STAT 33B Workbook 4

Yuanrui Zhu (3034615728)

Sep 24, 2020

This workbook is due **Sep 24, 2020** by 11:59pm PT.

The workbook is organized into sections that correspond to the lecture videos for the week. Watch a video, then do the corresponding exercises *before* moving on to the next video.

Workbooks are graded for completeness, so as long as you make a clear effort to solve each problem, you'll get full credit. That said, make sure you understand the concepts here, because they're likely to reappear in homeworks, quizzes, and later lectures.

As you work, write your answers in this notebook. Answer questions with complete sentences, and put code in code chunks. You can make as many new code chunks as you like.

In the notebook, you can run the line of code where the cursor is by pressing `Ctrl` + `Enter` on Windows or `Cmd` + `Enter` on Mac OS X. You can run an entire code chunk by clicking on the green arrow in the upper right corner of the code chunk.

Please do not delete the exercises already in this notebook, because it may interfere with our grading tools.

You need to submit your work in two places:

- Submit this Rmd file with your edits on bCourses.
- Knit and submit the generated PDF file on Gradescope.

R Graphics Overview

Watch the "R Graphics Overview" lecture video.

No exercises for this section. Keep at it!

The Tidyverse

Watch the "The Tidyverse" lecture video.

Exercise 1

How many packages are in the Tidyverse? Explore the website to find out. You can count the tidymodels packages as a single package.

YOUR ANSWER GOES HERE:

There are about 7 packages in the Tidyverse, they are: -readr -tibble -purrr -forcats -ggplot2 -dplyr -stringr

Tibbles

Watch the "Tibbles" lecture video.

Exercise 2

Read the documentation for the tibble package on the website.

1. Create a tibble with 4 rows and 3 columns from vectors. You can make up the data in the vectors. Use a different data type for each column.
2. Show how to convert the tibble to an ordinary data frame.

YOUR ANSWER GOES HERE:

```
#install.packages("tidyr")
library(tibble)
column1 = c(1, 2, 3, 4)
column2 = c("a", "b", "c", "d")
column3 = c(TRUE, FALSE, TRUE, FALSE)
tibble(column1, column2, column3)
```

```
## # A tibble: 4 x 3
##   column1 column2 column3
##     <dbl> <chr>   <lgl>
## 1       1 a       TRUE
## 2       2 b       FALSE
## 3       3 c       TRUE
## 4       4 d       FALSE
```

The Grammar of Graphics

Watch the "The Grammar of Graphics" lecture video.

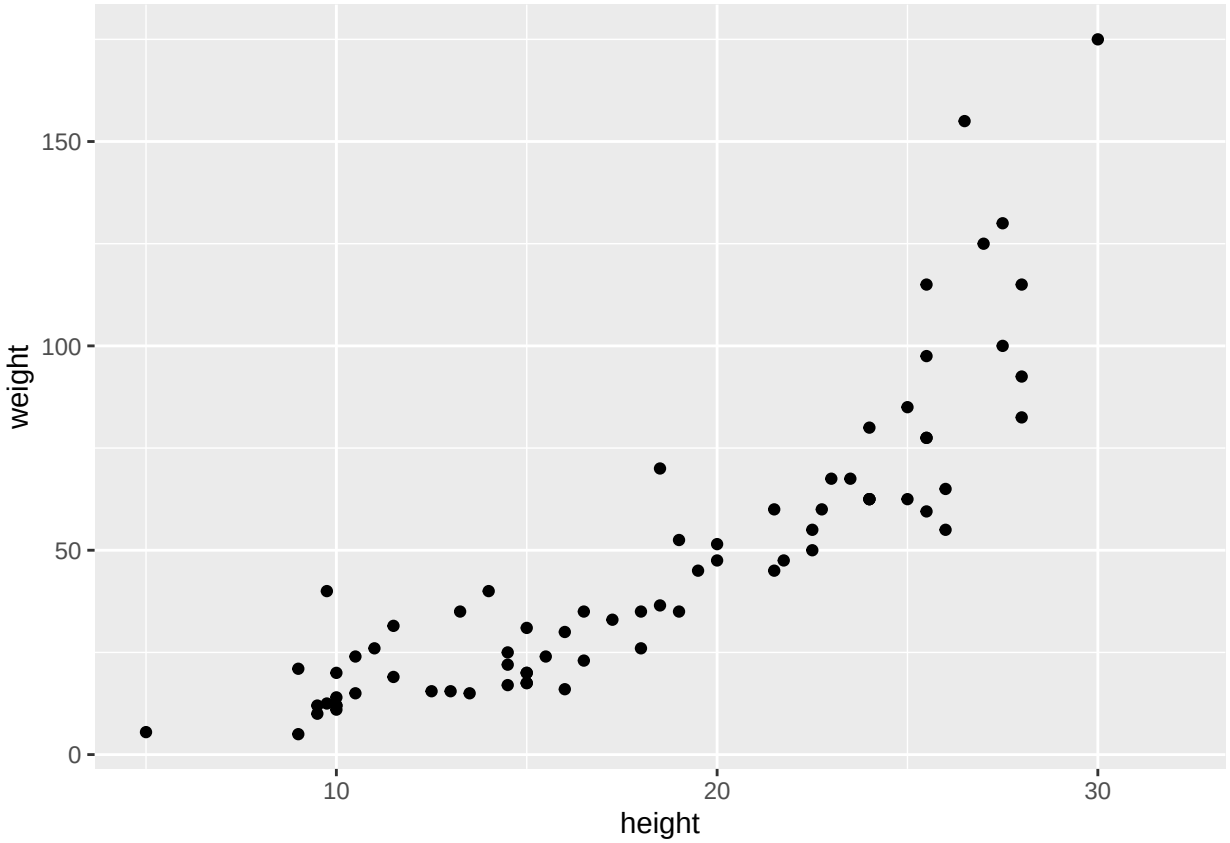
Exercise 3

Use ggplot2 and the dogs data to make a scatterplot that shows the relationship between height and weights. In 2-3 sentences, discuss any patterns you see in the plot.

YOUR ANSWER GOES HERE:

```
dogs = readRDS("dogs.rds")
library(ggplot2)
ggplot(dogs, aes(x = height, y = weight)) + geom_point()
```

```
## Warning: Removed 98 rows containing missing values (geom_point).
```



I discover that there is a strong positive correlation between a dog's height and its weight. That is, the taller the dog, the heavier it becomes.

Exercise 4

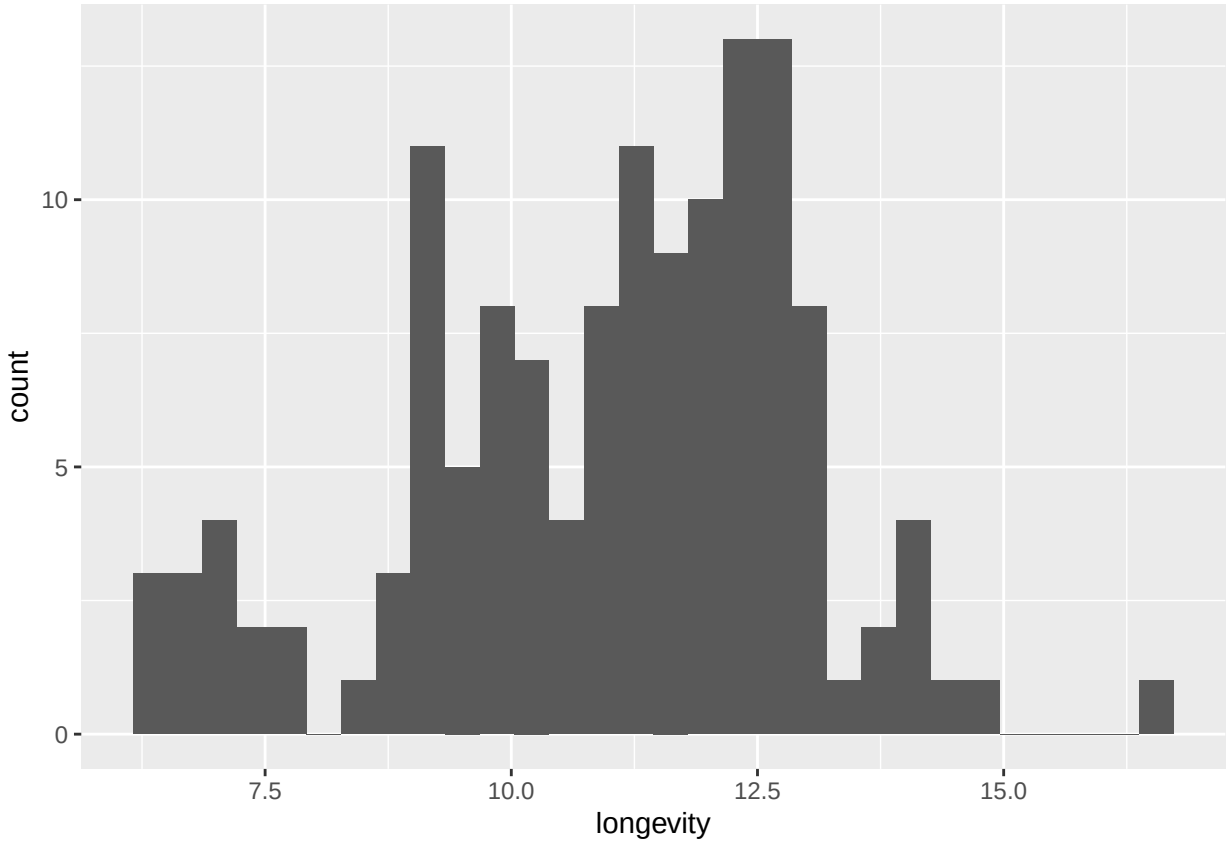
Use ggplot2 and the dogs data to make a histogram of longevity. How long do most dogs typically live? How spread out is the distribution of lifespans?

YOUR ANSWER GOES HERE:

```
ggplot(dogs, aes(x = longevity)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 37 rows containing non-finite values (stat_bin).
```



From the histogram, we can see most dogs lives between 10-12,5 years old. The spread is quite large, from less than 3 to approximately 20. The distribution is roughly normal, with two pinnacles around 8 and 12.5.

Saving Plots

Watch the "Saving Plots" lecture video.

No exercises for this section. Almost done!

Customizing Plots

Watch the "Customizing Plots" lecture video.

Exercise 5

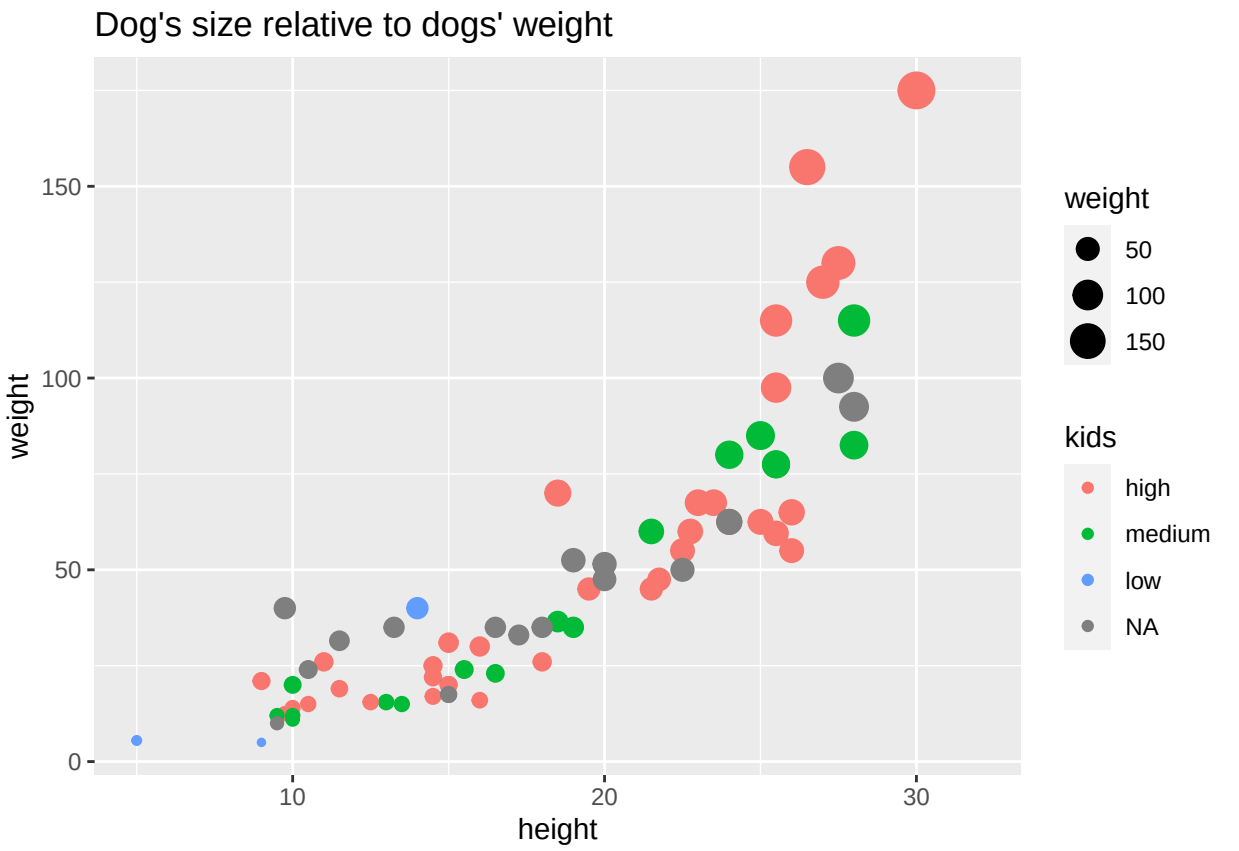
Revisit your scatterplot from Exercise 3. Make the size of each point correspond to the size of the dog. Make the color of each point correspond to how good the dog is with kids. Add an appropriate title and labels.

Hint: A "low" score for kids means the dog is bad with kids.

YOUR ANSWER GOES HERE:

```
ggplot(dogs, aes(x = height, y = weight)) + geom_point(aes(color = kids, size = weight)) +
  labs(title = "Dog's size relative to dogs' weight")
```

```
## Warning: Removed 98 rows containing missing values (geom_point).
```



Exercise 6

Use ggplot2 and the dogs data to answer each of the following questions.

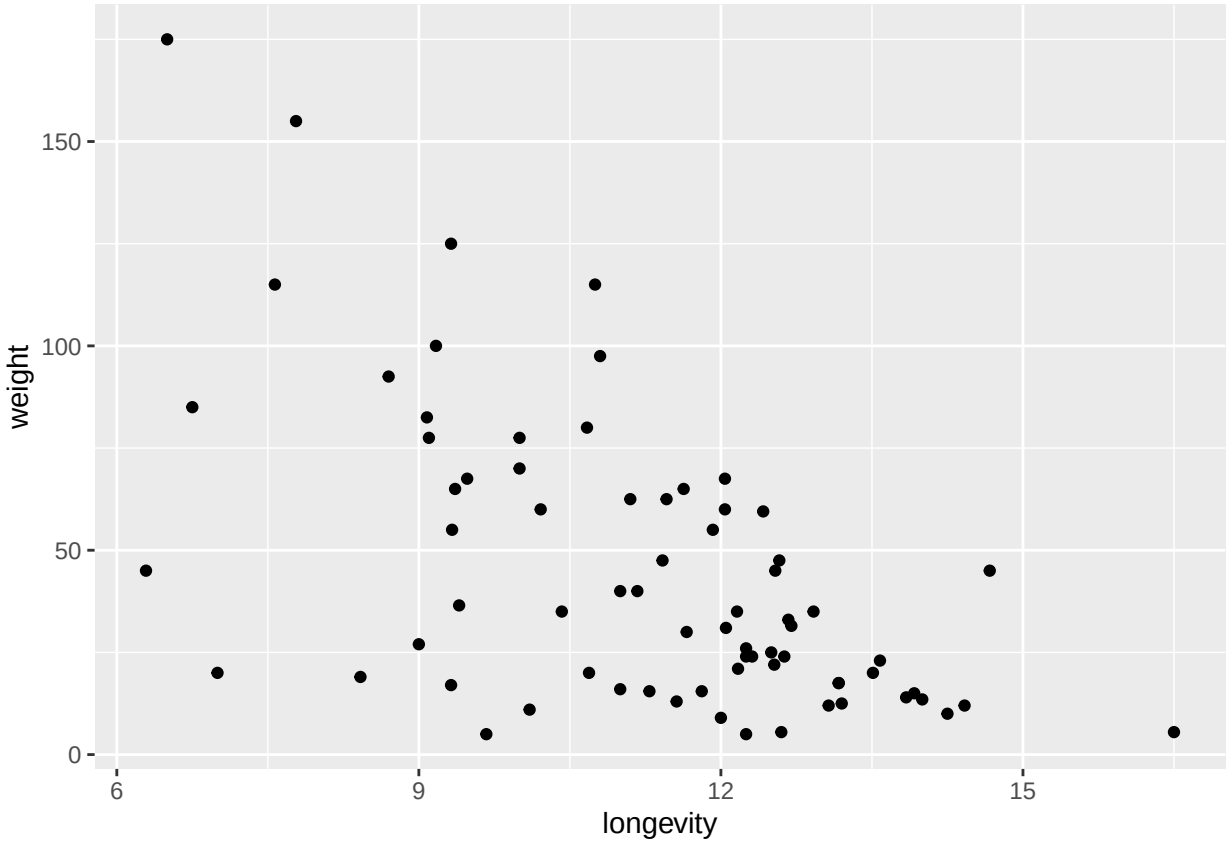
1. Is there a relationship between how long dogs live and how big (in any sense) they are?
2. Do small dogs tend to be good with kids? If not, does size seem to be related to how good dogs are with kids at all?
3. Is there a relationship between size and grooming needs?

Hint: The table from the notes for the next lecture video might help you decide which plots to use for these questions.

YOUR ANSWER GOES HERE:

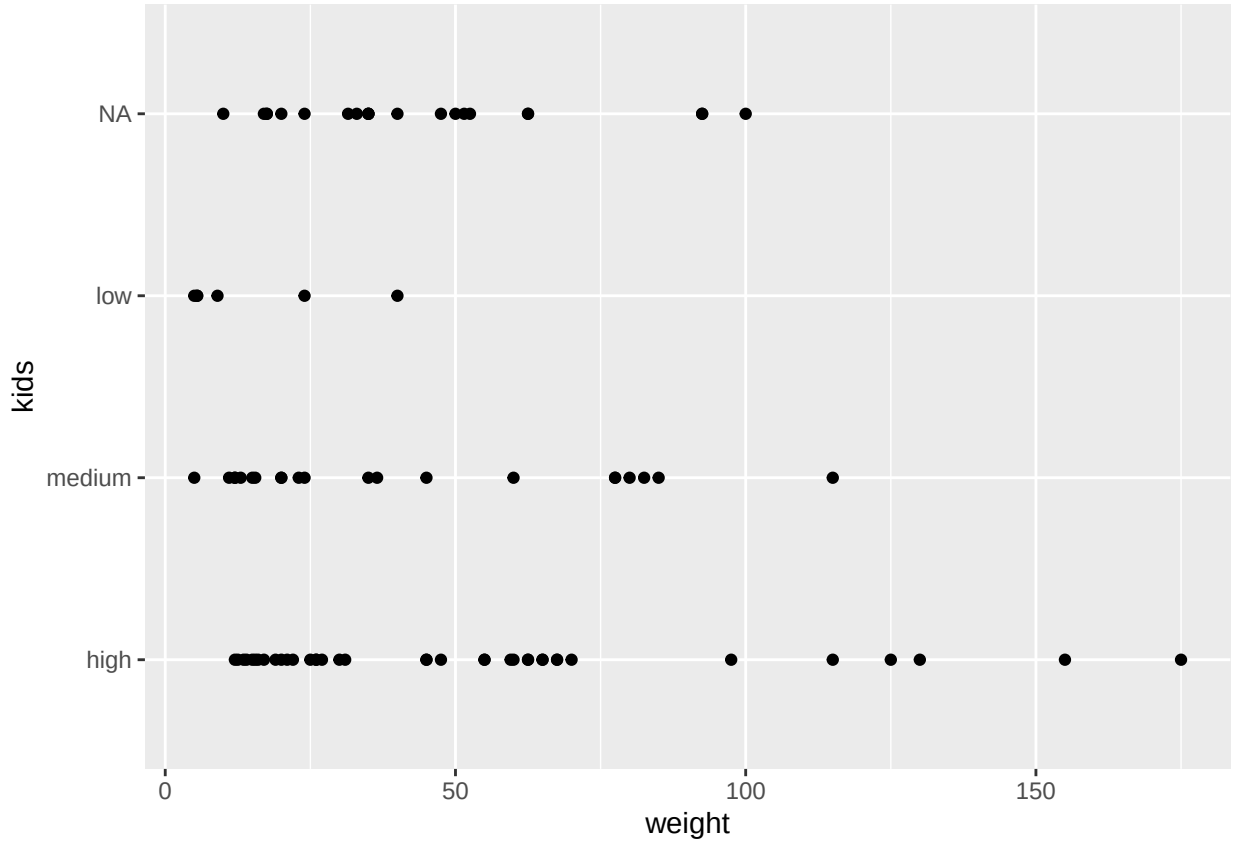
```
ggplot(dogs, aes(x = longevity, y = weight)) + geom_point()
```

```
## Warning: Removed 99 rows containing missing values (geom_point).
```



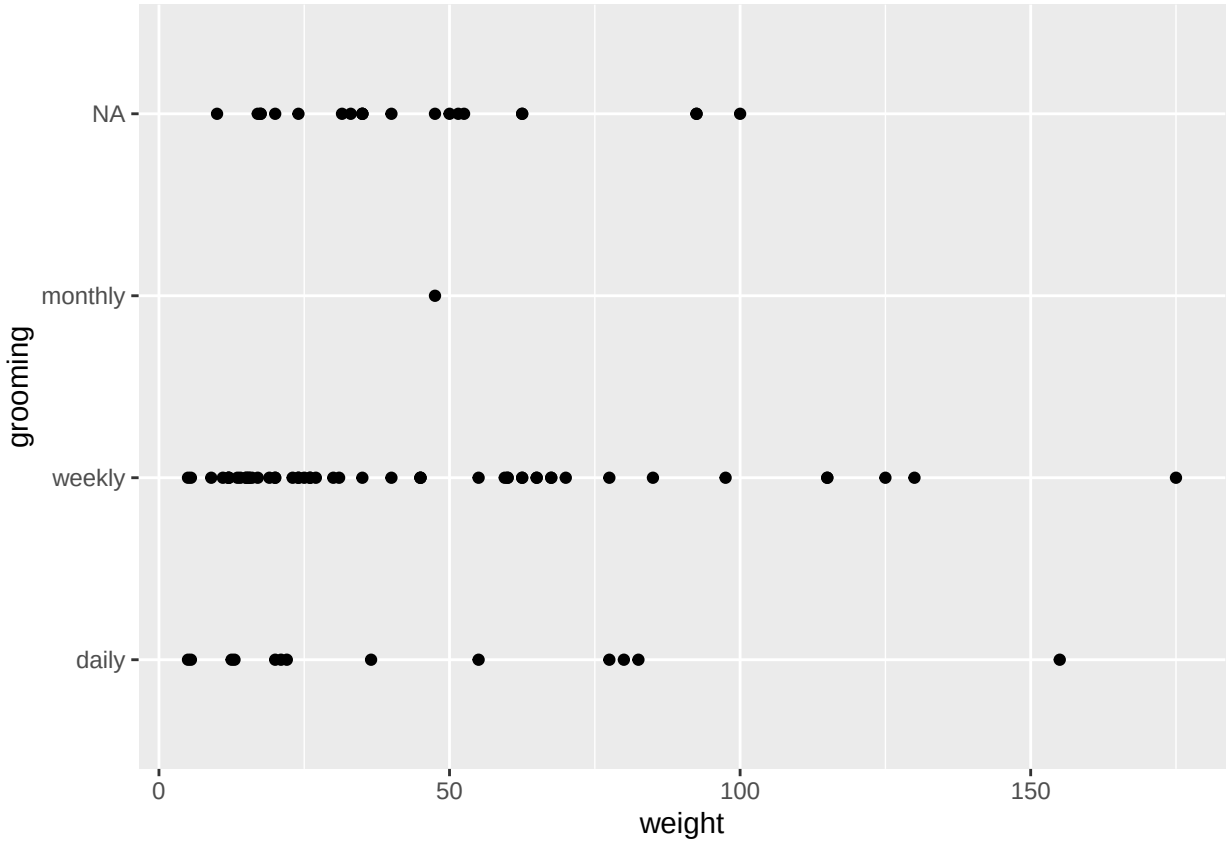

```
ggplot(dogs, aes(x = weight, y = kids)) + geom_point()
```

```
## Warning: Removed 86 rows containing missing values (geom_point).
```




```
ggplot(dogs, aes(x = weight, y = grooming)) + geom_point()
```

```
## Warning: Removed 86 rows containing missing values (geom_point).
```



There is a strong negative relationship between the dog size and its longevity. Small dogs don't tend to be good with kids (it seems kids and weight do not correlate to each other very much). There is barely no relationship between size and grooming needs.

Exploratory Data Analysis

This lecture video is **OPTIONAL**. You can skip it if you like.

The "Exploratory Data Analysis" video discusses how to choose an appropriate plot for a given set of columns.

No exercises for this section.