Visualizing data

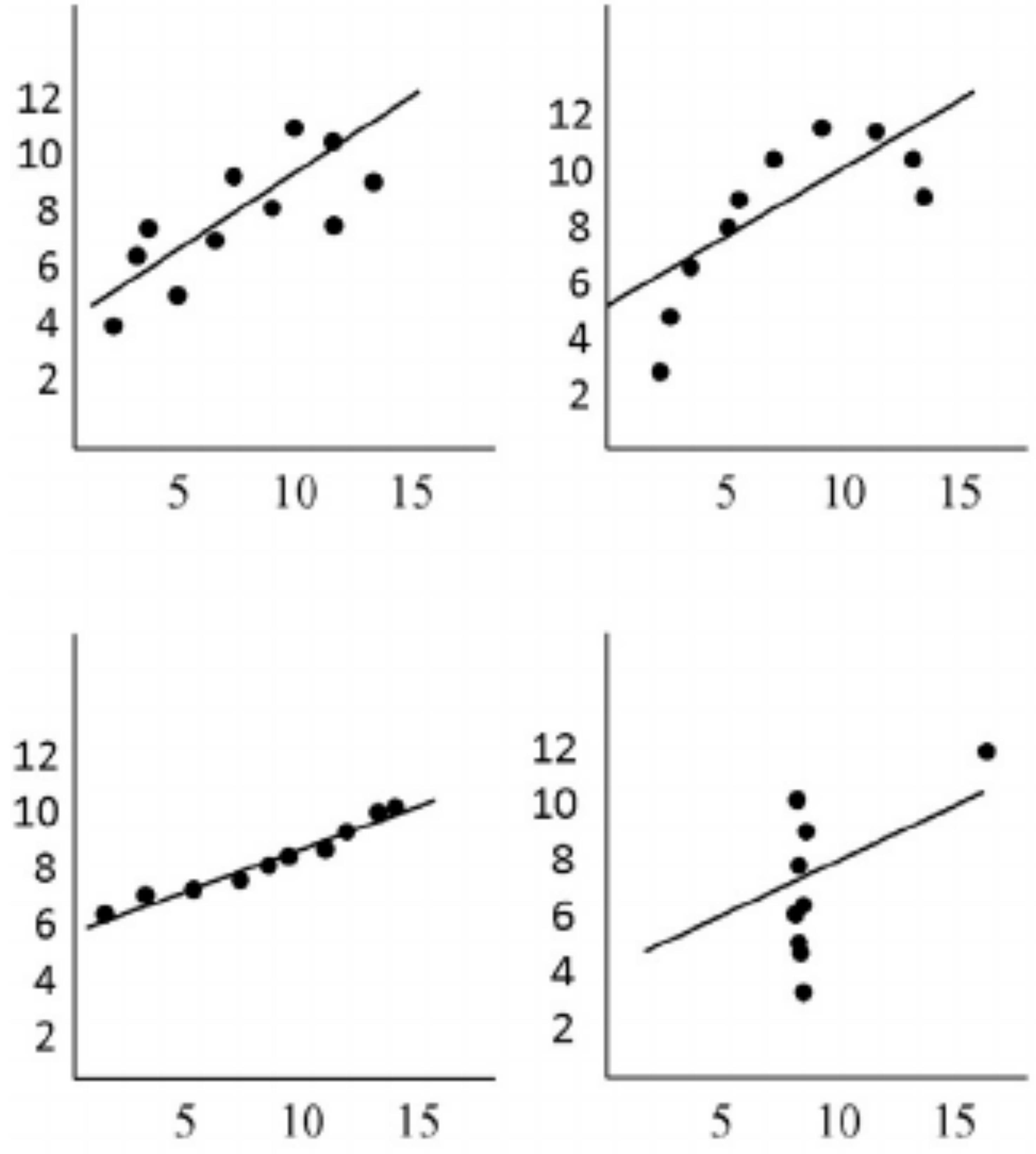
Experimental Research

Professor Velez (11/3/22)

Last class

- Described several data summaries
  - Mean, median, mode
  - Variance, standard deviation
  - Correlation, covariance

Anscombe's Quartet



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

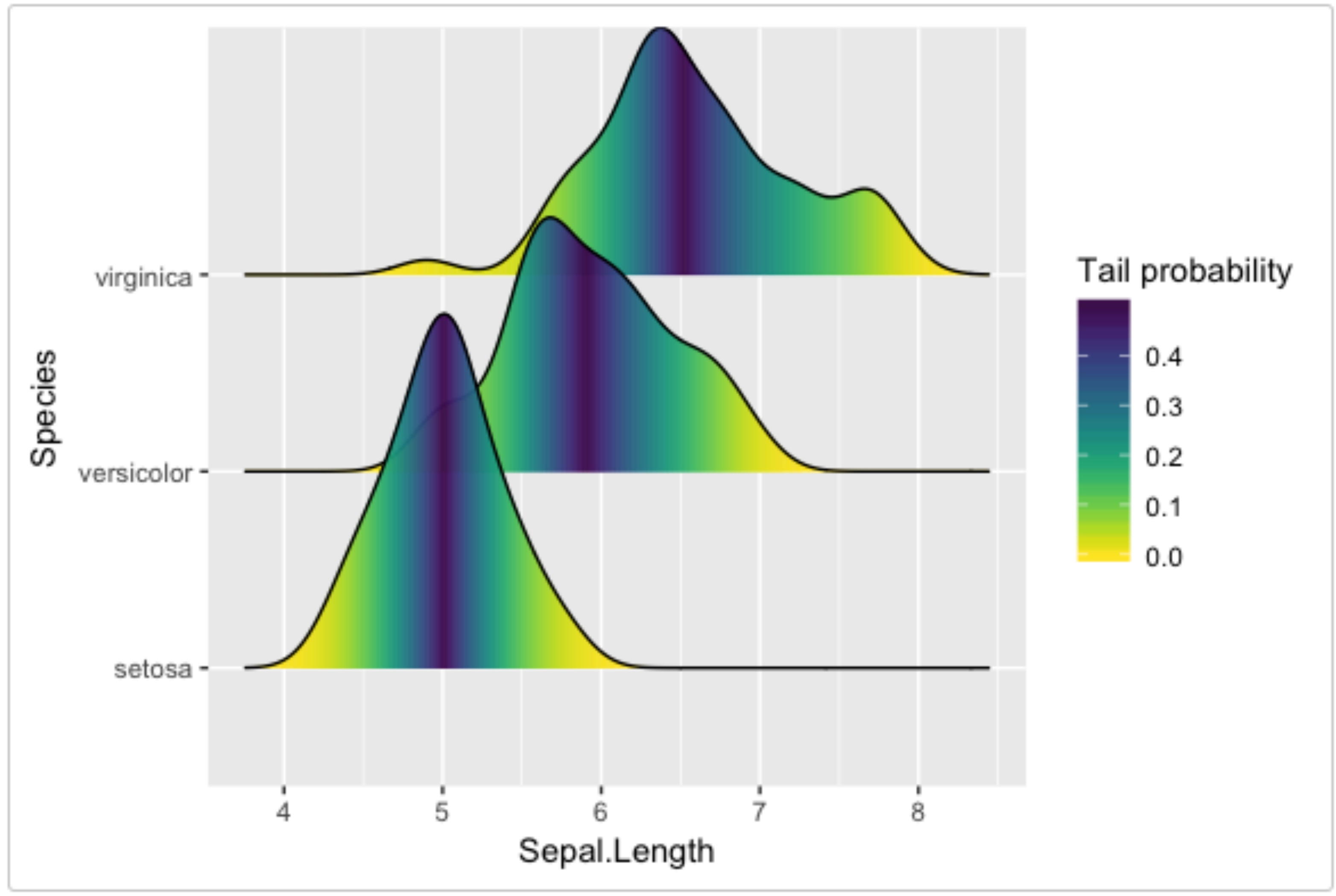
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

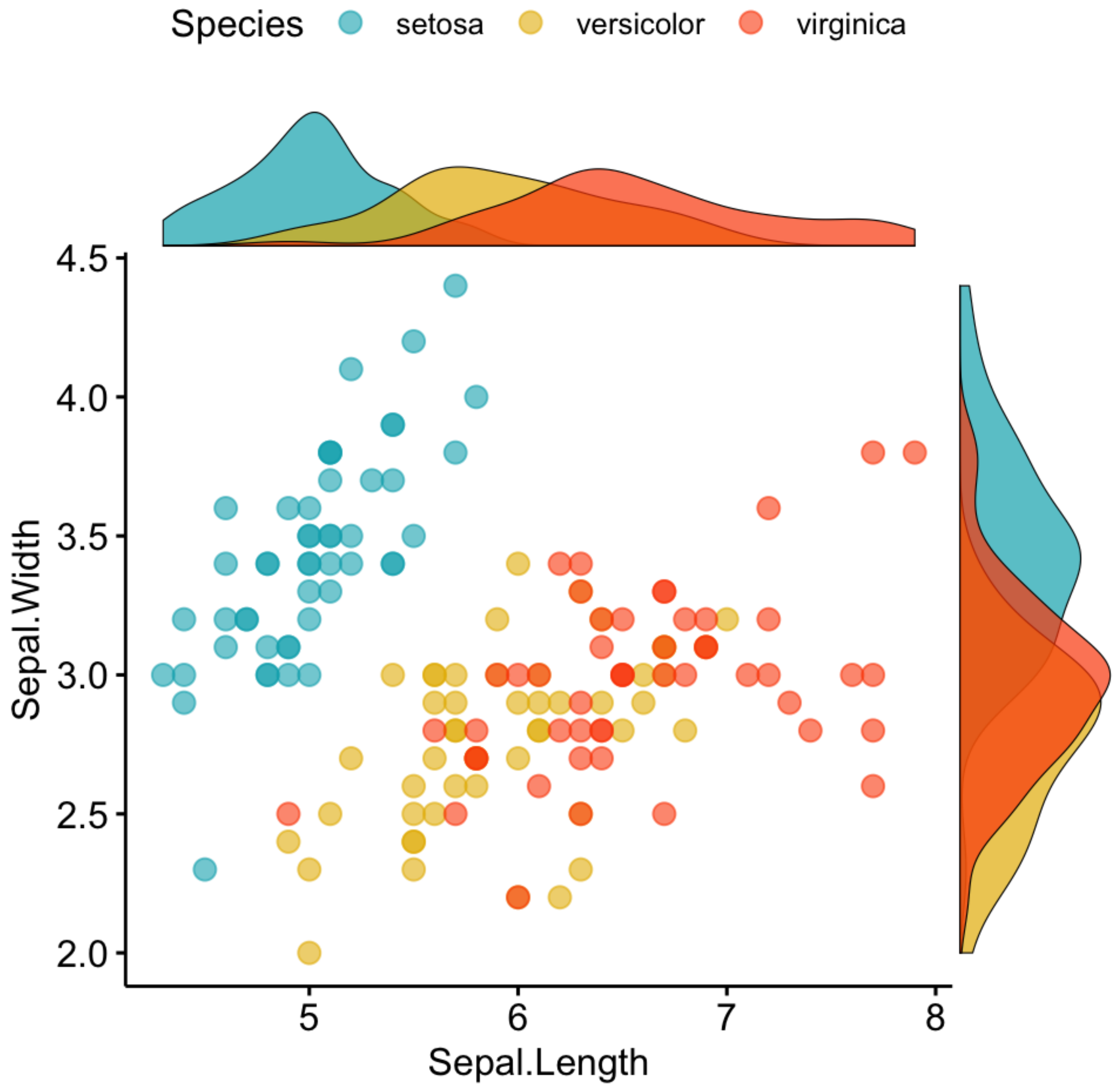
Last class

- Described several data summaries
- Discussed the Tufte philosophy
  - Maximize the data-ink/total-ink ratio
  - Minimize “chart junk”
- Introduce ggplot2
  - One of the most versatile data visualization libraries

ggplot2

- Successor to ggplot
- A data visualization package in *R*
- Powerful library that allows you to create customizable figures

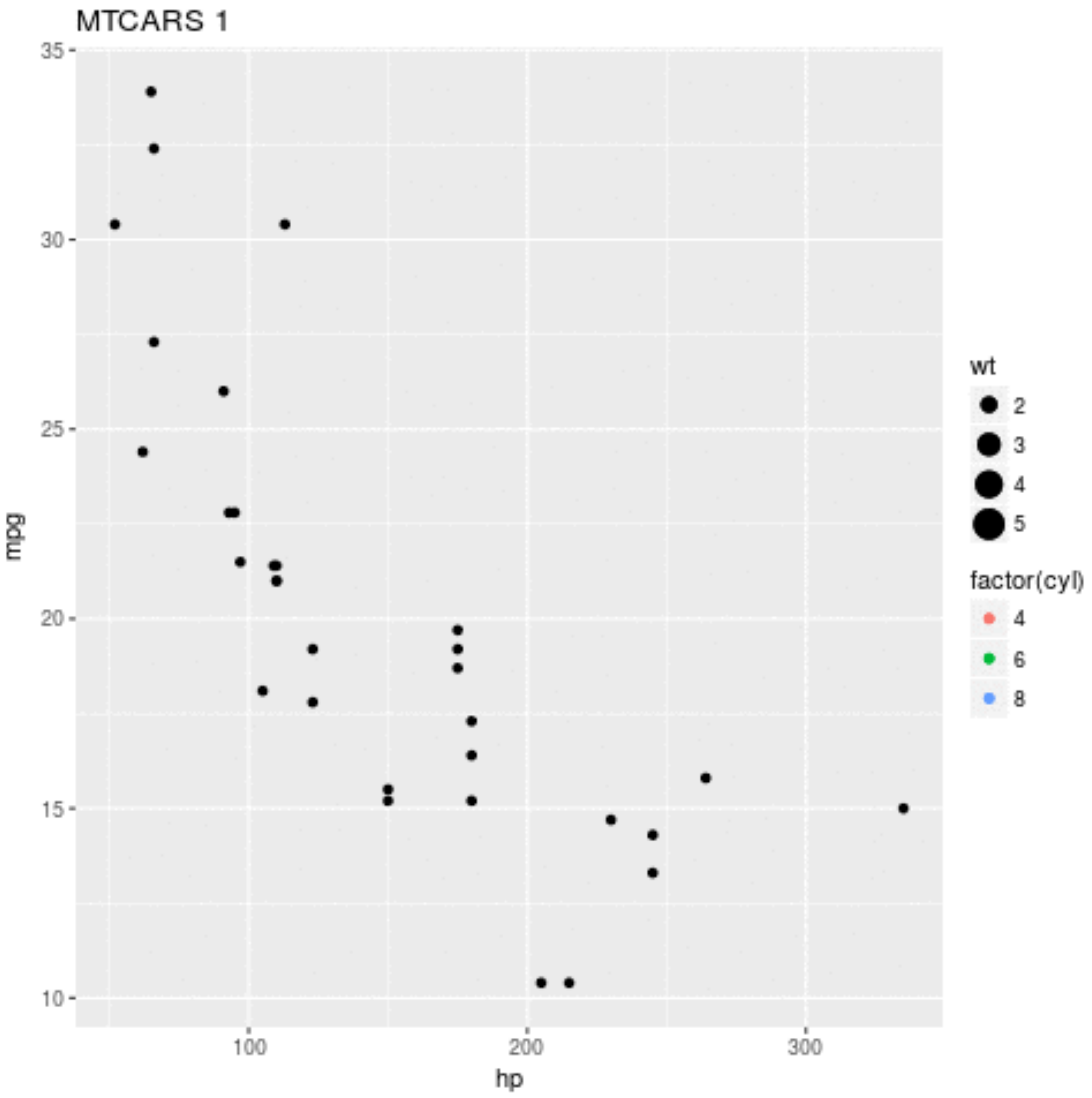




Power comes at a cost

- Involves learning a new grammar
- “The grammar of graphics”
  - Framework which follows a layered approach to describe and construct visualizations
- Some unintuitive aspects of ggplot2, but once you get the hang of it, virtually every aspect of a data visualization is within your control

Customizability



General structure

```
ggplot(data = ---, mapping = aes(x = ---, y = ---)) + geom_----()
```

General structure

```
ggplot(data = ---, mapping = aes(x = ---, y = ---)) + geom_----()
```

ggplot function

General structure

```
ggplot(data = ---, mapping = aes(x = ---, y = ---)) + geom_----()
```

data parameter

General structure

```
ggplot(data = ---, mapping = aes(x = ---, y = ---)) + geom_----()
```

mapping parameter

General structure

```
ggplot(data = ---, mapping = aes(x = ---, y = ---)) + geom_----()
```

layers

General structure

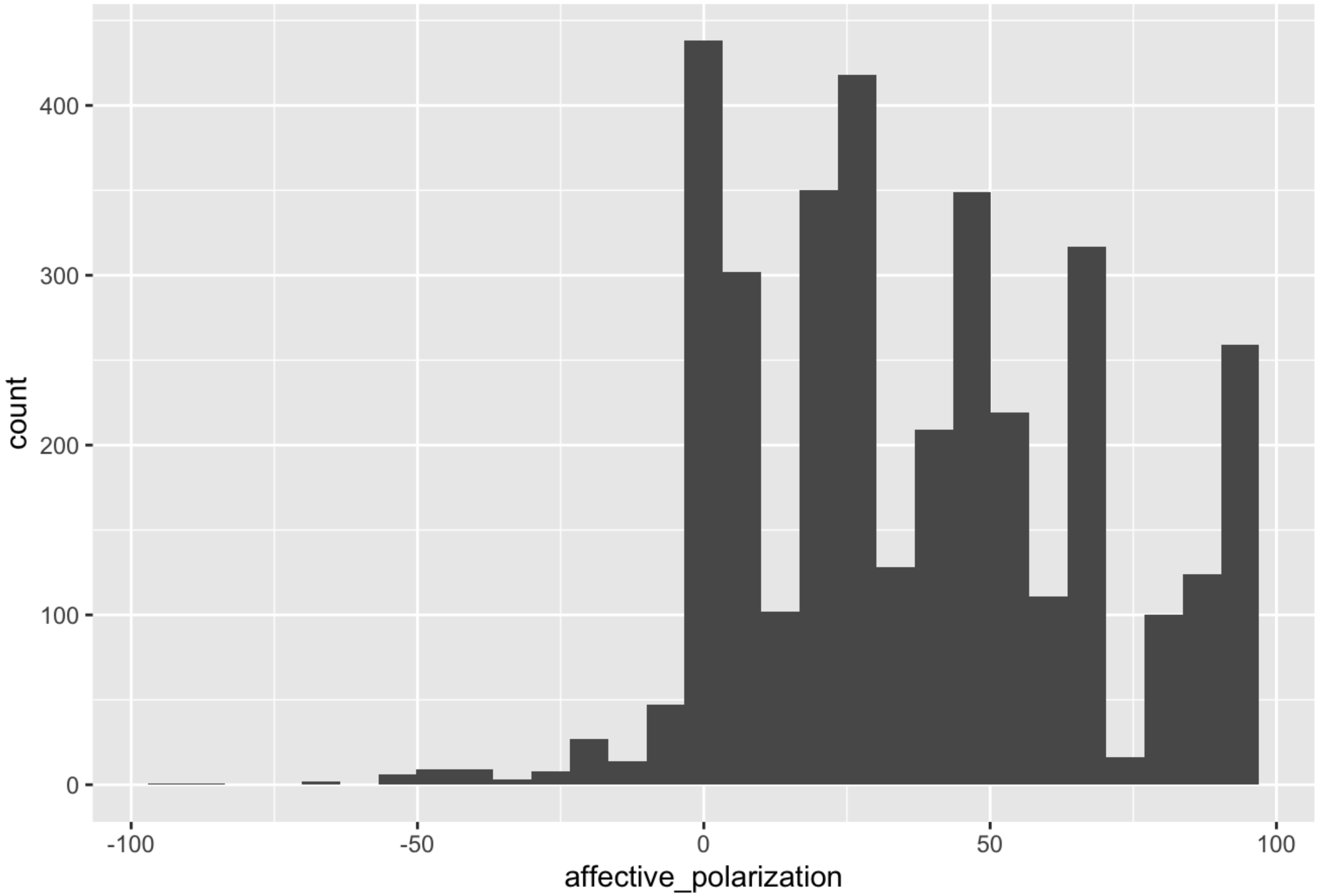
```
ggplot(data = ---, mapping = aes(x = ---, y = ---)) + geom_----()
```

layers

Your first plot

```
ggplot(data = anes, mapping = aes(x = affective_polarization)) + geom_histogram()
```

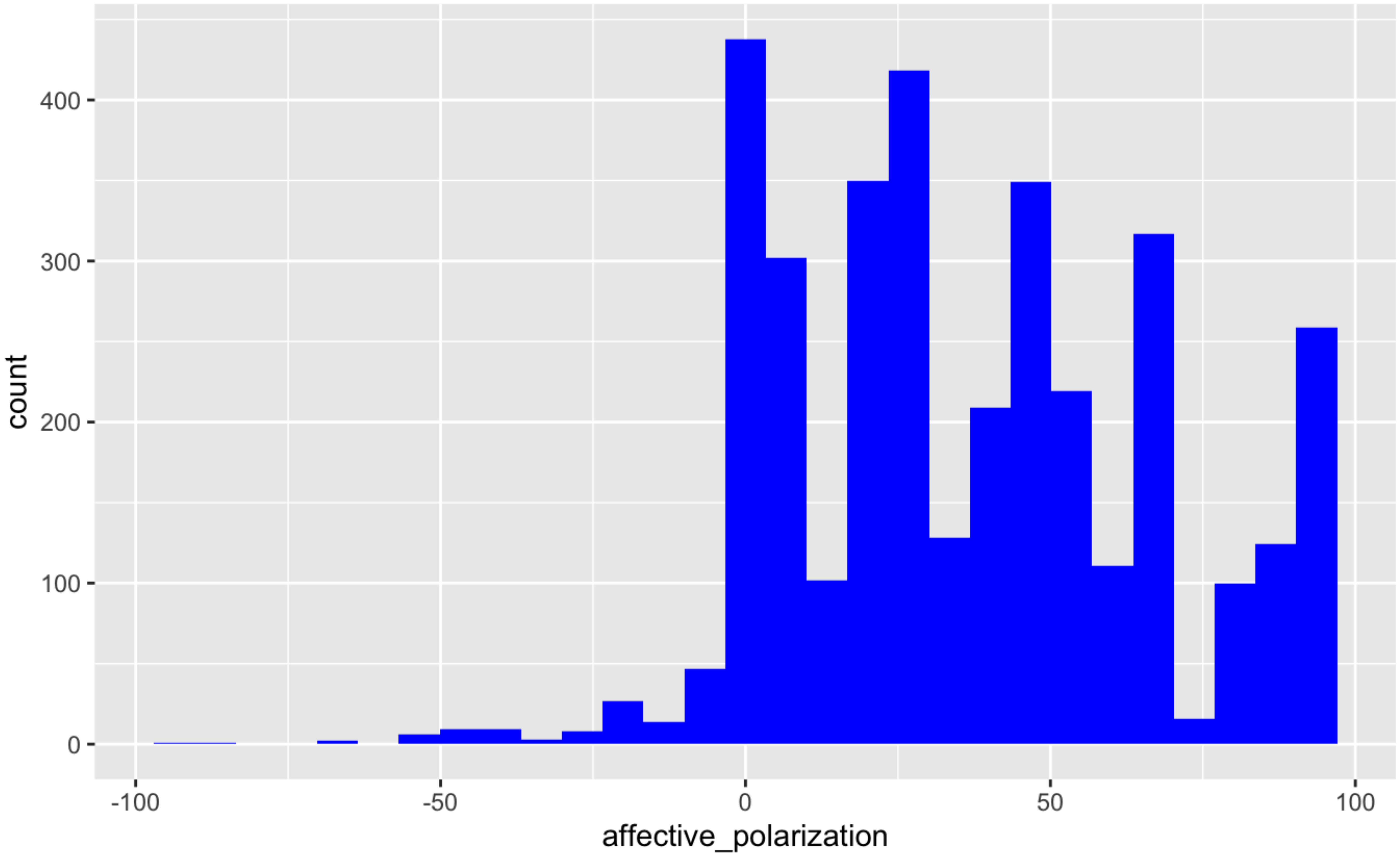
Your first plot



Your first modification

```
ggplot(data = anes, mapping = aes(x = affective_polarization)) +
geom_histogram(fill = “blue”)
```

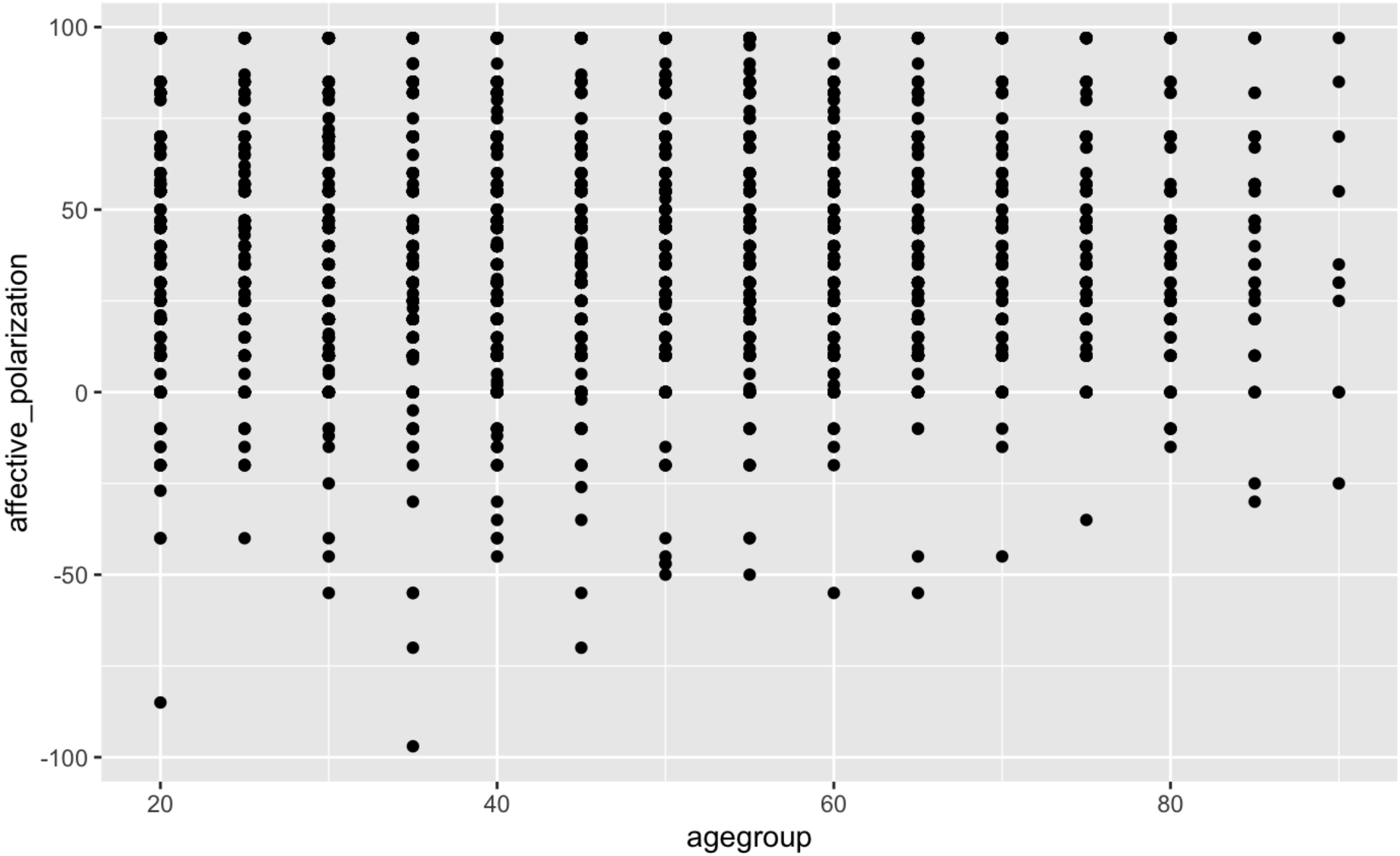
Your first modification



Your first scatterplot

```r
ggplot(data = anes, mapping = aes(x = age, y = affective_polarization)) +
geom_point()
```

Your first scatterplot



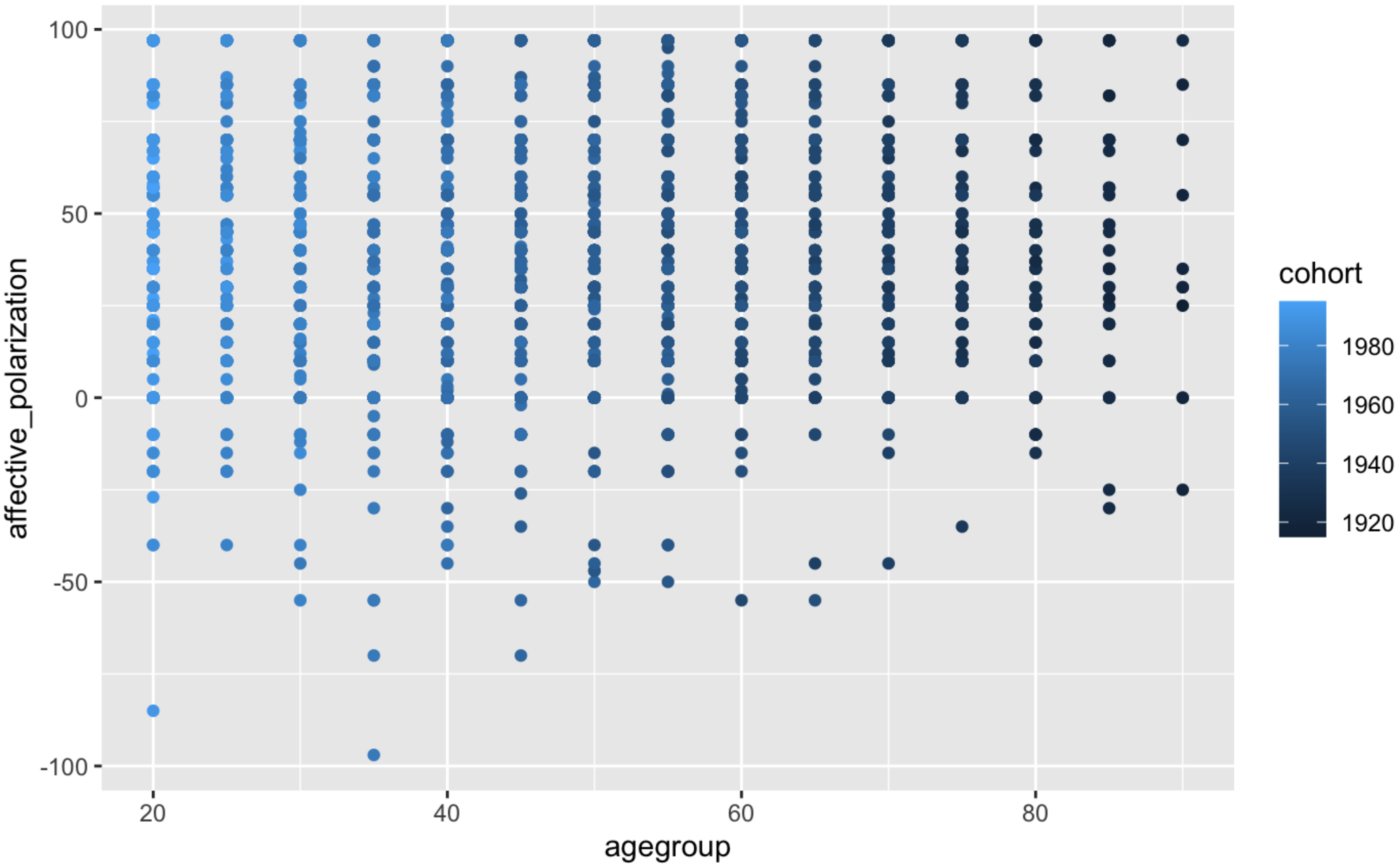
Aesthetics

fill
color
size
linetype
opacity (alpha)
shape

These can be set to fixed values (e.g., fill = “blue”) or they can represent values of variables

Your first data mapping

```
ggplot(data = anes, mapping = aes(x = age, y = affective_polarization, color =
cohort)) + geom_point()
```


affective_polarization
100
50
0
-50
-100
20
40
60
80
agegroup
cohort
1980
1960
1940
1920

Your turn

Experiment

Add color, size, alpha, and shape arguments to your aes() function
Map different aesthetics to different variables
How do continuous and discrete variables affect the aesthetic mappings?

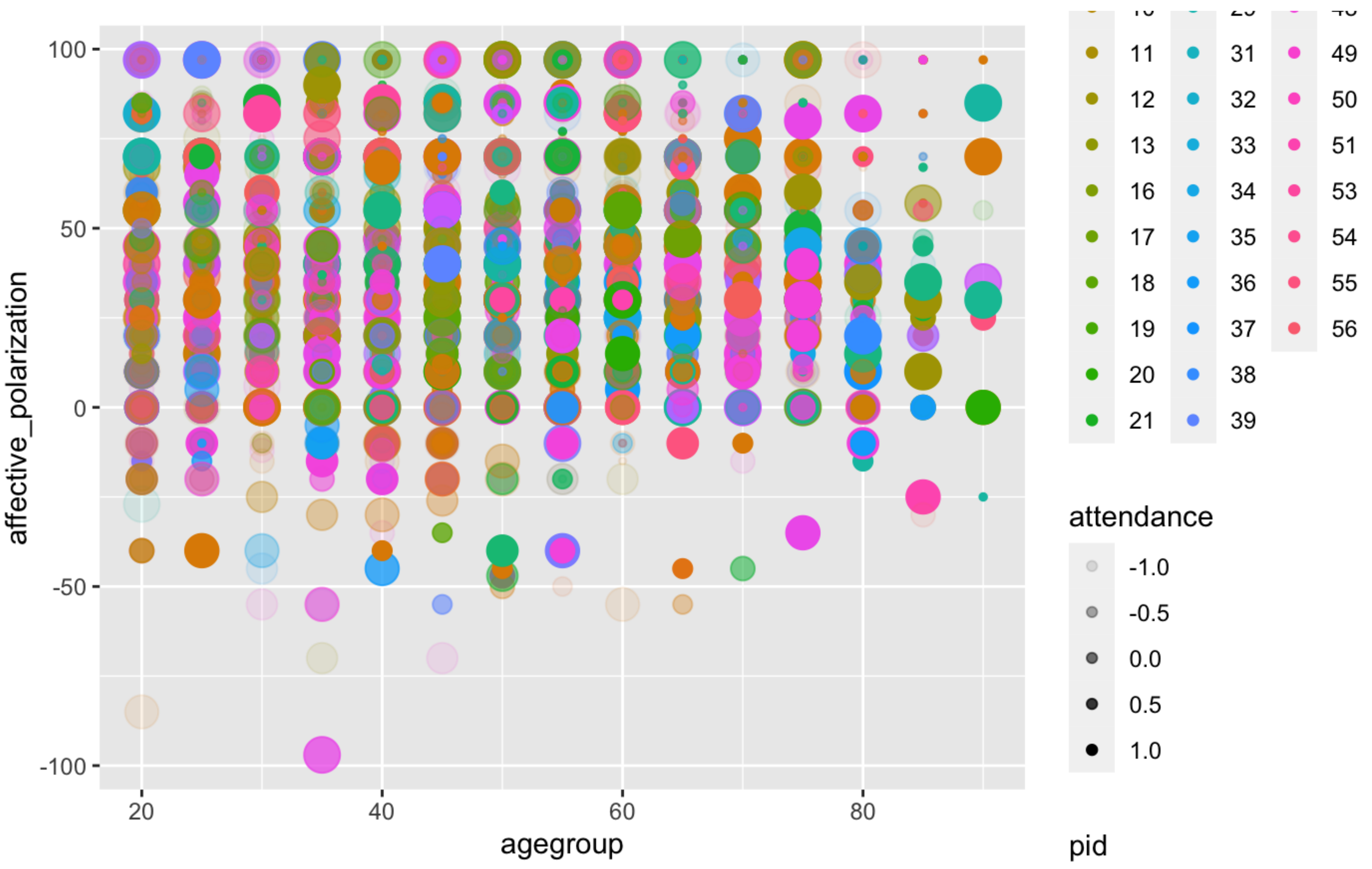
Your turn

Experiment

Add color, size, alpha, and shape arguments to your aes() function
Map different aesthetics to different variables
How do continuous and discrete variables affect the aesthetic mappings?

```
ggplot(data = anes, mapping = aes(x = age, y = affective_polarization, color = …,
shape = …, alpha = …, size = …)) + geom_point()
```

Your turn


affective_polarization
100
50
0
-50
-100
20
40
60
80
agegroup
11
12
13
16
17
18
19
20
21
31
32
33
34
35
36
37
38
39
49
50
51
53
54
55
56
attendance
-1.0
-0.5
0.0
0.5
1.0
pid

Discrete v Continuous

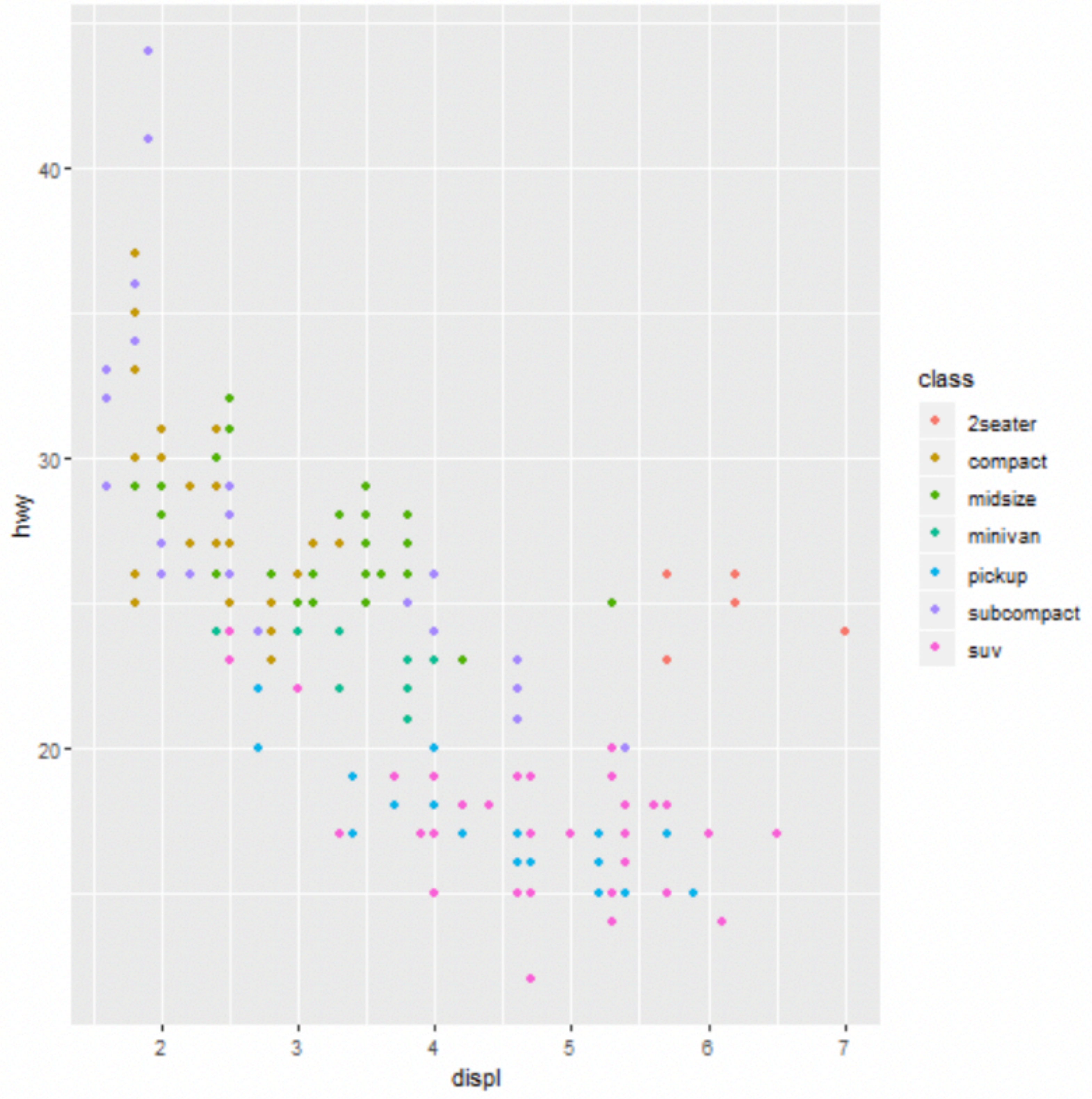
Aesthetics arguments (color, size, shape, etc.) are affected by mode and type of data

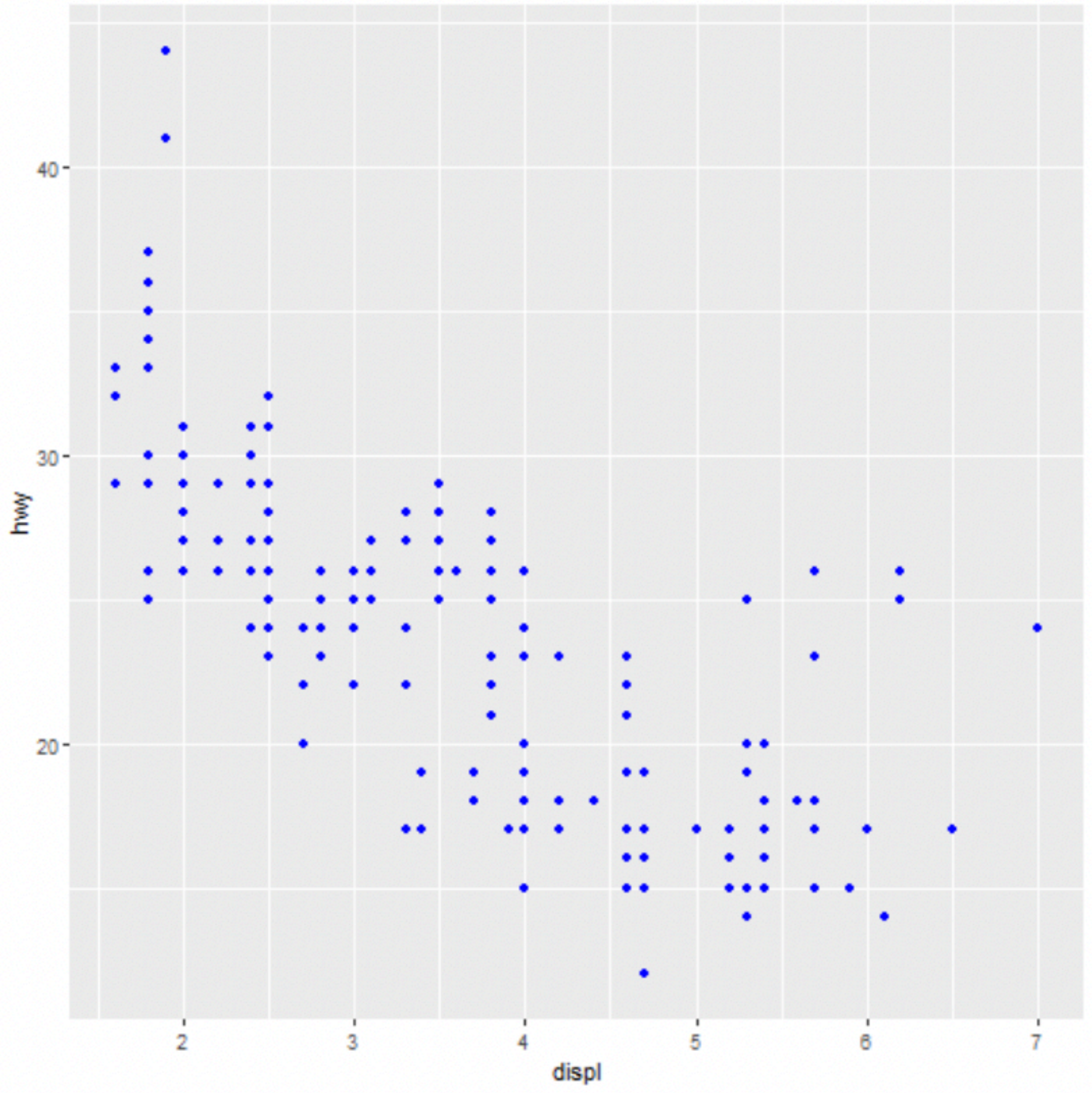
For example, shape argument will not work with continuous data.

Mapping vs. Setting

```
mpg %>%
  ggplot(aes(displ, hwy)) +
geom_point(aes(color = class))
```

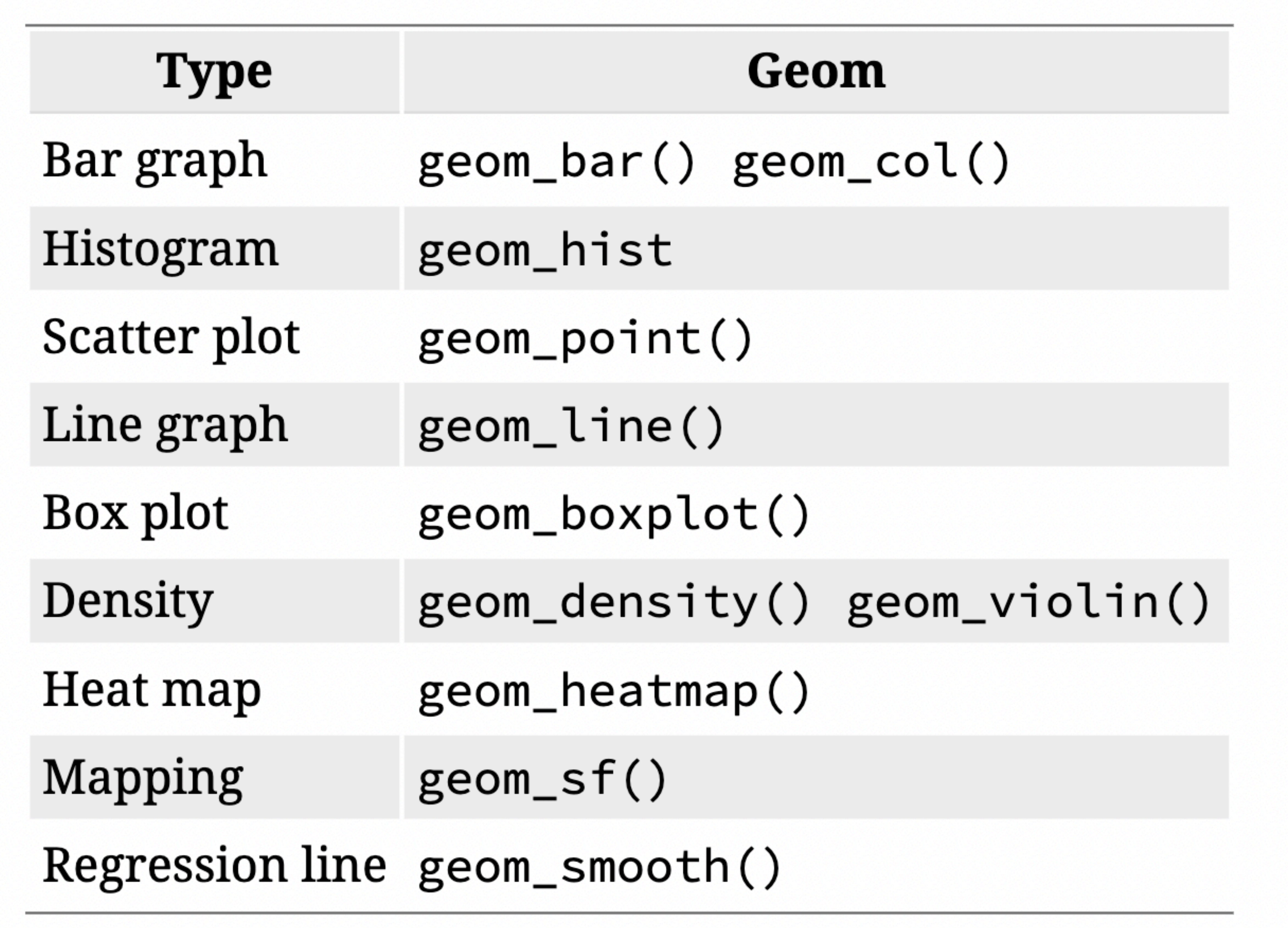
```
mpg %>%
  ggplot(aes(displ, hwy)) +
  geom_point(color = "blue")
```





Geoms

- “Geometric patterns” that can be used to represent the data



Type	Geom
Bar graph	`geom_bar()` `geom_col()`
Histogram	`geom_hist`
Scatter plot	`geom_point()`
Line graph	`geom_line()`
Box plot	`geom_boxplot()`
Density	`geom_density()` `geom_violin()`
Heat map	`geom_heatmap()`
Mapping	`geom_sf()`
Regression line	`geom_smooth()`

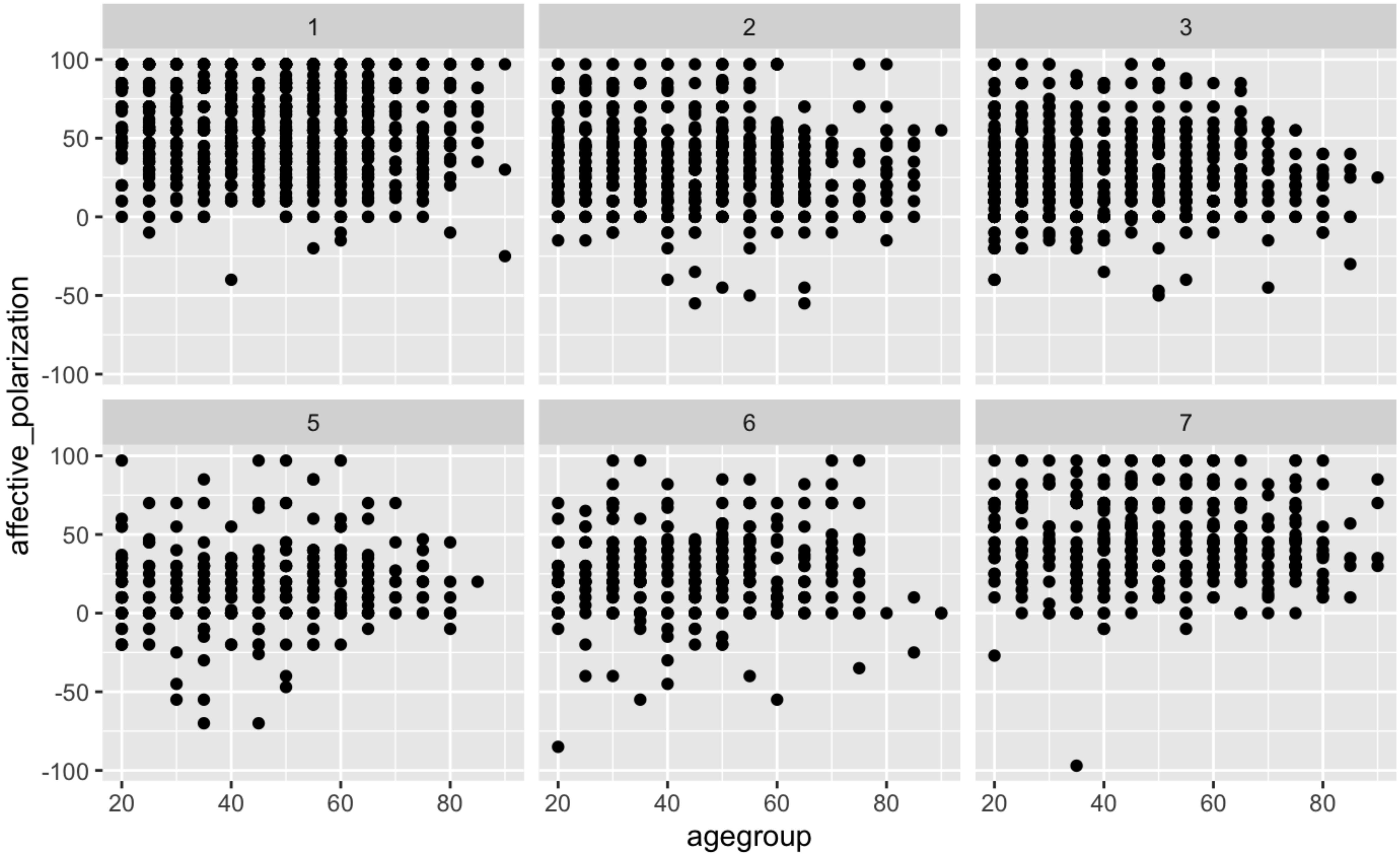
Facet wraps

- Suppose you'd like to create subplots, with each plot determined by the value of a variable
- Example: relationship between age and affective polarization across levels of partisanship

```
ggplot(data = anes, mapping = aes(x = age, y = affective_polarization)) +
geom_point() + facet_wrap()
```

Facet wraps

- Suppose you'd like to create subplots, with each plot determined by the value of a variable
- Example: relationship between age and affective polarization across levels of partisanship



Labels

- Most times you’ll want to customize the labels
- Add a layer!

```
ggplot(data = anes, mapping = aes(x = age, y = affective_polarization)) +
geom_point() + facet_wrap() + labs(x = “Age”, y = “Affective Polarization”, title
= “Scatterplot of Age and Affective Polarization”)
```