# L02 ggplot I

Data Visualization (STAT 302)

Shay Lebovitz

## Contents

### Overview

The goal of this lab is to begin the process of unlocking the power of `ggplot2` through constructing and experimenting with a few basic plots.

### Datasets

We'll be using data from the `blue_jays.rda` dataset which is already in the `/data` subdirectory in our **data_vis_labs** project. Below is a description of the variables contained in the dataset.

- `BirdID` - ID tag for bird
- `KnownSex` - Sex coded as F or M
- `BillDepth` - Thickness of the bill measured at the nostril (in mm)
- `BillWidth` - Width of the bill (in mm)
- `BillLength` - Length of the bill (in mm)
- `Head` - Distance from tip of bill to back of head (in mm)
- `Mass` - Body mass (in grams)
- `Skull` - Distance from base of bill to back of skull (in mm)
- `Sex` - Sex coded as `0 = female` or `1 = male`

We'll also be using a subset of the BRFSS (Behavioral Risk Factor Surveillance System) survey collected annually by the Centers for Disease Control and Prevention (CDC). The data can be found in the provided `cdc.txt` file — place this file in your `/data` subdirectory. The dataset contains 20,000 complete observations/records of 9 variables/fields, described below.

- `genhlth` - How would you rate your general health? (excellent, very good, good, fair, poor)
- `exerany` - Have you exercised in the past month? (`1 = yes`, `0 = no`)
- `hlthplan` - Do you have some form of health coverage? (`1 = yes`, `0 = no`)
- `smoke100` - Have you smoked at least 100 cigarettes in your life time? (`1 = yes`, `0 = no`)
- `height` - height in inches
- `weight` - weight in pounds
- `wtdesire` - weight desired in pounds
- `age` - in years
- `gender` - `m` for males and `f` for females

Notice we are setting a seed. This signifies we will be doing something that relies on a random process (e.g., random sampling). In order for our results to be reproducible we set the seed. This ensures that every time you run the code or someone else does, it will produce the exact same output. It is good coding etiquette to set the seed towards the top of your document/code.

```
# Set the seed for reproducibility
set.seed(31412718)

# Load package(s)
```

## Exercises

Complete the following exercises.

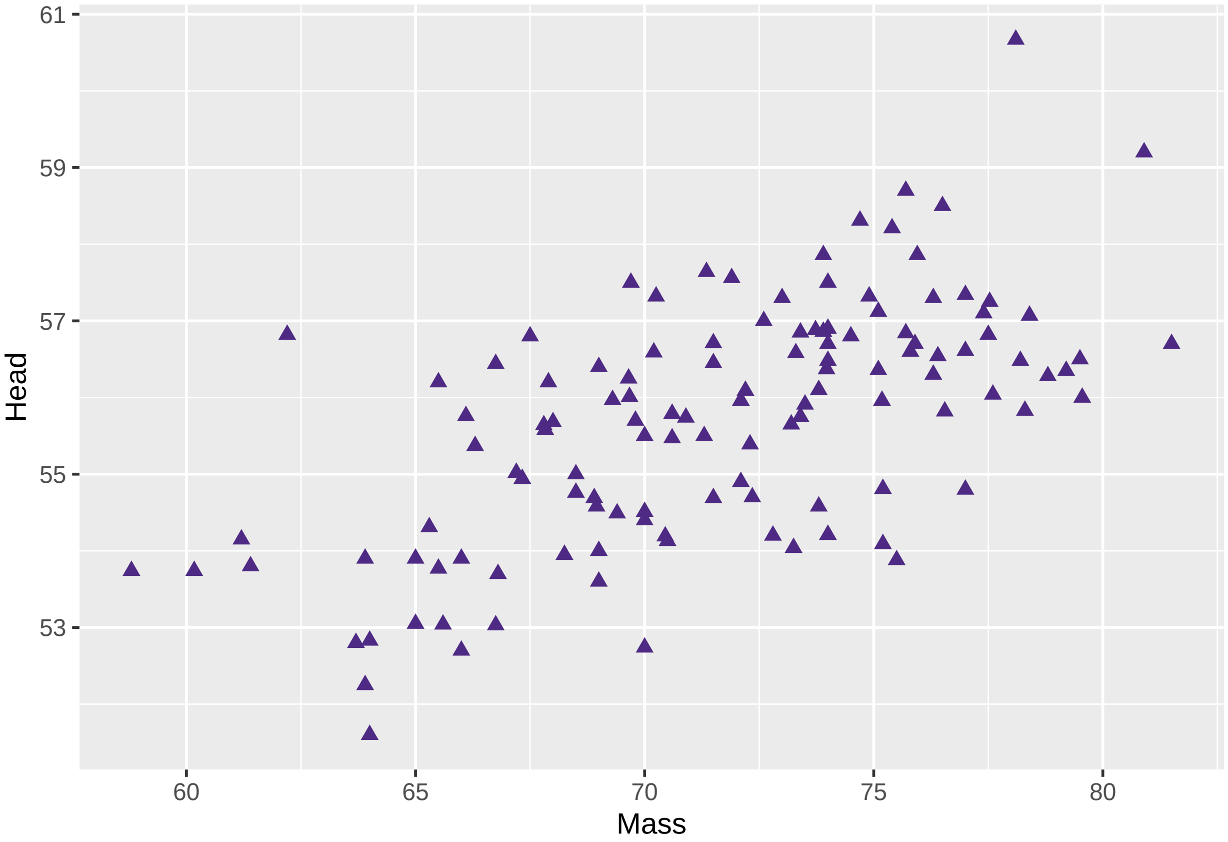
### Exercise 1

Using `blue_jay` dataset, construct the following scatterplots of `Head` by `Mass`:

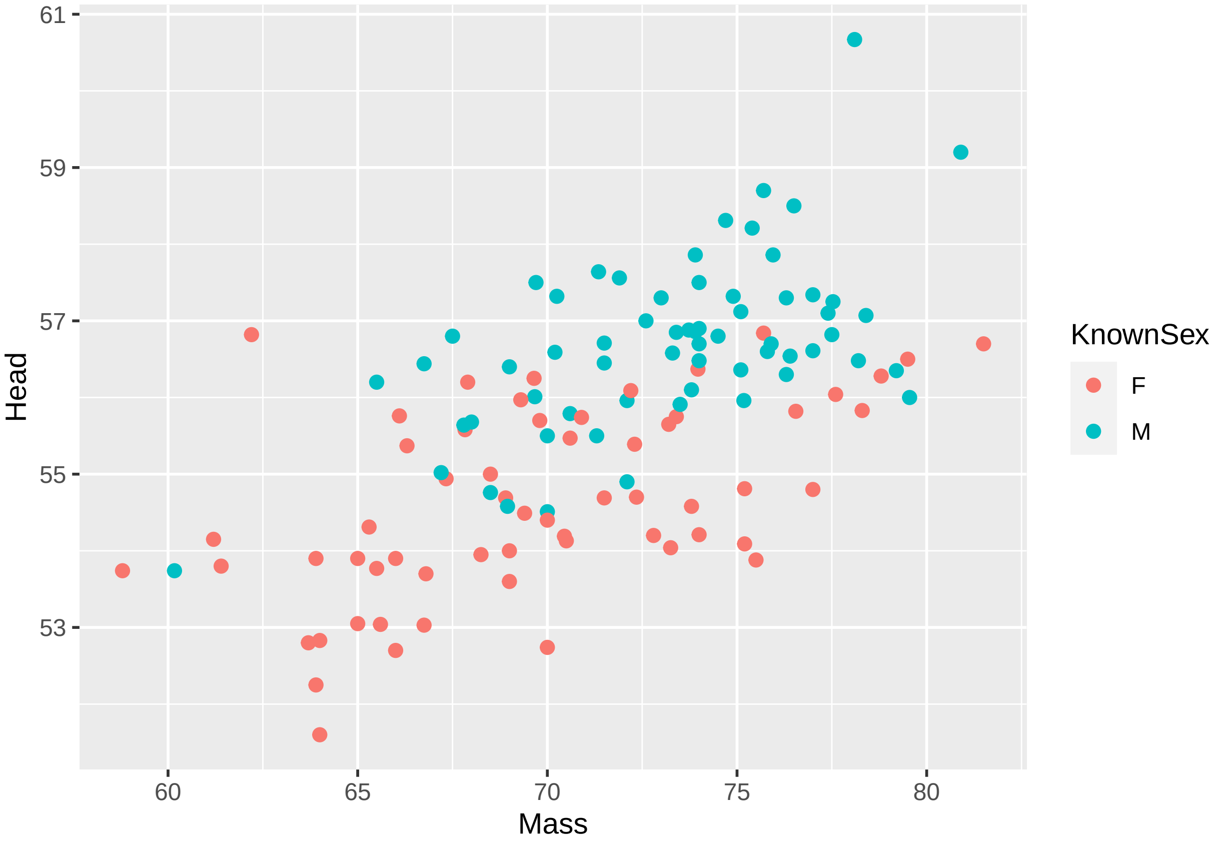
1. One with the `color` aesthetic set to Northwestern purple (`#4E2A84`), `shape` aesthetic set a solid/filled triangle, and `size` aesthetic set to 2.
2. One using `Sex` or `KnownSex` mapped to the `color` aesthetic. That is, determine which is more appropriate and explain why. Also set the `size` aesthetic to 2.

```
load('data/blue_jays.rda')

blue_jays %>%
  ggplot(aes(x = Mass, y = Head)) +
  geom_point(color = '#4E2A84', shape = 'triangle', size = 2)
```



```
blue_jays %>%
  ggplot(aes(x = Mass, y = Head)) +
  geom_point(aes(color = KnownSex), size = 2)
```



**`KnownSex` is better, as `Sex` would give the legend in 0 or 1 instead of F or M.**

Consider the `color` aesthetic in the plots for (1) and (2). Explain why these two usages of the `color` aesthetic are meaningfully different. **The reason the color aesthetic is different in these two is because in the first, it is setting a universal color to all points, whereas in the second, it is using color as a type of facet for another (categorical) variable.**

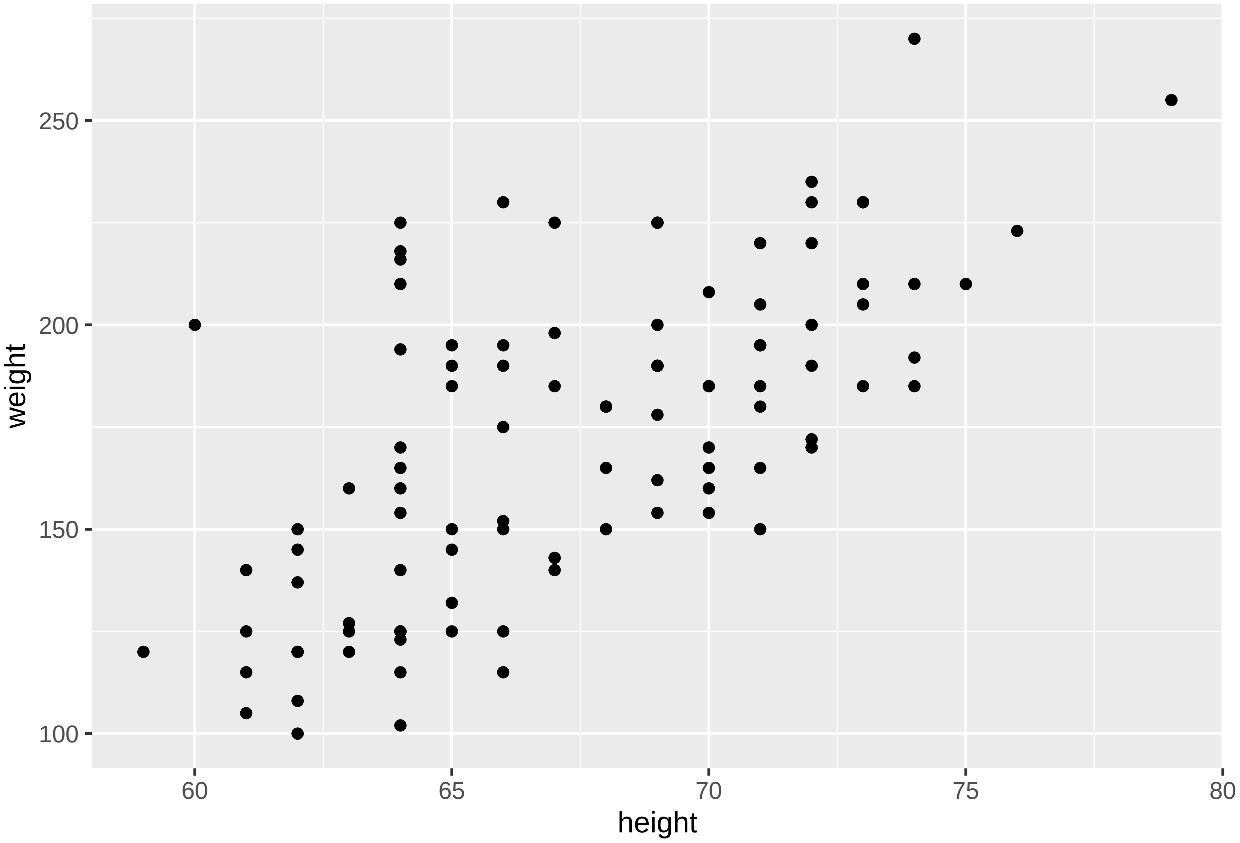
### Exercise 2

Using a random subsample of size 100 from the `cdc` dataset (code provided below), construct a scatterplot of `weight` by `height`. Construct 5 more scatterplots of `weight` by `height` that make use of aesthetic attributes `color` and `shape` (maybe `size` too). You can define both aesthetics at the same time in each plot or one at a time. Just experiment. — Should be six total plots.

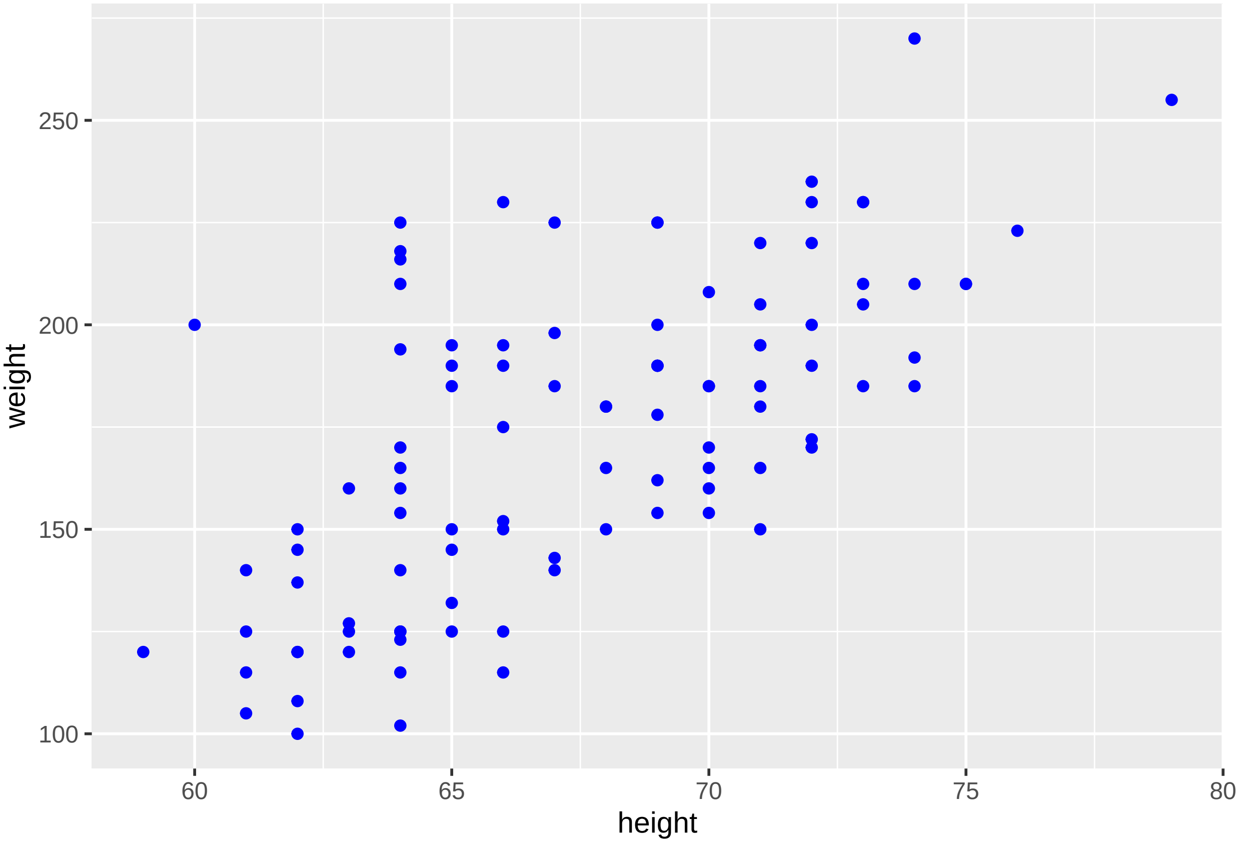
```
# Read in the cdc dataset
cdc <- read_delim(file = "data/cdc.txt", delim = "|") %>%
  mutate(genhlth = factor(genhlth,
    levels = c("excellent", "very good", "good", "fair", "poor")
  ))

# Selecting a random subset of size 100
cdc_small <- cdc %>% sample_n(100)
```

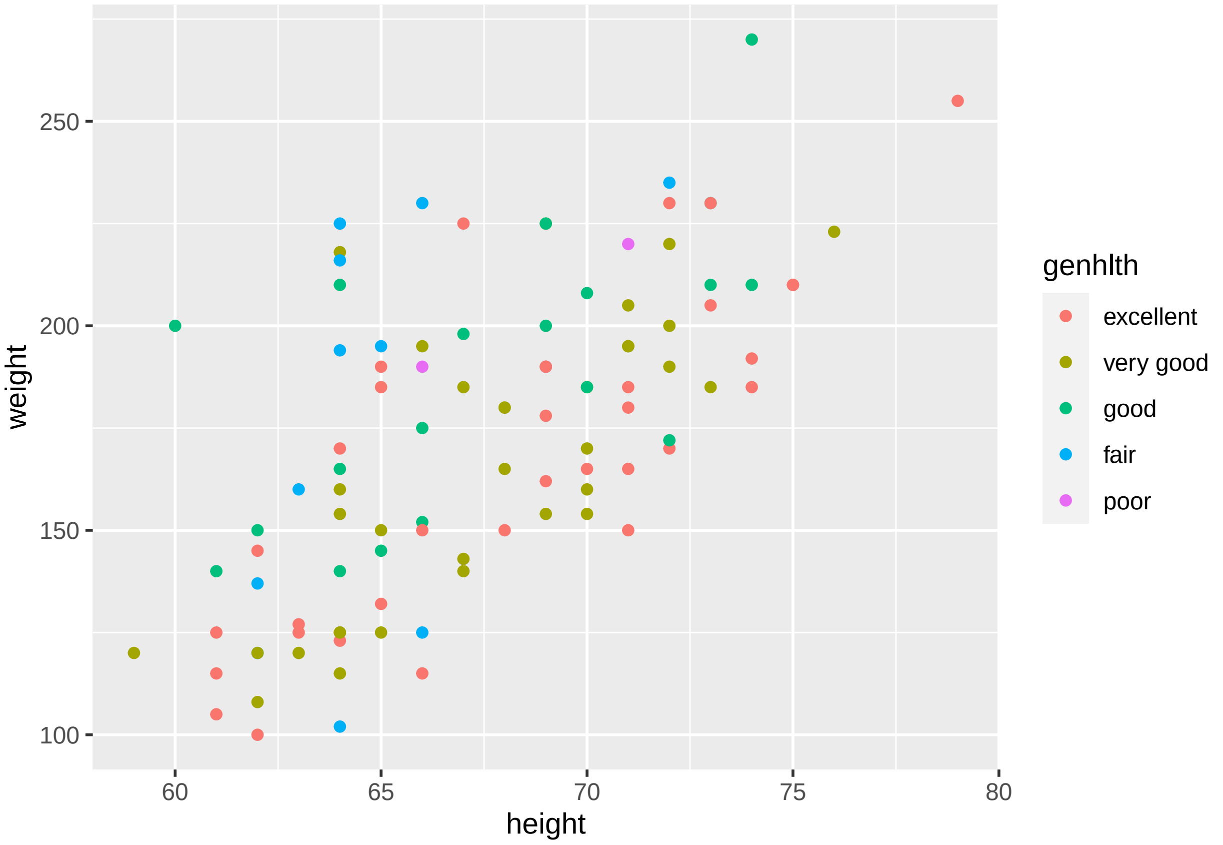
```
#1
cdc_small %>%
  ggplot(aes(height, weight)) +
  geom_point()
```



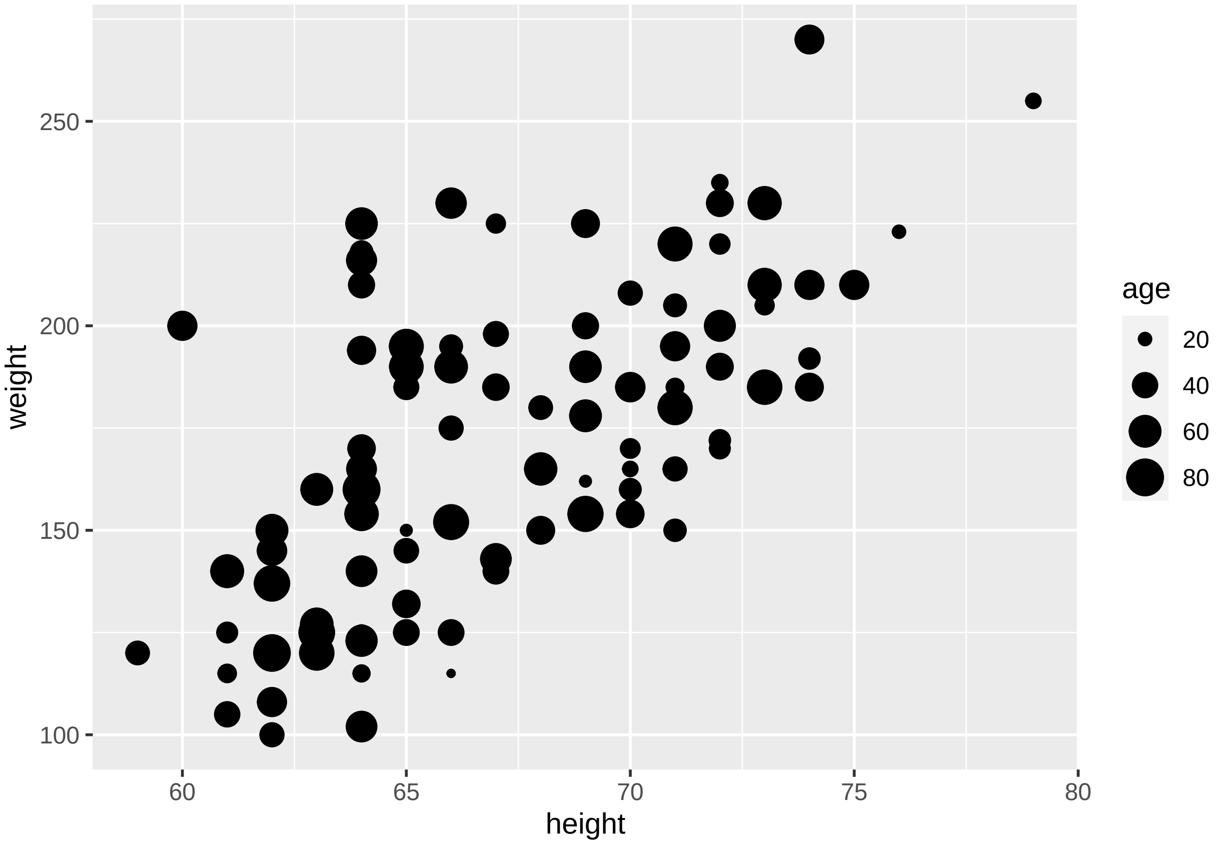
```
#2
cdc_small %>%
  ggplot(aes(height, weight)) +
  geom_point(color = 'blue')
```



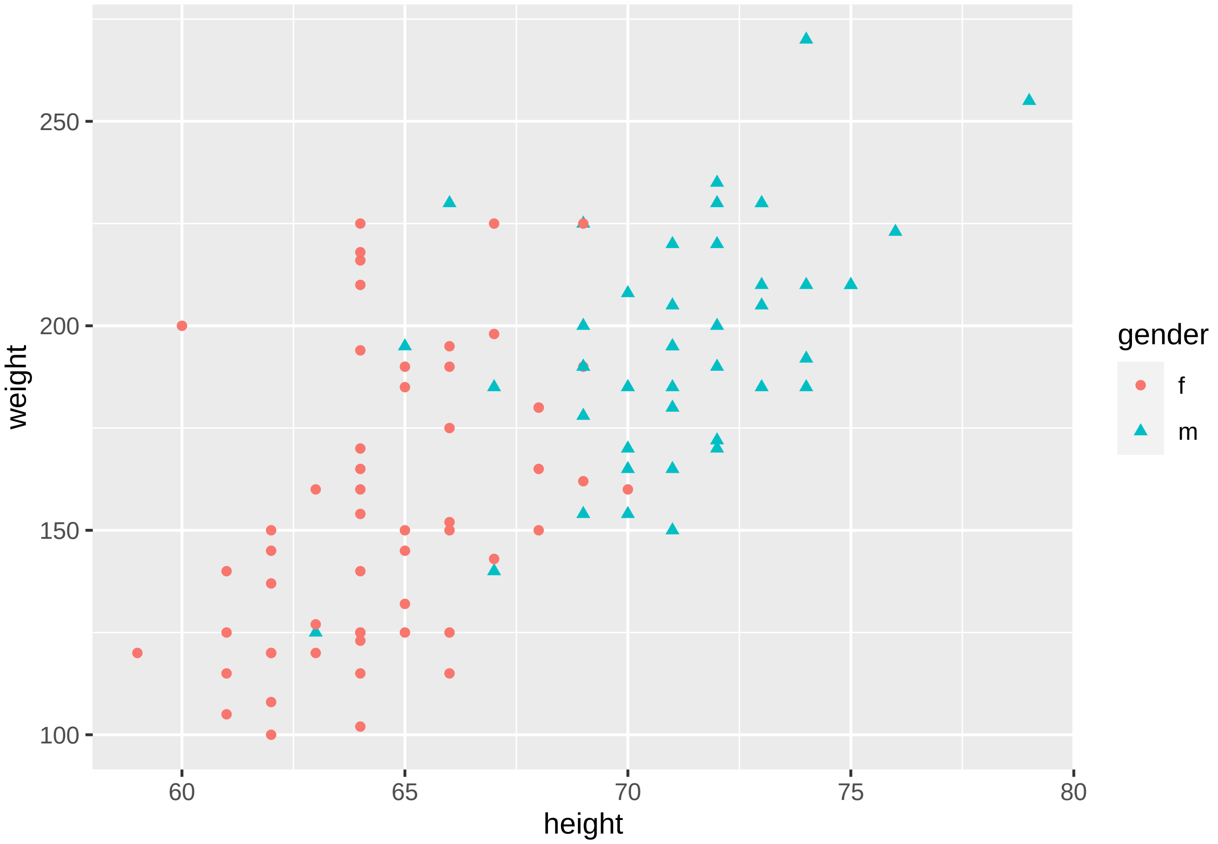
```
#3
cdc_small %>%
  ggplot(aes(height, weight)) +
  geom_point(aes(color = genhlth))
```



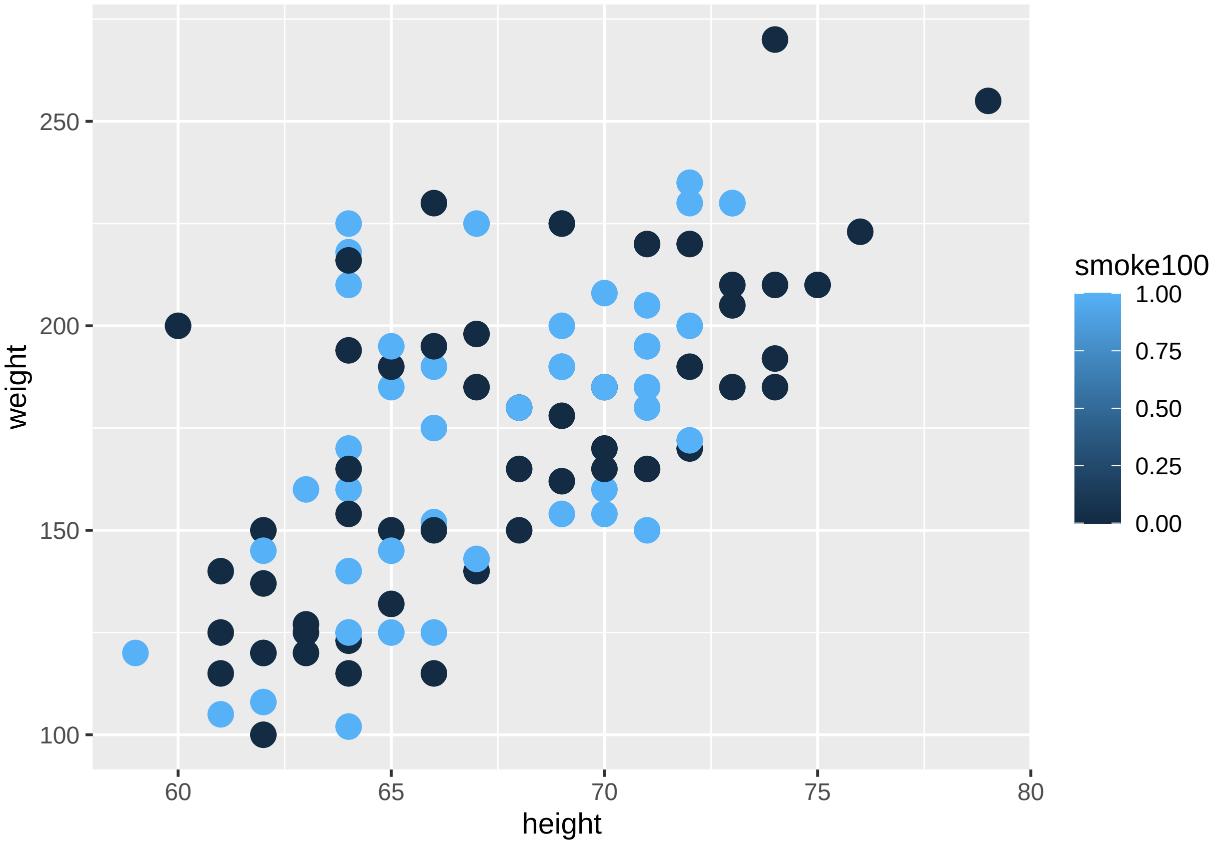
```
#4
cdc_small %>%
  ggplot(aes(height, weight)) +
  geom_point(aes(size = age))
```



```
#5
cdc_small %>%
  ggplot(aes(height, weight)) +
  geom_point(aes(shape = gender, color = gender))
```



```
#6
cdc_small %>%
  ggplot(aes(height, weight)) +
  geom_point(aes(color = smoke100), size = 4)
```


250
200
150
100
weight
60
65
70
75
80
height
smoke100
1.00
0.75
0.50
0.25
0.00