

Handling experimental data with R

Day 3

SALOS 2025

Alena Witzlack-Makarevich

Overview

- Today
 - Further expanding your GECO notebook
 - Doing summaries with `dplyr`
 - Gaining new insights into statistical visualization
 - Producing some visualization with `ggplot2`
- In the next days
 - Some inferential statistics and modeling
 - Applying what you learned to other datasets

Getting ready

- You have the GECO R notebook open
- It is relatively clean
- You loaded tidyverse with `library("tidyverse")`
- You have the GECO dataset loaded into R
- You downloaded the slides, just in case
- And you are all excited about the visualization

Building pipes



Build a pipe with %>%

- All `dplyr` verbs work similarly:
`verb(df, variable)`
- But you want to do several things with a data frame in one go:
pipe it!

```
newdf <- df %>%  
  verb1(var) %>%  
  verb2(var) %>%  
  verb3(var)
```

- Or non-idiomatically:

```
df %>%  
  verb1(var) %>%  
  verb2(var) %>%  
  verb3(var) -> newdf
```



Coming next: `summrize()`

Learn on your own:

- The five `dplyr` verbs:

- ✓ `select()`

- ✓ `filter()`

- `mutate()`

- `summarize()`

- `arrange()`

```
summarize() / summarise()  
group_by()
```



Aggregate values with `summarize()`

- `summarize()` aggregate many values down to a single summary

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

total
15

```
df %>% summarize(total = sum(value))
```

Aggregate values with `summarize()`

- `summarize()` the `RT` variable to calculate the mean reaction time

A

- name the new variable `mean_RT`, use the function `mean()`
- Check on your neighbor whether they know what “mean” means and how is it different (or not) from average

```
df %>% summarize(total = sum(value))
```

- The dataset contains 32,400 responses.
- Use `summarize()` to add up (`sum()`) all the correct answers

B

- (`ACC` of 1), name the new variable `sum_ACC`
- Run a quick sanity check with `filter()` for `ACC==1`.
You should get the same number. Do you?

Aggregate values with `summarize()`

- `summarize()` the `RT` variable to calculate the mean reaction time
- name the new variable `mean_RT`, use the function `mean()`
- Check on your neighbor whether they know what “mean” means and how is it different (or not) from average

A

```
92 # summarize()
93
94 Summarize() the RT variable to calculate the mean reaction time
95
96 ```{r message = FALSE}
97 GECO %>% summarize(mean_RT = mean(RT))
98 ```
```

A tibble: 1 × 1

mean_RT
<dbl>
715.867

1 row

no need to save
this as a new variable

Aggregate values with `summarize()`

B

- The dataset contains 32,400 responses.
- Use `summarize()` to add up (`sum()`) all the correct answers (`ACC` of 1), name the new variable `sum_ACC`
- Run a quick sanity check with `filter()` for `ACC==1`.
You should get the same number. Do you?

```

```{r message = FALSE}
GECO %>% summarize(sum_ACC = sum(ACC))
```

```

A tibble: 1 × 1

| sum_ACC
<dbl> |
|------------------|
| 29161 |

1 row

```

```{r message = FALSE}
GECO %>% filter(ACC == 1)
```

```

A tibble: 29,161 × 11

| PPNR
<dbl> | AGE
<dbl> | SEX
<chr> | HANDEDNESS
<chr> |
|---------------|--------------|--------------|---------------------|
| 1 | 20 | female | right |
| 1 | 20 | female | right |

group_by() + summarize()

- `summarize()` is not very useful on its own, pair it with `group_by()`
- This changes the unit of analysis from the complete dataset to individual groups
- `summarize()` is then apply to the dataset “by group”:

| df | |
|-------|-------|
| color | value |
| blue | 1 |
| black | 2 |
| blue | 3 |
| blue | 4 |
| black | 5 |

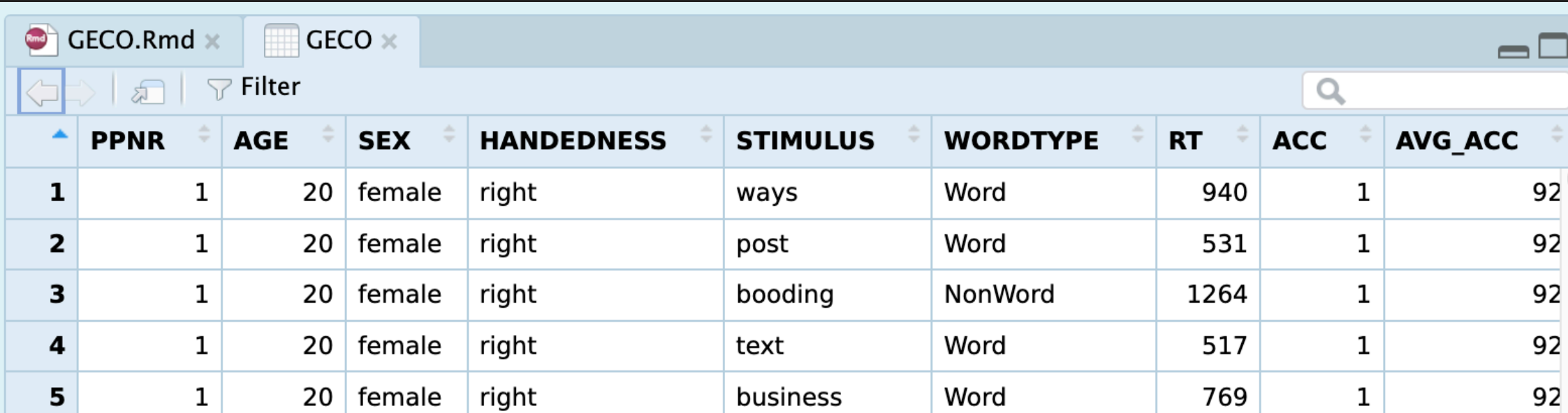
→

| color | total |
|-------|-------|
| blue | 8 |
| black | 7 |

```
df %>%  
  group_by(color) %>%  
  summarize(total = sum(value))
```

`group_by()` + `summarize()`

- What kind of `group_by()` and `summarize()` operations would be useful for the GECO dataset?



The screenshot shows an RStudio window with two tabs: 'GECO.Rmd' and 'GECO'. The 'GECO' tab is active, displaying a data table. The table has a search bar and a 'Filter' button. The columns are: PPNR, AGE, SEX, HANDEDNESS, STIMULUS, WORDTYPE, RT, ACC, and AVG_ACC. The first five rows are visible, showing data for different stimuli: 'ways', 'post', 'booding', 'text', and 'business'.

| | PPNR | AGE | SEX | HANDEDNESS | STIMULUS | WORDTYPE | RT | ACC | AVG_ACC |
|---|------|-----|--------|------------|----------|----------|------|-----|---------|
| 1 | 1 | 20 | female | right | ways | Word | 940 | 1 | 92 |
| 2 | 1 | 20 | female | right | post | Word | 531 | 1 | 92 |
| 3 | 1 | 20 | female | right | booding | NonWord | 1264 | 1 | 92 |
| 4 | 1 | 20 | female | right | text | Word | 517 | 1 | 92 |
| 5 | 1 | 20 | female | right | business | Word | 769 | 1 | 92 |

`group_by()` + `summarize()`

```
df %>%  
  group_by(color) %>%  
  summarize(total = sum(value))
```

- Group the dataset by **SEX** and calculate the mean reaction time **RT**.

A

Which group (**male** or **female**) is faster?

- What about accuracy? Which group is more accurate?
Calculate mean accuracy in the basis of the variable **ACC**.

- Are the participants slower in reacting to **NonWord** than to **Word**?

Group by **WORDTYPE** and calculate the mean reaction time for each group?

B

- Are there sex differences? Group by **WORDTYPE** and **SEX** and compare the average reaction time. Discuss the results with your neighbor.

`group_by()` + `summarize()`

```
df %>%  
  group_by(color) %>%  
  summarize(total = sum(value))
```

- Group the dataset by **SEX** and calculate the mean reaction time **RT**.

A

Which group (**male** or **female**) is faster?

```
114 ▾ ````{r message = FALSE, echo = FALSE}  
115 GECO %>% group_by(SEX) %>% summarize(mean_AGE = mean(AGE))  
116 ▲ ````
```

A tibble: 2 × 2

| SEX
<chr> | mean_AGE
<dbl> |
|---------------------|--------------------------|
| female | 18.36923 |
| male | 18.75000 |

2 rows

`group_by()` + `summarize()`

```
df %>%  
  group_by(color) %>%  
  summarize(total = sum(value))
```

- What about accuracy? Which group is more accurate?
Calculate mean accuracy in the basis of the variable **ACC**.

A

```
117 ▾ ```{r message = FALSE, echo = FALSE}  
118 GECO %>% group_by(SEX) %>% summarize(mean_ACC = mean(ACC))  
119 ▲ ```
```

A tibble: 2 × 2

| SEX
<chr> | mean_ACC
<dbl> |
|---------------------|--------------------------|
| female | 0.8931538 |
| male | 0.9279687 |

2 rows

group_by() + summarize()

- What could we do next with these results?

```
114 ▾ ```{r message = FALSE, echo = FALSE}
115 GECO %>% group_by(SEX) %>% summarize(mean_AGE = mean(AGE))
116 ▲ ```
```

A tibble: 2 × 2

| SEX
<chr> | mean_AGE
<dbl> |
|--------------|-------------------|
| female | 18.36923 |
| male | 18.75000 |

2 rows

```
117 ▾ ```{r message = FALSE, echo = FALSE}
118 GECO %>% group_by(SEX) %>% summarize(mean_ACC = mean(A))
119 ▲ ```
```

A tibble: 2 × 2

| SEX
<chr> | mean_ACC
<dbl> |
|--------------|-------------------|
| female | 0.8931538 |
| male | 0.9279687 |

2 rows

What is worth visualizing?

What is worth visualizing tomorrow?

- Stand up and join someone you haven't had a chance to talk to
- In groups of three look at the GECO dataset
- What variables and what relations should we inspect visually?
- What kind of plots would make sense?
- Draft on a piece of paper what those plots should look.
- Come up with at least two different plot types and try to use different variables

Visualization basics (1)

<https://www.1843magazine.com/travel/what-the-numbers-say/come-fly-with-me>

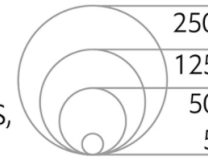
What's in a plot?

- The article considered data on 18 major international carriers: Asia's airlines are excellent, Europe's are competent, America's are awful
- Used flight-volume data from FlightStats.com and customer-satisfaction data from Skytrax (users were asked to rate carriers on a 1 to 5 scale quality of their food, service, entertainment, etc.)
- How many variables are represented in this plot?

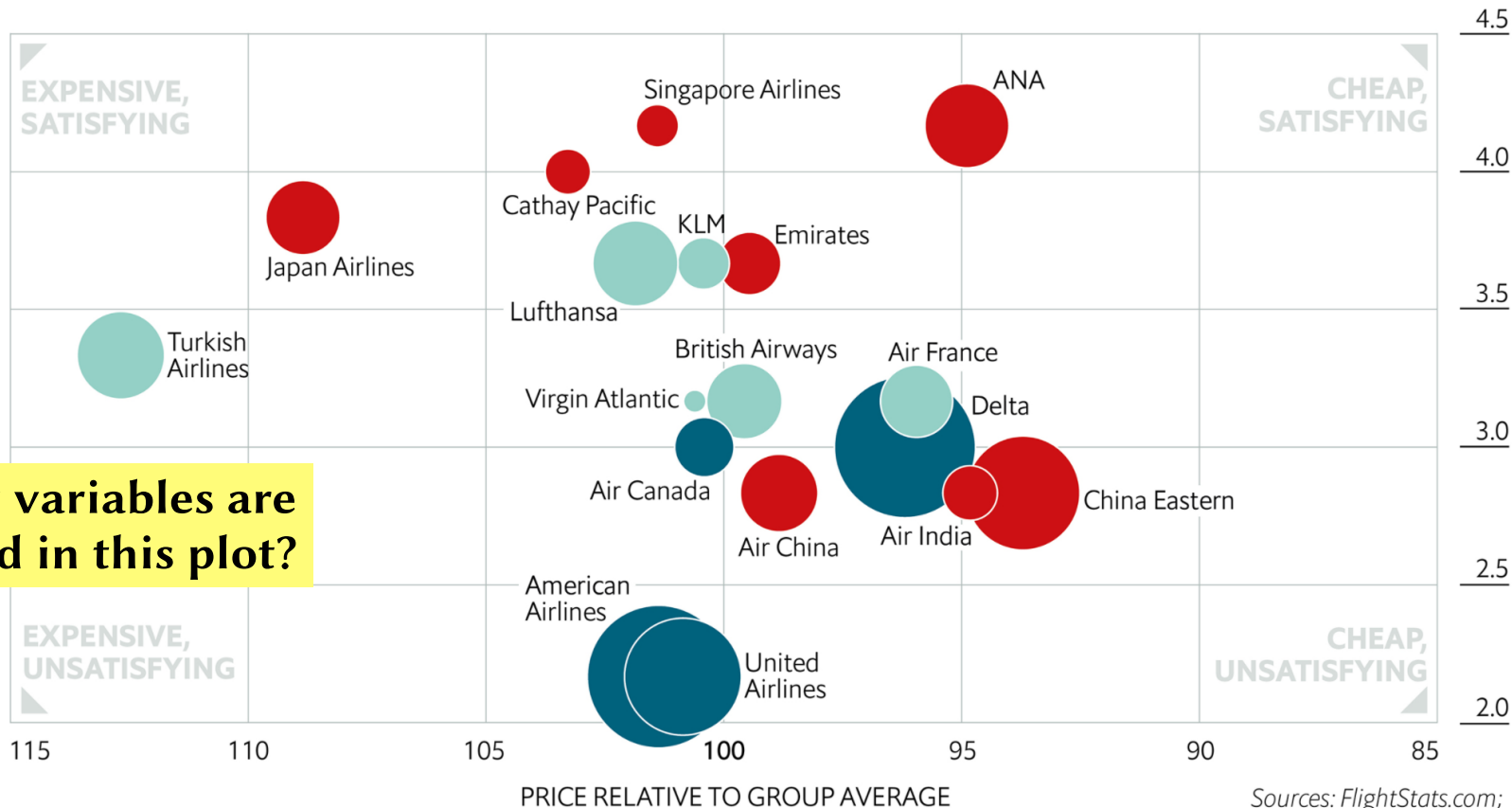
INTERNATIONAL AIRLINES PRICE V SERVICE

● NORTH AMERICAN ● EUROPEAN ● ASIAN

TOTAL FLIGHTS, Q4 2015, '000



AVERAGE CUSTOMER RATING (OUT OF 5)



How many variables are represented in this plot?

Sources: FlightStats.com; Skytrax; Google Flights; airlines

Aesthetics

- What does this word mean to you?

Aesthetics

- What does this word mean to you?
Probably, something along these lines

aes·thet·ics | es'THediks | (also **esthetics**)

plural noun [*usually treated as singular*]

a set of principles concerned with the nature and appreciation of beauty, especially in art: *these could definitely be some of the best wireless earbuds for those with an eye for aesthetics.*

- the branch of philosophy that deals with the principles of beauty and artistic taste: *she tries to impress her audience with abstruse references to modern philosophies or theories of aesthetics.*

Aesthetics

aesthetic < Gr. *aisthētikos* < *aisthēta* ‘perceptible things’
< *aisthēsthai* ‘perceive’

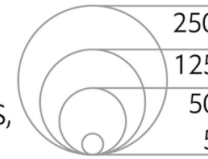
In statistical graphics (specifically in the ggplot2 sense) this old usage is meant: aesthetics are principles for relating sensory attributes (color, shape, etc.) to variables. In modern usage (since mid 18th c., first in Ger. and then in Eng.), *aesthetics* can also mean taste or beauty.

- In a plot one aesthetic attribute can represent one variable
- Sophisticated plots combine various aesthetic attributes (color, shape, line type, size, position, transparency, text, etc.) to represent multiple variables at ones

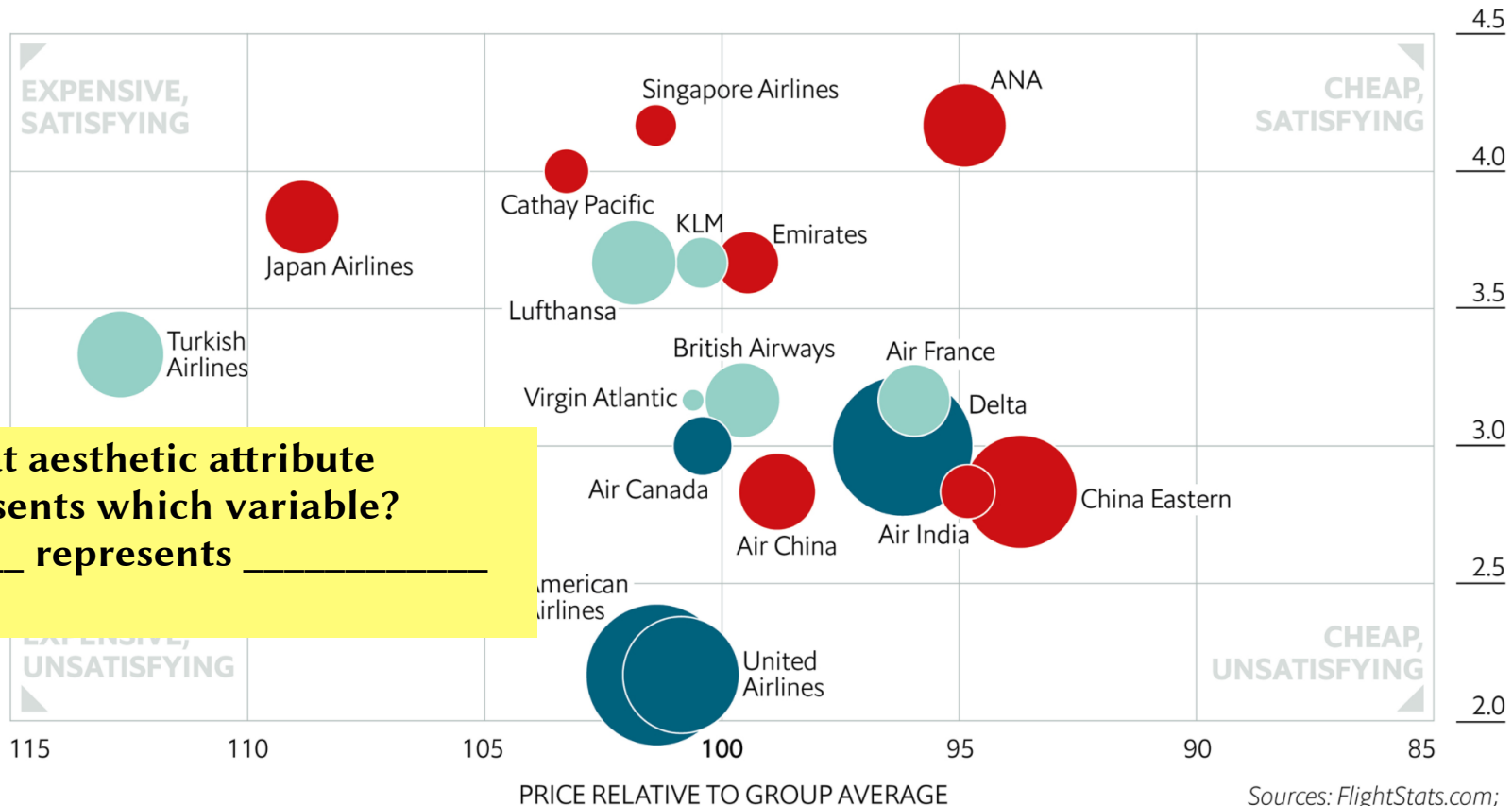
INTERNATIONAL AIRLINES PRICE V SERVICE

● NORTH AMERICAN ● EUROPEAN ● ASIAN

TOTAL FLIGHTS,
Q4 2015, '000



AVERAGE CUSTOMER
RATING (OUT OF 5)



What aesthetic attribute represents which variable?
_____ represents _____

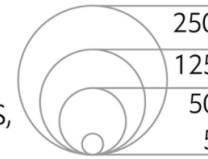
Sources: FlightStats.com; Skytrax; Google Flights; airlines

<https://www.1843magazine.com/travel/what-the-numbers-say/come-fly-with-me>

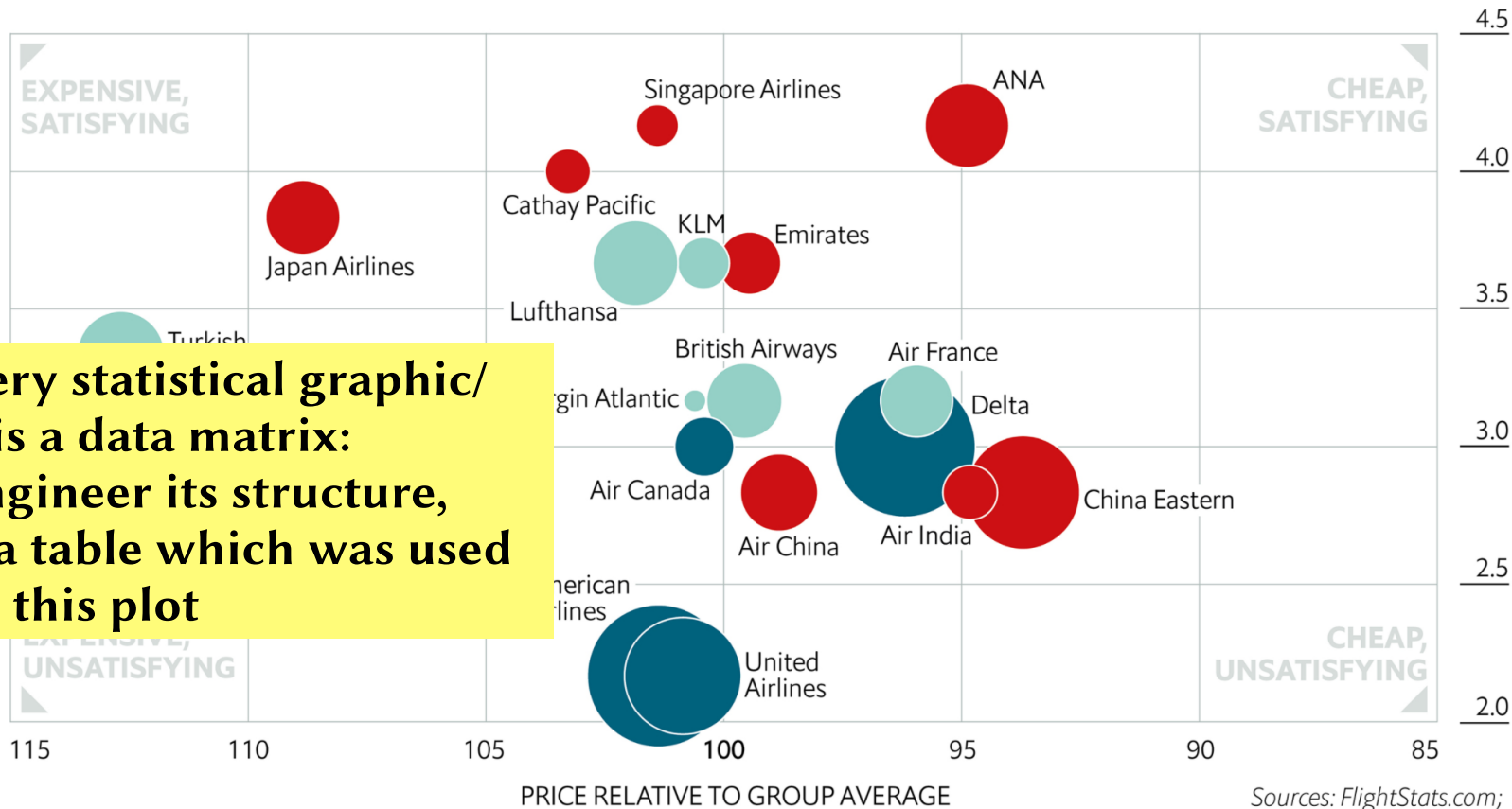
INTERNATIONAL AIRLINES PRICE V SERVICE

● NORTH AMERICAN ● EUROPEAN ● ASIAN

TOTAL FLIGHTS, Q4 2015, '000



AVERAGE CUSTOMER RATING (OUT OF 5)



**Behind every statistical graphic/
plot there is a data matrix:
Reverse-engineer its structure,
i.e. sketch a table which was used
to produce this plot**

Sources: FlightStats.com;
Skytrax; Google Flights; airlines

Behind every plot ...

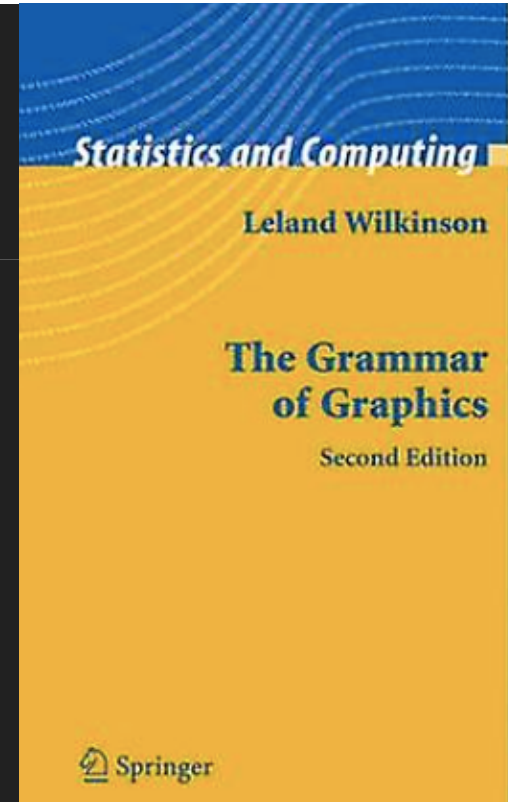
| Airline | Total flights | Price relative ... | Customer rating |
|---------|---------------|--------------------|-----------------|
| ANA | 80 | 94 | 4.2 |
| Delta | ? | ? | ? |
| Turkish | ? | ? | ? |
| ... | ... | ... | ... |

- every column represents a particular variable
- each row/record corresponds to a given member of the data set in question
- Tabular data are inherently rectangular and cannot have “ragged rows”
- If any row is lacking information for a particular column a missing value (NA) must be stored in that cell

ggplot2

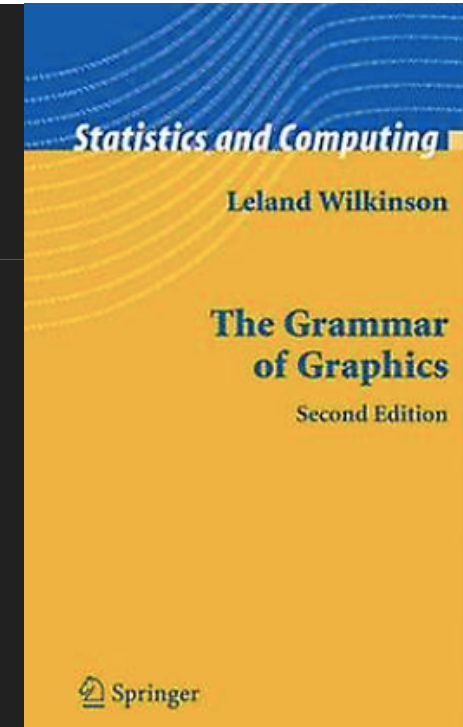
The grammar of graphics

- The package ggplot2 is based on the principles in *The Grammar of Graphics* (this is why gg) by Leland Wilkinson (1999)



The grammar of graphics

- The package ggplot2 is based on the principles in *The Grammar of Graphics* (this is why gg) by Leland Wilkinson (1999)
- A statistical graphic is understood as a mapping from **data** to **aesthetic** attributes (colour, shape, size) of **geometric** objects (points, lines, bars)
- The plot may also contain statistical transformations of the data
- Faceting can be used to generate the same plot for different subsets of the dataset



Layers of graphics in ggplot2

- The *layered* Grammar of Graphics by Wickham (2009) adjusts Wilkinson's principles to R
- Each layer/component of the *Grammar of Graphics* has a special name in `ggplot2`

Theme

Coordinates

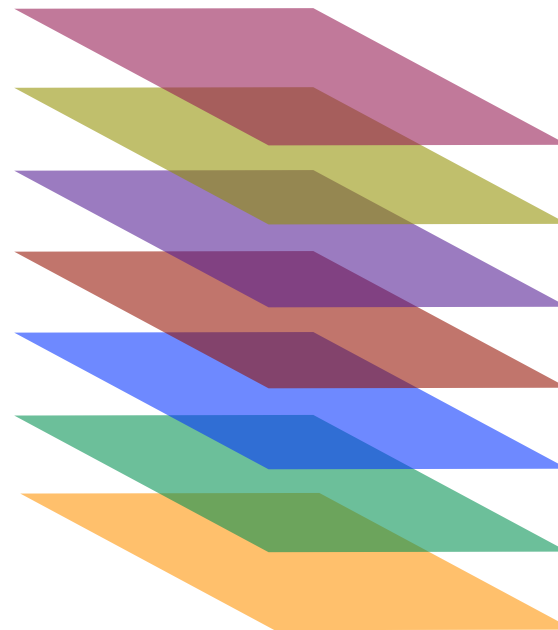
Statistics

Facets

Geometries

Aesthetics

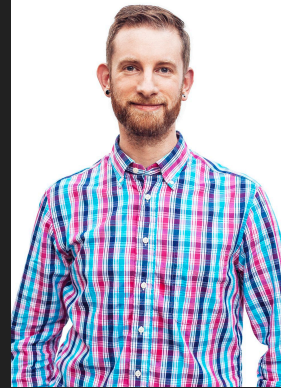
Data



https://en.wikipedia.org/wiki/Hadley_Wickham

People behind the tools

- Hadley Wickham is an adjunct professor of statistics at the University of Auckland, Stanford University, and Rice University
- Doctoral thesis (2008) on *Practical tools for exploring data and models*



The layered grammar of graphics in ggplot2

- The layered Grammar of Graphics by Wickham (2009) adjusts Wilkinson's principles to R
- Each layer/component of the Grammar of Graphics has a special name in ggplot2

Theme

Coordinates

Statistics

Facets

Geometries

Aesthetics

Data



`ggplot()`

- The basic ggplot2 function:
`ggplot(df, aes(x = var)) +
 geom_histogram()`

The symbol + at the end of each line
if there is something to follow

Theme
Coordinates
Statistics
Facets
Geometries
Aesthetics
Data



ggplot()

- The basic ggplot2 function:
`ggplot(df, aes(x = var)) +
 geom_histogram()`

- Let's plot RT
`ggplot(GECO, aes(x = RT)) +
 geom_histogram()`

- Alternative notation:
`GECO %>% ggplot(aes(x = RT)) +
 geom_histogram()`

Theme
Coordinates
Statistics
Facets
Geometries
Aesthetics
Data



ggplot()

- The basic ggplot2 function:

```
ggplot(df, aes(x = var)) +  
  geom_histogram()
```

- Let's plot RT

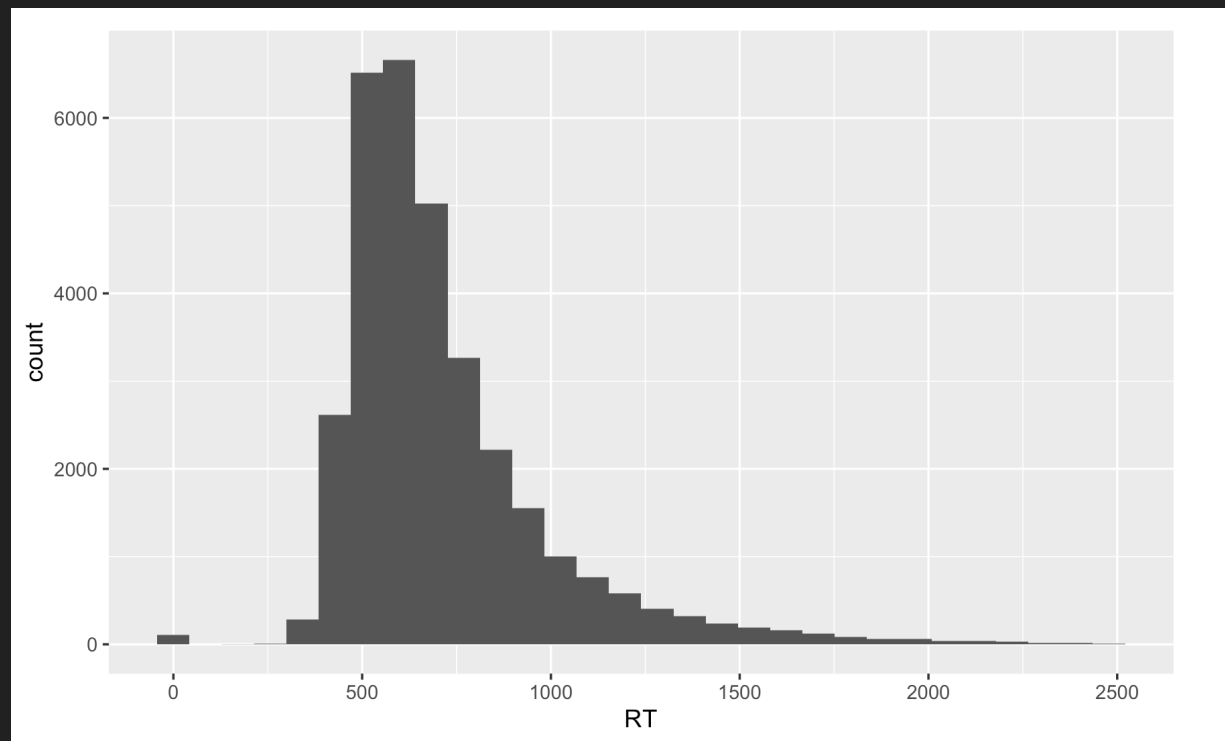
```
ggplot(GECO, aes(x = RT, fill = SEX)) +  
  geom_histogram()
```

- Alternative notation:

```
GECO %>% ggplot(aes(x = RT)) +  
  geom_histogram()
```

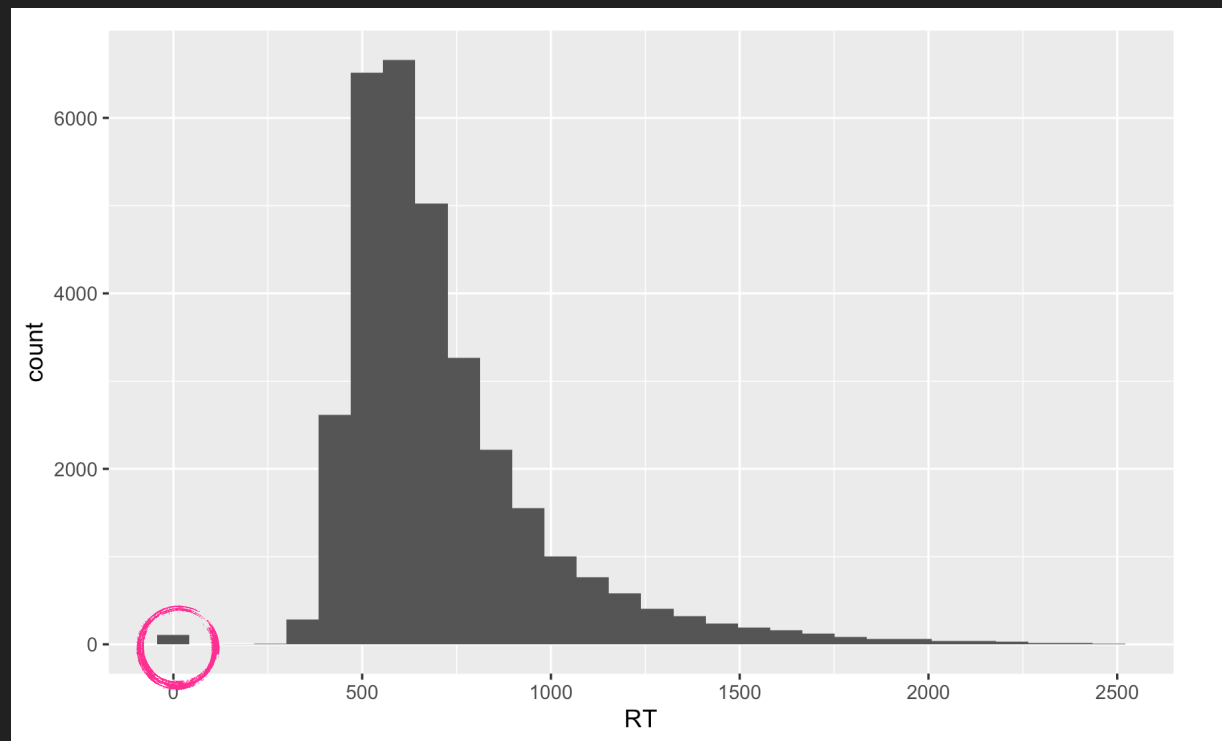
Histogram

- Are there any obvious issues with the dataset? Does it need some cleaning?



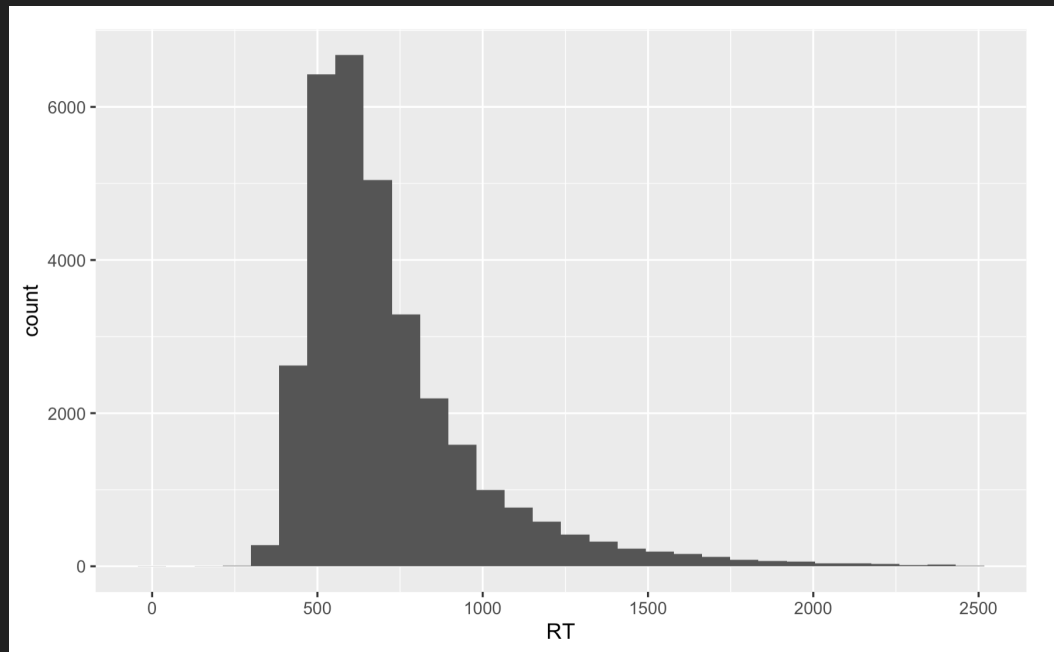
Histogram

- Are there any obvious issues with the dataset? Does it need some cleaning?



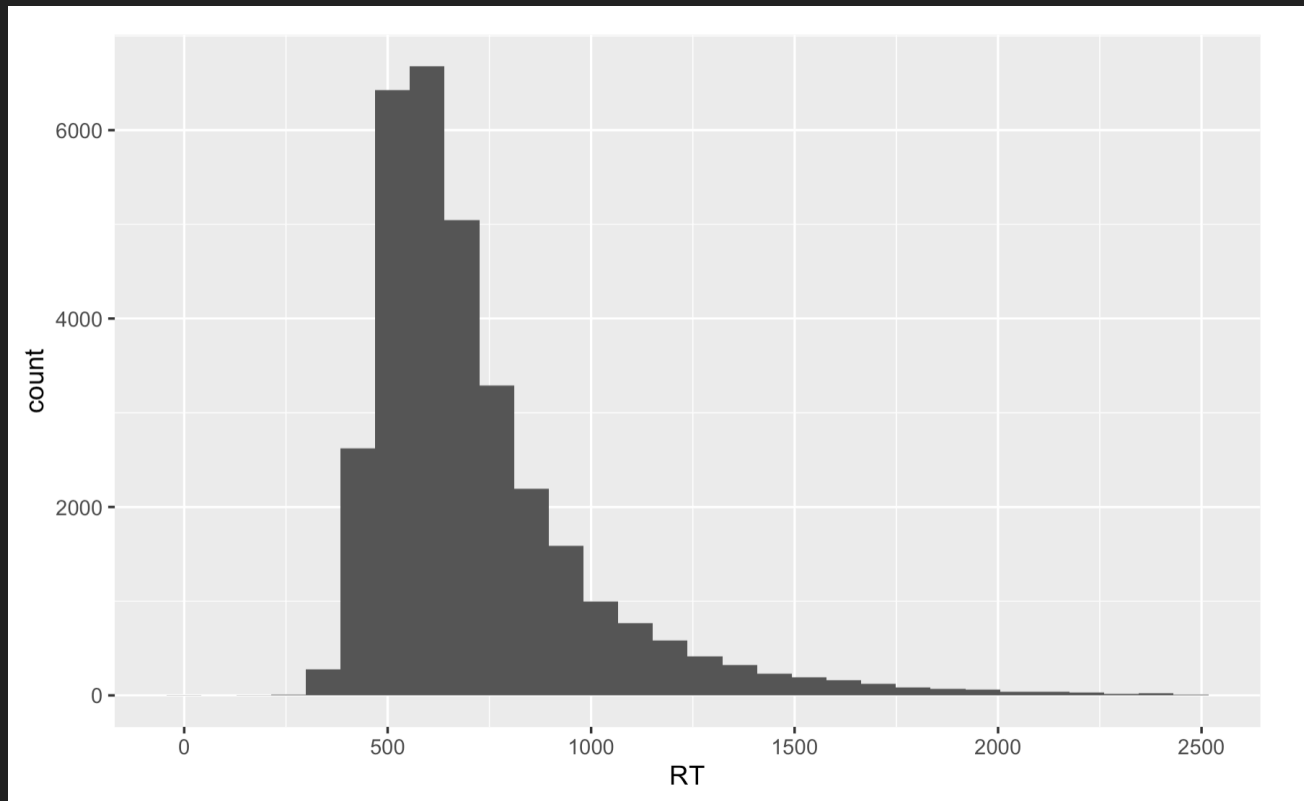
Remove 0's in RT

```
GECO <- GECO %>% filter(RT>0)  
GECO %>% ggplot(aes(x = RT)) +  
  geom_histogram()
```



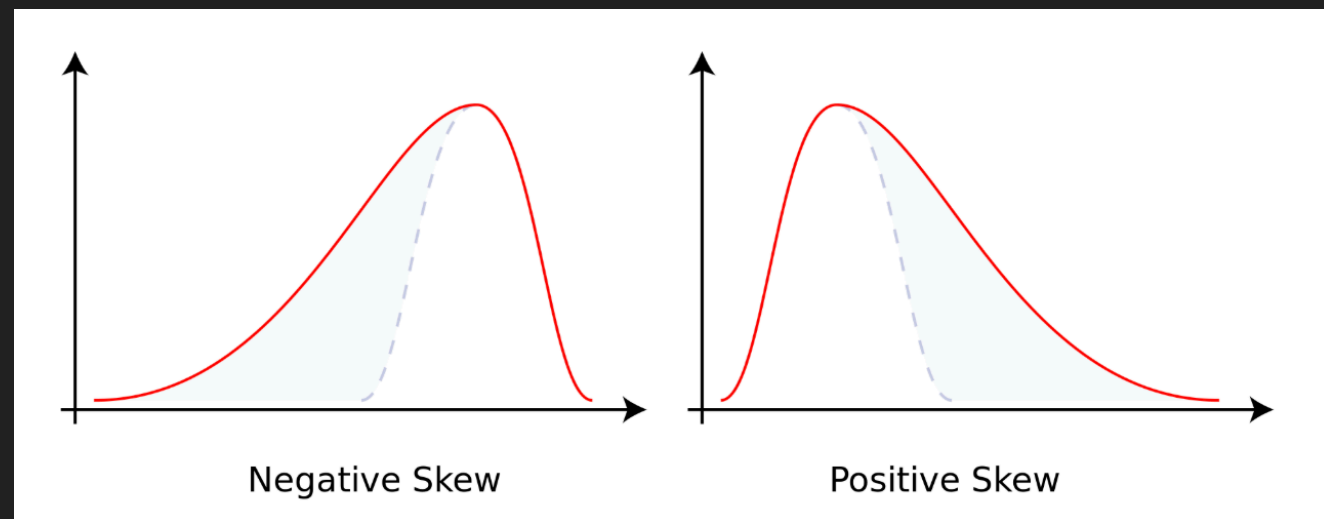
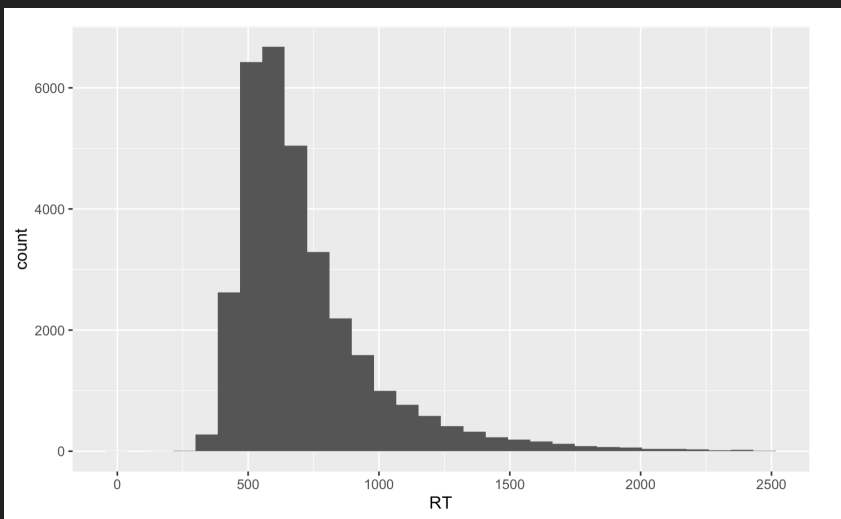
The shape of the distribution

- What is the shape of the distribution? Is it a normal distribution?



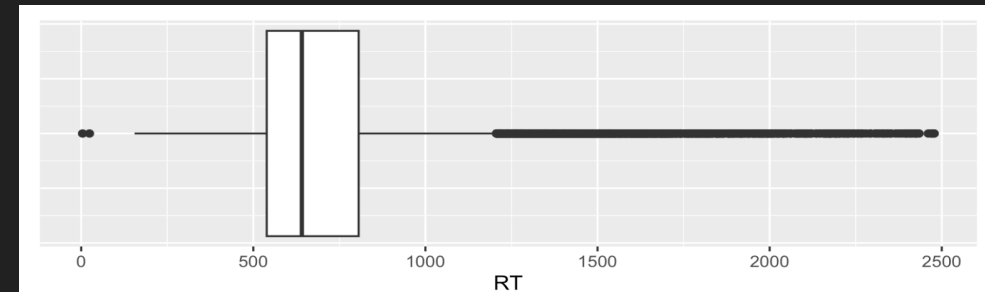
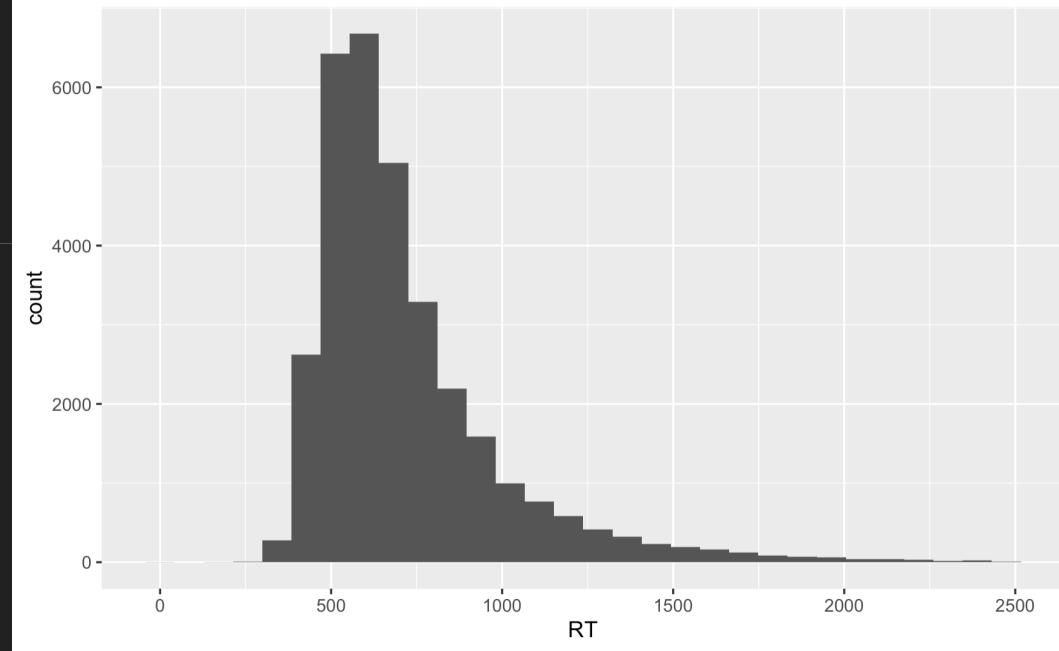
The shape of the distribution

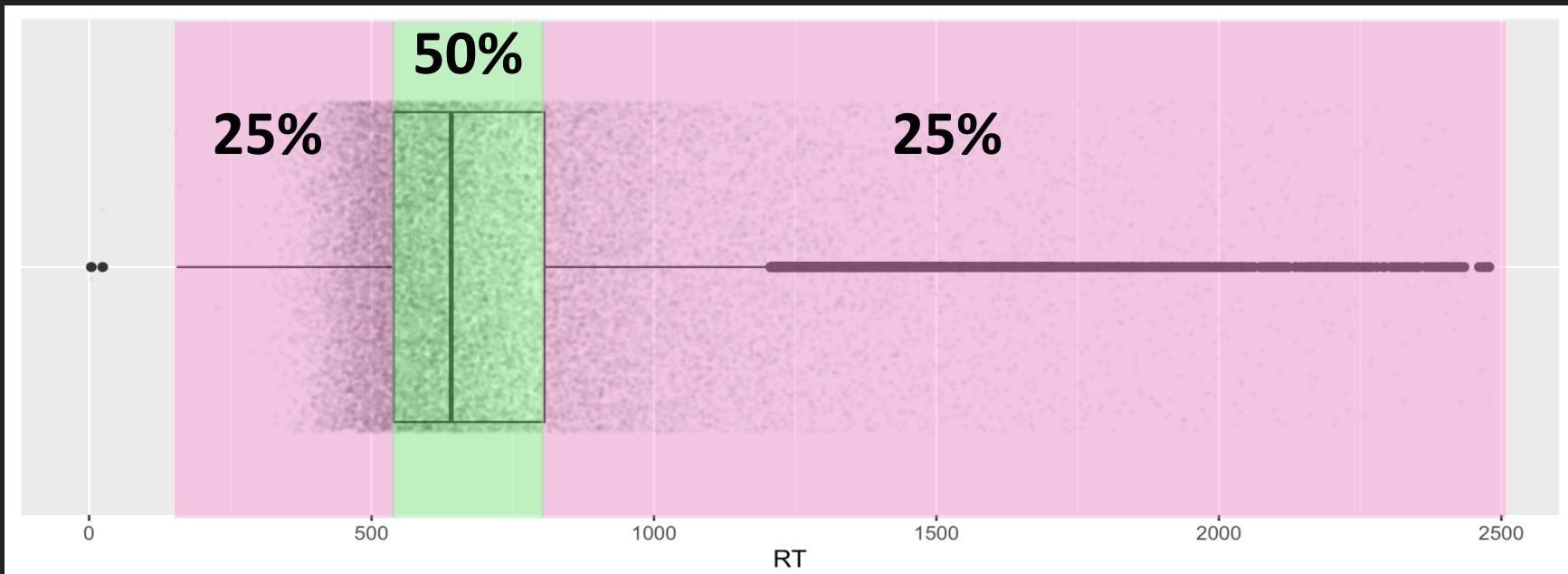
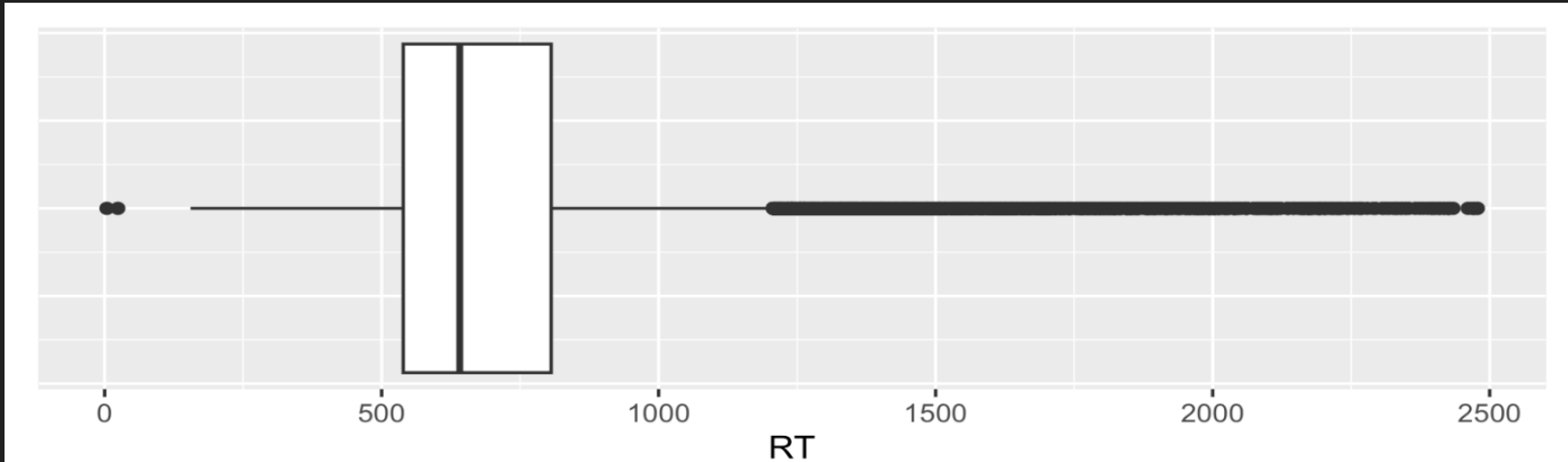
- What is the shape of the distribution? Is it a normal distribution?
- Where is the long tail?



Same data, different visualization

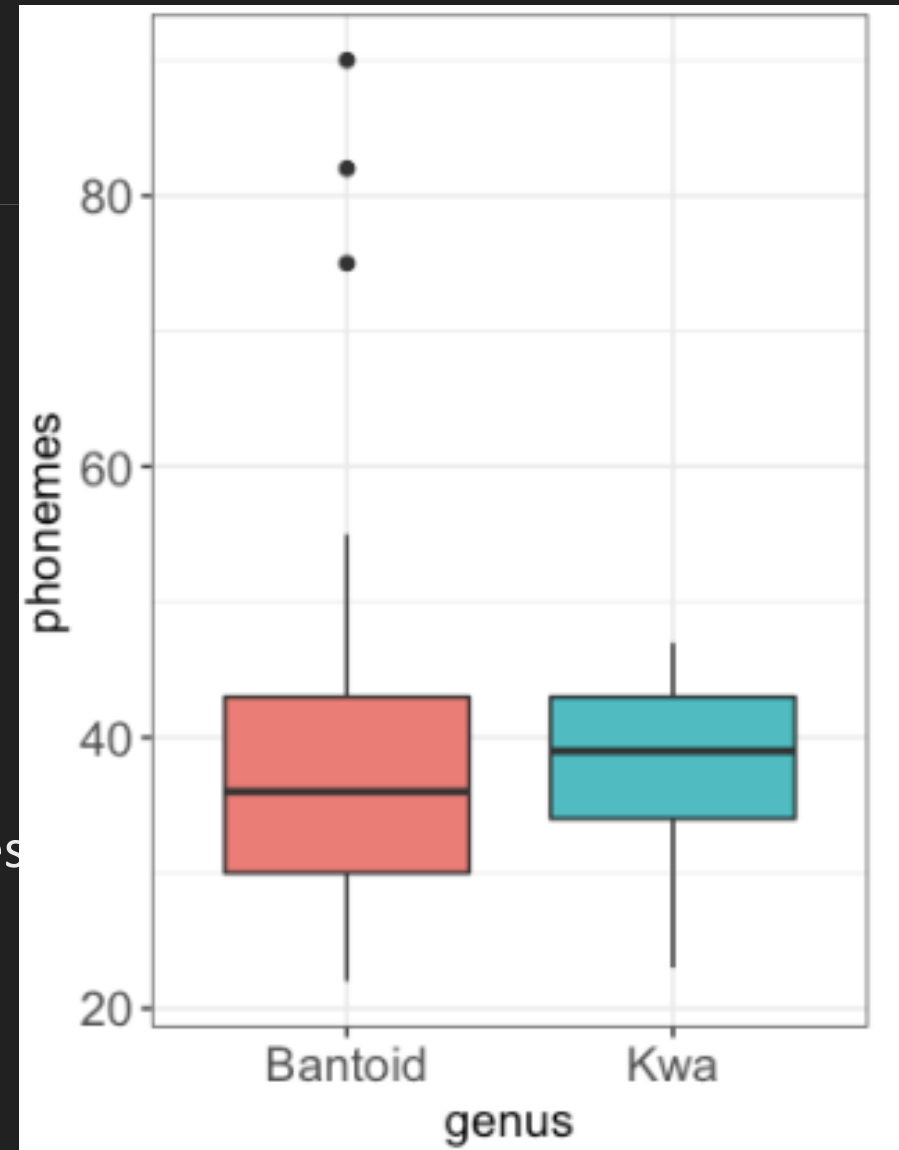
- Same data can be also visualized
- with a boxplot
- Explain to your neighbor what is the box and what are the whiskers
- What does the line in the middle represent?





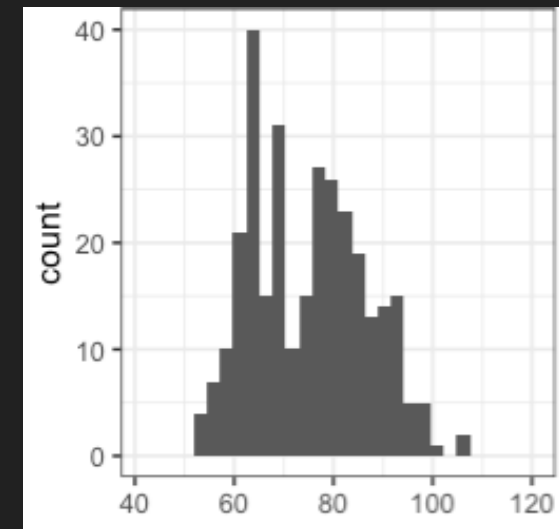
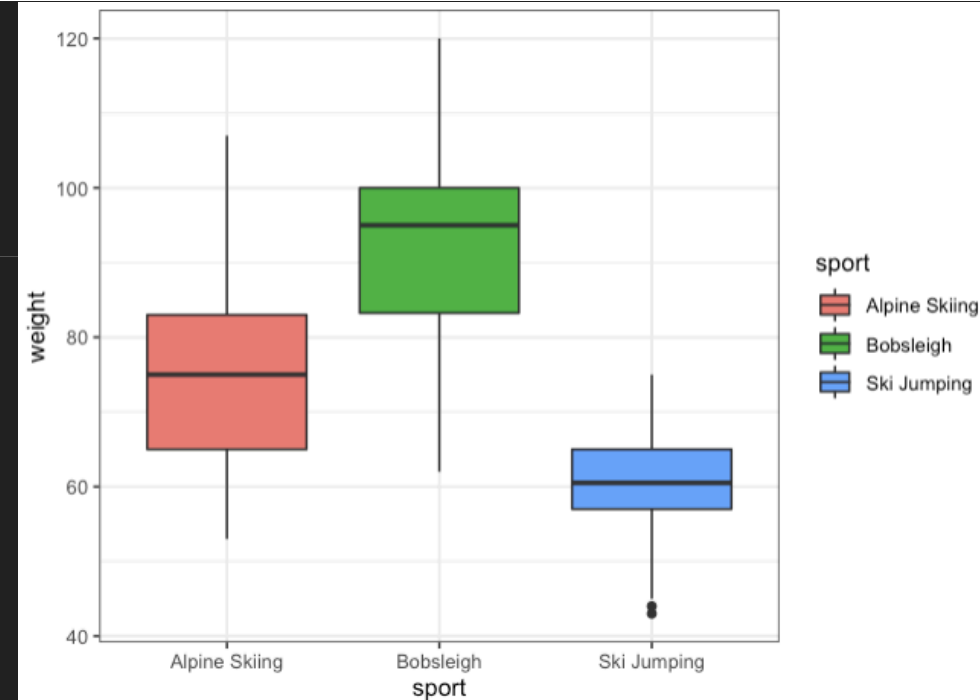
Reading boxplots

- Consider the two boxplots for the phoneme inventories of the Bantoid and Kwa languages
- Which distribution has **outliers**?
How many?
- Which distribution is more compact?
- Which distribution has the largest **range** (with and without the **outliers**)?
- What is the range of Kwa phoneme inventories
- Which distribution is **negatively skewed**?
- How many languages are in each boxplot?



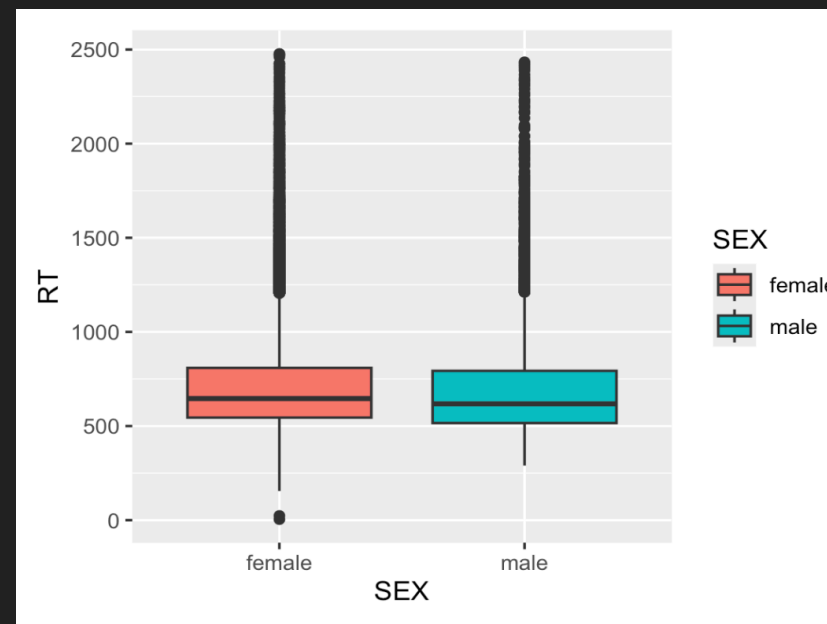
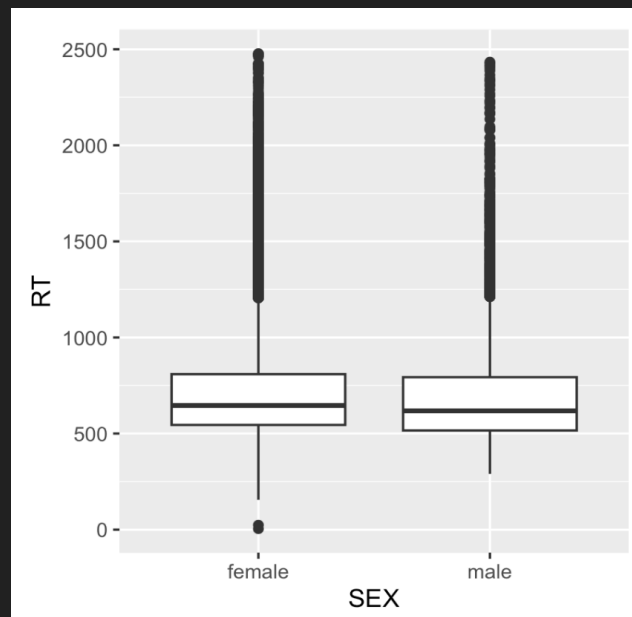
Reading boxplots

- Consider the three boxplots for the weight (kg) of different athletes
- Which distribution has outliers? How many?
- Which athletes are the heaviest on average? How do you know?
- Which distribution has the heaviest athlete? How heavy is this athlete?
- Which distribution has the smallest range?
- Which boxplot corresponds to this histogram?



Side-by-side boxplots

- Great to compare groups: e.g. **RT** by **SEX** or **RT** by **HANDEDNESS**
- Try it out: you need `x =` and `y =` in `aes()`
- You can add `fill = SEX` to fill the boxes with color by grouping variable



Side-by-side boxplots

- Great to compare groups: e.g. RT by SEX or RT by HANDEDNESS
- Try it out: you need `box` `x =` and `y =` in `aes()`
- You can add `fill = SEX` to fill the boxes with color by grouping variable:

```
GECO %>%
```

```
ggplot(aes(x = SEX, y = RT, fill = SEX)) +  
  geom_boxplot()
```

