52	-98	59	48	34
5	79	66	-29	-35	-95	39	-80	-39	61	96	-56	74	-95	-30	75	-12	98	58	9
-46	-6	21	93	47	47	24	-46	39	11	-92	22	50	-88	-35	52	64	7	36	-40

Types of statistics. Those that measure

a) Sameness / Central Tendency

- Mean, Median, Mode

b) Differentness / Variance

- range
- variance
- standard deviation
- Mean absolute deviation
- IQR

Descriptive Statistics / Summary statistics

- To go from lots of numbers to a few numbers

84	48	-74	-71	-57	27	-16	12	-33	-48	25	-30	6	83	45	29	-72	57	23	-35
-16	61	32	86	27	96	83	-41	27	-57	-76	60	-13	83	38	-75	-58	-30	-58	16
-94	-100	-80	-22	-63	-84	15	52	75	-58	88	-51	54	-73	-50	-80	-19	-9	36	26
77	86	84	-99	-47	-94	-4	-60	-69	74	32	-7	98	0	3	8	31	38	38	-46
88	70	89	33	-53	13	44	22	98	32	70	92	90	-41	52	3	62	-65	-93	-26
29	-89	-94	-83	5	97	17	-59	-88	-61	25	-84	72	-22	-51	99	41	-37	21	65
73	51	-22	21	71	18	-68	92	-82	-20	33	95	95	26	33	60	39	-29	-62	18
6	43	17	70	-68	-12	51	3	73	-29	1	-66	82	-72	26	-52	-98	59	48	34
5	79	66	-29	-35	-95	39	-80	-39	61	96	-56	74	-95	-30	75	-12	98	58	9
-46	-6	21	93	47	47	24	-46	39	11	-92	22	50	-88	-35	52	64	7	36	-40

Types of statistics. Those that measure

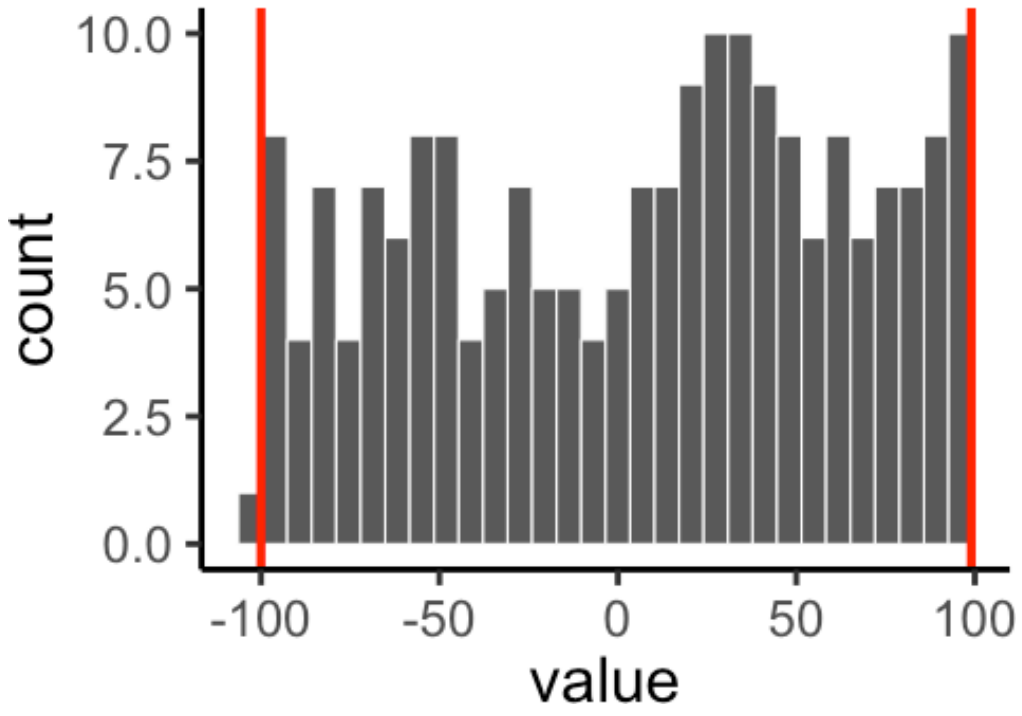
a) Sameness / Central Tendency

b) Differentness / Variance

c) Skewness/ Kurtosis

..

The Range



Thinking about the range

- Pros: Great way to find out the largest possible difference
- Cons?

Some numbers

Here are two sets of numbers. What is the range? Does it do a good job showing the average differences?

```
1 5 6 5 4 5 6 5 4 5 6 5 4 100
```

```
1 2 1 2 1 1 1 1 2 2 2 2 1 2
```

Thinking about the range

- Pros: Great way to find out the largest possible difference
- Cons: The biggest possible difference is probably not representative of all the differences in the numbers

Average differences

It would be nice if we could find a way to measure the average amount of differences.

This average could be a **representative** value that summarizes the differences between the numbers

Average differences

What should the average difference for these numbers be?

```
1 2 1 2 1 1 1 2 2 2 1 2
```

- All of the numbers are 1s or 2s.
- The difference between 1 and 2 is 1
- It seems the average difference should be 1 (+ or -)

Differences between numbers

Consider these 10 numbers:

```
1 3 4 5 5 6 7 8 9 24
```

- We can see there are some differences, they are not all the same.
- We can measure the differences, by finding the difference between each score, and every other scores
- e.g., 1-3 = 2, 1-4 = 3, etc.

Problem: The sum = 0

	1	3	4	5	5	6	7	8	9	24
1	0	2	3	4	4	5	6	7	8	23
3	-2	0	1	2	2	3	4	5	6	21
4	-3	-1	0	1	1	2	3	4	5	20
5	-4	-2	-1	0	0	1	2	3	4	19
5	-4	-2	-1	0	0	1	2	3	4	19
6	-5	-3	-2	-1	-1	0	1	2	3	18
7	-6	-4	-3	-2	-2	-1	0	1	2	17
8	-7	-5	-4	-3	-3	-2	-1	0	1	16
9	-8	-6	-5	-4	-4	-3	-2	-1	0	15
24	-23	-21	-20	-19	-19	-18	-17	-16	-15	0

Summarizing the difference scores

1. We can find the differences between scores
2. There are lots of difference scores
3. Even though we can see the difference scores have different values, we can't summarize them in the normal fashion
4. The sum adds up to 0...
5. How can we solve the problem?

Difference scores from the mean

Consider these numbers:

```
1 6 4 2 6 8
```

1. We can compute the mean to describe the central tendency of the numbers

2. How far off is the mean for each number? This is the amount of error

3. The difference scores from the mean show how far off (different) each score is from the mean

difference score = $\bar{X} - x_i$

The mean minimizes the deviations

The mean is the only number that minimizes the sum of the deviations (difference scores)

$$\sum_{i=1}^{i=N} (\bar{X} - x_i) = 0$$

SS (sum of squared deviations)

The formula for the sum of squared deviations (SS, also called sum of squares) is:

$$SS = \sum_{i=1}^{i=N} (\bar{X} - x_i)^2$$

Variance

The average of the sum of the squared difference scores from the mean

$$Variance = SS/N$$

$$Variance = \frac{\sum_{i=1}^{i=N} (\bar{X} - x_i)^2}{N}$$

Usefulness

Pros: The variance provides us with one summary number about the average differences

Cons? We squared the differences, so the variance doesn't directly relate to size of the original differences

Standard deviation = sqrt(variance)

When we took the square root of the variance, we also did something else, called computing the **standard deviation**.

$$\text{standard deviation} = \sqrt{\text{variance}}$$

$$\text{standard deviation} = \sqrt{\frac{SS}{N}}$$

$$\text{standard deviation} = \sqrt{\frac{\sum_{i=1}^{i=N} (\bar{X} - x_i)^2}{N}}$$

The standard deviation is a summary of the variability in the data that is in the same scale as the original differences

Formula notation

- Basic Structure: In R, formulas typically take the form: y ~ x
- Read as "y is modeled as a function of x"
- Operators:
  - Add terms to the model
    - Bus_wait_time ~ Average_wait_time + delay_time
  - Remove terms from the model
  - Include main effects and all interactions

Ggformula vs. ggplot2

- gf_point(y ~ x, data= “dino.csv”)

⇔

- ggplot (data= “dino.csv”, aes(x=x, y=y))+ geom_point()

Ggformula vs. ggplot2

- gf_point(y ~ x, data= “dino.csv”)

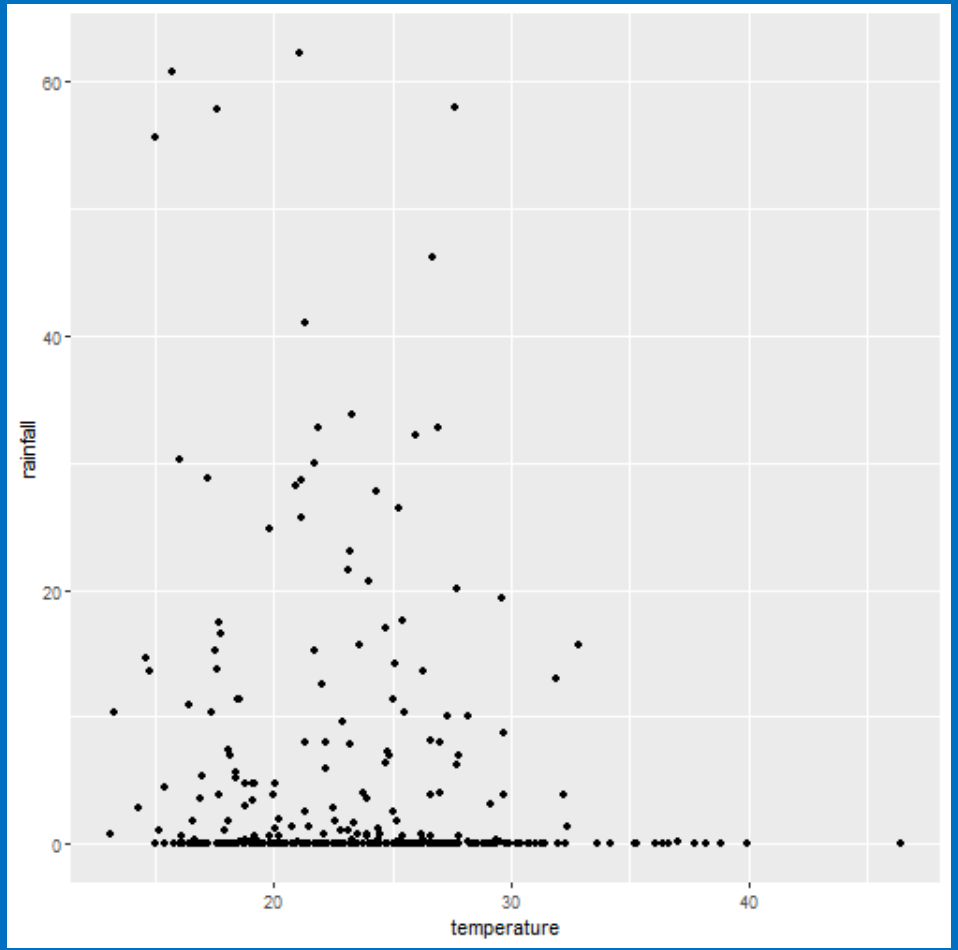
⇔

- ggplot (data= “dino.csv”, aes(x=x, y=y))+ geom_point()

gf_point is just a wrapper for ggplot2, you can mix and match (using the +)

GGPlot (a primer)

```
ggplot(
   beaches,
   aes(temperature, rainfall)
) +
geom_point()
```

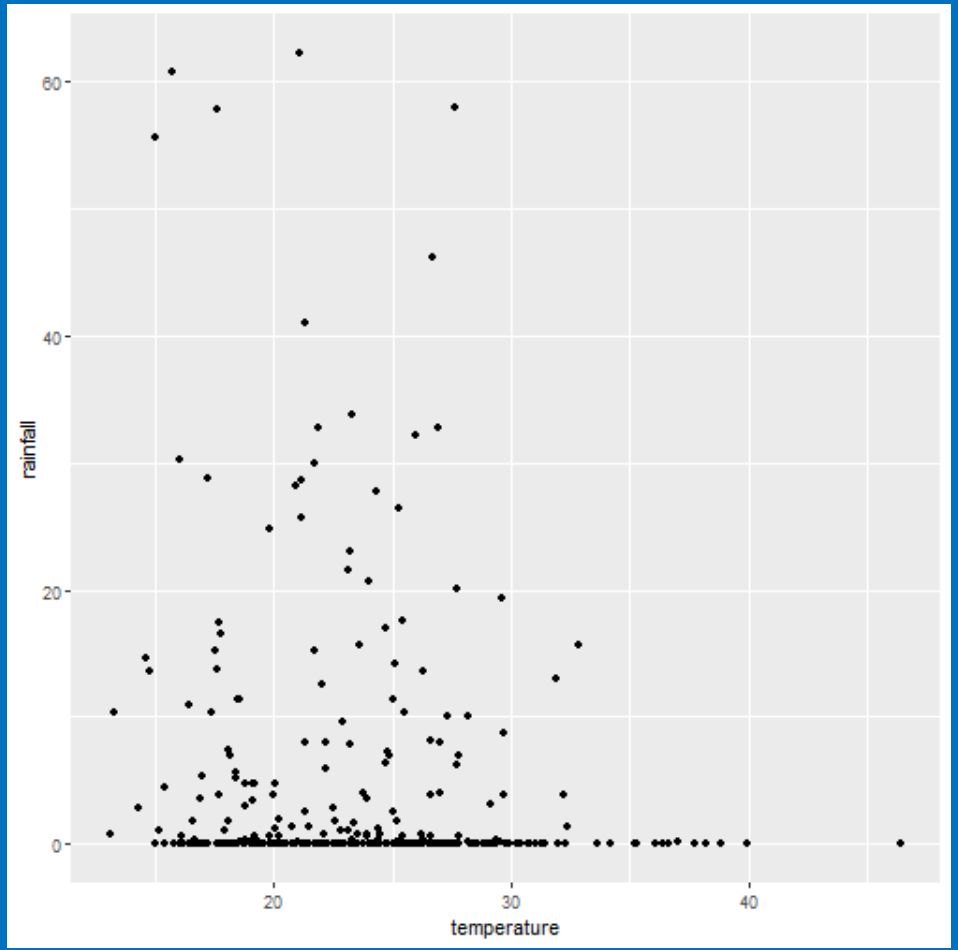


GGPlot (a primer)

```
ggplot(
   beaches,
   aes(temperature, rainfall)
) +
geom_point()
```

- Verbose?
- Why is the code odd?

  ‘+’ is rarely used this way.



GGPlot (a primer)

- “GG” refers to a “Grammar of Graphics”

A grammar..

- composes & reuses small parts
- Allows for complex structures from simple units

.. Of graphics

- Uses the “painters’ model”
- A plot is built in layers
- Each layer is drawn on top of the last
  - starting from an empty canvas

A blank canvas

```
ggplot()
```

```r
# load data
beaches <- read_csv(here("data","sydneybeaches3.csv"))

# show data
beaches
```

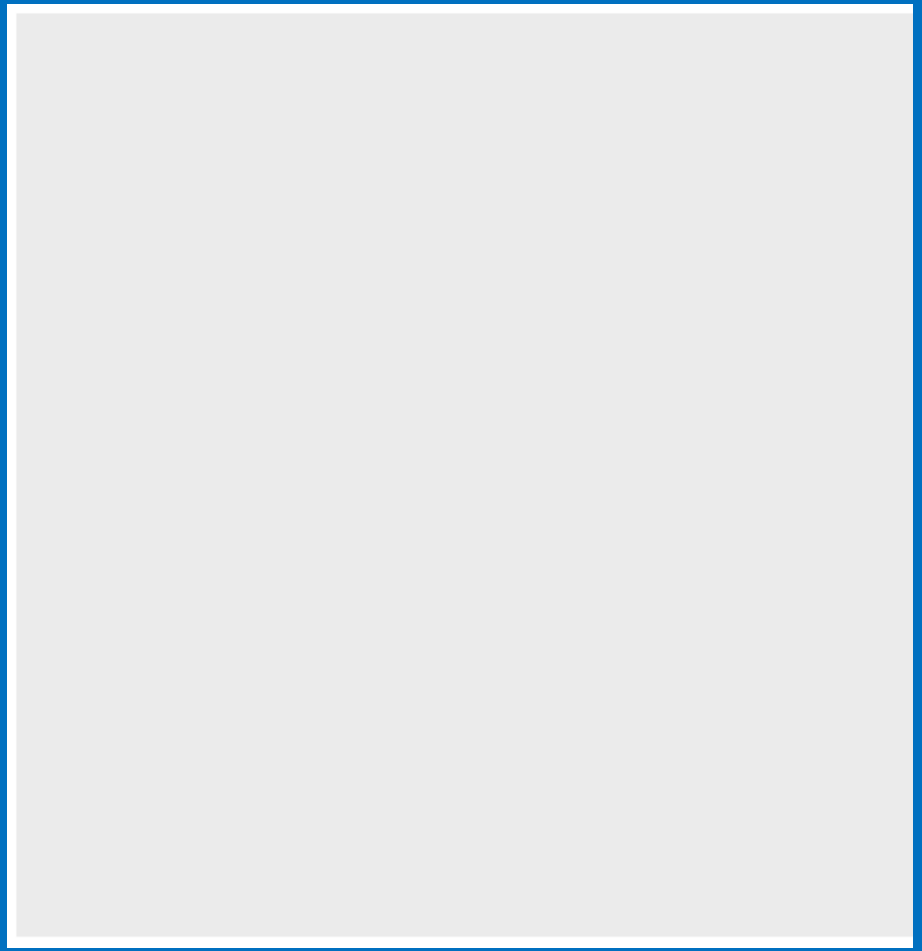
```
## # A tibble: 344 x 12
##    date        year month   day season rainfall temperature enterococci
##    <date>     <dbl> <dbl> <dbl>  <dbl>    <dbl>       <dbl>        <dbl>
##  1 2013-01-02  2013     1     2      1        0        23.4          6.7
##  2 2013-01-06  2013     1     6      1        0        30.3          2
##  3 2013-01-12  2013     1    12      1        0        31.4         69.1
##  4 2013-01-18  2013     1    18      1        0        46.4          9
##  5 2013-01-24  2013     1    24      1        0        27.5         33.9
##  6 2013-01-30  2013     1    30      1      0.6        26.6         26.5
##  7 2013-02-05  2013     2     5      1      0.1        25.7         66.9
##  8 2013-02-11  2013     2    11      1        8        22.2        118.
##  9 2013-02-17  2013     2    17      1     13.6        26.3         75
## 10 2013-02-23  2013     2    23      1      7.2        24.8        311.
## # ... with 334 more rows, and 4 more variables: day_num <dbl>,
## #   month_num <dbl>, month_name <chr>, season_name <chr>
```

~~A blank canvas~~

Specify data

```
ggplot(

)
```



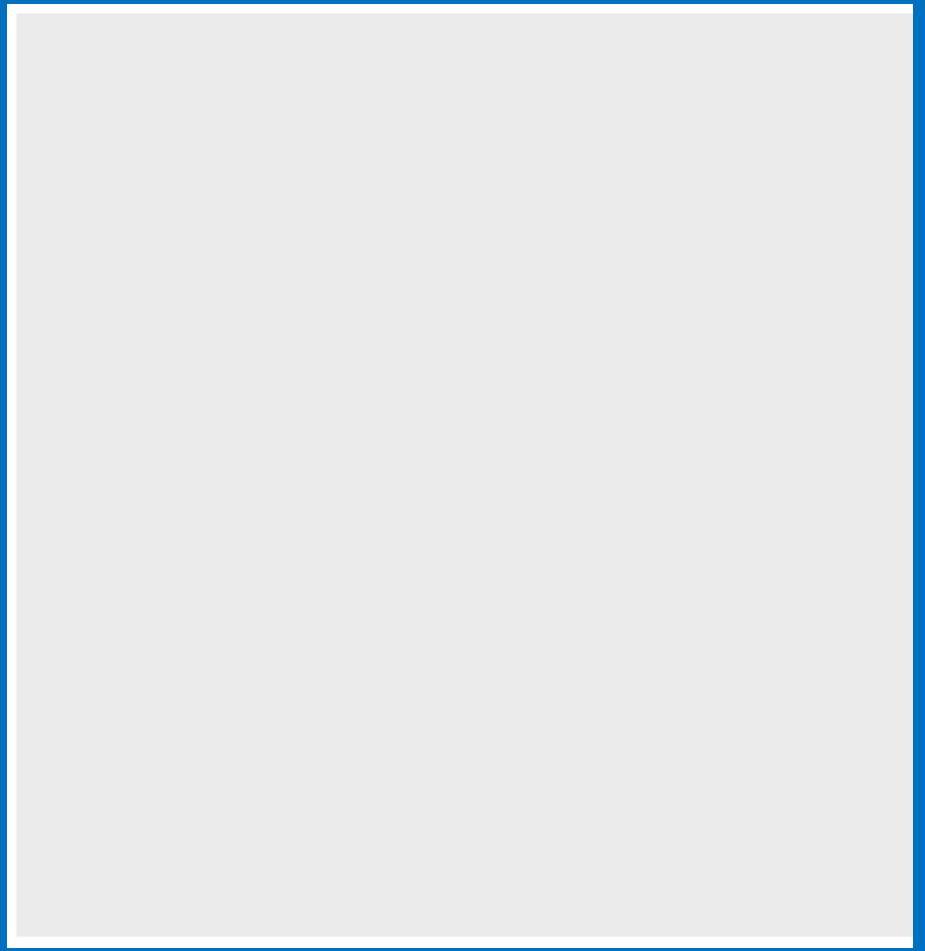
~~A blank canvas~~

Specify data

```
ggplot(
  data = beaches,



)
```



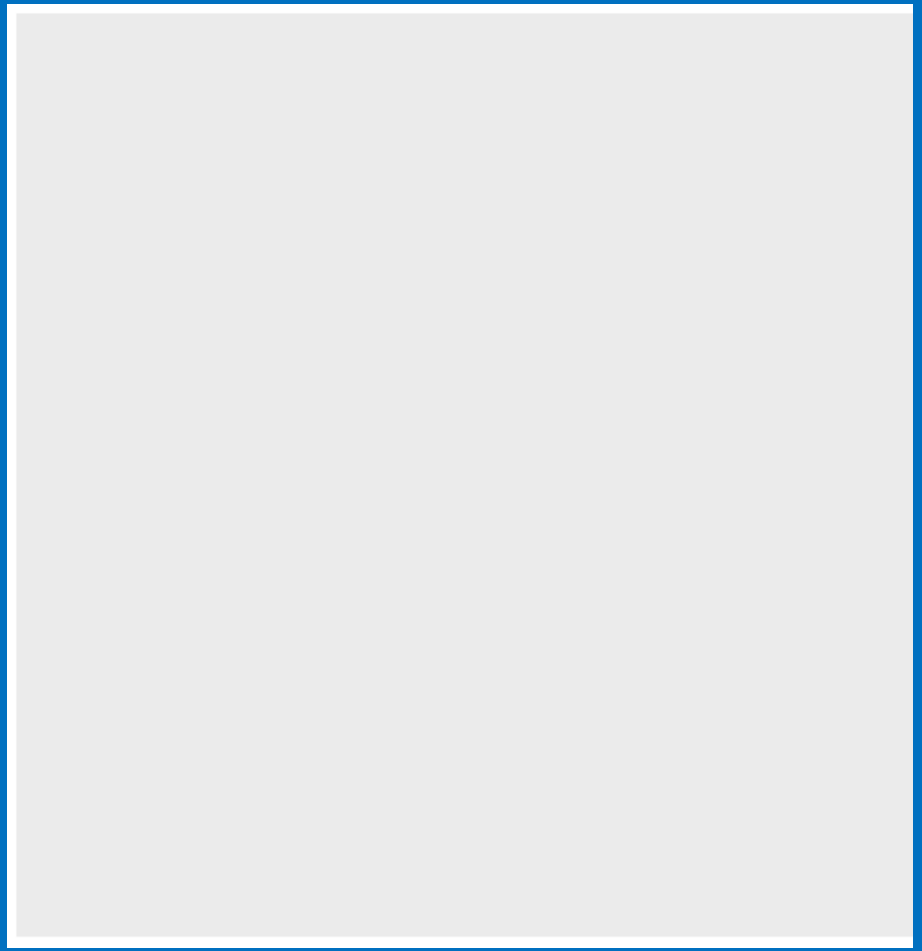
~~A blank canvas~~
Specify data

```
ggplot(
  data = beaches,



)
```

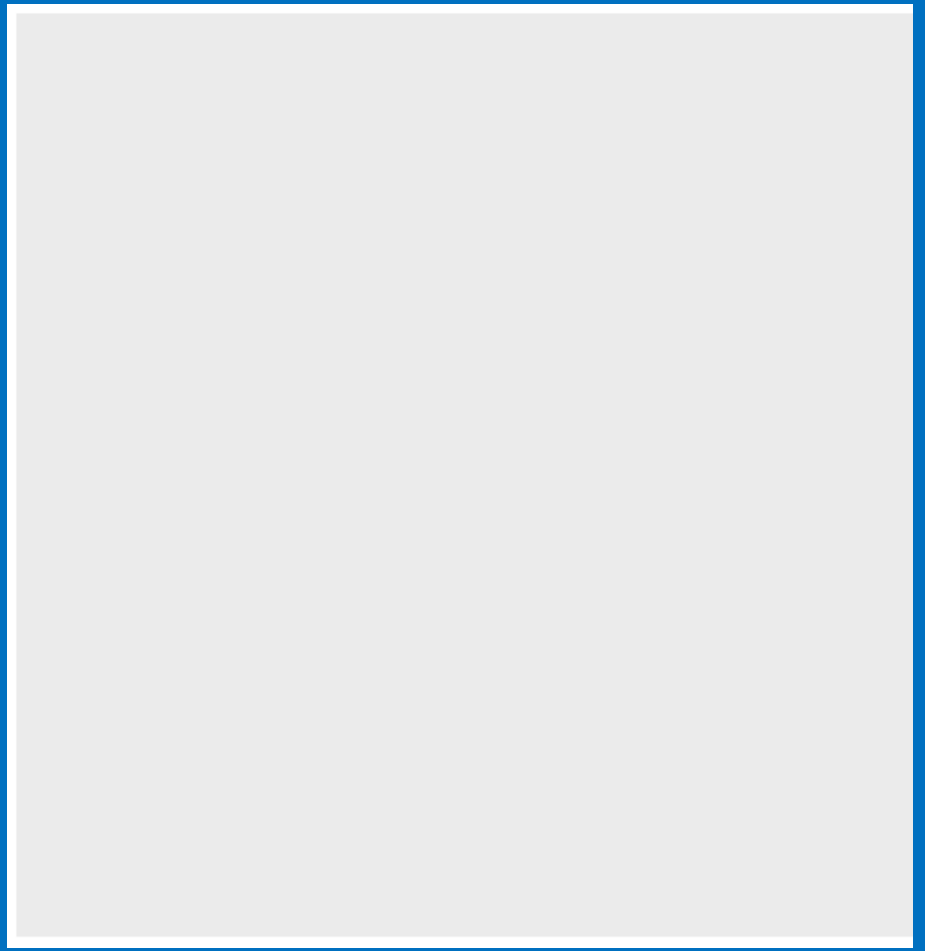
- We want to specify more about plotting aesthetics
  - X-axis
  - Y-axis
  - Shape of marker
  - Color of marker
  - , etc.



Specify data & plotting layer

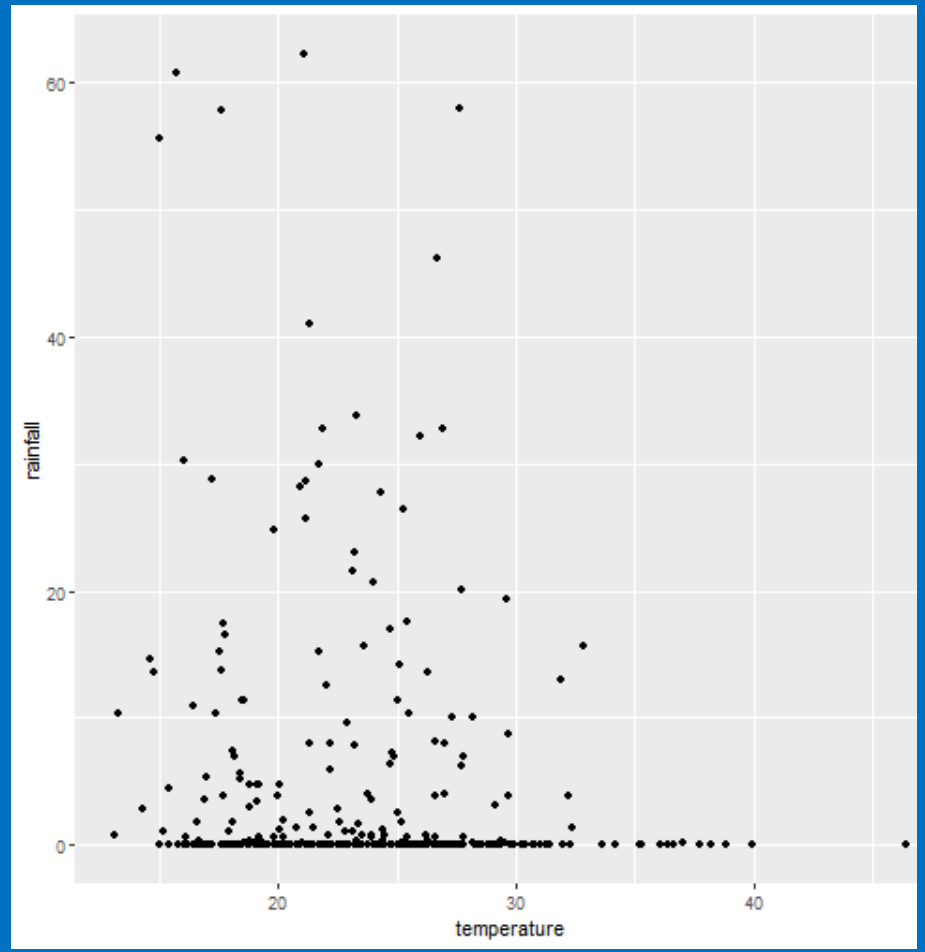
```
ggplot(
  data = beaches,
  mapping = aes(
    x = temperature,
    y = rainfall
  )
) +
```

- We have mappings now, but still need to specify the type of plot we want.
  - Data points?
  - Lines?
  - Histograms?
  - etc.



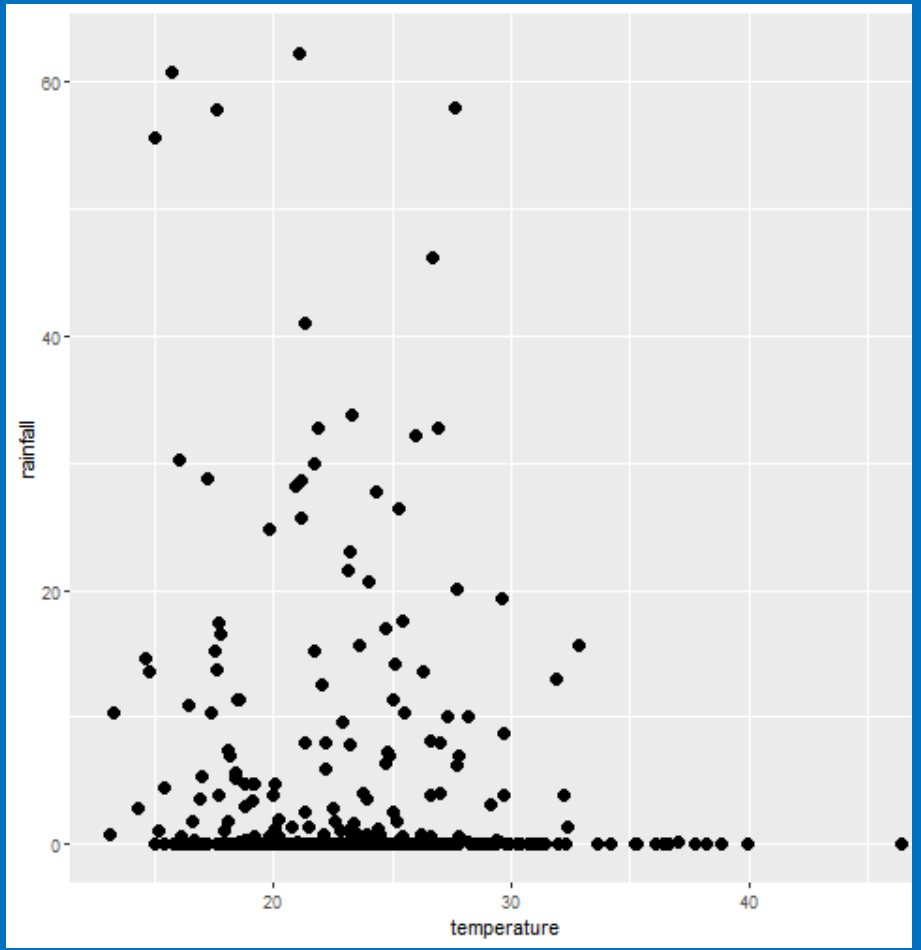
Specify data & plotting layer

```
ggplot(
  data = beaches,
  mapping = aes(
    x = temperature,
    y = rainfall
  )
) + geom_point()
```



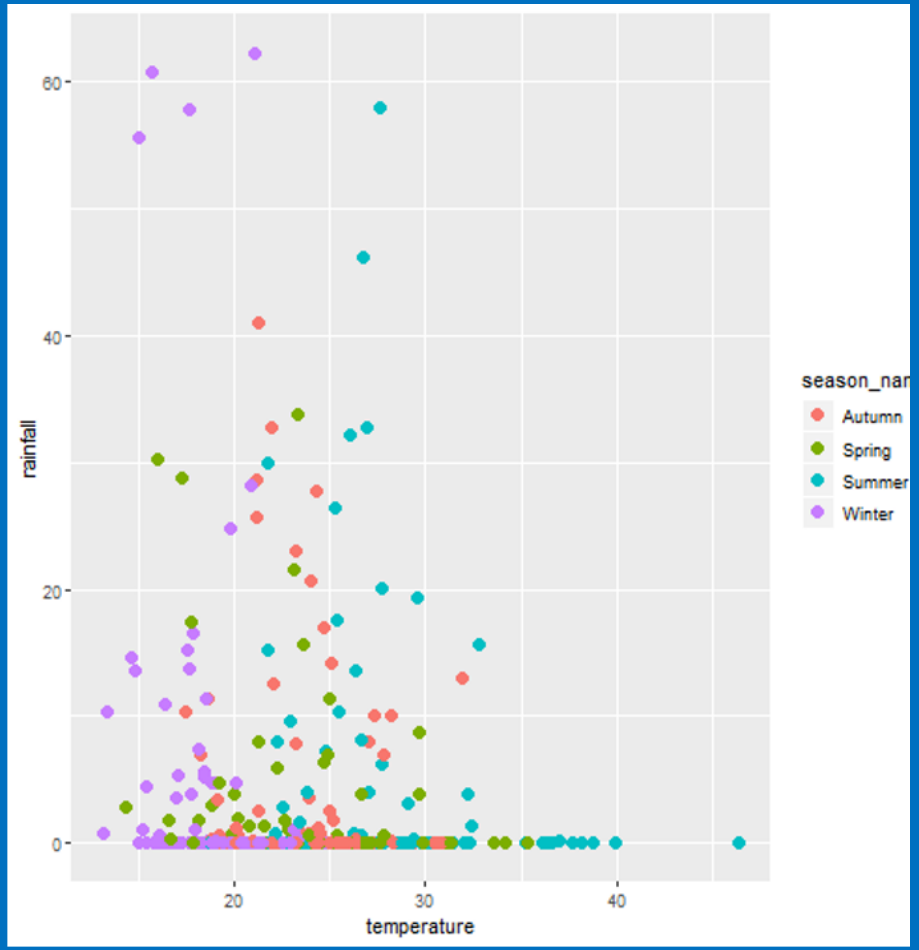
Specify data & plotting layer

```
ggplot(
  data = beaches,
  mapping = aes(
    x = temperature,
    y = rainfall
  )
) + geom_point(size =3)
```



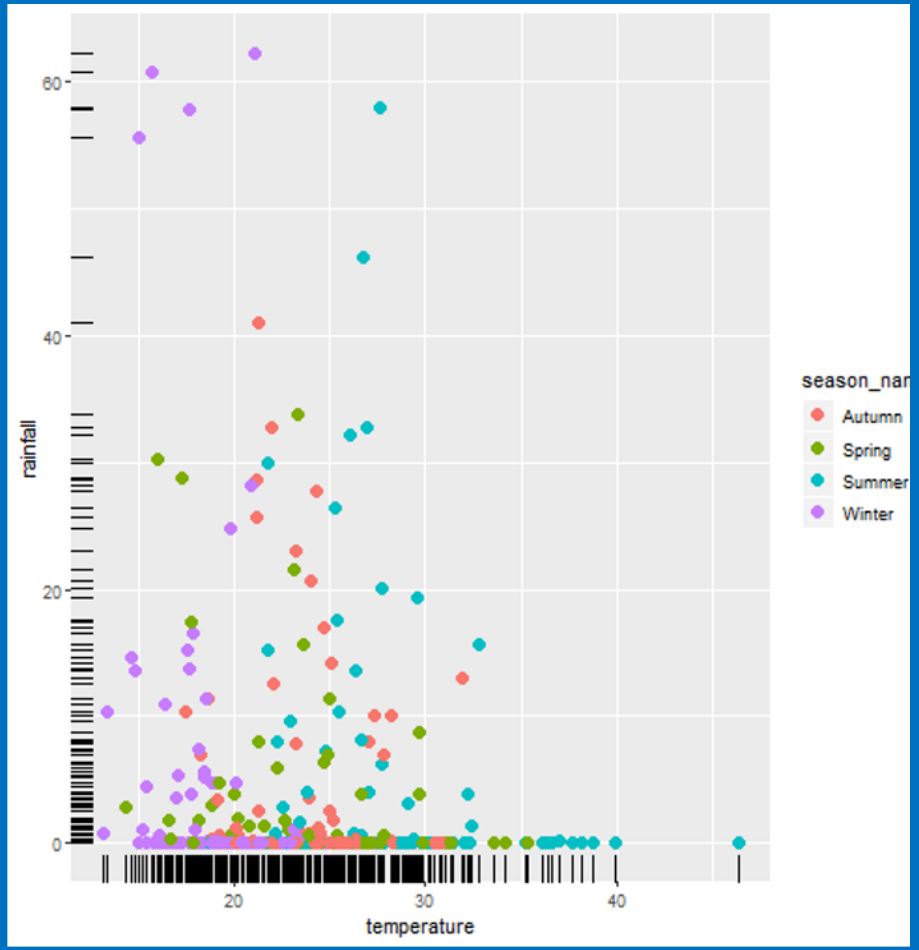
Specify data & plotting layer (with layer mappings)

```
ggplot(
  data = beaches, mapping
  = aes(
    x = temperature,
    y = rainfall
          )
  )
  + geom_point(
      mapping = aes(colour = season_name),
      size = 3)
```



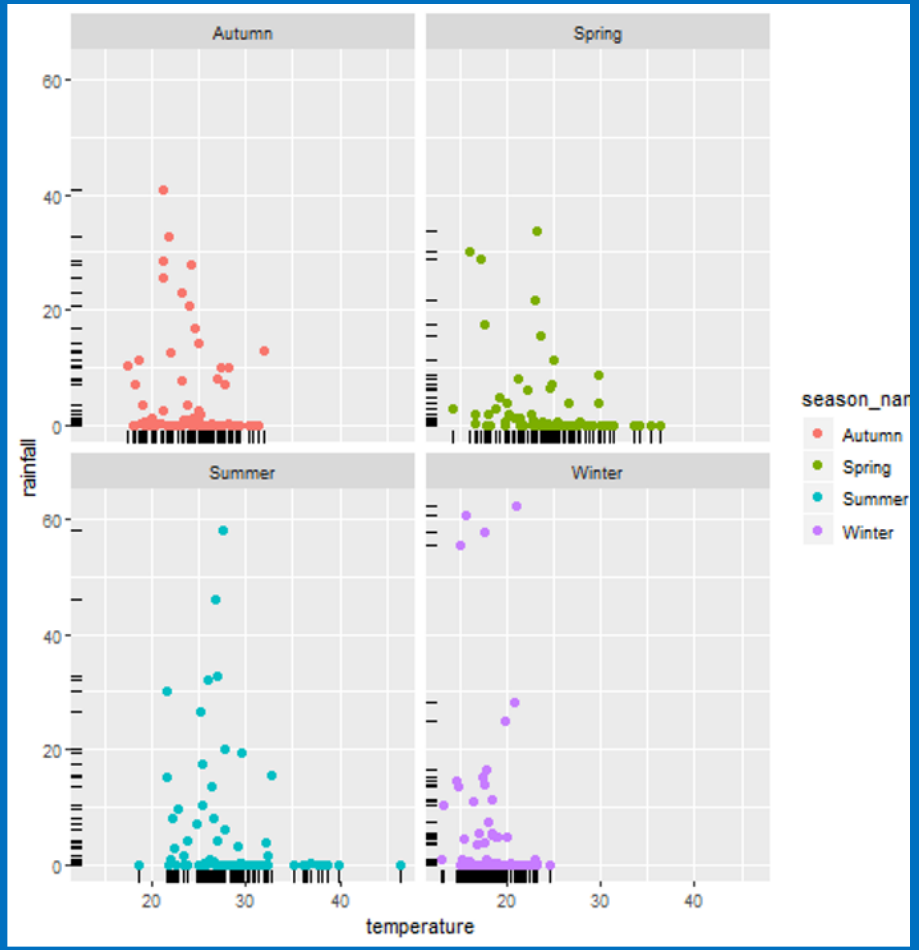
Add more layers

```
ggplot(
  data = beaches, mapping
  = aes(
    x = temperature,
    y = rainfall
          )
  )
  + geom_point(
      mapping = aes(colour = season_name),
      size = 3)
  + geom_rug()
```



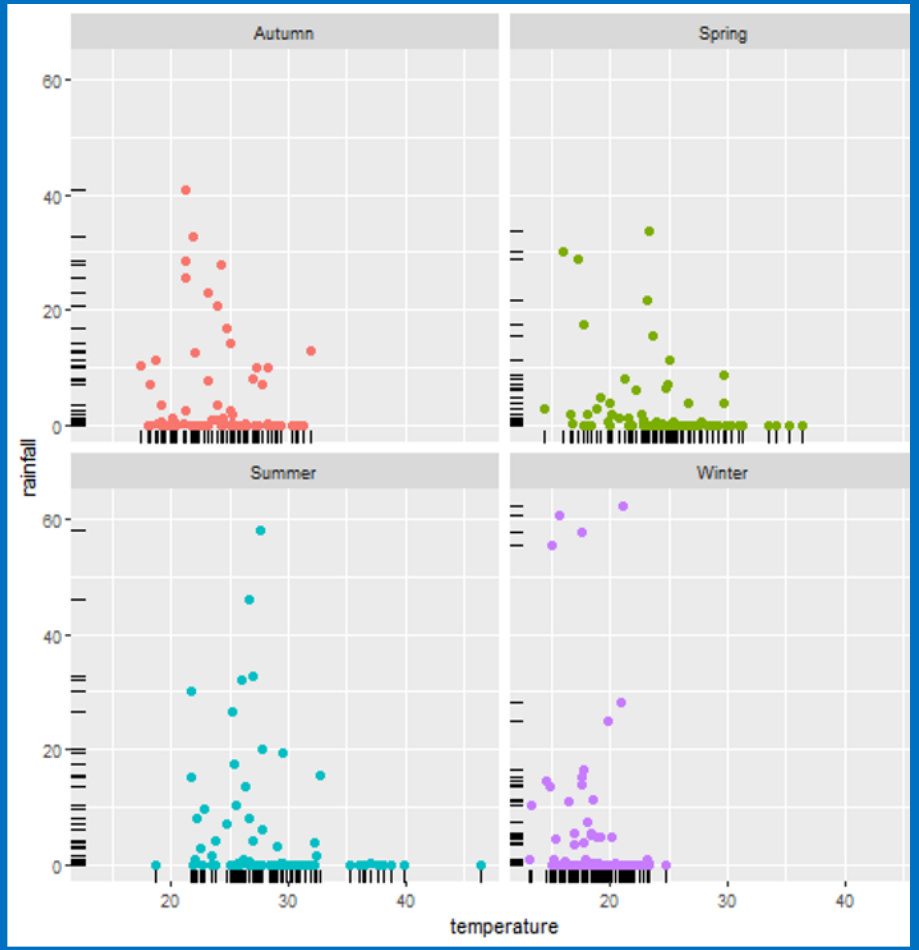
Add more layers

```
ggplot(
  data = beaches, mapping
  = aes(
    x = temperature,
    y = rainfall
          )
  )
  + geom_point(
      mapping = aes(colour = season_name),
      size = 2)
  + geom_rug()
  + facet_wrap(vars(season_name))
```



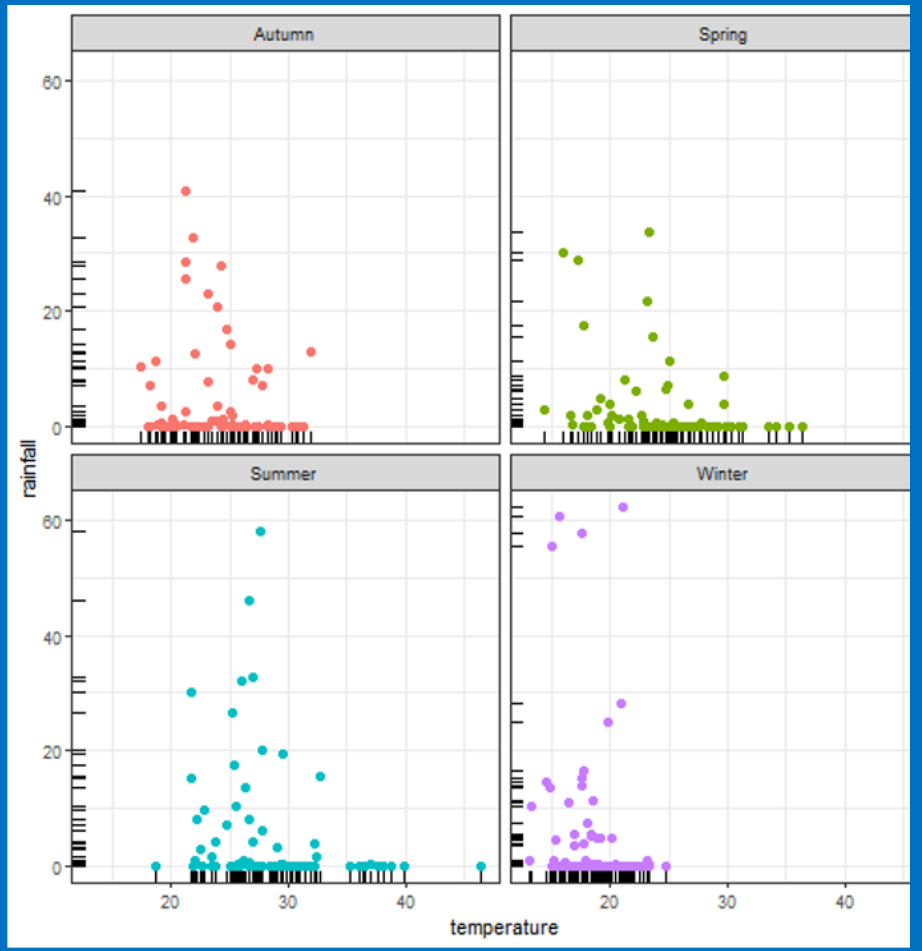
Add more layers

```
ggplot(
  data = beaches, mapping
  = aes(
    x = temperature,
    y = rainfall
          )
  )
  + geom_point(
      mapping = aes(colour = season_name),
      size = 2,
      show.legend = FALSE)
  + geom_rug()
  + facet_wrap(vars(season_name))
```



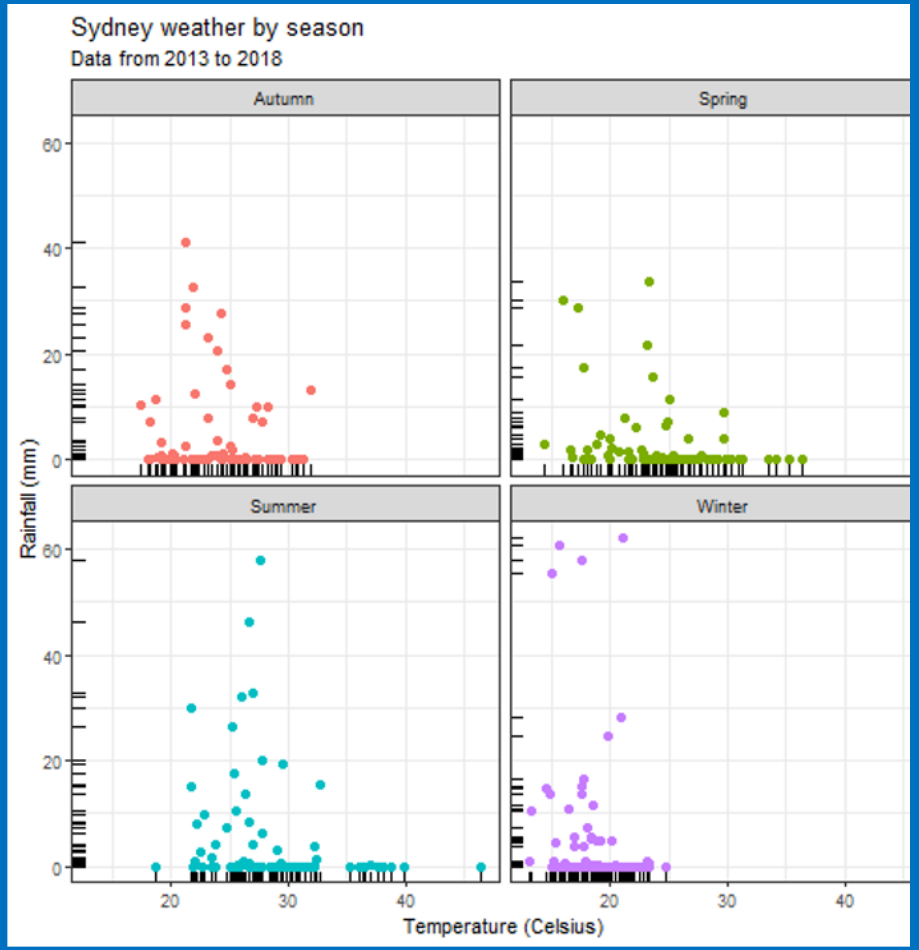
Add more layers

```
ggplot(
  data = beaches, mapping
  = aes(
    x = temperature,
    y = rainfall
          )
  )
  + geom_point(
      mapping = aes(colour = season_name),
      size = 2,
      show.legend = FALSE)
  + geom_rug()
  + facet_wrap(vars(season_name))
  + theme_bw()
```



Add more layers

```
ggplot(
  data = beaches, mapping
  = aes(
    x = temperature,
    y = rainfall
          )
  )
  + geom_point(
      mapping = aes(colour = season_name),
      size = 2,
      show.legend = FALSE)
  + geom_rug()
  + facet_wrap(vars(season_name))
  + theme_bw()
  +labs(
  title = "Sydney weather by season",
  subtitle = "Data from 2013 to 2018", x =
  "Temperature (Celsius)",
  y = "Rainfall (mm)"
   )
```



Defaults

- These are all needed
  - Data
  - Aesthetics
  - Layers
  - Facets
  - Theme
  - (Coordinates)
  - (Scales)

Defaults

- These are all needed
  - Data
  - Aesthetics
  - Layers
  - Facets
  - Theme
  - (Coordinates)
  - (Scales)

```
ggplot(
  beaches,
  aes(temperature, rainfall)
) +
geom_point()
```

Defaults

- These are all needed
  - Data
  - Aesthetics
  - Layers
  - Facets
  - Theme
  - (Coordinates)
  - (Scales)

```
ggplot(
   beaches,
   aes(temperature, rainfall)
) +
geom_point()
```

is internally represented as

```
ggplot(
   data = beaches,
   mapping = aes(
      x = temperature, y = rainfall)) +
layer(
   geom = "point",
   stat = "identity",
   position = "identity") +
facet_null() +
theme_grey() +
coord_cartesian() +
scale_x_continuous() +
scale_y_continuous()
```

Graphs are composed of layers (“GG”)