

# Data collection and data wrangling

The ASTA team

## Contents

<b>1</b>	<b>Data</b>	<b>1</b>
1.1	Data example . . . . .	1
<b>2</b>	<b>Summaries and plots of qualitative variables</b>	<b>2</b>
2.1	Tables of qualitative variables . . . . .	2
2.2	Plots of qualitative variables . . . . .	2
<b>3</b>	<b>Target population and random sampling</b>	<b>4</b>
3.1	Population parameters . . . . .	4
3.2	Aim of statistics . . . . .	4
3.3	Random sampling schemes . . . . .	5
<b>4</b>	<b>Biases</b>	<b>5</b>
4.1	Types of biases . . . . .	5
4.2	Example of sample bias: United States presidential election, 1936 . . . . .	5
4.3	Example of response bias: Wording matters . . . . .	6
4.4	Example of response bias: Order of questions matter . . . . .	6
4.5	Example of survivor bias: Bullet holes of honor . . . . .	6
4.6	Example of selection bias . . . . .	7

## 1 Data

### 1.1 Data example

We use data about penguins from the R package palmerpenguins

```
pingviner <- palmerpenguins::penguins
pingviner

## # A tibble: 344 x 8
##   species island  bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>          <dbl>          <dbl>          <int>        <int>
## 1 Adelie  Torgersen         39.1           18.7            181          3750
## 2 Adelie  Torgersen         39.5           17.4            186          3800
## 3 Adelie  Torgersen         40.3           18              195          3250
## 4 Adelie  Torgersen         NA             NA              NA           NA
## 5 Adelie  Torgersen         36.7           19.3            193          3450
## 6 Adelie  Torgersen         39.3           20.6            190          3650
## 7 Adelie  Torgersen         38.9           17.8            181          3625
## 8 Adelie  Torgersen         39.2           19.6            195          4675
## 9 Adelie  Torgersen         34.1           18.1            193          3475
## 10 Adelie Torgersen         42             20.2            190          4250
## # i 334 more rows
```

```
## # i 2 more variables: sex <fct>, year <int>
```

## 2 Summaries and plots of qualitative variables

### 2.1 Tables of qualitative variables

- The main function to make tables from a data frame of observations is `tally()` which tallies (counts up) the number of observations within a given category. E.g:

```
tally(~species, data = pingviner)
```

```
## species
##   Adelie Chinstrap   Gentoo
##     152      68     124
```

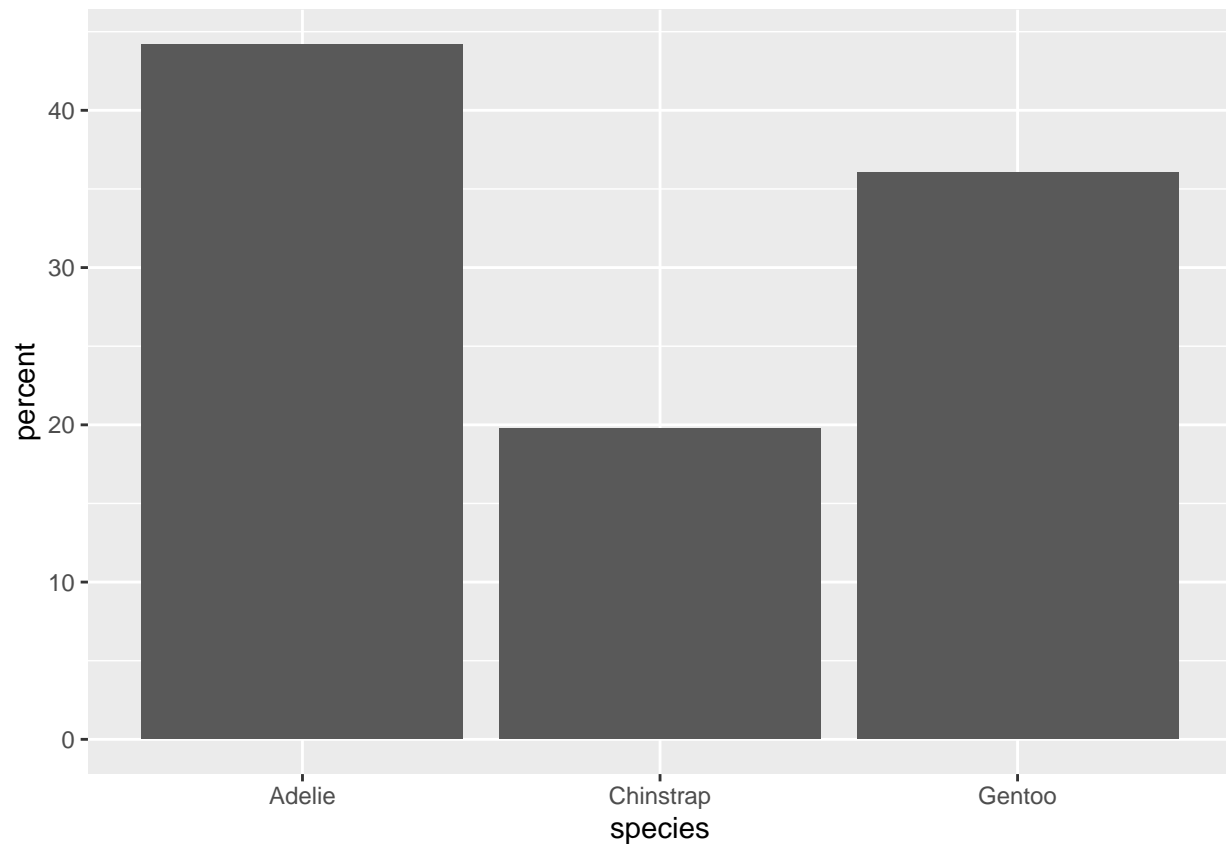
```
tally(species ~ island, data = pingviner)
```

```
##           island
## species   Biscoe Dream Torgersen
##   Adelie      44   56     52
##   Chinstrap    0   68     0
##   Gentoo     124    0     0
```

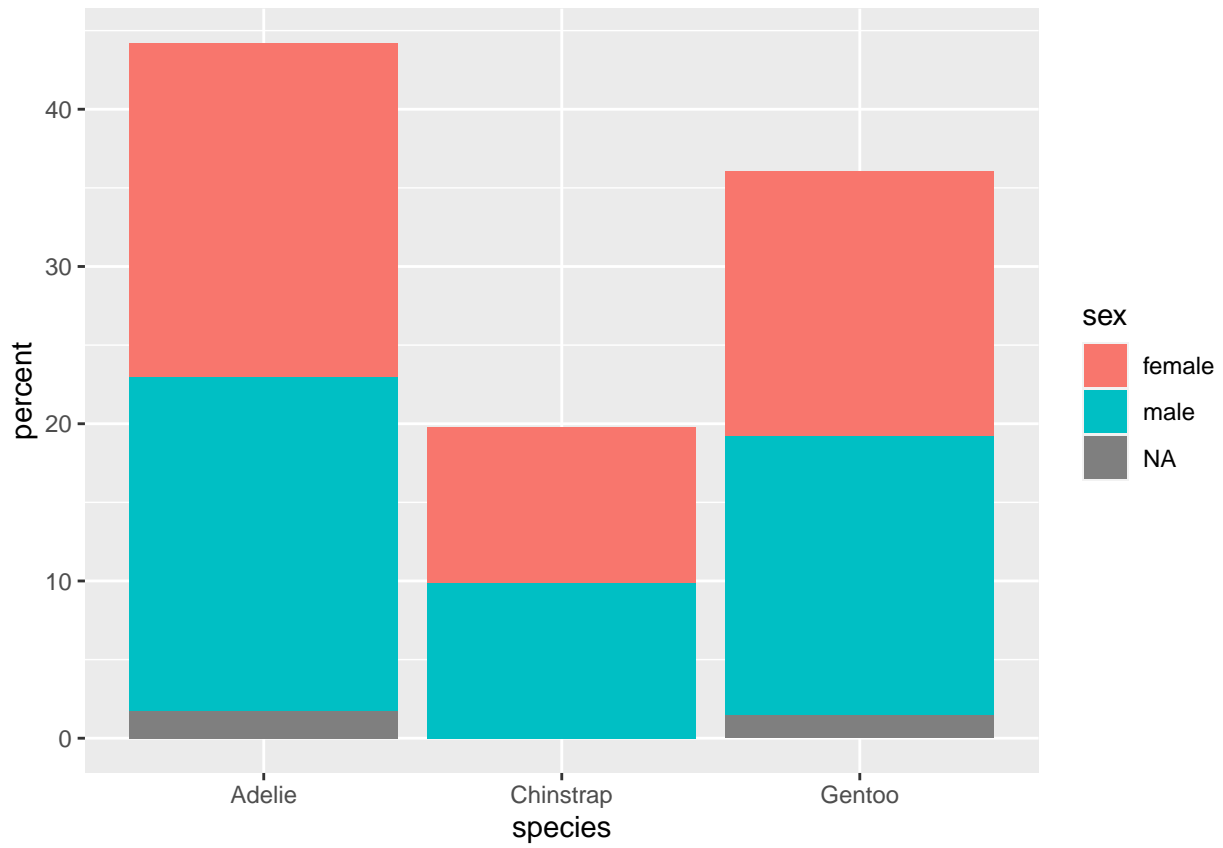
### 2.2 Plots of qualitative variables

- The main plotting functions for qualitative variables are `gf_percents()` and `gf_bar()`. E.g:

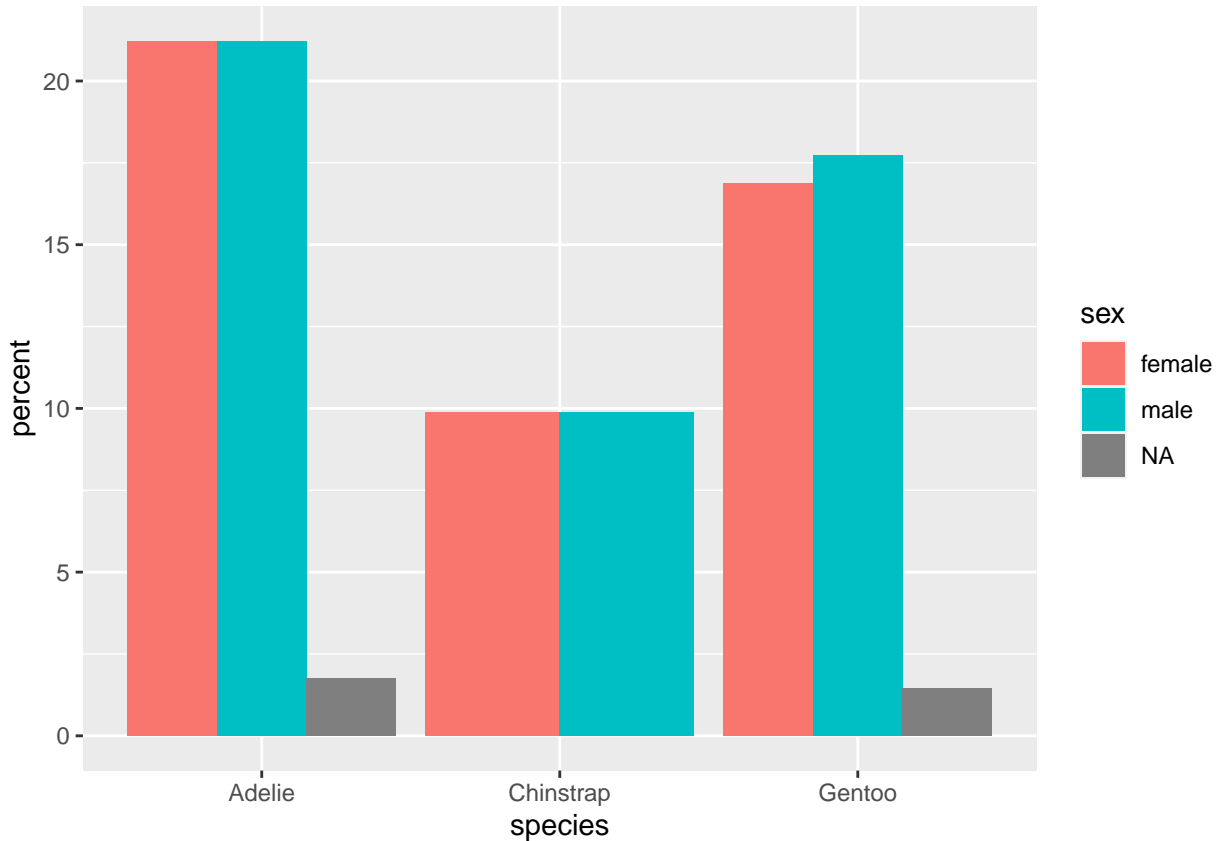
```
gf_percents(~species, data = pingviner)
```



```
gf_percents(~species, fill = ~sex, data = pingviner)
```



```
gf_percents(~species, fill = ~sex, data = pingviner, position = position_dodge())
```



### 3 Target population and random sampling

#### 3.1 Population parameters

- When the sample size grows, then e.g. the mean of the sample,  $\bar{y}$ , will stabilize around a fixed value,  $\mu$ , which is usually unknown. The value  $\mu$  is called the **population mean**.
- Correspondingly, the standard deviation of the sample,  $s$ , will stabilize around a fixed value,  $\sigma$ , which is usually unknown. The value  $\sigma$  is called the **population standard deviation**.
- Notation:
  - $\mu$  (mu) denotes the population mean.
  - $\sigma$  (sigma) denotes the population standard deviation.

Population	Sample
$\mu$	$\bar{y}$
$\sigma$	$s$

##### 3.1.1 A word about terminology

- **Standard deviation:** a measure of variability of a population or a sample.
- **Standard error:** a measure of variability of an estimate. For example, a measure of variability of the sample mean.

#### 3.2 Aim of statistics

- Statistics is all about “saying something” about a population.

- Typically, this is done by taking a random sample from the population.
- The sample is then analysed and a statement about the population can be made.
- The process of making conclusions about a population from analysing a sample is called **statistical inference**.

### 3.3 Random sampling schemes

Possible strategies for obtaining a random sample from the target population are explained in Agresti section 2.4:

- **Simple sampling: each possible sample of equal size equally probable**
- Systematic sampling
- Stratified sampling
- Cluster sampling
- Multistage sampling
- ...

## 4 Biases

### 4.1 Types of biases

Agresti section 2.3:

- Sampling/selection bias
  - Probability sampling: each sample of size  $n$  has same probability of being sampled
    - \* Still problems: undercoverage, groups not represented (inmates, homeless, hospitalized, ...)
  - Non-probability sampling: probability of sample not possible to determine
    - \* E.g. volunteer sampling
- Response bias
  - E.g. poorly worded, confusing or even order of questions
  - Lying if think socially unacceptable
- Non-response bias
  - Non-response rate high; systematic in non-responses (age, health, believes)

### 4.2 Example of sample bias: United States presidential election, 1936

(Based on Agresti, this and this.)

- Current president: Franklin D. Roosevelt
- Election: Franklin D. Roosevelt vs Alfred Landon (Republican governor of Kansas)
- Literary Digest: magazine with history of accurately predicting winner of past 5 presidential elections

---

#### 4.2.1 Results

- Literary Digest poll: Landon: 57%; Roosevelt: 43%
- Actual results: Landon: 38%; Roosevelt: 62%
- Sampling error:  $57\% - 38\% = 19\%$ 
  - Practically all of the sampling error was the result of **sample bias**
  - Poll size of > 2 mio. individuals participated – extremely large poll

---

#### 4.2.2 Problems (biases)

- Mailing list of about 10 mio. names was created

- Based on every telephone directory, lists of magazine subscribers, rosters of clubs and associations, and other sources
- Each one of 10 mio. received a mock ballot and asked to return the marked ballot to the magazine
- “respondents who returned their questionnaires represented only that subset of the population with a relatively intense interest in the subject at hand, and as such constitute in no sense a random sample ... it seems clear that the minority of anti-Roosevelt voters felt more strongly about the election than did the pro-Roosevelt majority” (*The American Statistician*, 1976)
- Biases:
  - Sample bias
    - \* List generated towards middle- and upper-class voters (e.g. 1936 and telephones)
    - \* Many unemployed (club memberships and magazine subscribers)
  - Non-response bias
    - \* Only responses from 2.3/2.4 mio out of 10 million people

### 4.3 Example of response bias: Wording matters

New York Times/CBS News poll on attitude to increased fuel taxes

- “Are you in favour of a new gasoline tax?” - 12% said yes.
- “Are you in favour of a new gasoline tax to decrease US dependency on foreign oil?” - 55% said yes.
- “Do you think a new gas tax would help to reduce global warming?” - 59% said yes.

### 4.4 Example of response bias: Order of questions matter

US study during cold war asked two questions:

1 “Do you think that US should let Russian newspaper reporters come here and sent back whatever they want?”

2 “Do you think that Russia should let American newspaper reporters come in and sent back whatever they want?”

The percentage of yes to question 1 was 36%, if it was asked first and 73%, when it was asked last.

### 4.5 Example of survivor bias: Bullet holes of honor

(Based on this.)

- World War II
  - Royal Air Force (RAF), UK
    - Lost many planes to German anti-aircraft fire
  - Armor up!
    - Where?
    - Count up all the bullet holes in planes that returned from missions
      - \* Put extra armor in the areas that attracted the most fire
- 
- Hungarian-born mathematician Abraham Wald:
    - If a plane makes it back safely with a bunch of bullet holes in its wings: holes in the wings aren't very dangerous
      - \* **Survivorship bias**
    - Armor up the areas that (on average) don't have any bullet holes
      - \* They never make it back, apparently dangerous

Section of plane	Bullet holes per square foot
Engine	1.11

---

Section of plane	Bullet holes per square foot
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.80

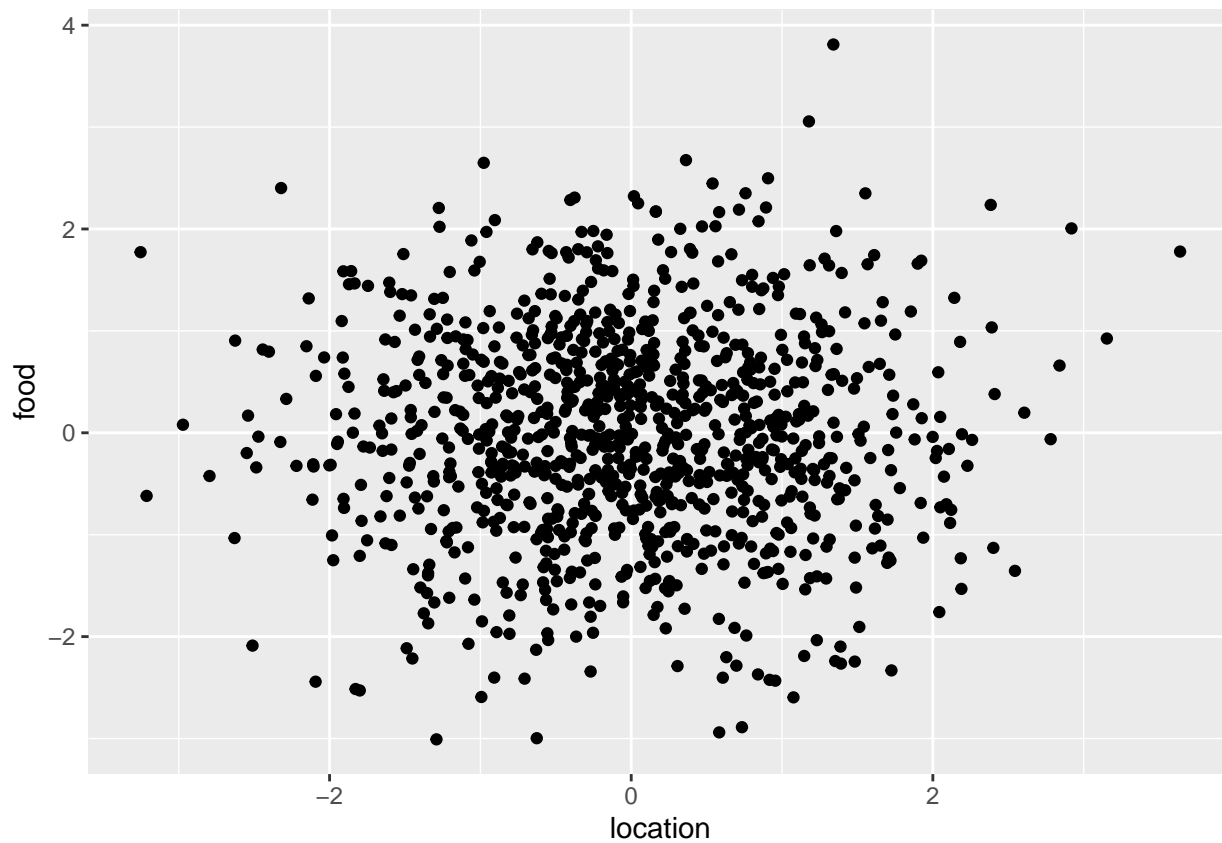
---

(See also this xkcd)

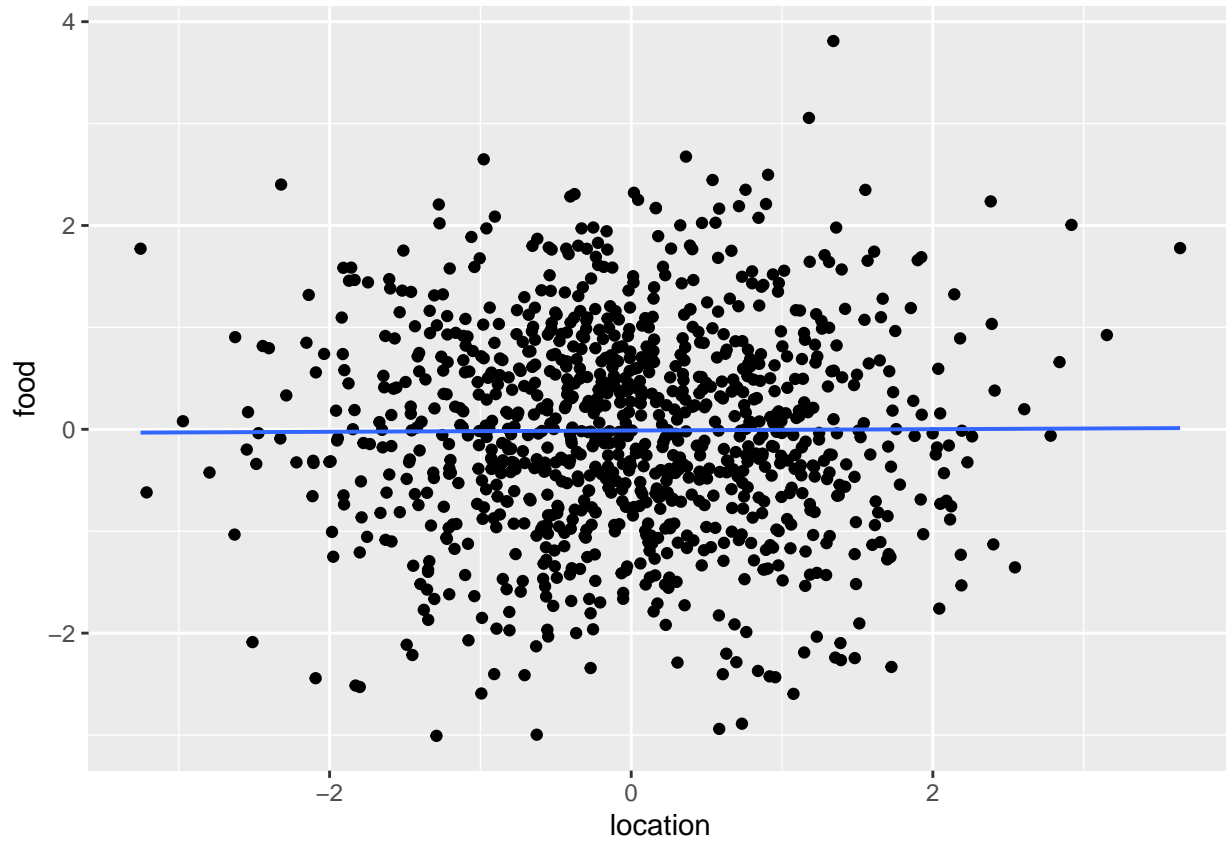
## 4.6 Example of selection bias

All restaurants:

```
set.seed(1)
n <- 1000
food <- rnorm(n, mean = 0, sd = 1)
location <- rnorm(n, mean = 0, sd = 1)
gf_point(food ~ location)
```



```
gf_point(food ~ location) %>% gf_lm()
```



```
cor.test(food, location)
```

```
##  
## Pearson's product-moment correlation  
##  
## data: x and y  
## t = 0.2, df = 998, p-value = 0.8  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.056 0.068  
## sample estimates:  
## cor  
## 0.0064
```

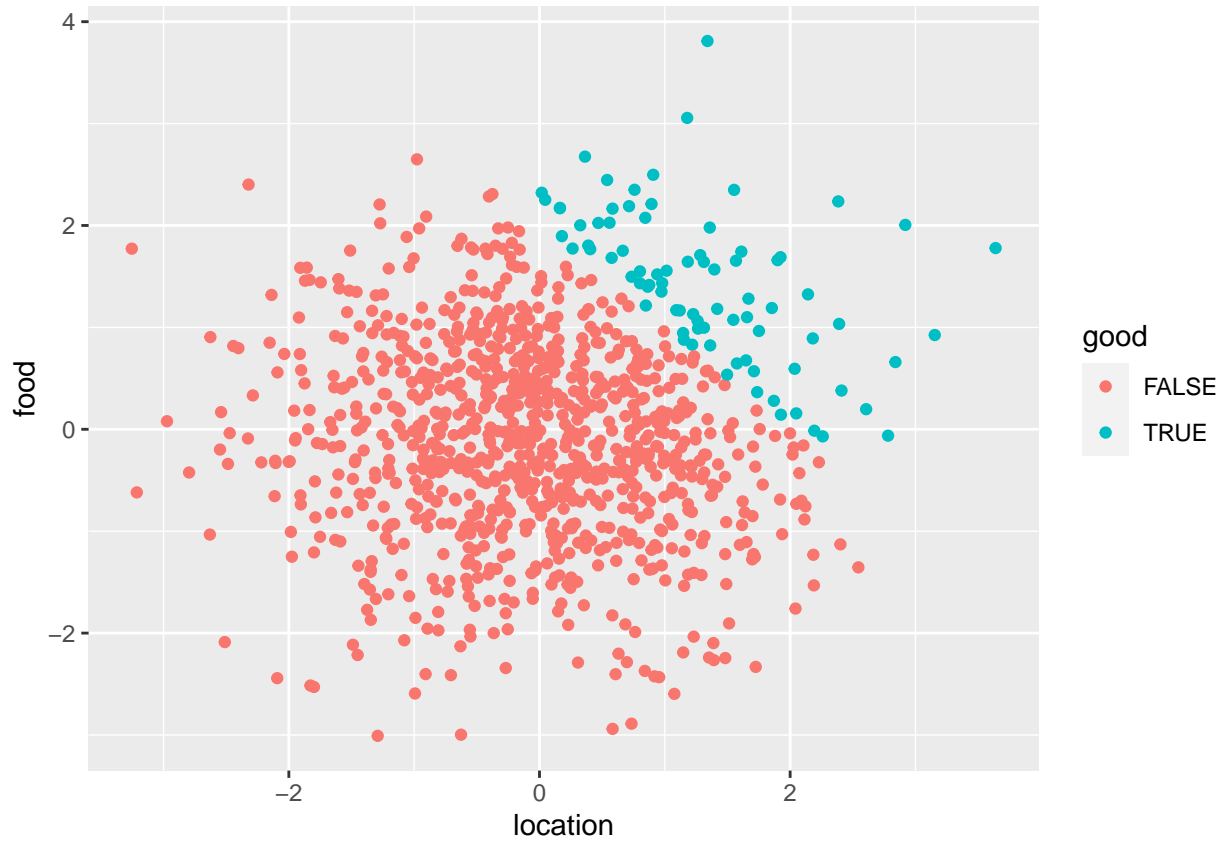
---

Total score = food + location

Good review if score > 2

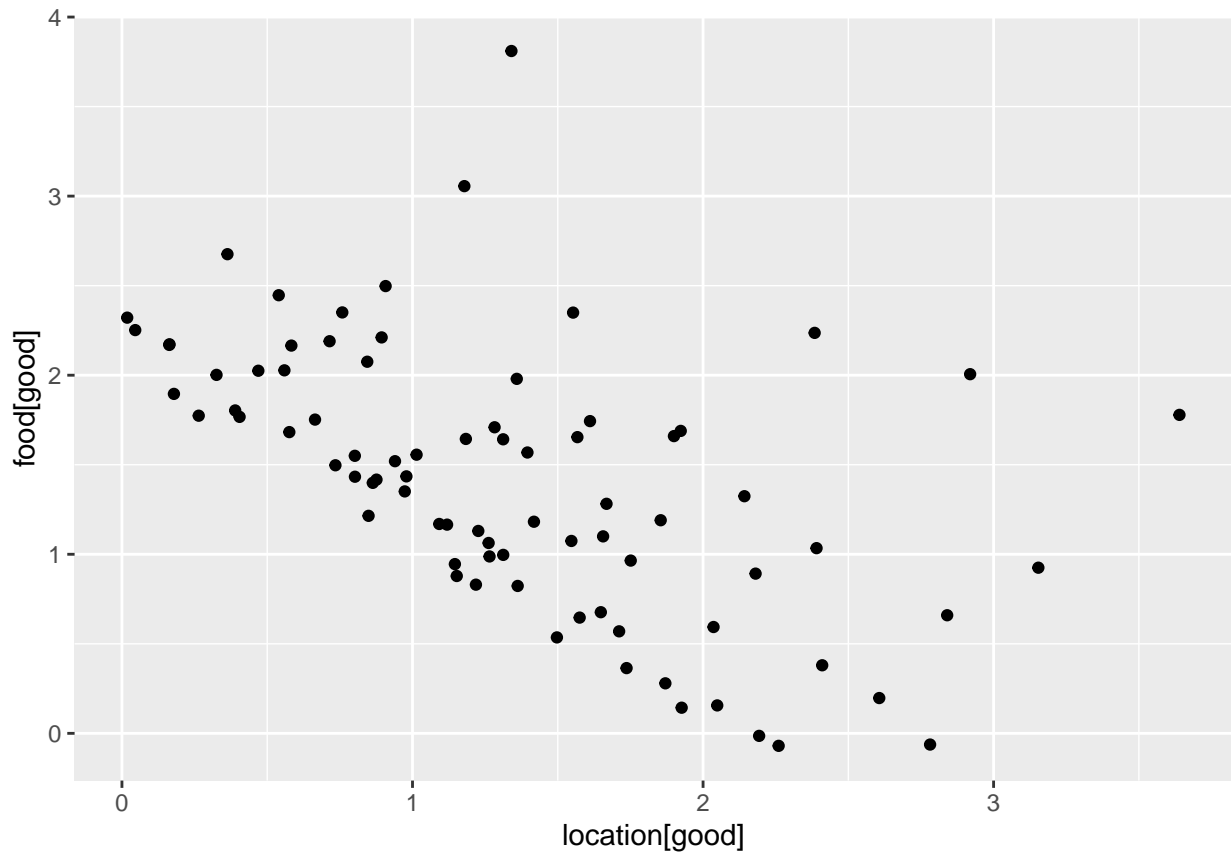
```
score <- food + location  
good <- score > 2  
gf_point(food ~ location, color = ~ good)
```



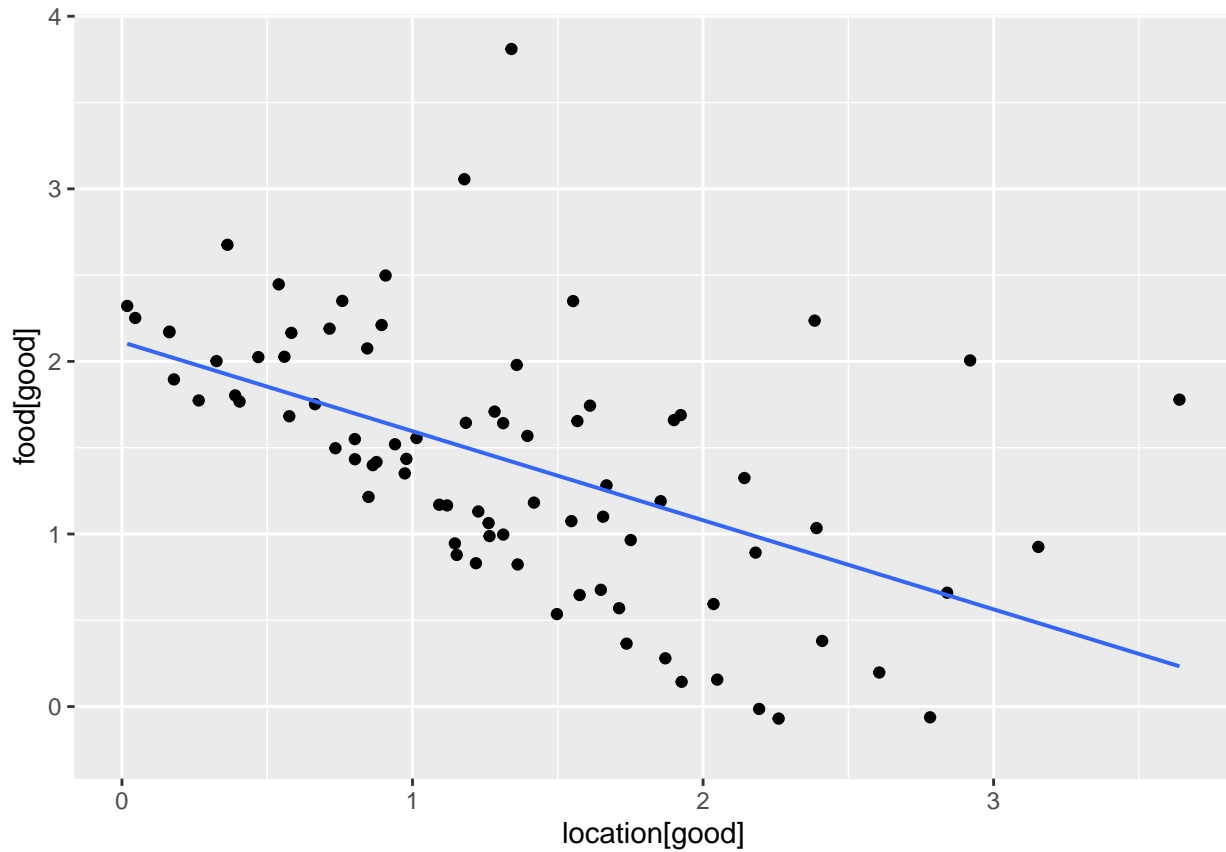


#### 4.6.1 Focusing on “good” restaurants

```
gf_point(food[good] ~ location[good])
```



```
gf_point(food[good] ~ location[good]) %>%  
  gf_lm()
```



```
cor.test(food[good], location[good])
```

```
##  
## Pearson's product-moment correlation  
##  
## data: x and y  
## t = -6, df = 79, p-value = 4e-07  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.67 -0.35  
## sample estimates:  
## cor  
## -0.53
```