4 & 5: Exploring Data (Part 2)

Packages Needed To Rereate Code on Slides

```
library(tidyverse) # for ggplot functions/plotting
library(openintro) # for data set
```

Warning: It is not expected that you understand all the R code in this presentation right now. You will go over more R code in SDS 100 in the coming weeks. However, you are welcome to try to make these plots on your own.

Data Frame We Will Be Using

```
# Preview the data
head(cars93)
```

```
# A tibble: 6 x 6
  type    price mpg_city drive_train passengers weight
  <fct>   <dbl>    <int> <fct>             <int>  <int>
1 small    15.9       25 front                 5   2705
2 midsize  33.9       18 front                 5   3560
3 midsize  37.7       19 front                 6   3405
4 midsize  30         22 rear                  4   3640
5 midsize  15.7       22 front                 6   2880
6 large    20.8       19 front                 6   3470
```

Data Frame We Will Be Using

These cars represent a random sample for 1993 models that were in both *Consumer Reports* and *PACE Buying Guide*.

- `type`: The vehicle type
- `price`: Vehicle price (USD).
- `mpg_city`: Vehicle mileage in city (miles per gallon).
- `drive_train`: Vehicle drive train with levels `4WD`, `front`, and `rear`.
- `passengers`: The vehicle passenger capacity.
- `weight`: Vehicle weight (lbs).

54 cases (rows) and 6 variables (columns)

Two Categorical Variables

Two Categorical Variables - Summary Statistics

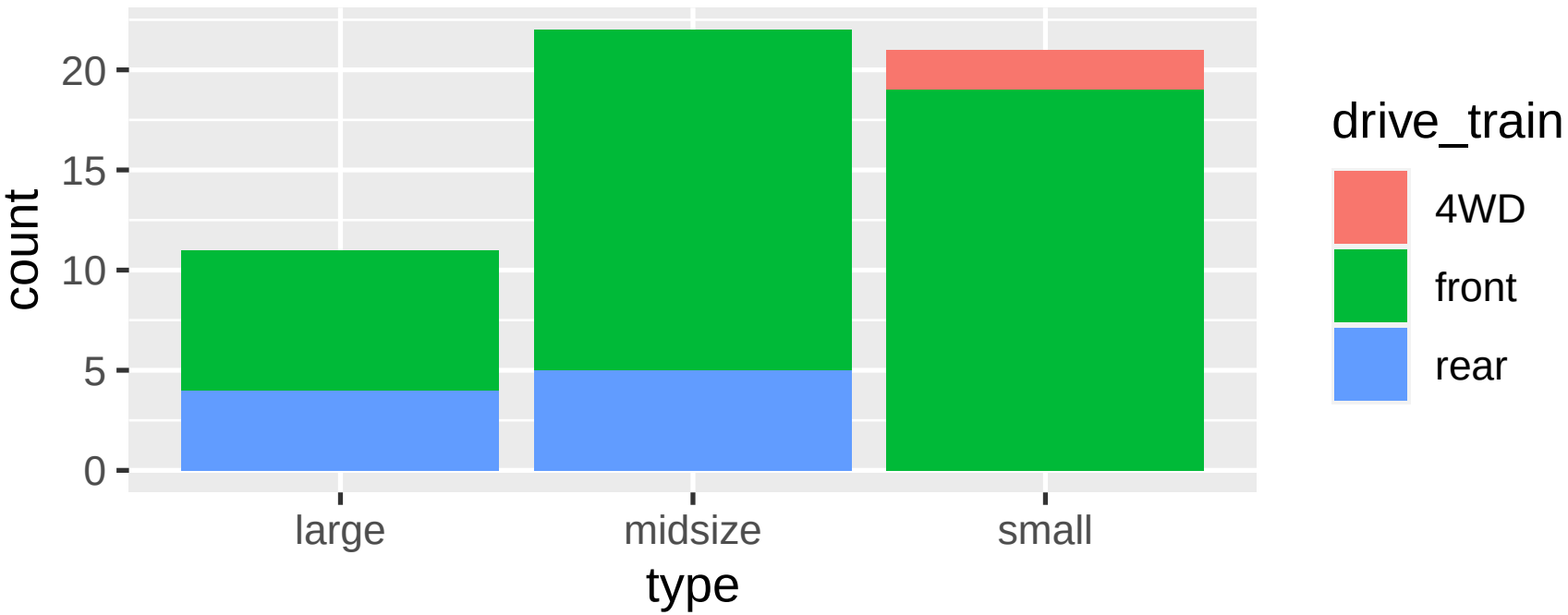
- A table that summarizes count data two categorical variables is called a **contingency table**.
- Each value in the table represents the number of times a particular combination of variable outcomes occurred.

```
table(cars93$type, cars93$drive_train)
```

```
         4WD front rear
large      0     7    4
midsize    0    17    5
small      2    19    0
```

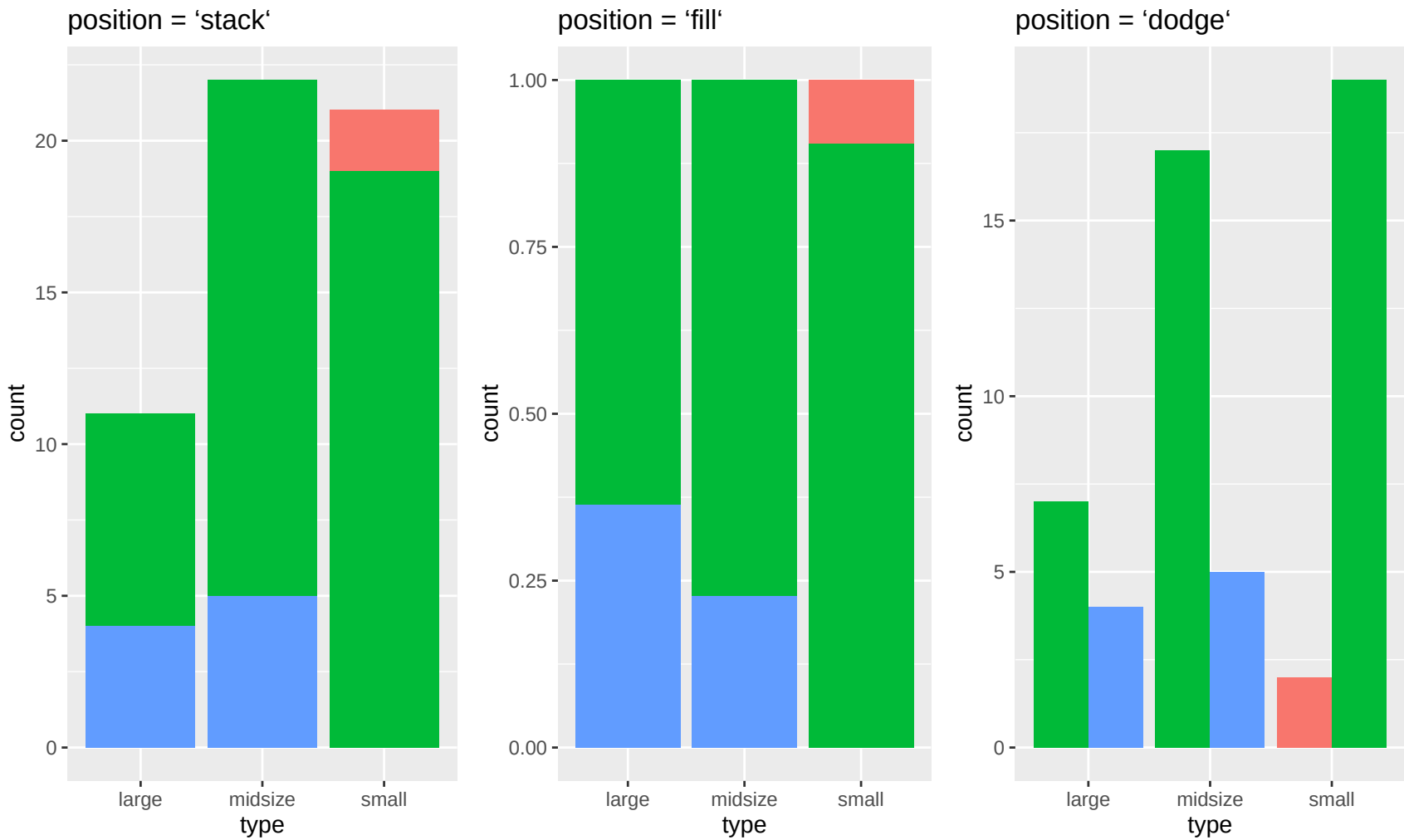
Two Categorical Variables - Bar Charts

```
ggplot(cars93, aes(x = type, fill = drive_train)) +
  geom_bar(position = 'stack')
```



Two Categorical Variables - Bar Charts

Different ways to create bar charts. What are some pros and cons of each?



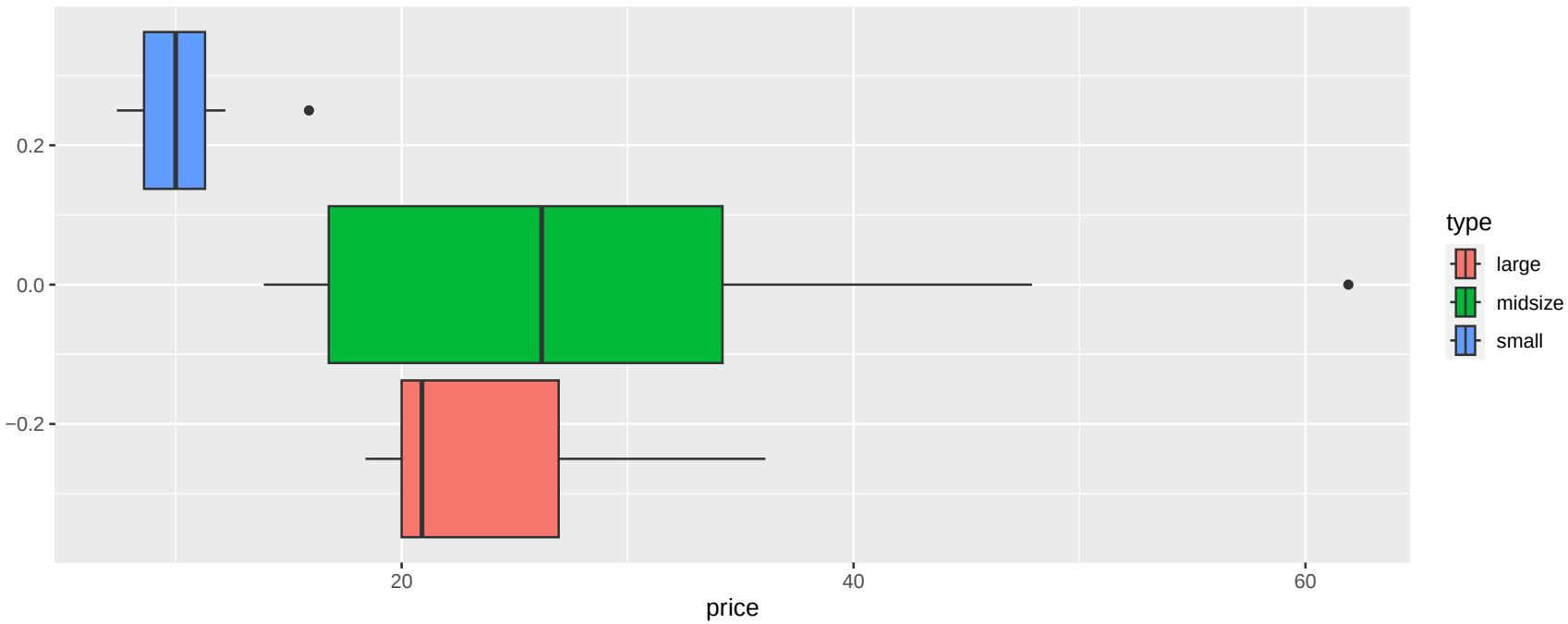
One Categorical and One Numerical Variable

One Categorical and One Numerical

- Typically comparing a numerical variable across groups.
- Ultimately want to see how something changes across groups.

Boxplots Across Groups

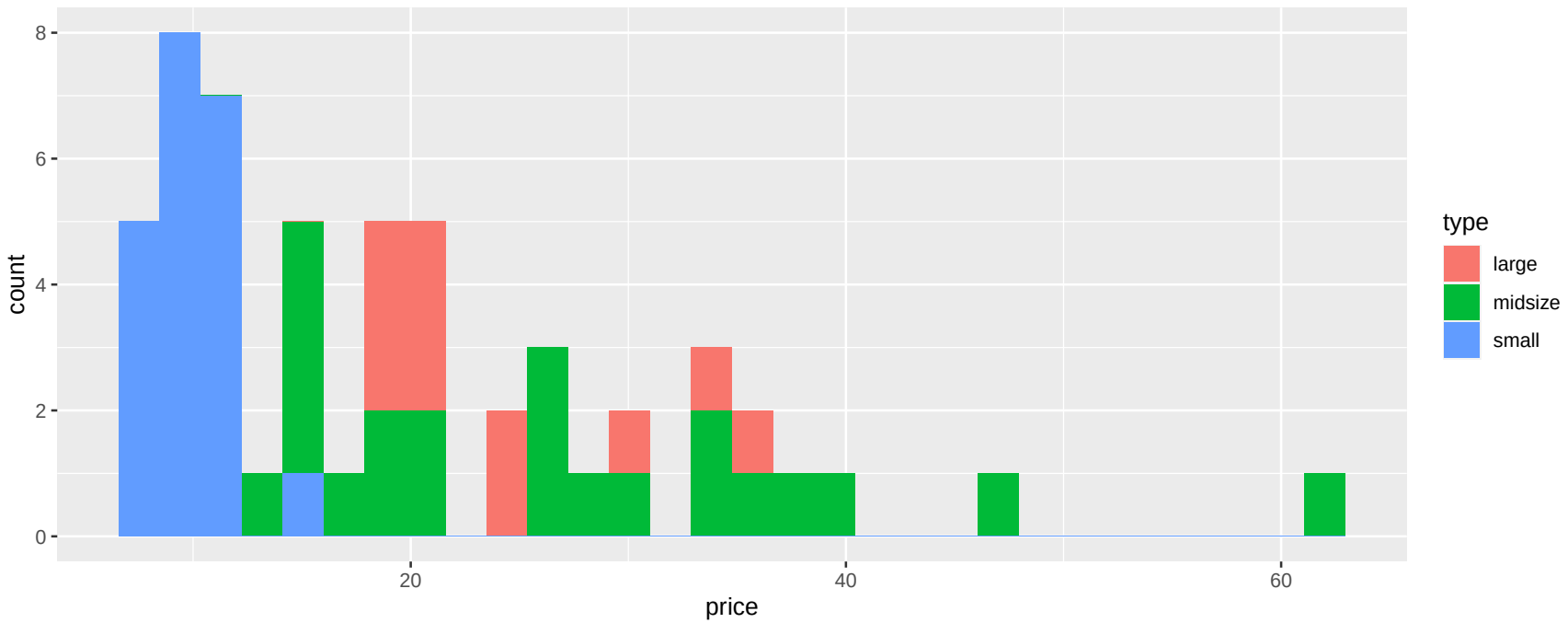
```
ggplot(cars93, aes(x = price, fill = type))+
  geom_boxplot()
```



Histograms Across Groups

Can be hard to read if overlaid.

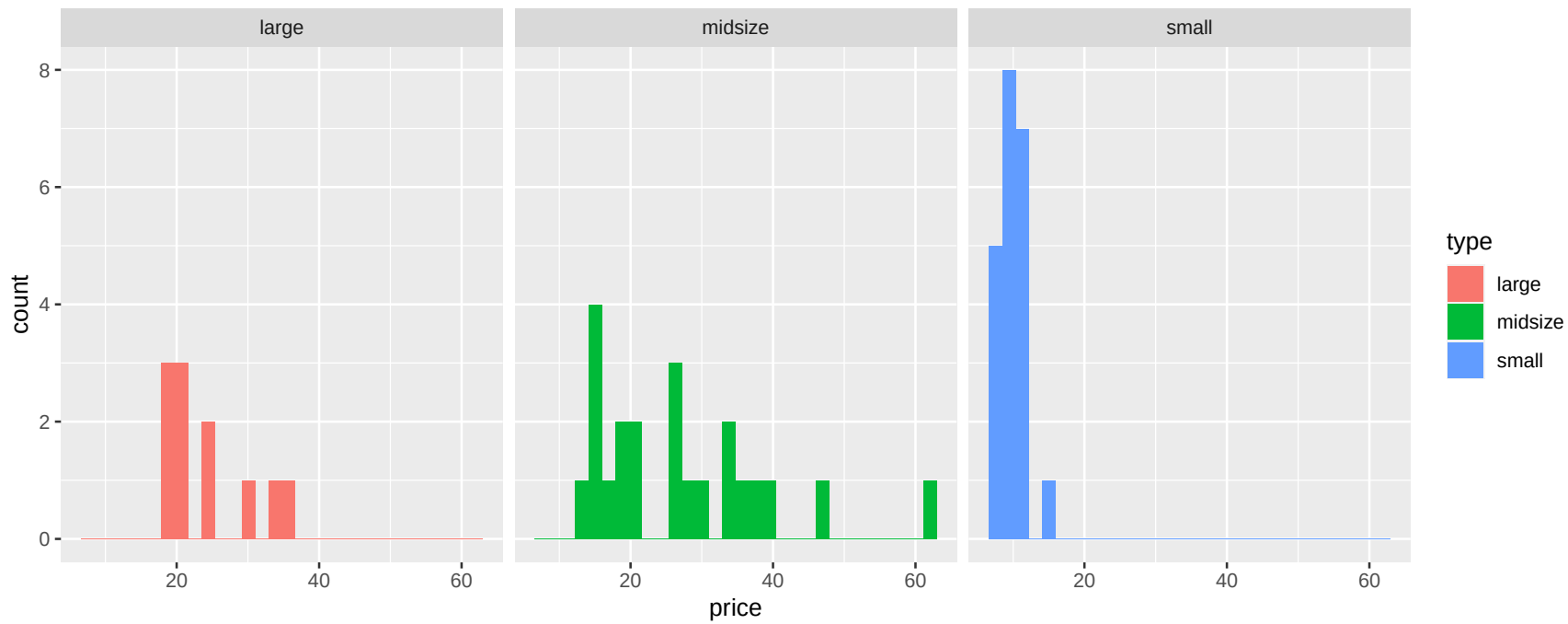
```
ggplot(cars93, aes(x = price, fill = type))+
  geom_histogram()
```



Histograms Across Groups

Side-by-side plots using *faceting*.

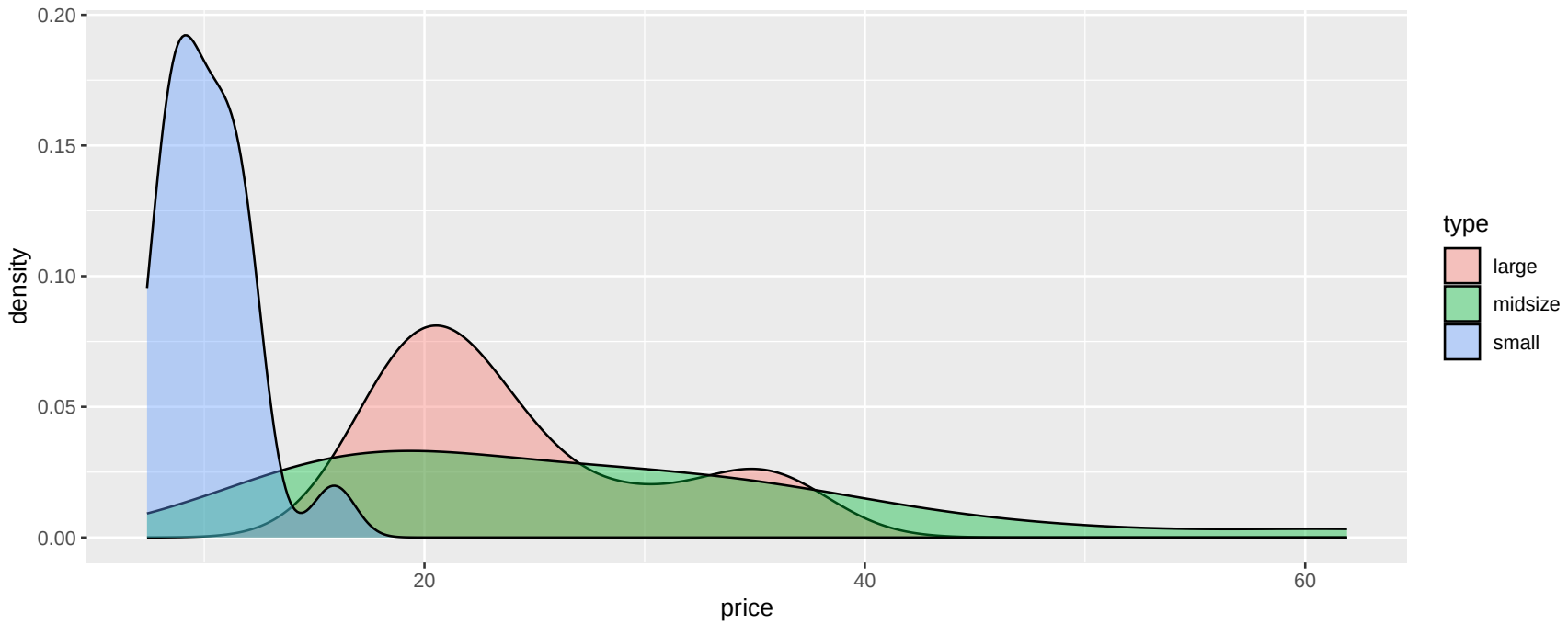
```
ggplot(cars93, aes(x = price, fill = type))+
  geom_histogram()+
  facet_grid(~type)
```



Density Plot Across Groups

Typically easier to read.

```
ggplot(cars93, aes(x = price, fill = type))+
  geom_density(alpha = 0.4)
```

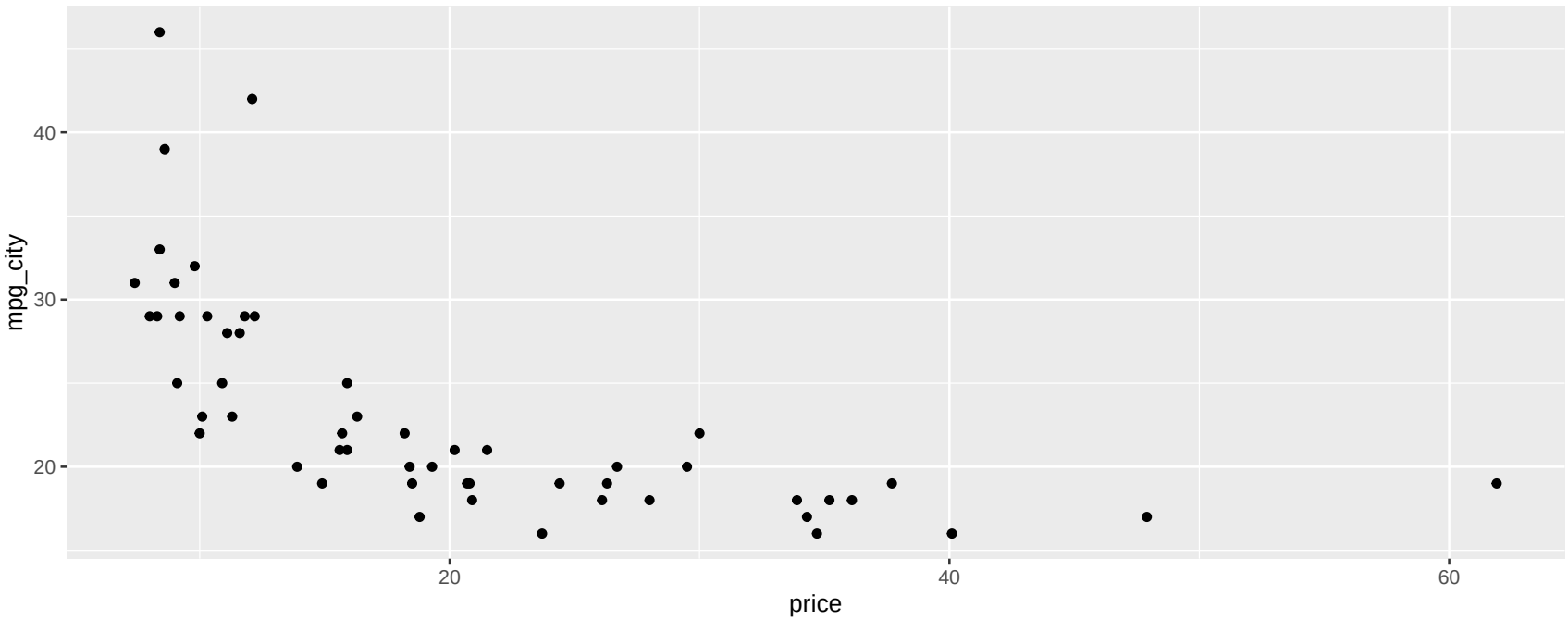


Two Numerical Variables

Scatterplot

A **scatterplot** provides a case-by-case view of data for two numerical variables

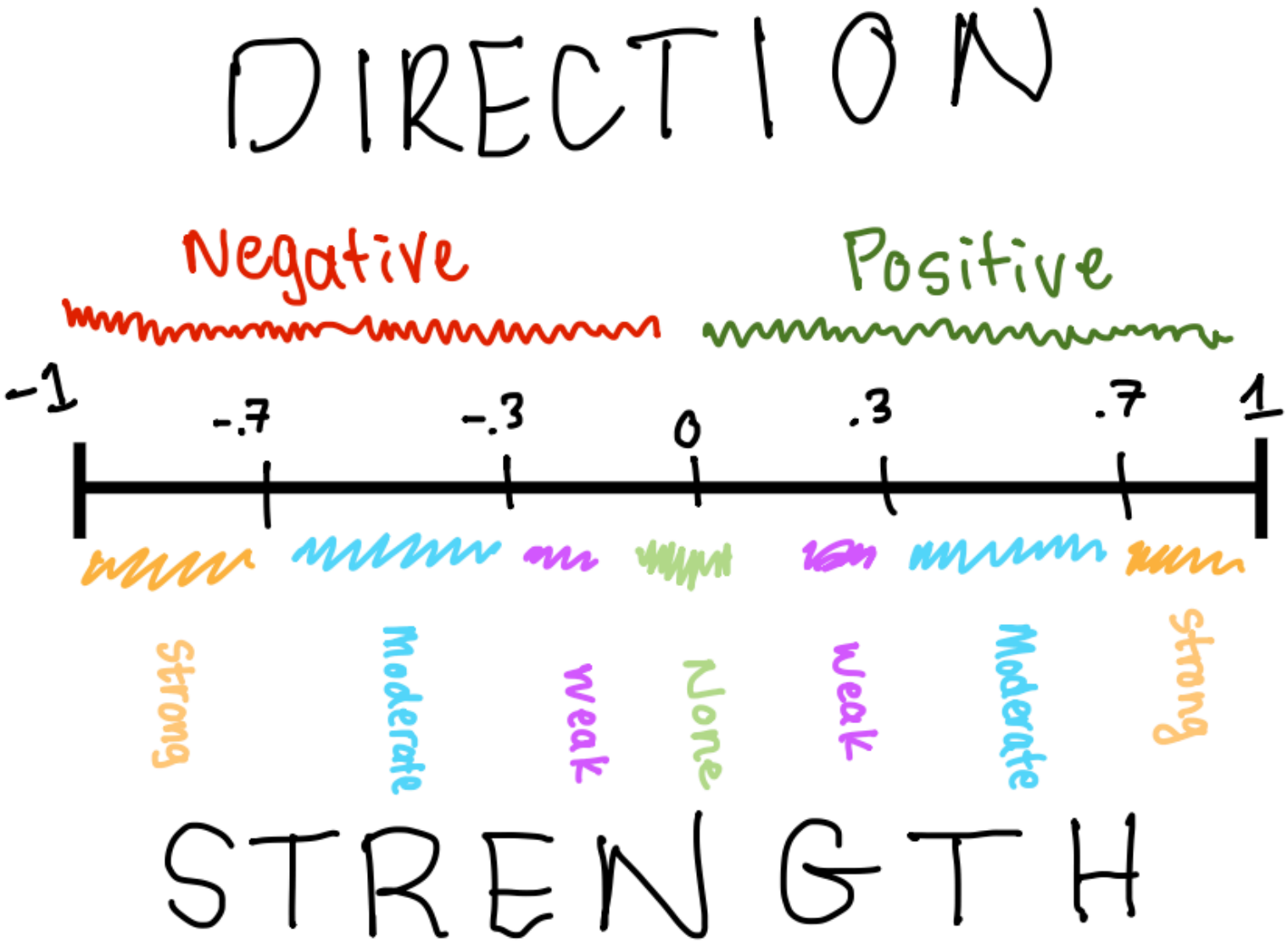
```
ggplot(cars93, aes(x = price, y = mpg_city))+
  geom_point()
```



Summary Statistics - Correlation

- Use **correlation** to describes the strength and direction of the linear relationship between two variables.
- Always between -1 and 1.
- Denoted by r.
  - $+/-$ describes the **direction**. Are points trending upwards or downwards?
  - Distance from zero describes the **strength.** How much do points deviate from a straight line?
- Has no units and will not be affected by a linear change in the units (e.g., going from inches to centimeters).

Summary Statistics - Correlation



Summary Statistics - Correlation

We often say:

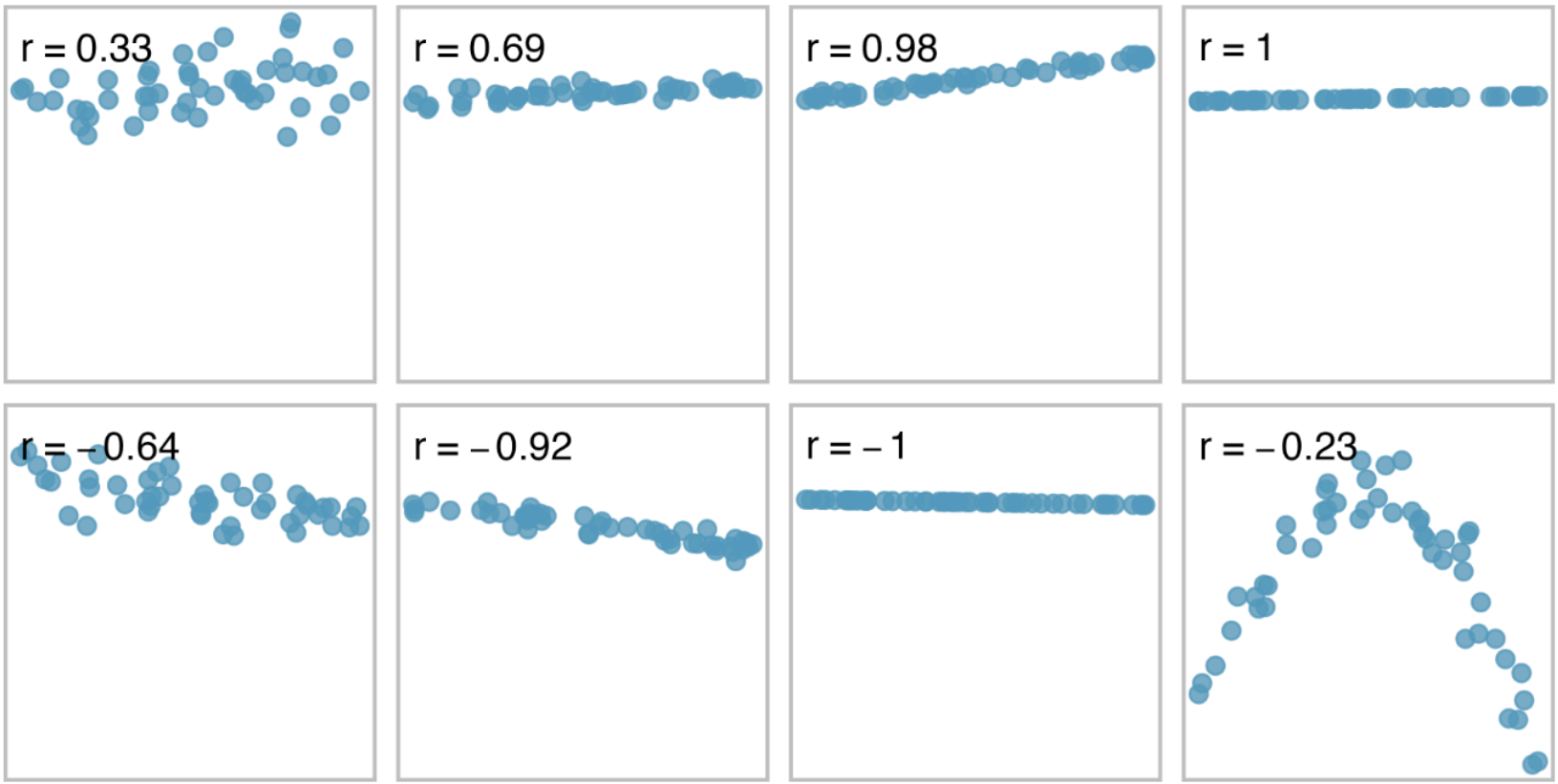
"X and Y are strength direction correlated."

For Example:

1. Height and shoe size are strongly positively correlated.
2. Income and average commute time are weakly negatively correlated.

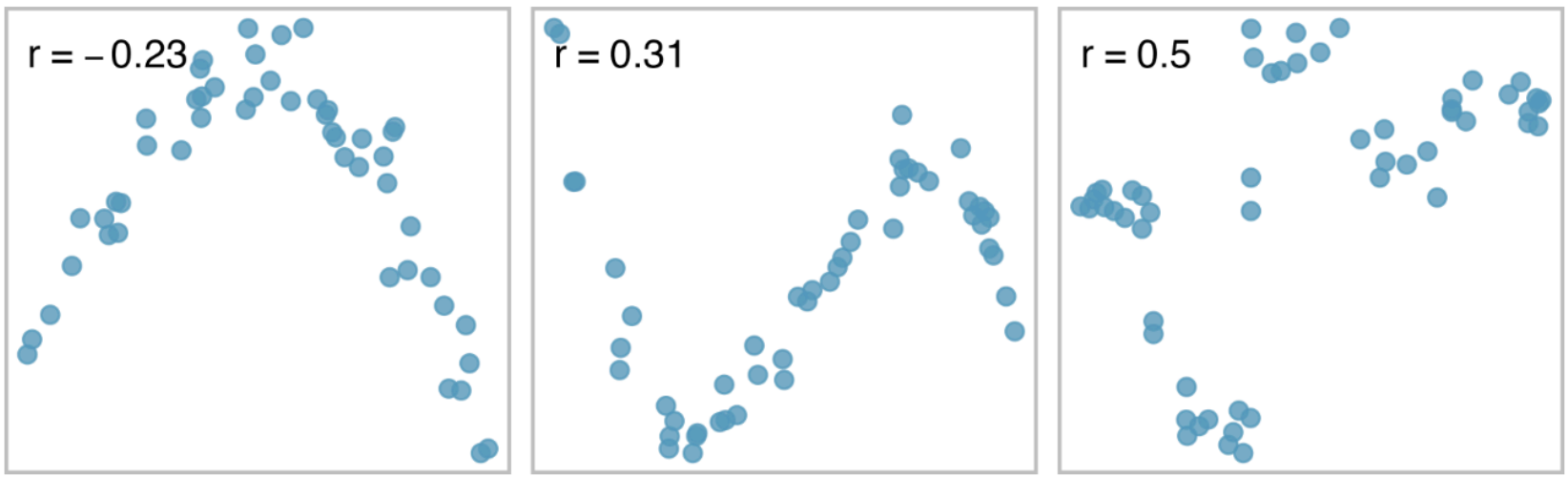
Play with the wording so it feels right! This is just a loose guide.

Correlation Examples



Correlation and Nonlinear Relationships

Correlation does not capture nonlinear relationships.



Calculate the Correlation in R

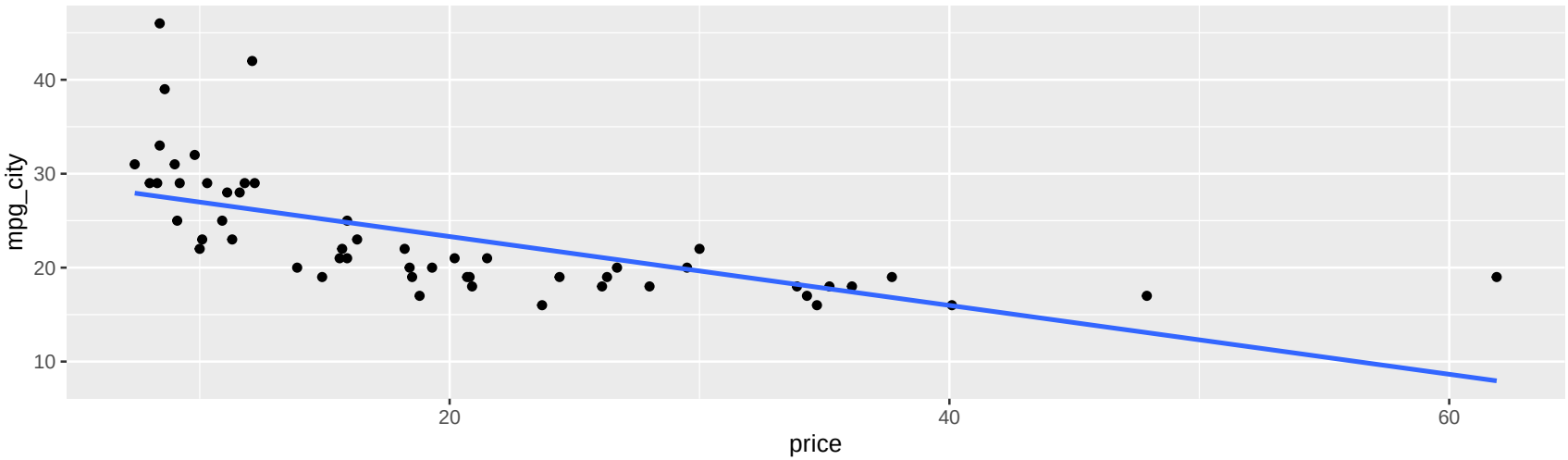
```
cor(cars93$price, cars93$mpg_city)
```

```
[1] -0.6368445
```

How would we describe the linear relationship using the correlation alone?

See (Linear) Line of Best fit in R

```
ggplot(cars93, aes(x = price, y = mpg_city))+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Does this relationship look linear?

More than 2 Variables

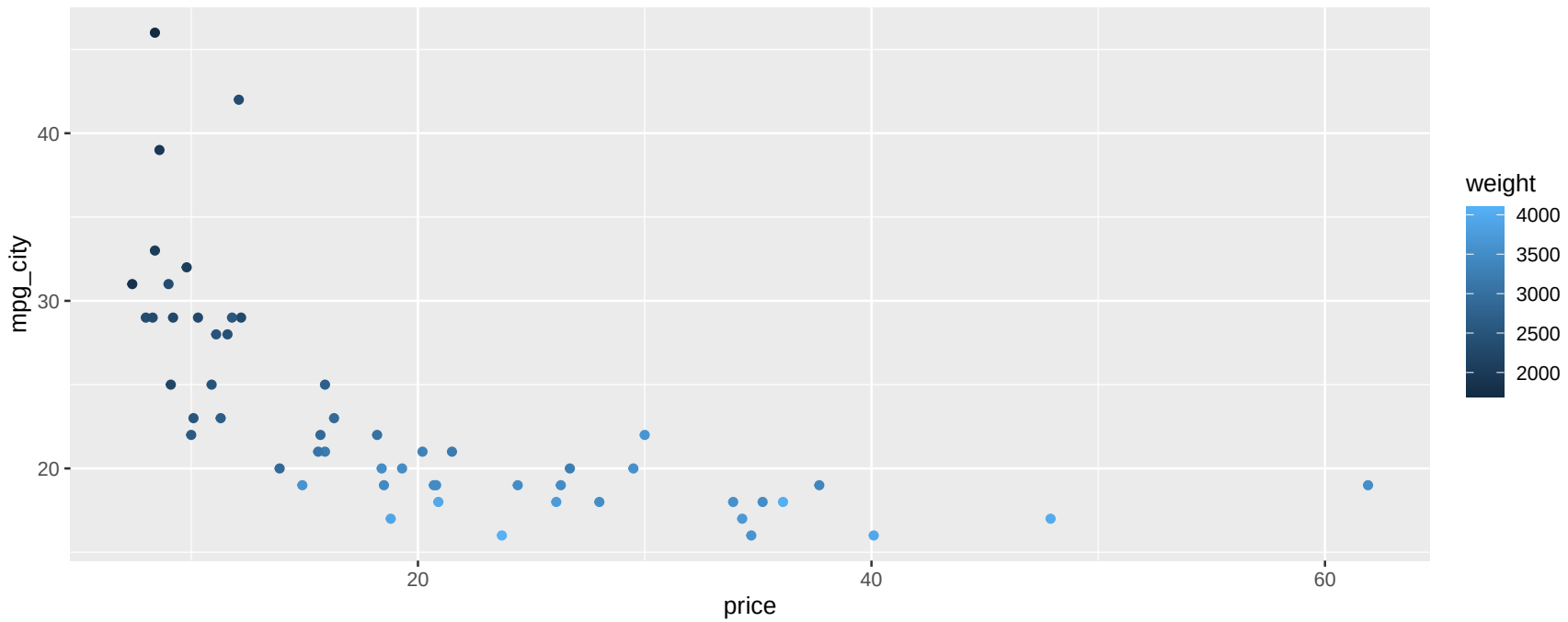
A Few Options At A Glance

- At least two numerical variables $\rightarrow$ Scatterplots with
  - different point colors
  - different point shapes
  - faceting
- Three categorical variables $\rightarrow$ Boxplots with facetting
- Two categorical variables, one numerical $\rightarrow$ Tile plots

And more!

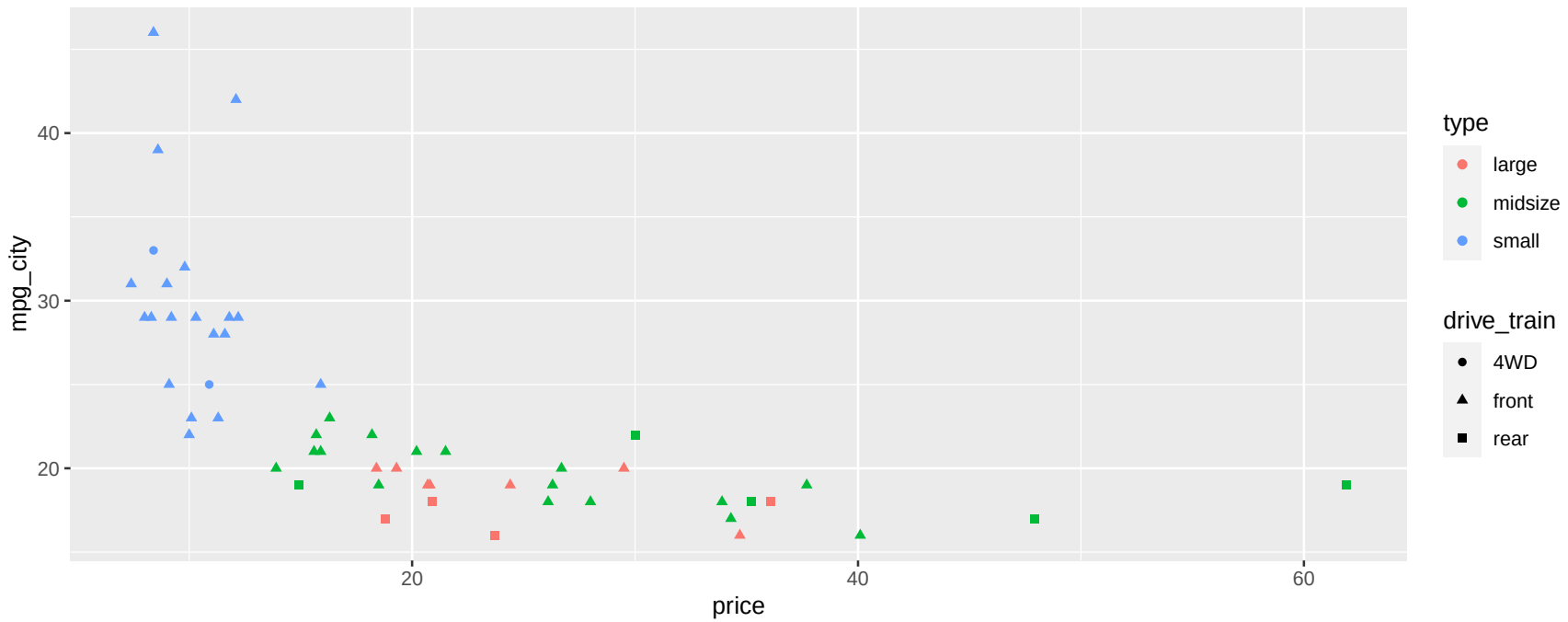
Scatterplots with 3 or more variables

```
ggplot(cars93, aes(x = price, y = mpg_city,
                   col = weight))+
  geom_point()
```



Scatterplots with 3 or more variables

```
ggplot(cars93, aes(x = price, y = mpg_city,
                   shape = drive_train,
                   col = type))+
  geom_point()
```



3 Categorical Variables with Barplots

```
ggplot(cars93, aes(x = type, fill = drive_train)) +
  geom_bar(position = 'stack')+
  facet_wrap(~passengers)
```