

Ziling Zhen Functions and tidy evaluation

Based on Chapter 25 from *R for Data Science*

You can download this .qmd file from [here](#). Just hit the Download Raw File button.

Introduction (from Ch 25 of R4DS)

One of the best ways to improve your reach as a data scientist is to write functions. Functions allow you to automate common tasks in a more powerful and general way than copy-and-pasting. Writing a function has four big advantages over using copy-and-paste:

- You can give a function an evocative name that makes your code easier to understand.
- As requirements change, you only need to update code in one place, instead of many.
- You eliminate the chance of making incidental mistakes when you copy and paste (i.e. updating a variable name in one place, but not in another).
- It makes it easier to reuse work from project-to-project, increasing your productivity over time.

A good rule of thumb is to consider writing a function whenever you’ve copied and pasted a block of code more than twice (i.e. you now have three copies of the same code). We’ll learn about three useful types of functions:

- Vector functions take one or more vectors as input and return a vector as output.
- Data frame functions take a data frame as input and return a data frame as output.
- Plot functions that take a data frame as input and return a plot as output.

```
# Initial packages required (we'll be adding more)
library(tidyverse)
library(nycflights13)
```

Do not Repeat Yourself: Also known as DRY, if you copy or paste code more than twice, you should write a function instead.

When writing a function, it is usually best to start with the code you know works for one instance, and then “function-ize” it.

Vector functions

Example 1: Rescale variables from 0 to 1.

This code creates a 10 x 4 tibble filled with random values taken from a normal distribution with mean 0 and SD 1

```
df <- tibble(  
  a = rnorm(10),  
  b = rnorm(10),  
  c = rnorm(10),  
  d = rnorm(10)  
)  
df
```

```
# A tibble: 10 x 4  
      a      b      c      d  
  <dbl> <dbl> <dbl> <dbl>  
1 -0.156  2.15  -0.452  1.84  
2 -2.30   0.482  0.114  -0.0149  
3 -0.696 -0.590  0.734   0.856  
4 -0.726 -0.163  1.35    1.22  
5 -0.609 -0.743 -1.04    0.183  
6 -0.760 -0.554 -0.118  -0.0894  
7 -1.51   1.37  -1.34   -1.43  
8  1.45  -0.973  0.596  -0.760  
9  0.693  0.0496  0.00146  0.551  
10 0.0720 0.286  -0.425  -0.253
```

This code below for rescaling variables from 0 to 1 is ripe for functions... we did it four times!

It's easiest to start with working code and turn it into a function.

```
df$a <- (df$a - min(df$a)) / (max(df$a) - min(df$a))  
df$b <- (df$b - min(df$b)) / (max(df$b) - min(df$b))  
df$c <- (df$c - min(df$c)) / (max(df$c) - min(df$c))  
df$d <- (df$d - min(df$d)) / (max(df$d) - min(df$d))  
df
```

```
# A tibble: 10 x 4  
      a      b      c      d  
  <dbl> <dbl> <dbl> <dbl>
```

```

1 0.572 1      0.329 1
2 0      0.466 0.539 0.434
3 0.428 0.123 0.770 0.700
4 0.420 0.259 1      0.810
5 0.451 0.0739 0.110 0.494
6 0.411 0.134 0.453 0.411
7 0.210 0.751 0      0
8 1      0      0.719 0.206
9 0.798 0.327 0.497 0.607
10 0.633 0.403 0.339 0.361

```

Notice first what changes and what stays the same in each line. Then, if we look at the first line above, we see we have one value we're using over and over: `df$a`. So our function will have one input. We'll start with our code from that line, then replace the input (`df$a`) with `x`. We should give our function a name that explains what it does. The name should be a verb.

```

# I'm going to show you how to write the function in class!
# I have it in the code already below, but don't look yet!
# Let's try to write it together first!
rescale01 <- function(x){
  (x - min(x)) / (max(x) - min(x))
}

rescale01(df$a)

```

```

[1] 0.5720994 0.0000000 0.4279657 0.4198576 0.4512378 0.4108775 0.2101299
[8] 1.0000000 0.7982893 0.6327888

```

.....

```

# Our function (first draft!)
rescale01 <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

```

Note the **general form of a function**:

```

name <- function(arguments) {
  body
}

```

Every function contains 3 essential components:

- A name. The name should clearly evoke what the function does; hence, it is often a verb (action). Here we'll use `rescale01` because this function rescales a vector to lie between 0 and 1. `snake_case` is good; `CamelCase` is just okay.
- The arguments. The arguments are things that vary across calls and they are usually nouns - first the data, then other details. Our analysis above tells us that we have just one; we'll call it `x` because this is the conventional name for a numeric vector, but you can use any word.
- The body. The body is the code that's repeated across all the calls. By default a function will return the last statement; use `return()` to specify a return value

Summary: Functions should be written for both humans and computers!

Once we have written a function we like, then we need to test it with different inputs!

```
temp <- c(4, 6, 8, 9)
rescale01(temp)
```

```
[1] 0.0 0.4 0.8 1.0
```

```
temp0 <- c(4, 6, 8, 9, NA)
rescale01(temp0)
```

```
[1] NA NA NA NA NA
```

OK, so NA's don't work the way we want them to.

```
rescale01 <- function(x) {
  (x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
}
rescale01(temp)
```

```
[1] 0.0 0.4 0.8 1.0
```

```
rescale01(temp0)
```

```
[1] 0.0 0.4 0.8 1.0 NA
```

We can continue to improve our function. Here is another method, which uses the existing `range` function within R to avoid 3 max/min executions:

```
rescale01 <- function(x) {
  rng <- range(x, na.rm = TRUE)
  (x - rng[1]) / (rng[2] - rng[1])
}
rescale01(temp)
```

```
[1] 0.0 0.4 0.8 1.0
```

```
rescale01(c(0, 5, 10))
```

```
[1] 0.0 0.5 1.0
```

```
rescale01(c(-10, 0, 10))
```

```
[1] 0.0 0.5 1.0
```

```
rescale01(c(1, 2, 3, NA, 5))
```

```
[1] 0.00 0.25 0.50 NA 1.00
```

We should continue testing unusual inputs. Think carefully about how you want this function to behave... the current behavior is to include the Inf (infinity) value when calculating the range. You get strange output everywhere, but it's pretty clear that there is a problem right away when you use the function. In the example below (`rescale1`), you ignore the infinity value when calculating the range. The function returns Inf for one value, and sensible stuff for the rest. In many cases this may be useful, but it could also hide a problem until you get deeper into an analysis.

```
x <- c(1:10, Inf)
rescale01(x)
```

```
[1] 0 0 0 0 0 0 0 0 0 0 NaN
```

```
rescale1 <- function(x) {
  rng <- range(x, na.rm = TRUE, finite = TRUE)
  (x - rng[1]) / (rng[2] - rng[1])
}
rescale1(x)
```

```
[1] 0.0000000 0.1111111 0.2222222 0.3333333 0.4444444 0.5555556 0.6666667
[8] 0.7777778 0.8888889 1.0000000          Inf
```

Now we've used functions to simplify original example. We will learn to simplify further in iterations (Ch 26)

```
df <- tibble(
  a = rnorm(10),
  b = rnorm(10),
  c = rnorm(10),
  d = rnorm(10)
)
# add a little noise
df$a[5] = NA
df$b[6] = Inf
df
```

```
# A tibble: 10 x 4
      a         b         c         d
  <dbl>   <dbl> <dbl>   <dbl>
1  1.75    0.0475 -1.80    1.59
2  1.20   -2.35    0.364  0.657
3 -1.38    0.779  -0.492 -2.08
4 -0.609   1.28    0.412 -0.755
5 NA       1.07    1.19    1.52
6  0.0317 Inf     -2.45    0.781
7  0.0815  1.31    1.18    0.212
8 -0.874   0.0610  0.665 -0.582
9 -0.691   0.255  -0.639  1.31
10 -0.394  0.139  -2.26  -1.76
```

```
df$a_new <- rescale1(df$a)
df$b_new <- rescale1(df$b)
df$c_new <- rescale1(df$c)
df$d_new <- rescale1(df$d)
df
```

```
# A tibble: 10 x 8
      a         b         c         d a_new  b_new c_new d_new
  <dbl>   <dbl> <dbl>   <dbl> <dbl>   <dbl> <dbl> <dbl>
1  1.75    0.0475 -1.80    1.59    1     0.655 0.178    1
```

2	1.20	-2.35	0.364	0.657	0.824	0	0.772	0.746
3	-1.38	0.779	-0.492	-2.08	0	0.855	0.537	0
4	-0.609	1.28	0.412	-0.755	0.245	0.991	0.785	0.361
5	NA	1.07	1.19	1.52	NA	0.935	1	0.982
6	0.0317	Inf	-2.45	0.781	0.450	Inf	0	0.780
7	0.0815	1.31	1.18	0.212	0.466	1	0.996	0.625
8	-0.874	0.0610	0.665	-0.582	0.161	0.659	0.854	0.409
9	-0.691	0.255	-0.639	1.31	0.219	0.712	0.496	0.924
10	-0.394	0.139	-2.26	-1.76	0.314	0.680	0.0508	0.0882

```
df %>%
  mutate(a_new = rescale1(a),
         b_new = rescale1(b),
         c_new = rescale1(c),
         d_new = rescale1(d))
```

A tibble: 10 x 8

	a	b	c	d	a_new	b_new	c_new	d_new
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1.75	0.0475	-1.80	1.59	1	0.655	0.178	1
2	1.20	-2.35	0.364	0.657	0.824	0	0.772	0.746
3	-1.38	0.779	-0.492	-2.08	0	0.855	0.537	0
4	-0.609	1.28	0.412	-0.755	0.245	0.991	0.785	0.361
5	NA	1.07	1.19	1.52	NA	0.935	1	0.982
6	0.0317	Inf	-2.45	0.781	0.450	Inf	0	0.780
7	0.0815	1.31	1.18	0.212	0.466	1	0.996	0.625
8	-0.874	0.0610	0.665	-0.582	0.161	0.659	0.854	0.409
9	-0.691	0.255	-0.639	1.31	0.219	0.712	0.496	0.924
10	-0.394	0.139	-2.26	-1.76	0.314	0.680	0.0508	0.0882

Even better - from Chapter 26

```
df |> mutate(across(a:d, rescale1))
```

A tibble: 10 x 8

	a	b	c	d	a_new	b_new	c_new	d_new
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	0.655	0.178	1	1	0.655	0.178	1
2	0.824	0	0.772	0.746	0.824	0	0.772	0.746
3	0	0.855	0.537	0	0	0.855	0.537	0
4	0.245	0.991	0.785	0.361	0.245	0.991	0.785	0.361
5	NA	0.935	1	0.982	NA	0.935	1	0.982

6	0.450	Inf	0	0.780	0.450	Inf	0	0.780
7	0.466	1	0.996	0.625	0.466	1	0.996	0.625
8	0.161	0.659	0.854	0.409	0.161	0.659	0.854	0.409
9	0.219	0.712	0.496	0.924	0.219	0.712	0.496	0.924
10	0.314	0.680	0.0508	0.0882	0.314	0.680	0.0508	0.0882

Options for handling NAs in functions

Before we try some practice problems, let's consider various options for handling NAs in functions. We used the `na.rm` option within functions like `min`, `max`, and `range` in order to take care of missing values. But there are alternative approaches:

- filter/remove the NA values before rescaling
- create an if statement to check if there are NAs; return an error if NAs exist
- create a `removeNAs` option in the function we are creating

Let's take a look at each alternative approach in turn:

Filter/remove the NA values before rescaling

```
df <- tibble(
  a = rnorm(10),
  b = rnorm(10),
  c = rnorm(10),
  d = rnorm(10)
)
df$a[5] = NA
df
```

```
# A tibble: 10 x 4
      a      b      c      d
  <dbl> <dbl> <dbl> <dbl>
1 -0.0426 0.861 -0.0971 0.718
2 -0.519 0.586 0.192 -0.141
3 -3.08 -0.459 0.166 -1.87
4 -1.10 0.303 -0.364 -1.65
5 NA -0.389 1.75 -0.725
6 1.71 -0.190 -0.119 1.22
7 0.105 1.04 0.109 1.27
8 1.00 -1.28 0.163 0.240
9 1.15 0.246 -0.839 0.00947
10 0.165 -1.35 -0.567 -0.194
```



```
rescale_basic <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}
```

```
df %>%
  filter(!is.na(a)) %>%
  mutate(new_a = rescale_basic(a))
```

```
# A tibble: 9 x 5
      a      b      c      d new_a
  <dbl> <dbl> <dbl> <dbl> <dbl>
1 -0.0426 0.861 -0.0971 0.718 0.635
2 -0.519 0.586 0.192 -0.141 0.536
3 -3.08 -0.459 0.166 -1.87 0
4 -1.10 0.303 -0.364 -1.65 0.415
5 1.71 -0.190 -0.119 1.22 1
6 0.105 1.04 0.109 1.27 0.666
7 1.00 -1.28 0.163 0.240 0.853
8 1.15 0.246 -0.839 0.00947 0.884
9 0.165 -1.35 -0.567 -0.194 0.678
```

[Pause to Ponder:] Do you notice anything in the output above that gives you pause?

Questioning if it is necessary to remove the entire row because if we wanted to calculate a new_b in the same mutate, we wouldn't have those weighted values in row 5.

Create an if statement to check if there are NAs; return an error if NAs exist

First, here's an example involving weighted means:

```
# Create function to calculate weighted mean
wt_mean <- function(x, w) {
  sum(x * w) / sum(w)
}
wt_mean(c(1, 10), c(1/3, 2/3))
```

```
[1] 7
```

```
wt_mean(1:6, 1:3)
```

```
[1] 7.666667
```

[Pause to Ponder:] Why is the answer to the last call above 7.67? Aren't we taking a weighted mean of 1-6, all of which are below 7?

It repeats the list of 1, 2, and 3 so it matches the same size as 1:6, so in theory the w vector will repeat itself to be 1, 2, 3, 1, 2, 3. Thus when it calculates the wt_mean function it will calculate $1 + 4 + + 4 + 10 + 18 = 46$, then since that is the $\text{sum}(x * w)$ then it will divide it by the $\text{sum}(w)$ in our case is $1+2+3$, thus we have $46 / 6$ which is equal to 7.666667.

```
# update function to handle cases where data and weights of unequal length
wt_mean <- function(x, w) {
  if (length(x) != length(w)) {
    stop("`x` and `w` must be the same length", call. = FALSE)
  } else {
    sum(w * x) / sum(w)
  }
}
wt_mean(1:6, 1:3)
```

Error: `x` and `w` must be the same length

```
wt_mean(1:6, 1:6)
```

```
[1] 4.333333
```

```
# should produce an error now if weights and data different lengths
# - nice example of if and else
```

[Pause to Ponder:] What does the call. option do?

For `call. = TRUE` when an error happens it will say “Error in `wt_mean(1:6, 1:3)`”, then give out your error message. For `call. = FALSE` when an error happens it will say “Error:” and give you your error message. Basically indicates if the function you’re trying to run will be in the error message in our case this is `wt_mean(1:6, 1:3)`

Now let's apply this to our rescaling function

```
rescale_w_error <- function(x) {  
  if (is.na(sum(x))) {  
    stop("`x` cannot have NAs", call. = FALSE)  
  } else {  
    (x - min(x)) / (max(x) - min(x))  
  }  
}  
  
temp <- c(4, 6, 8, 9)  
rescale_w_error(temp)
```

```
[1] 0.0 0.4 0.8 1.0
```

```
temp <- c(4, 6, 8, 9, NA)  
rescale_w_error(temp)
```

Error: `x` cannot have NAs

[Pause to Ponder:] Why can't we just use `if (is.na(x))` instead of `is.na(sum(x))`?

It returns a vector instead of a scalar. The `If` needs a single value that can be either true or false.

Create a `removeNAs` option in the function we are creating

```
rescale_NAoption <- function(x, removeNAs = FALSE) {  
  (x - min(x, na.rm = removeNAs)) /  
    (max(x, na.rm = removeNAs) - min(x, na.rm = removeNAs))  
}  
  
temp <- c(4, 6, 8, 9)  
rescale_NAoption(temp)
```

```
[1] 0.0 0.4 0.8 1.0
```

```
temp <- c(4, 6, 8, 9, NA)  
rescale_NAoption(temp, removeNAs = TRUE)
```

```
[1] 0.0 0.4 0.8 1.0 NA
```

OK, but all the other summary stats functions use `na.rm` as the input, so to be consistent, it's probably better to do something slightly awkward like this:

```
rescale_NAoption <- function(x, na.rm = FALSE) {  
  (x - min(x, na.rm = na.rm)) /  
    (max(x, na.rm = na.rm) - min(x, na.rm = na.rm))  
}  
  
temp <- c(4, 6, 8, 9, NA)  
rescale_NAoption(temp, na.rm = TRUE)
```

```
[1] 0.0 0.4 0.8 1.0 NA
```

`wt_mean()` is an example of a “summary function (single value output) instead of a”mutate function” (vector output) like `rescale01()`. Here’s another summary function to produce the mean absolute percentage error:

```
mape <- function(actual, predicted) {  
  sum(abs((actual - predicted) / actual)) / length(actual)  
}  
  
y <- c(2,6,3,8,5)  
yhat <- c(2.5, 5.1, 4.4, 7.8, 6.1)  
mape(actual = y, predicted = yhat)
```

```
[1] 0.2223333
```

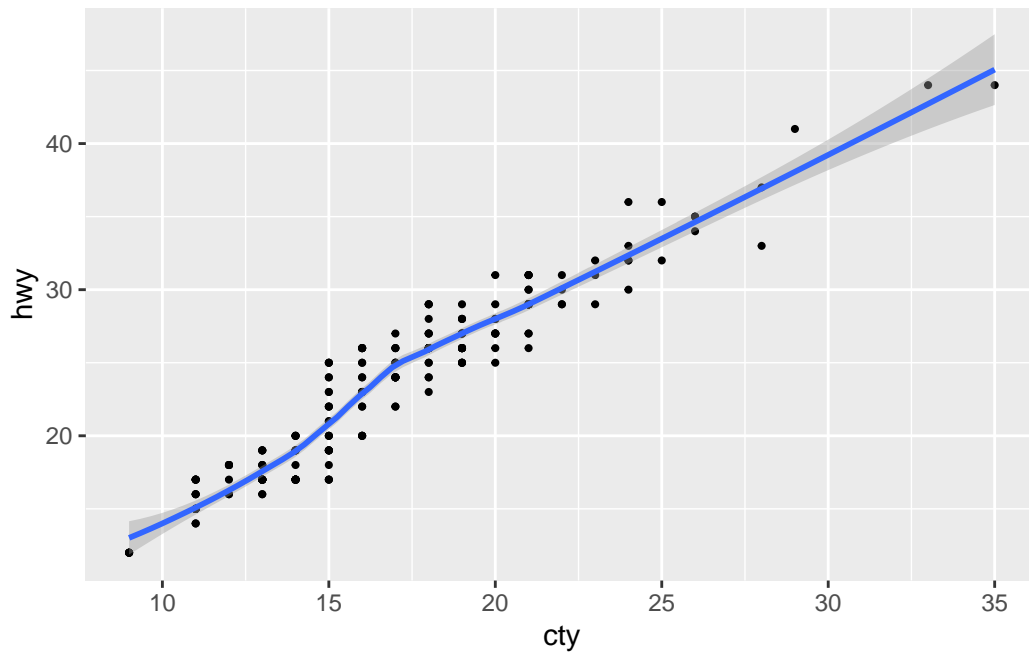
Data frame functions

These work like dplyr verbs, taking a data frame as the first argument, and then returning a data frame or a vector.

Demonstration of tidy evaluation in functions

```
# Start with working code then functionize
ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +
  geom_point(size = 0.75) +
  geom_smooth()
```

``geom_smooth()`` using method = 'loess' and formula = 'y ~ x'



```
make_plot <- function(dataset, xvar, yvar, pt_size = 0.75) {
  ggplot(data = dataset, mapping = aes(x = xvar, y = yvar)) +
    geom_point(size = pt_size) +
    geom_smooth()
}

make_plot(dataset = mpg, xvar = cty, yvar = hwy) # Error!
```

Error in ``geom_point()``:
! Problem while computing aesthetics.
i Error occurred in the 1st layer.

Caused by error:
! object 'cty' not found

The problem is tidy evaluation, which makes most common coding easier, but makes some less common things harder. Key terms to understand tidy evaluation:

- env-variables = live in the environment (mpg)
- data-variables = live in data frame or tibble (cty)
- data masking = tidyverse use data-variables as if they are env-variables. That is, you don't always need `mpg$cty` to access `cty` in tidyverse

The key idea behind data masking is that it blurs the line between the two different meanings of the word “variable”:

- env-variables are “programming” variables that live in an environment. They are usually created with `<-`.
- data-variables are “statistical” variables that live in a data frame. They usually come from data files (e.g. .csv, .xls), or are created manipulating existing variables.

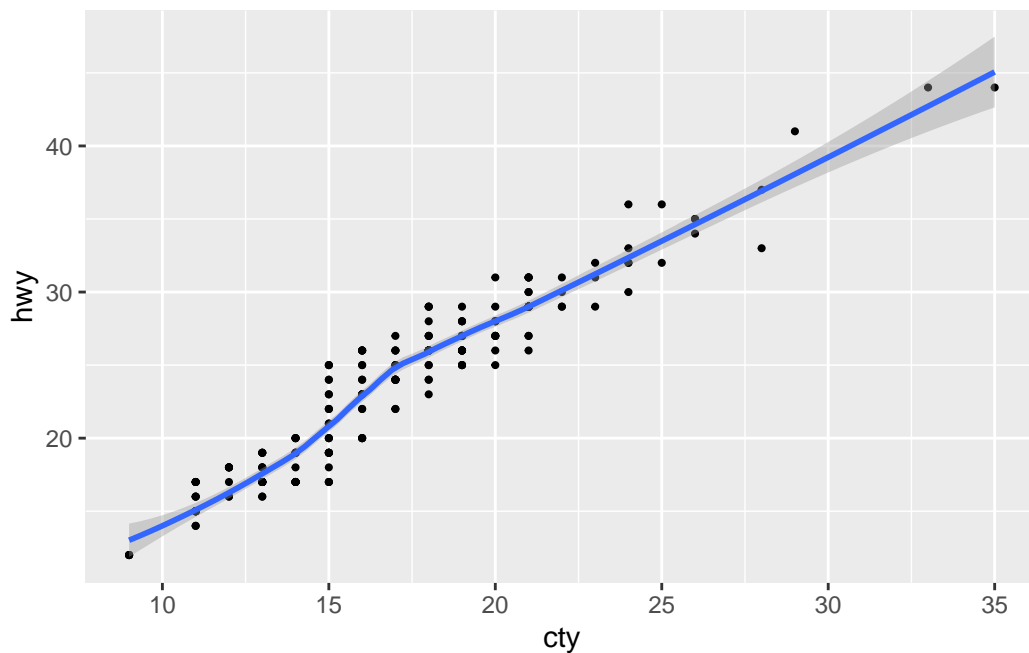
The solution is to embrace `{{ }}` data-variables which are user inputs into functions. One way to remember what's happening, as suggested by our book authors, is to think of `{{ }}` as looking down a tunnel — `{{ var }}` will make a dplyr function look inside of `var` rather than looking for a variable called `var`. Thus, embracing a variable tells dplyr to use the value stored inside the argument, not the argument as the literal variable name.

See Section 25.3 of R4DS for more details (and there are plenty!).

```
# This will work to make our plot!
make_plot <- function(dataset, xvar, yvar, pt_size = 0.75) {
  ggplot(data = dataset, mapping = aes(x = {{ xvar }}, y = {{ yvar }})) +
    geom_point(size = pt_size) +
    geom_smooth()
}

make_plot(dataset = mpg, xvar = cty, yvar = hwy)
```

``geom_smooth()`` using method = 'loess' and formula = 'y ~ x'



I often wish it were easier to get my own custom summary statistics for numeric variables in EDA rather than using `mosaic:favstats()`. Using `group_by()` and `summarise()` from the tidyverse reads clearly but takes so many lines, but if I only had to write the code once...

```
summary6 <- function(data, var) {
  data |> summarize(
    min = min({{ var }}, na.rm = TRUE),
    mean = mean({{ var }}, na.rm = TRUE),
    median = median({{ var }}, na.rm = TRUE),
    max = max({{ var }}, na.rm = TRUE),
    n = n(),
    n_miss = sum(is.na({{ var }})),
    .groups = "drop"    # to leave the data in an ungrouped state
  )
}
```

```
mpg |> summary6(hwy)
```

```
# A tibble: 1 x 6
   min mean median  max    n n_miss
<int> <dbl> <dbl> <int> <int> <int>
1    12  23.4    24    44   234     0
```

Even cooler, I can use my new function with `group_by()`!

```
mpg |>
  group_by(drv) |>
  summary6(hwy)
```

```
# A tibble: 3 x 7
  drv      min mean median   max     n n_miss
<chr> <int> <dbl> <dbl> <int> <int> <int>
1 4         12 19.2     18    28   103     0
2 f         17 28.2     28    44   106     0
3 r         15 21      21    26    25     0
```

You can even pass conditions into a function using the embrace:

[Pause to Ponder:] Predict what the code below will do, and (only) then run it to check. Think about: why do we have `sort = sort`? why not embrace `df`? why didn't we need `n` in the arguments?

It will create a new function, that filters by the specified condition, counts the variable and then finds the proportion for each variable. `sort = sort`, sort on the left hand is the option in count, and the sort on the right hand side is our name for it. So in our function it'll output it in numerical order.

We didn't need `n` in the argument because it is being created in our count.

```
new_function <- function(df, var, condition, sort = TRUE) {
  df |>
    filter({{ condition }}) |>
    count({{ var }}, sort = sort) |>
    mutate(prop = n / sum(n))
}

mpg |> new_function(var = manufacturer,
                   condition = manufacturer %in% c("audi", "honda", "hyundai", "nissan", "s
```

Data-masking vs. tidy-selection (Section 25.3.4)

Why doesn't the following code work?


```
count_missing <- function(df, group_vars, x_var) {
  df |>
    group_by({{ group_vars }}) |>
    summarize(
      n_miss = sum(is.na({{ x_var }})),
      .groups = "drop"
    )
}

flights |>
  count_missing(c(year, month, day), dep_time)
```

```
Error in `group_by()` :
i In argument: `c(year, month, day)`.
Caused by error:
! `c(year, month, day)` must be size 336776 or 1, not 1010328.
```

The problem is that `group_by()` uses data-masking rather than tidy-selection; it is selecting certain variables rather than evaluating values of those variables. These are the two most common subtypes of tidy evaluation:

- Data-masking is used in functions like `arrange()`, `filter()`, `mutate()`, and `summarize()` that compute with variables. Data masking is an R feature that blends programming variables that live inside environments (env-variables) with statistical variables stored in data frames (data-variables).
- Tidy-selection is used for functions like `select()`, `relocate()`, and `rename()` that select variables. Tidy selection provides a concise dialect of R for selecting variables based on their names or properties.

More detail can be found [here](#).

The error above can be solved by using the `pick()` function, which uses tidy selection inside of data masking:

```
count_missing <- function(df, group_vars, x_var) {
  df |>
    group_by(pick({{ group_vars }})) |>
    summarize(
      n_miss = sum(is.na({{ x_var }})),
      .groups = "drop"
    )
}
```

```
}

flights |>
  count_missing(c(year, month, day), dep_time)
```

```
# A tibble: 365 x 4
   year month   day n_miss
  <int> <int> <int> <int>
1  2013     1     1     4
2  2013     1     2     8
3  2013     1     3    10
4  2013     1     4     6
5  2013     1     5     3
6  2013     1     6     1
7  2013     1     7     3
8  2013     1     8     4
9  2013     1     9     5
10 2013     1    10     3
# i 355 more rows
```

[Pause to Ponder:] Here's another nice use of `pick()`. Predict what the function will do, then run the code to see if you are correct.

It will choose those the rows and columns we want to count, then take the values of column in our pivot wider is a differetn way of organizing counts, we will have manufacturer, model, and then columns with the number of cylinders

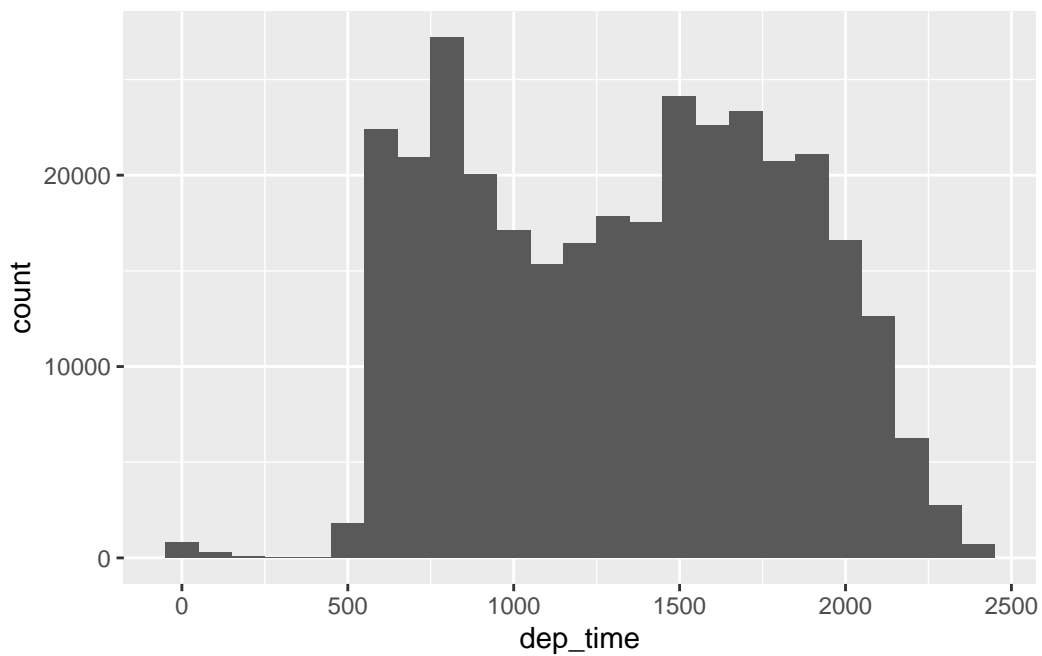
```
# Source: https://twitter.com/pollicipes/status/1571606508944719876
new_function <- function(data, rows, cols) {
  data |>
    count(pick(c({{ rows }}, {{ cols }}})) |>
    pivot_wider(
      names_from = {{ cols }},
      values_from = n,
      names_sort = TRUE,
      values_fill = 0
    )
}

mpg |> new_function(c(manufacturer, model), cyl)
```

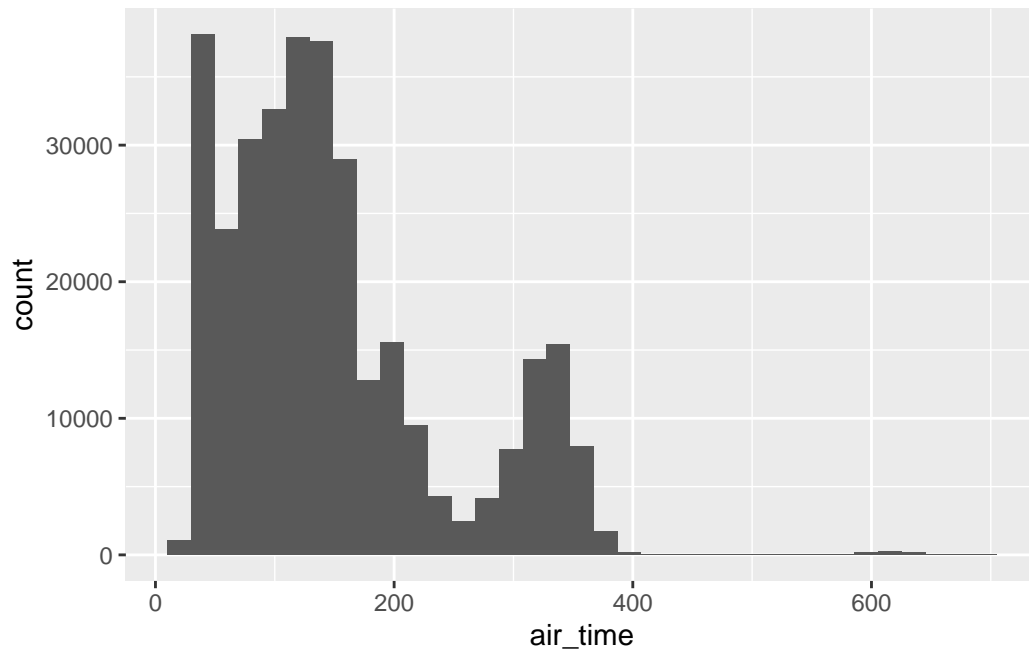
Plot functions

Let's say you find yourself making a lot of histograms:

```
flights |>  
  ggplot(aes(x = dep_time)) +  
  geom_histogram(bins = 25)
```

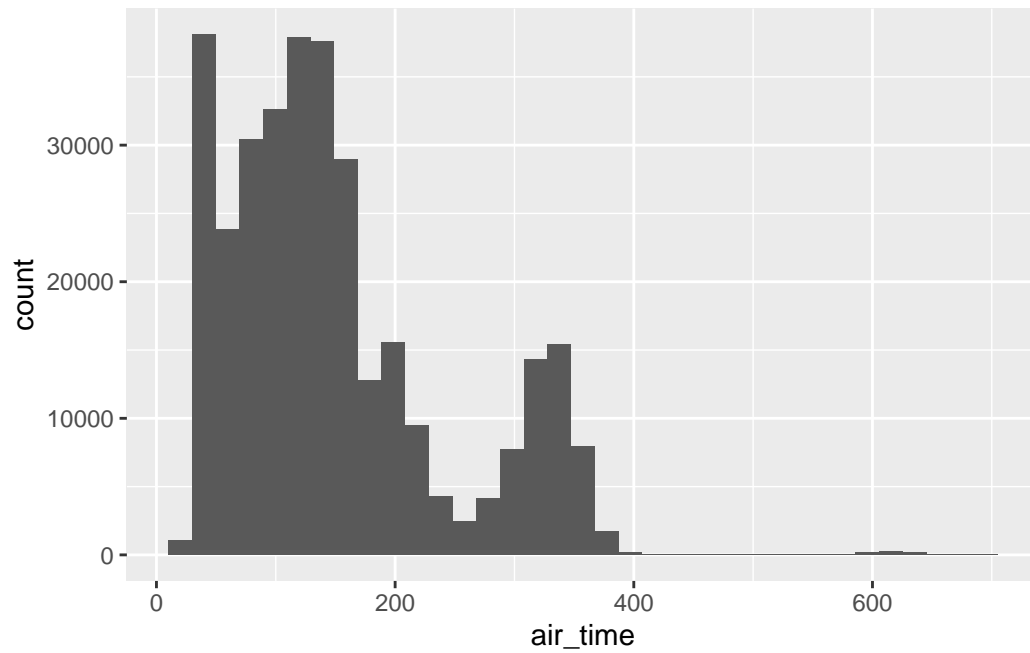


```
flights |>  
  ggplot(aes(x = air_time)) +  
  geom_histogram(bins = 35)
```



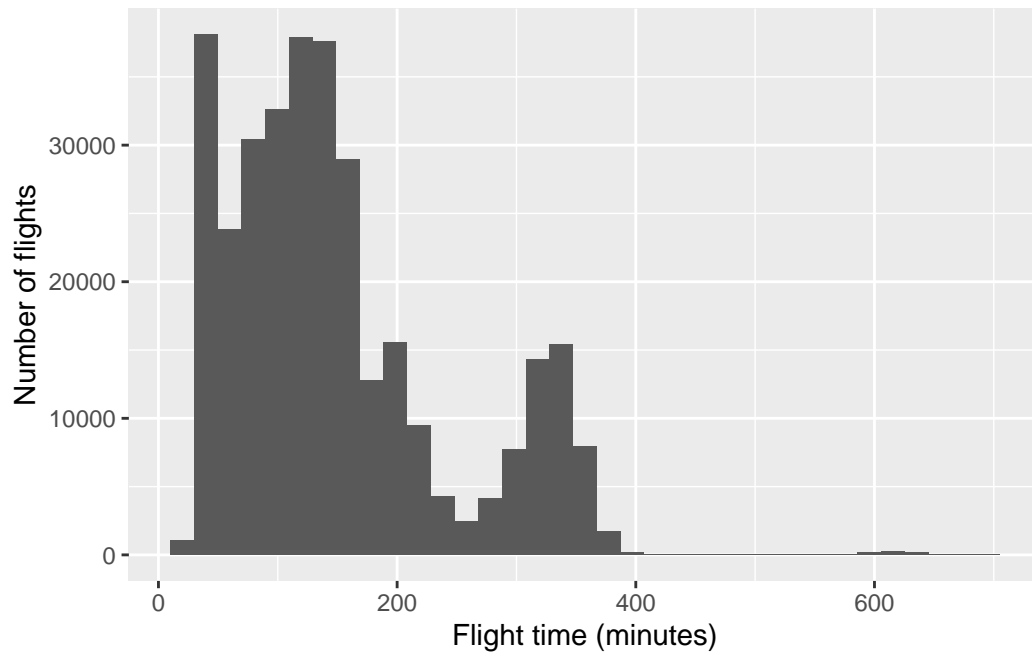
Just use embrace to create a histogram-making function

```
histogram <- function(df, var, bins = NULL) {  
  df |>  
    ggplot(aes(x = {{ var }})) +  
    geom_histogram(bins = bins)  
}  
  
flights |> histogram(air_time, 35)
```



Since `histogram()` returns a `ggplot`, you can add any layers you want

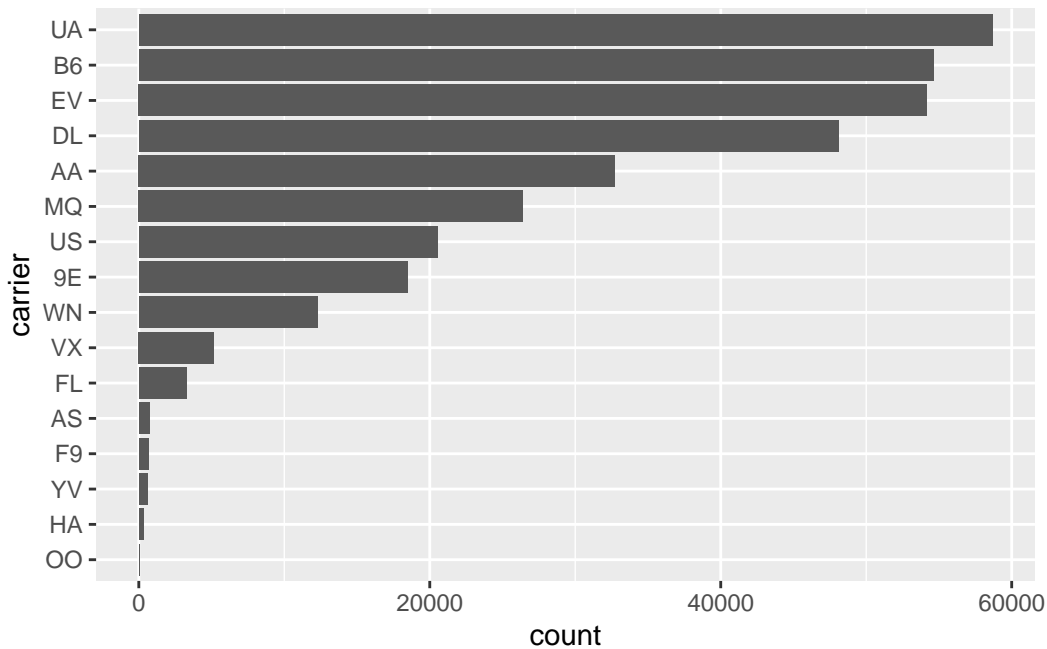
```
flights |>
  histogram(air_time, 35) +
  labs(x = "Flight time (minutes)", y = "Number of flights")
```



You can also combine data wrangling with plotting. Note that we need the “walrus operator” (`:=`) since the variable name on the left is being generated with user-supplied data.

```
# sort counts with highest values at top and counts on x-axis
sorted_bars <- function(df, var) {
  df |>
    mutate({{ var }} := fct_rev(fct_infreq({{ var }}})) |>
    ggplot(aes(y = {{ var }})) +
    geom_bar()
}

flights |> sorted_bars(carrier)
```



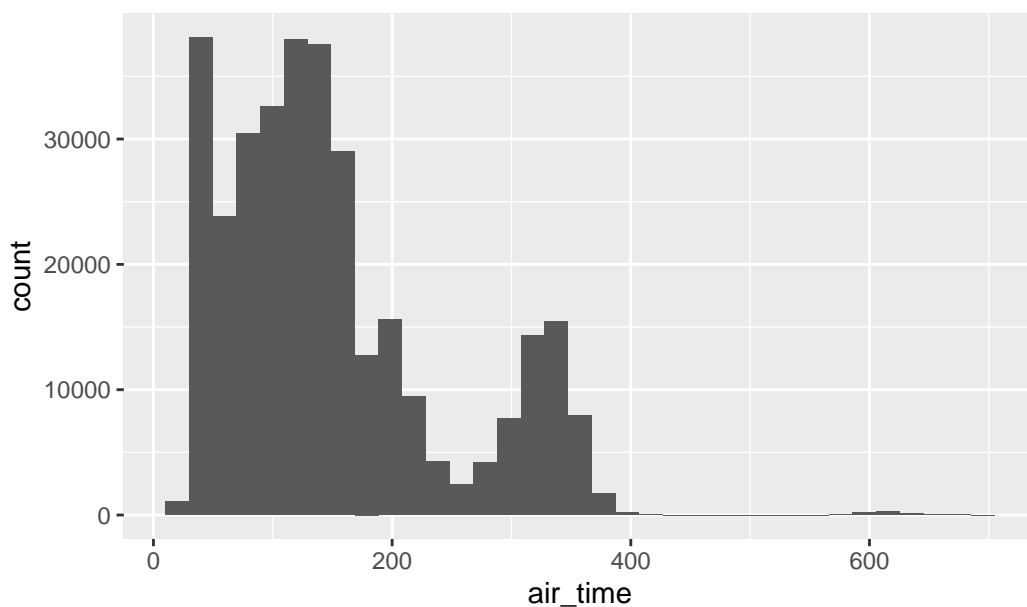
Finally, it would be really helpful to label plots based on user inputs. This is a bit more complicated, but still do-able!

For this, we'll need the `rlang` package. `rlang` is a low-level package that's used by just about every other package in the tidyverse because it implements tidy evaluation (as well as many other useful tools).

Check out the following update of our `histogram()` function which uses the `englue()` function from the `rlang` package:

```
histogram <- function(df, var, bins) {  
  label <- rlang::englue("A histogram of {{var}} with binwidth {bins}")  
  
  df |>  
    ggplot(aes(x = {{ var }})) +  
    geom_histogram(bins = bins) +  
    labs(title = label)  
}  
  
flights |> histogram(air_time, 35)
```

A histogram of air_time with binwidth 35



On Your Own

1. Rewrite this code snippet as a function: `x / sum(x, na.rm = TRUE)`. This code creates weights which sum to 1, where NA values are ignored. Test it for at least two different vectors. (Make sure at least one has NAs!)

```
weight_sum <- function(x) {
  x / sum(x, na.rm = TRUE)
}

temp1 <- c(2, 4, 6, 8, 10)
temp2 <- c(2, 4, 6, 8, 10, NA)

weight_sum(temp2)
```

```
[1] 0.06666667 0.13333333 0.20000000 0.26666667 0.33333333      NA
```

2. Create a function to calculate the standard error of a variable, where $SE = \sqrt{\text{variance} / \text{sample size}}$. Hint: start with a vector like `x <- 0:5` or `x <- gss_cat$age` and write code to find the SE of `x`, then turn it into a function to handle any vector `x`. Note: `var` is the function to find variance in R and `sqrt` does

square root. `length` may also be handy. Test your function on two vectors that do not include NAs (i.e. do **not** worry about removing NAs at this point).

```
standard_e <- function(x){  
  length = length(x)  
  sd = sd(x)  
  sd/sqrt(length)  
}  
  
standard_e(x <- 0:5)
```

```
[1] 0.7637626
```

3. Use your `se` function within `summarize` to get a table of the mean and s.e. of `hwy` and `cty` by class in the `mpg` dataset.

```
mpg |>  
  group_by(class) |>  
  summarize(mean_cty = mean(cty),  
            se_cty = standard_e(cty),  
            mean_hwy = mean(hwy),  
            se_hwy = standard_e(hwy))
```

```
# A tibble: 7 x 5  
  class      mean_cty se_cty mean_hwy se_hwy  
  <chr>      <dbl>  <dbl>   <dbl> <dbl>  
1 2seater      15.4  0.245    24.8  0.583  
2 compact      20.1  0.494    28.3  0.552  
3 midsize      18.8  0.304    27.3  0.334  
4 minivan      15.8  0.553    22.4  0.622  
5 pickup       13    0.356    16.9  0.396  
6 subcompact   20.4  0.778    28.1  0.909  
7 suv          13.5  0.307    18.1  0.378
```

4. Use your `se` function within `summarize` to get a table of the mean and s.e. of `arr_delay` and `dep_delay` by carrier in the `flights` dataset. Why does the output look like this?

```
flights |>  
  group_by(carrier) |>  
  summarize(mean_arr_delay = mean(arr_delay),  
            se_arr_delay = standard_e(arr_delay),  
            mean_dep_delay = mean(dep_delay),  
            se_dep_delay = standard_e(dep_delay))
```

```
# A tibble: 16 x 5
  carrier mean_arr_delay se_arr_delay mean_dep_delay se_dep_delay
  <chr>      <dbl>      <dbl>      <dbl>      <dbl>
1 9E         NA         NA         NA         NA
2 AA         NA         NA         NA         NA
3 AS         NA         NA         NA         NA
4 B6         NA         NA         NA         NA
5 DL         NA         NA         NA         NA
6 EV         NA         NA         NA         NA
7 F9         NA         NA         NA         NA
8 FL         NA         NA         NA         NA
9 HA        -6.92         4.06         4.90         4.01
10 MQ         NA         NA         NA         NA
11 OO         NA         NA         NA         NA
12 UA         NA         NA         NA         NA
13 US         NA         NA         NA         NA
14 VX         NA         NA         NA         NA
15 WN         NA         NA         NA         NA
16 YV         NA         NA         NA         NA
```

It looks like this because we have NA's for arr_delay and dep_delay and our function isn't built/designed to handle NA's.

5. Make your `se` function handle NAs with an `na.rm` option. Test your new function (you can call it `se` again) on a vector that doesn't include NA and on the same vector with an added NA. **Be sure to check that it gives the expected output with `na.rm = TRUE` and `na.rm = FALSE`.** Make `na.rm = FALSE` the default value. Repeat #4. (Hint: be sure when you divide by sample size you don't count any NAs)

```
standard_e <- function(x, na.rm = TRUE){
  length = length(x) - sum(is.na(x))
  sd = sd(x, na.rm = TRUE)
  sd/sqrt(length)
}
```

```
test <- c(0:5, NA)
```

```
standard_e(test, na.rm = TRUE)
```

```
[1] 0.7637626
```

```

flights |>
  drop_na(arr_delay, dep_delay) |>
  group_by(carrier) |>
  summarize(mean_arr_delay = mean(arr_delay, na.rm = TRUE),
            se_arr_delay = standard_e(arr_delay, na.rm = TRUE),
            mean_dep_delay = mean(dep_delay, na.rm = TRUE),
            se_dep_delay = standard_e(dep_delay, na.rm = TRUE))

```

```

# A tibble: 16 x 5
  carrier mean_arr_delay se_arr_delay mean_dep_delay se_dep_delay
  <chr>         <dbl>         <dbl>         <dbl>         <dbl>
1 9E             7.38           0.381          16.4           0.346
2 AA             0.364           0.238           8.57           0.209
3 AS            -9.93           1.37            5.83           1.18
4 B6             9.46           0.184          13.0           0.165
5 DL             1.64           0.203           9.22           0.182
6 EV            15.8           0.221          19.8           0.205
7 F9            21.9           2.36           20.2           2.24
8 FL            20.1           0.960          18.6           0.932
9 HA            -6.92           4.06            4.90           4.01
10 MQ            10.8           0.273          10.4           0.247
11 OO            11.9           9.02           12.6           8.00
12 UA             3.56           0.170          12.0           0.148
13 US             2.13           0.235           3.74           0.198
14 VX             1.76           0.699          12.8           0.615
15 WN             9.65           0.427          17.7           0.394
16 YV            15.6           2.27           18.9           2.11

```

6. Create `both_na()`, a function that takes two vectors of the same length and returns how many positions have an NA in both vectors. Hint: create two vectors like `test_x <- c(1, 2, 3, NA, NA)` and `test_y <- c(NA, 1, 2, 3, NA)` and write code that works for `test_x` and `test_y`, then turn it into a function that can handle any `x` and `y`. (In this case, the answer would be 1, since both vectors have NA in the 5th position.) Test it for at least one more combination of `x` and `y`.

```

test_x <- c(1, 2, 3, NA, NA)
test_y <- c(NA, 1, 2, 3, NA)

both_na <- function(x, y){
  sum(is.na(x) & is.na(y))
}

```

```
both_na(test_x, test_y)
```

```
[1] 1
```

7. Run your code from (6) with the following two vectors: `test_x <- c(1, 2, 3, NA, NA, NA)` and `test_y <- c(NA, 1, 2, 3, NA)`. Did you get the output you wanted or expected? Modify your function using `if`, `else`, and `stop` to print an error if `x` and `y` are not the same length. Then test again with `test_x`, `test_y` and the sets of vectors you used in (6).

```
test_x2 <- c(1, 2, 3, NA, NA, NA)
test_y2 <- c(NA, 1, 2, 3, NA)

bothna2 <- function(x,y){
  if (length(x) != length(y)) {
    stop("`x` and `y` ARE NOT THE SAME LENGTH, the same pls", call. = FALSE)
  } else {
    sum(is.na(x) & is.na(y))
  }
}

bothna2(test_x2, test_y2)
```

Error: `x` and `y` ARE NOT THE SAME LENGTH, the same pls

```
bothna2(test_x, test_y)
```

```
[1] 1
```

8. Here is a way to get `not_cancelled` flights in the `flights` dataset:

```
not_cancelled <- flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))
```

Is it necessary to check `is.na` for both departure and arrival? Using `summarize`, find the number of flights missing departure delay, arrival delay, and both. (Use your new function!)

```
flights |>
  summarize(dep_na = sum(is.na(dep_delay)),
            arr_na = sum(is.na(arr_delay)),
            both_na = bothna2(dep_delay, arr_delay))
```

```
# A tibble: 1 x 3
  dep_na arr_na both_na
  <int>  <int>   <int>
1   8255   9430    8255
```

Is it not necessary to check is.na forth both departure and arrival, we could only check for departure assuming that they're the same flights. However, there are more NA values for arr_na than dep_na.

9. Read the code for each of the following three functions, puzzle out what they do, and then brainstorm better names.

```
duration_mins <- function(time1, time2) {
  hour1 <- time1 %/% 100
  min1 <- time1 %% 100
  hour2 <- time2 %/% 100
  min2 <- time2 %% 100

  (hour2 - hour1)*60 + (min2 - min1)
}

area_inches <- function(lengthcm, widthcm) {
  (lengthcm / 2.54) * (widthcm / 2.54)
}

non_answer <- function(x) {
  fct_collapse(x, "non answer" = c("No answer", "Refused",
                                   "Don't know", "Not applicable"))
}
```

f1, function to calculate duration of event (like flights in our flights dataset), will output out a time in minutes. This function takes in two inputs being time EXPLAIN LATER REMEMBER

f2, function that calculates area of a square or rectangle/something with 2 sides with inputs being in centimeters, and output being inches

f3 function for looking at a vector/column and if the answers are “No answer”, “Refused”, “Don’t know”, “Not applicable” then it becomes non answer.

10. Explain what the following function does and demonstrate by running `foo1(x)` with a few appropriately chosen vectors `x`. (Hint: set `x` and run the “guts” of the function piece by piece.)

```
foo1 <- function(x) {  
  diff <- x[-1] - x[1:(length(x) - 1)]  
  sum(diff < 0)  
}
```

The function takes an input vector `x` and then slices the first number out so we have the every number after the first number then subtracts it from everything but the last number. For ex. the 2nd number minus the 1st and then the 3rd minus the 2nd and 4th minus the 3rd etc. until there are no more numbers in the list. It will then assign that to `diff`, if the sum of `diff` is less than zero then it will print out a 0 and if none of those differences are less than zero it will spit out a zero, if it is less than zero then it will give the number of digits in the vector that are less than zero.

11. The `foo1()` function doesn't perform well if a vector has missing values. Amend `foo1()` so that it produces a helpful error message and stops if there are any missing values in the input vector. Show that it works with appropriately chosen vectors `x`. Be sure you add `error = TRUE` to your R chunk, or else knitting will fail!

```
foo1 <- function(x) {  
  if(any(is.na(x))) {  
    stop ("Input Vector Contains NA Values", call. = FALSE)  
  }  
  diff <- x[-1] - x[1:(length(x) - 1)]  
  sum(diff < 0)  
}  
  
temp0
```

```
[1] 4 6 8 9 NA
```

```
foo1(temp0)
```

Error: Input Vector Contains NA Values

12. Write a function called `greet` using `if`, `else if`, and `else` to print out “good morning” if it’s before 12 PM, “good afternoon” if it’s between 12 PM and 5 PM, and “good evening” if it’s after 5 PM. Your function should work if you input a time like: `greet(time = "2018-05-03 17:38:01 CDT")` or if you input the current time with `greet(time = Sys.time())`. [Hint: check out the `hour` function in the `lubridate` package]

```
hour("2018-05-03 23:38:01 CDT")
```

```
[1] 23
```

```
greet <- function(time){
  hour = hour(time)
  if (hour < 12) {
    print("good morning")
  } else if (hour >= 12 & hour < 17){
    print("good afternoon")
  } else {
    print("good evening")
  }
}

greet(time = Sys.time())
```

```
[1] "good morning"
```

13. Modify the `summary6()` function from earlier to add an argument that gives the user an option to remove missing values, if any exist. Show that your function works for (a) the `hwy` variable in `mpg_tbl <- as_tibble(mpg)`, and (b) the `age` variable in `gss_cat`.

```
summary6 <- function(data, var, na.rm = TRUE) {
  data |> summarize(
    min = min({{ var }}, na.rm = na.rm),
    mean = mean({{ var }}, na.rm = na.rm),
    median = median({{ var }}, na.rm = na.rm),
    max = max({{ var }}, na.rm = na.rm),
    n = n(),
    n_miss = sum(is.na({{ var }})),
    .groups = "drop"      # to leave the data in an ungrouped state
  )
}

mpg |>
  summary6(hwy, na.rm = FALSE)
```

```
# A tibble: 1 x 6
  min mean median max n n_miss
<int> <dbl> <dbl> <int> <int> <int>
1    12  23.4    24    44   234     0
```

```
gss_cat |>
  summary6(age, na.rm = FALSE)
```

```
# A tibble: 1 x 6
  min mean median max n n_miss
<int> <dbl> <int> <int> <int> <int>
1    NA    NA    NA    NA 21483    76
```

```
gss_cat |>
  summary6(age, na.rm = TRUE)
```

```
# A tibble: 1 x 6
  min mean median max n n_miss
<int> <dbl> <int> <int> <int> <int>
1    18  47.2    46    89 21483    76
```

14. Add an argument to (13) to produce summary statistics by group for a second variable (you should now have 4 possible inputs to your function). Show that your function works for (a) the `hwy` variable in `mpg_tbl <- as_tibble(mpg)` grouped by `drv`, and (b) the `age` variable in `gss_cat` grouped by `partyid`.

```
summary6 <- function(data, group_var, var, na.rm = TRUE) {
  data |>
    group_by(pick({{ group_var }})) |>
    summarize(
      min = min({{ var }}, na.rm = na.rm),
      mean = mean({{ var }}, na.rm = na.rm),
      median = median({{ var }}, na.rm = na.rm),
      max = max({{ var }}, na.rm = na.rm),
      n = n(),
      n_miss = sum(is.na({{ var }})),
      sd = sd({{ var }}, na.rm = na.rm),
      iqr = iqr({{ var }}, na.rm = na.rm),
      .groups = "drop" # to leave the data in an ungrouped state
    )
}
```



```
summary6(mpg, drv, hwy)
```

```
# A tibble: 3 x 9
  drv      min mean median   max     n n_miss   sd   iqr
<chr> <int> <dbl>  <dbl> <int> <int>  <int> <dbl> <dbl>
1 4         12 19.2    18    28   103     0  4.08  4.08
2 f         17 28.2    28    44   106     0  4.21  4.21
3 r         15 21     21    26    25     0  3.66  3.66
```

```
new_party_id <- summary6(gss_cat, partyid, age)
```

15. Create a function that has a vector as the input and returns the last value. (Note: Be sure to use a name that does not write over an existing function!)

```
last_value <- function(x){
  last_num <- length(x)
  x[last_num]
}
```

```
temp
```

```
[1] 4 6 8 9 NA
```

```
last_value(temp)
```

```
[1] NA
```

```
temp1
```

```
[1] 2 4 6 8 10
```

```
last_value(temp1)
```

```
[1] 10
```

16. Save your final table from (14) and write a function to draw a scatterplot of a measure of center (mean or median - user can choose) vs. a measure of spread (sd or IQR - user can choose), with points sized by sample size, to see if there is constant variance. Each point should be labeled with partyid, and the plot title should reflect the variables chosen by the user.

```
library(ggplot2)
# Testing out what I need for my function
# new_party_id |>
# ggplot(aes(x = mean, y = sd)) +
#   geom_point(aes(size = n)) +
#   geom_smooth(method = lm, se = FALSE) +
#   geom_label_repel(aes(label = partyid)) +
#   labs(title = "title")

plot_summary6 <- function(tbl, center, spread) {
  label <- rlang::englu("A scatterplot of {{center}} vs {{spread}} with size being sample size")
  tbl |>
    ggplot(aes(x = {{center}}, y = {{spread}} )) +
    geom_point(aes(size = n)) +
    geom_smooth(method = lm, se = FALSE) +
    geom_label_repel(aes(label = partyid)) +
    labs(title = label)
}

plot_summary6(new_party_id, median, iqr)
```

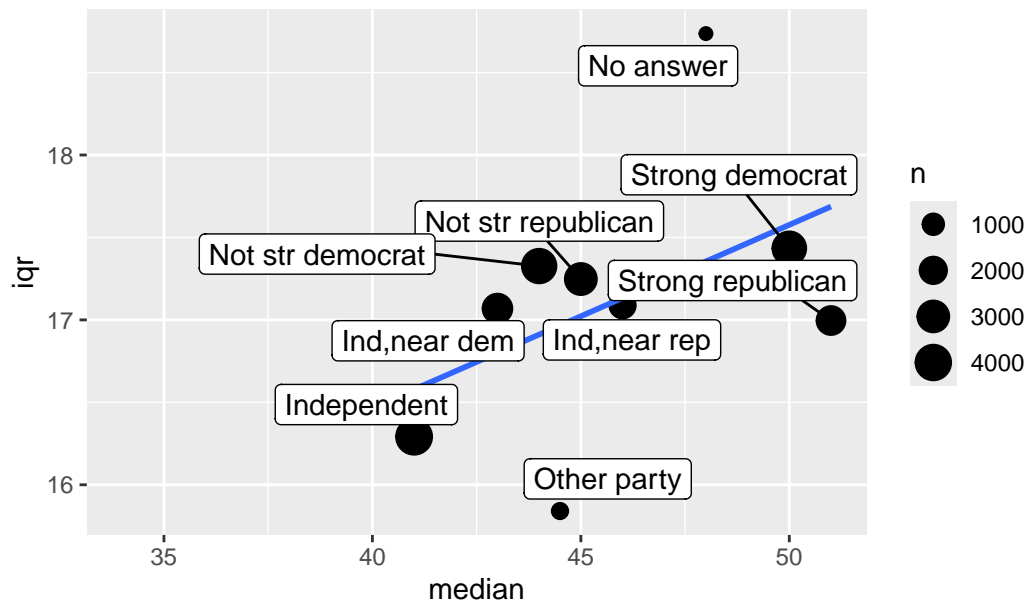
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_label_repel()`).

A scatterplot of median vs iqr with size being sample size



```
plot_summary6(new_party_id, mean, sd)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_label_repel()`).
```

A scatterplot of mean vs sd with size being sample size

