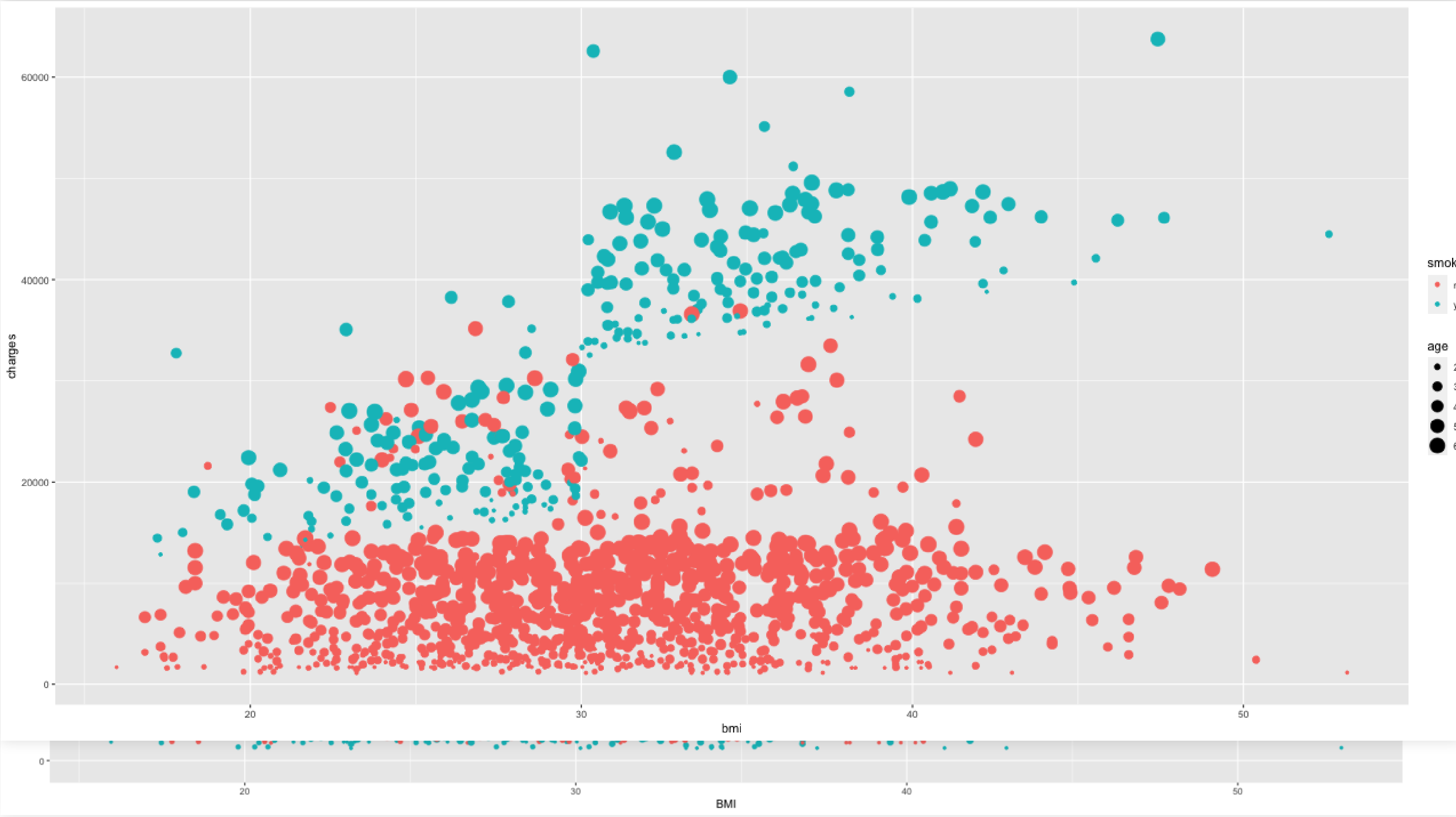
Module 6: Data



Data Visualization

1. Be able to plot in ggplot
2. Know what types of figures you want
3. Use plots to gain information

1. ggplot syntax
2. aesthetics mapping
3. geometry
4. labels

What is ggplot?

-
  - Grammar of graphics abstraction of graphics ideas “Shorten the distance from mind to page”
  - ggplot is a data visualization package, that is part of the tidyverse suite of packages
-

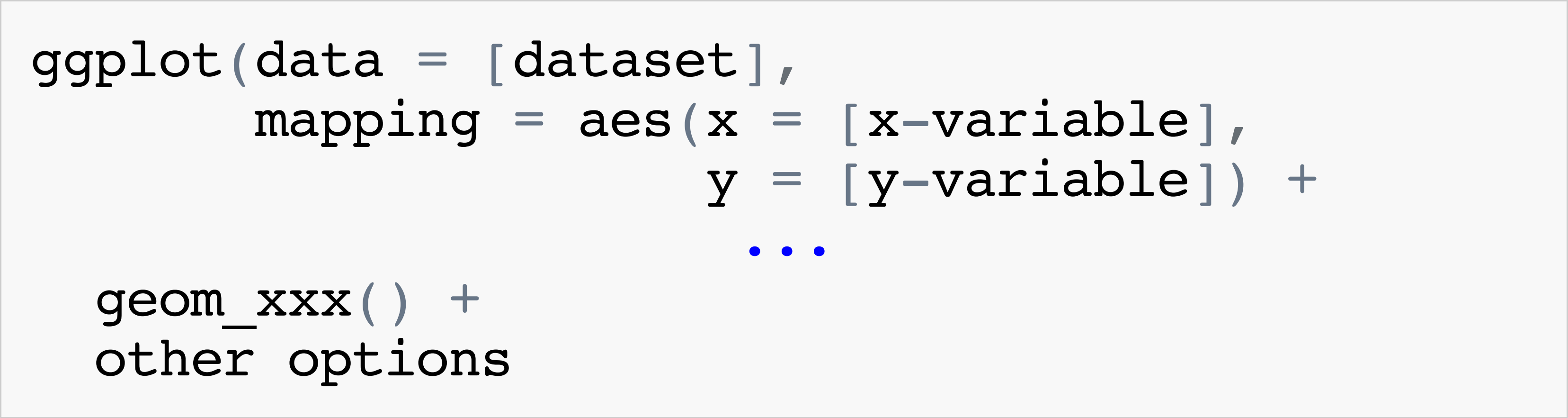
```r
library(tidyverse)
```

Basic Components of ggplot (Layers)

Basic Components of ggplot (Layers)

- Layer 1: Background layer (ggplot())
  - A data frame
  - Aesthetic mapping: how data are mapped to x-axis, y-axis, color, size, etc
- Layer 2: Geometry layer (geom_xxx())
  - geometric objects like points, lines, shapes
- Layer 3: Labels Layer (labs())
  - title, legend, etc
- Others…

Simplest ggplot Code Structure

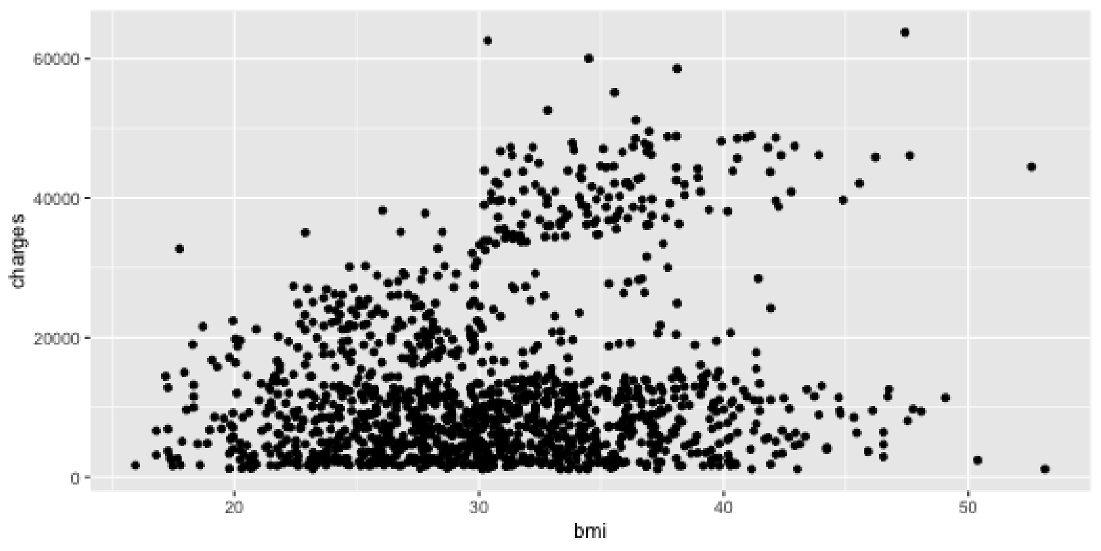


```
ggplot(data = [dataset],
       mapping = aes(x = [x-variable],
                     y = [y-variable]) +
                      ...
  geom_xxx() +
  other options
```

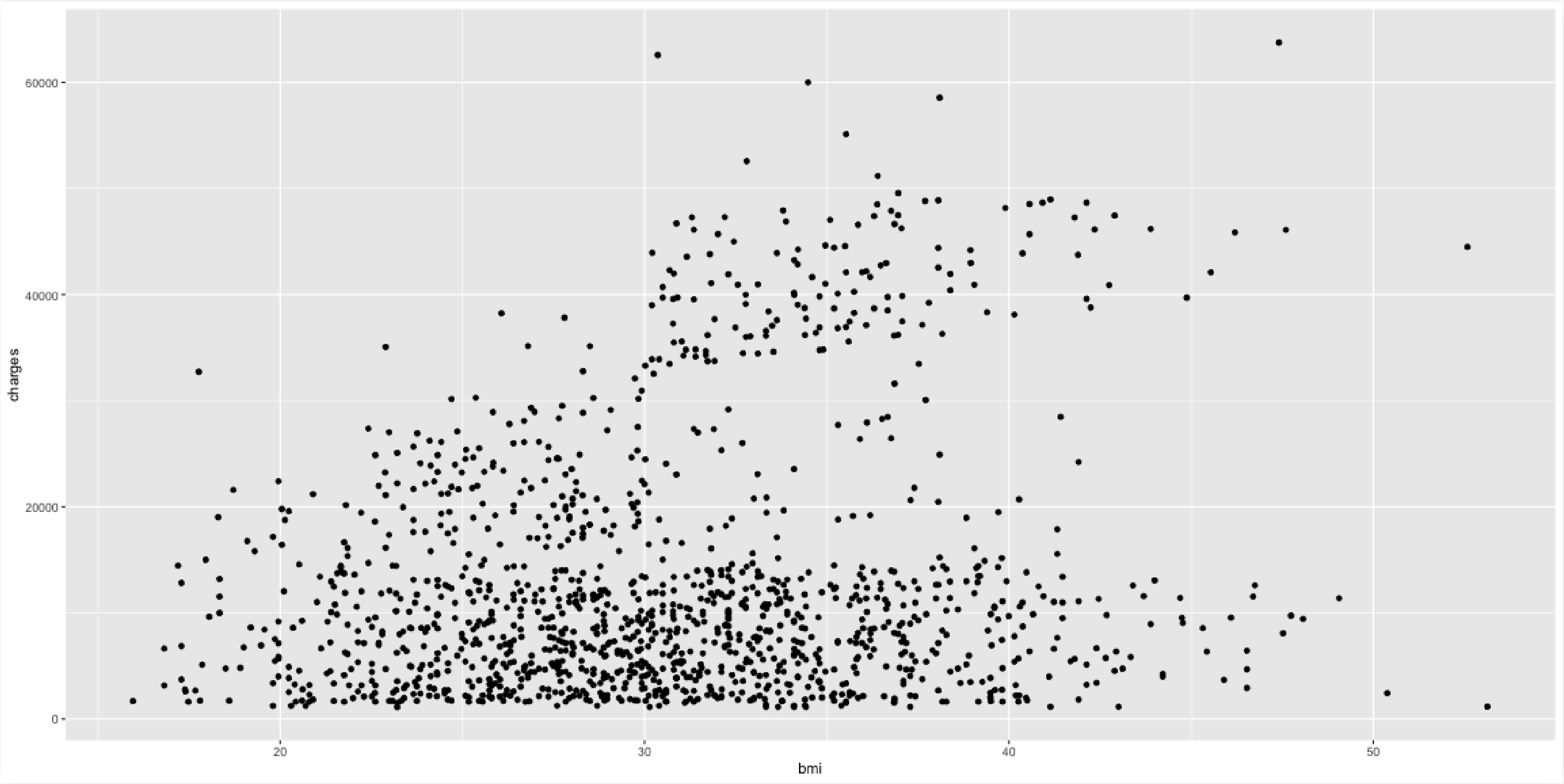
BMI vs. Charges

```
insurance = read.csv('insurance.csv')
```

```
ggplot(data = insurance,
       mapping = aes(x = bmi,
                     y = charges)) +
  geom_point()
```



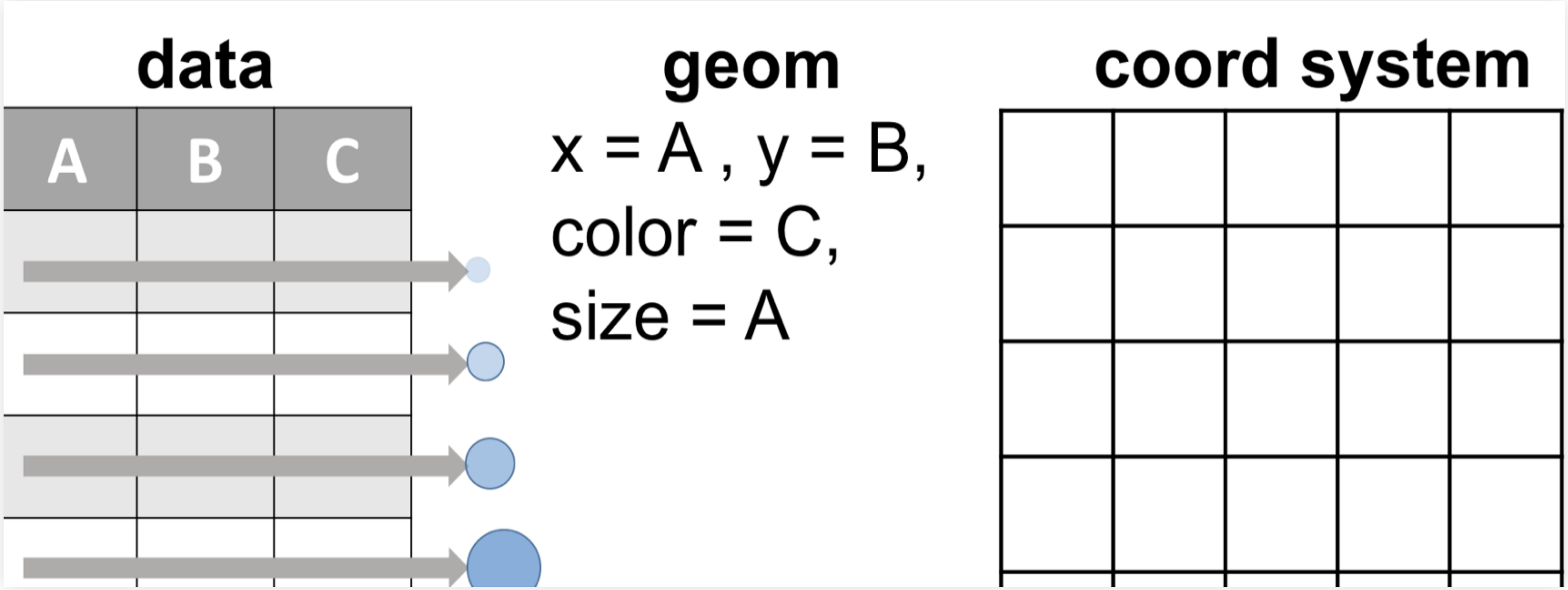
How would you describe this relationship? What other variables would help us understand data points that don't follow the overall trend?



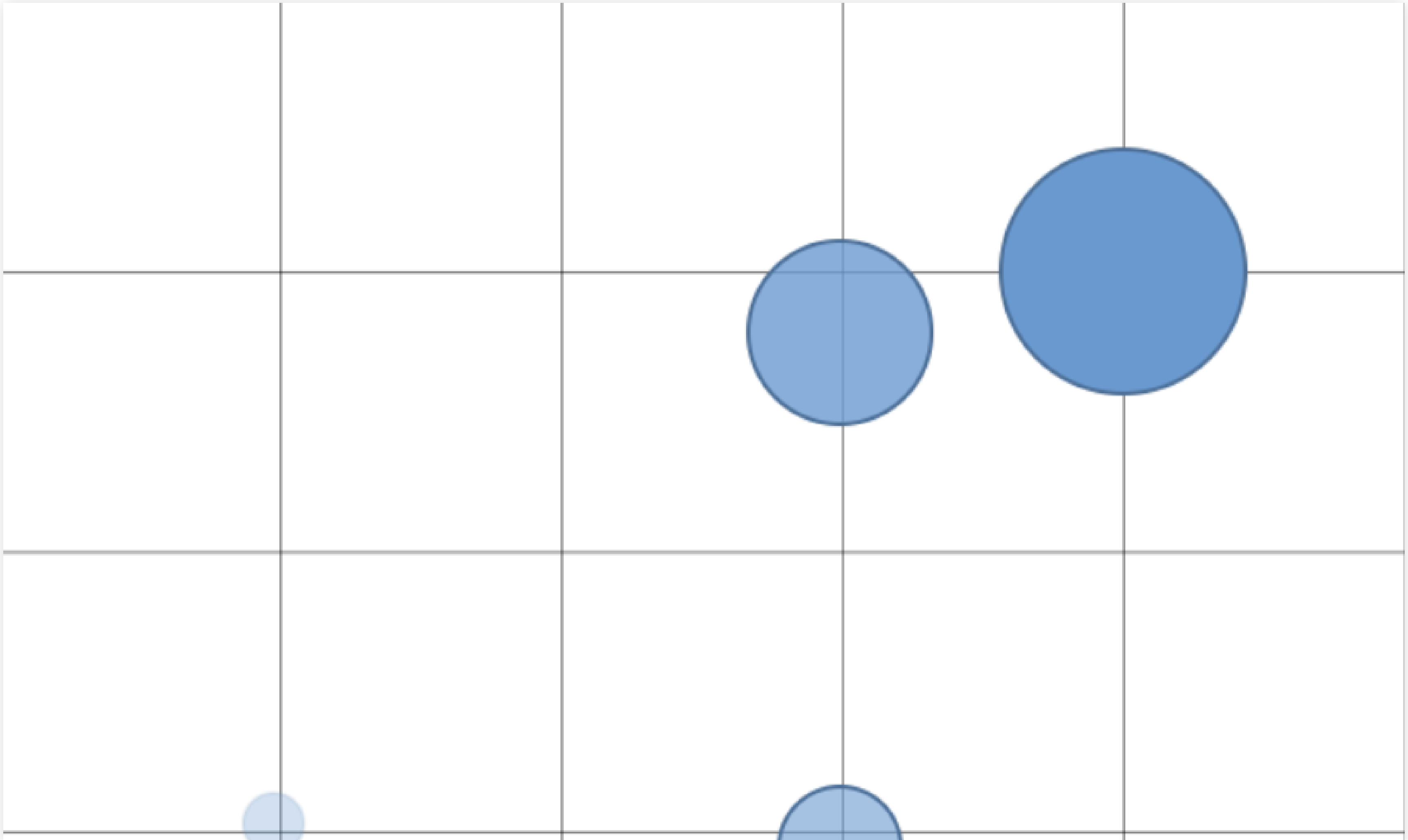
Aesthetic Mapping

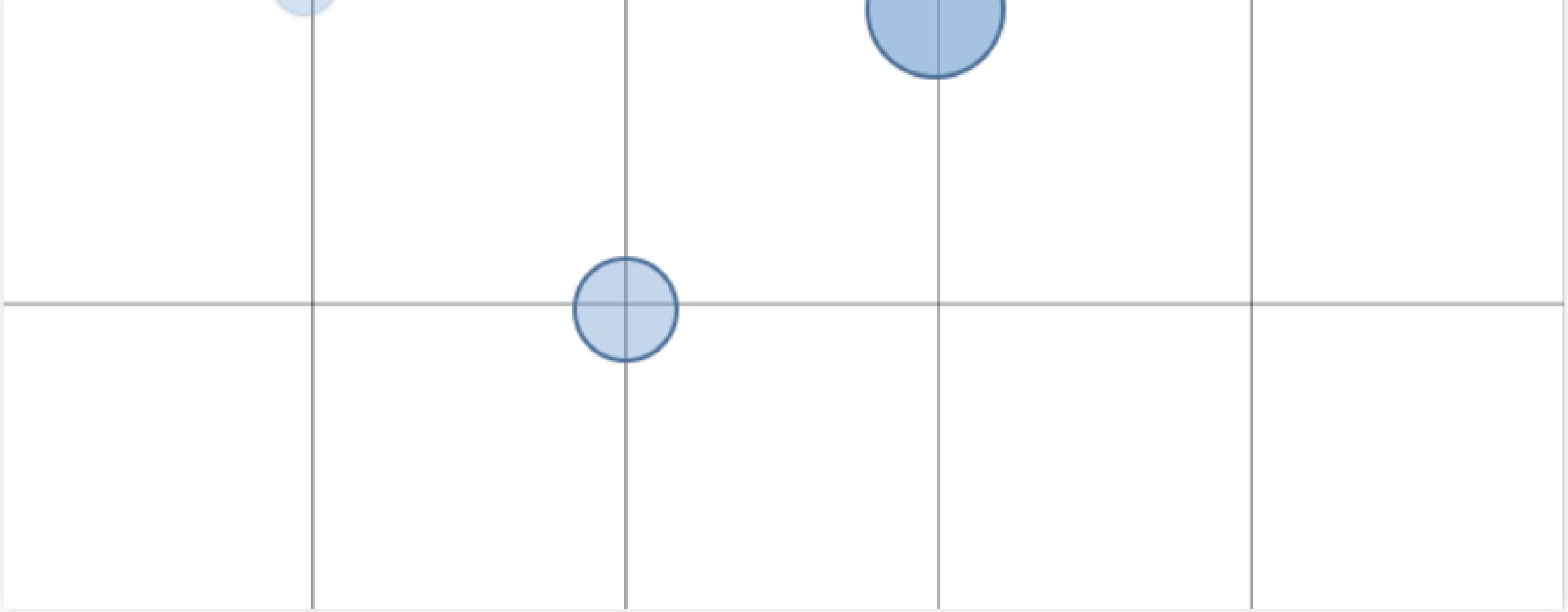
What is Aesthetic Mappings?

- To display values, map variables in the data to visual properties of the geom (aesthetics)
- An aesthetic is a visual property of the objects in your plot
- Including things like size, shape, color or x and y locations



Aesthetic Mappings

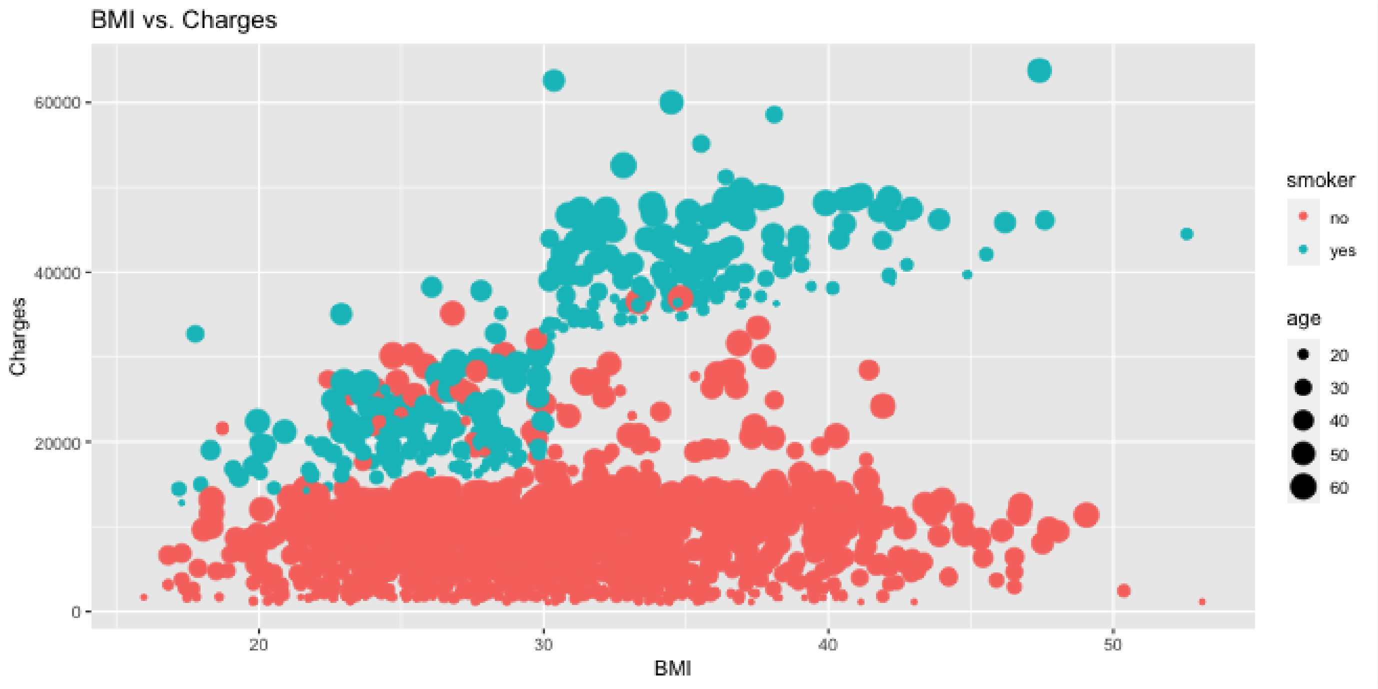




bmi vs. charges + smoker + age

Adding Labels

```
ggplot(data=insurance,
       mapping=aes(x=bmi,
                   y=charges,
                   color=smoker,
                   size=age)) +
  geom_point() +
  labs(title="BMI vs. Charges", x= "BMI",
y="Charges")
```


BMI vs. Charges
60000
40000
20000
0
Charges
20
30
40
50
BMI
smoker
no
yes
age
20
30
40
50
60

Geometry: Type of the Figures

- What types of figures can ggplot plot?
- How should we choose the type of figures?
- How to read different types of figures?

How to Choose Type of Figures?

- Number of variables
- Type of variables

Number of variables involved

-
    - distribution of single variable
    - bar plot, histogram, density plot, etc
-
    - relationship between two variables
    - scatter plot, line plot, boxplot, (segmented) bar plot, etc
-
    - relationship between many variables at once, usually focusing on the relationship between two while conditioning for others

Types of variables

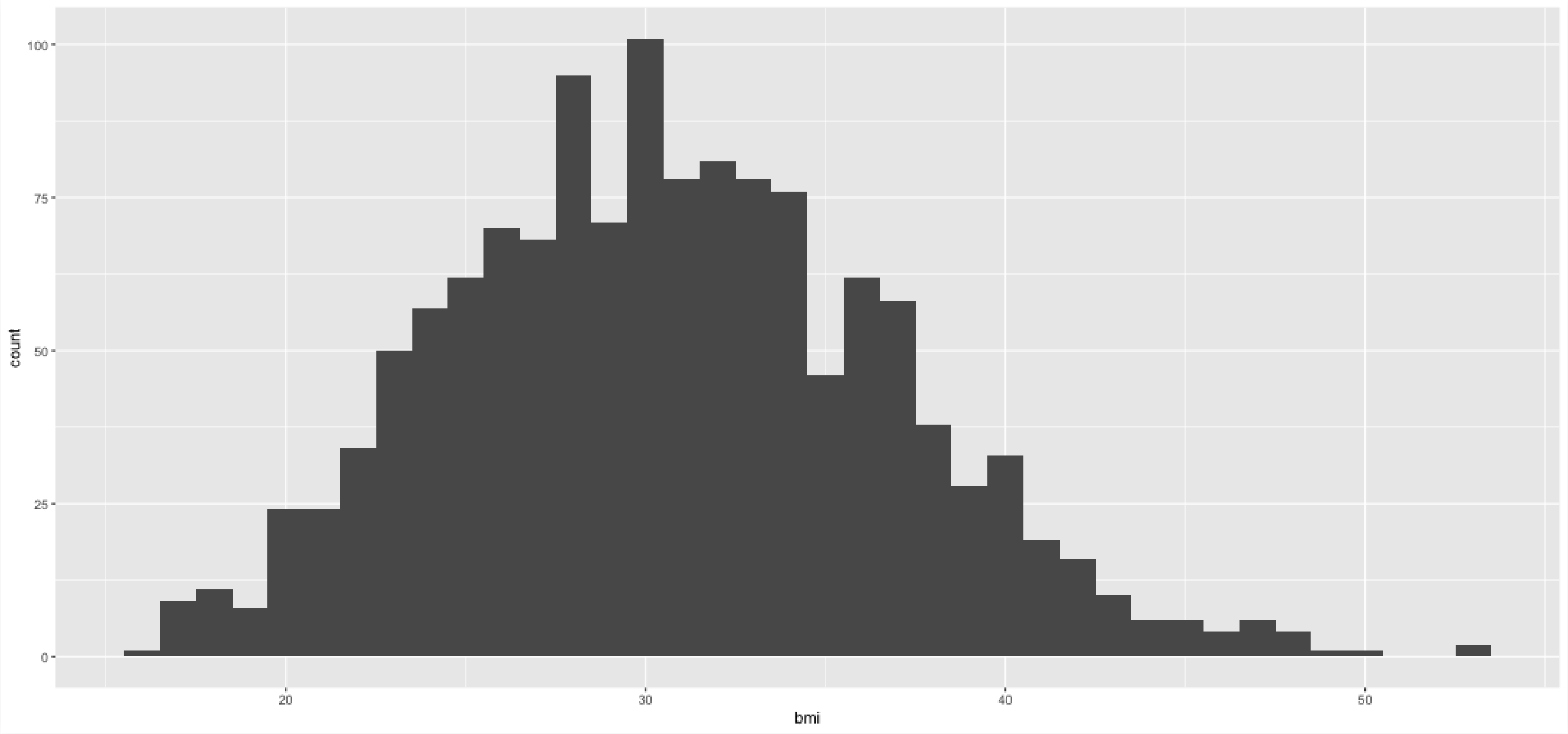
-
  - continuous: BMI
  - discrete: age
  - Some typical plot types include scatter plot, histogram, box plot, density plot
-
  - ordinal: education - highschool, some college, college degree
  - non-ordinal: gender
  - Some typical plot types include bar plots and ordered bar plots

Visualizing Univariant Data

-
  - histogram, density plot
  - geom_histogram(), geom_density()
-
  - bar plot
  - geom_bar()

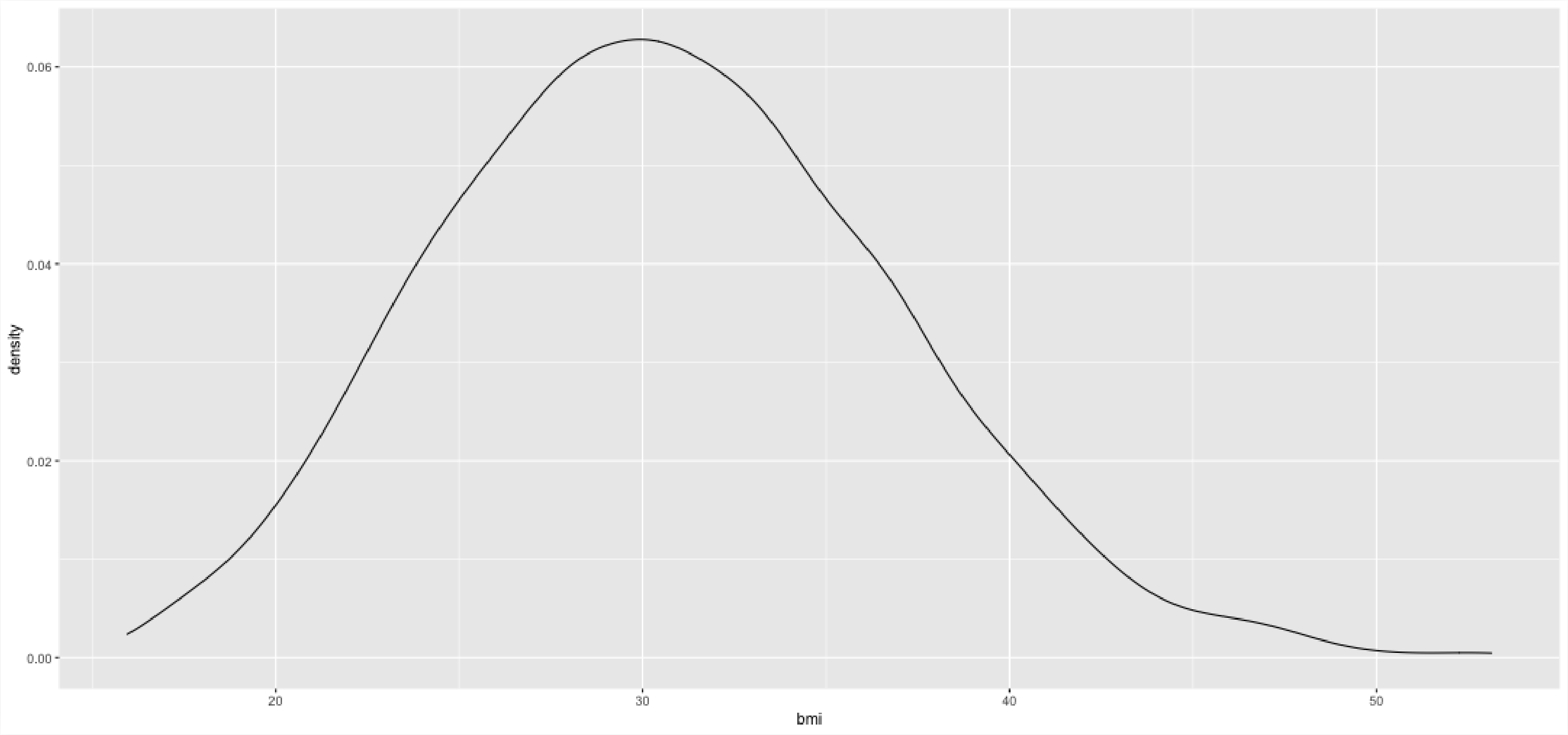
Histograms

```
ggplot(data = insurance, mapping = aes(x =
bmi)) +
  geom_histogram(binwidth = 1)
```



Density PLots

```
ggplot(data = insurance, mapping = aes(x =
bmi)) +
  geom_density()
```

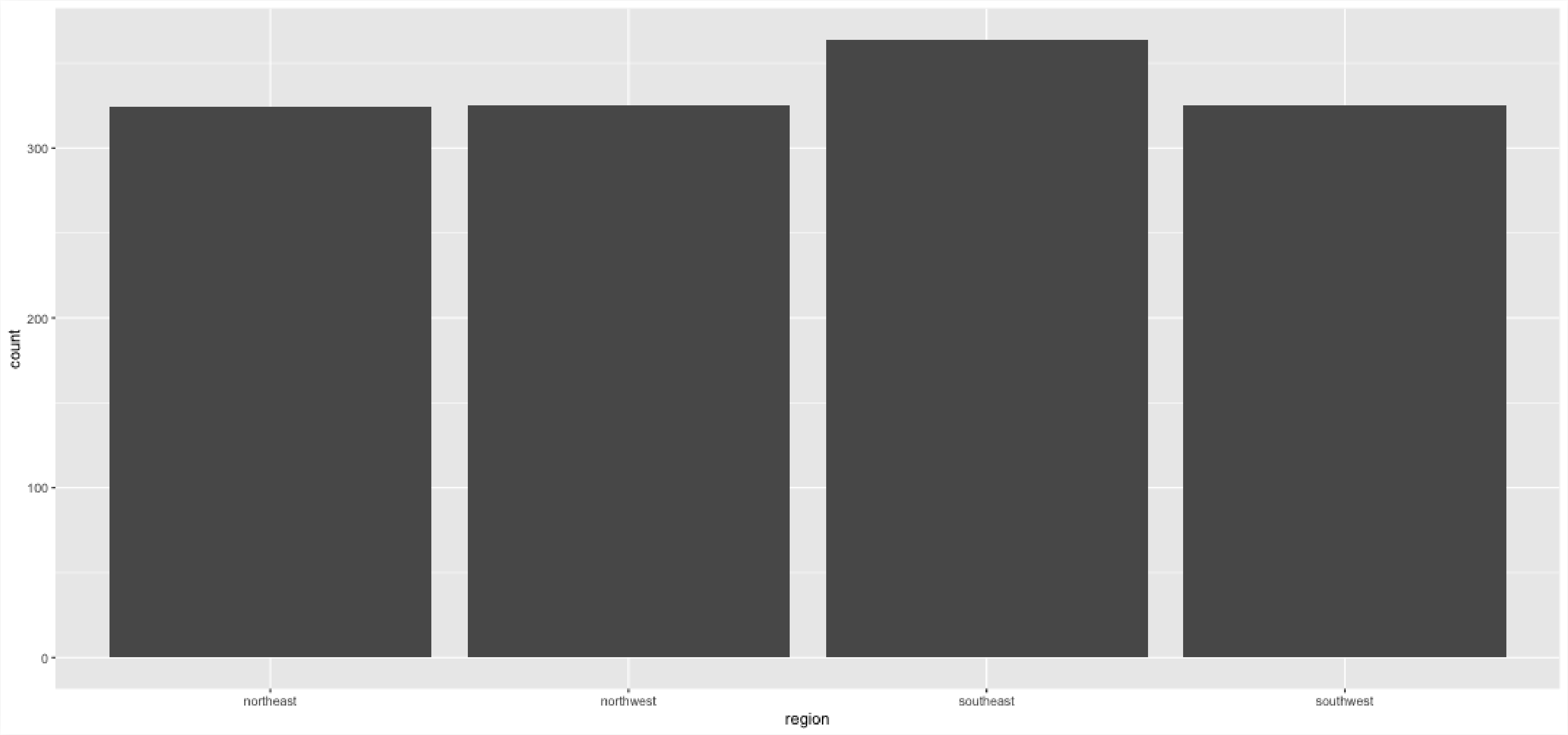


Describing shapes of numerical distributions

- modality: unimodal, bimodal, multimodal, uniform
- skewness: right-skewed, left-skewed, symmetric (skew is to the side of the longer tail)
- center: mean (mean), median (median), mode (not always useful)
- spread: range (range), standard deviation (sd)
- unusual observations

Visualizing Univariant Categorical Data:Bar Plots

```
ggplot(data = insurance, mapping = aes(x =
region)) +
  geom_bar()
```

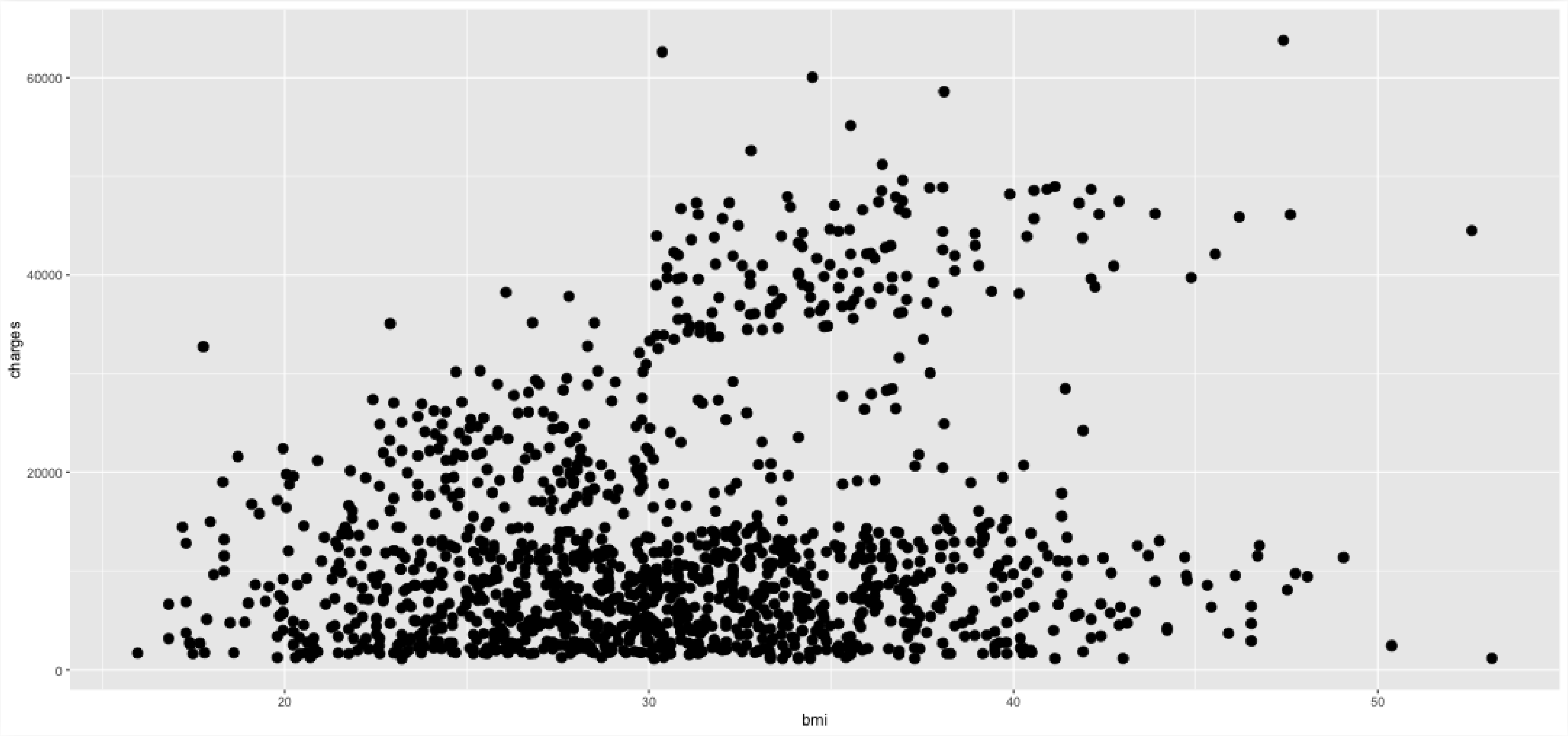


Visualizing Bivariate Data

-
  - scatter plot, line plot
  - geom_point(), geom_smooth()
-
  - box plot, bar plot
  - geom_boxplot(), geom_bar()
-
  - (segmented) bar plot
  - geom_bar()

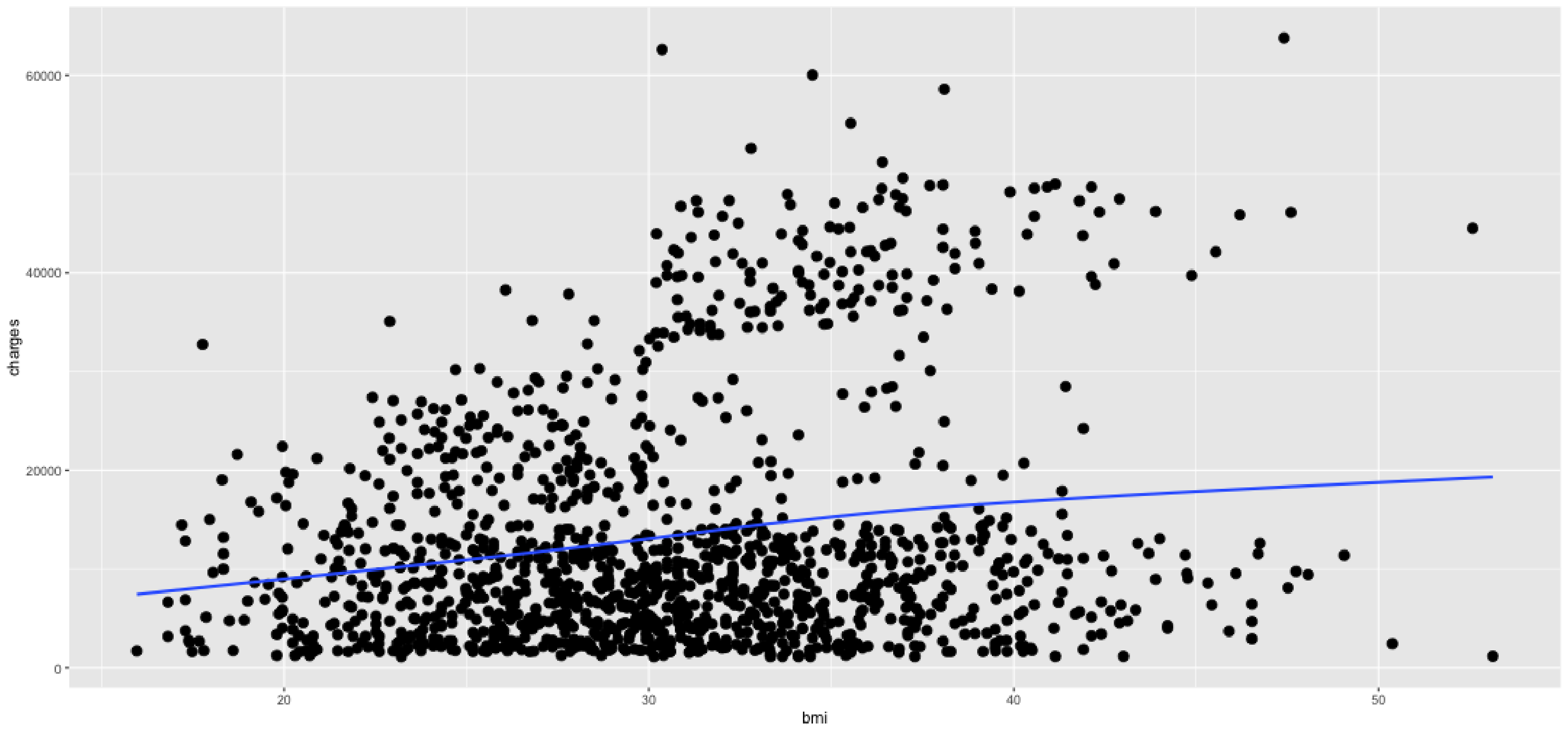
Num vs. Num Scatterplot

```
ggplot(data=insurance,
       mapping=aes(x=bmi,
                   y=charges)) +
  geom_point(size=3)
```



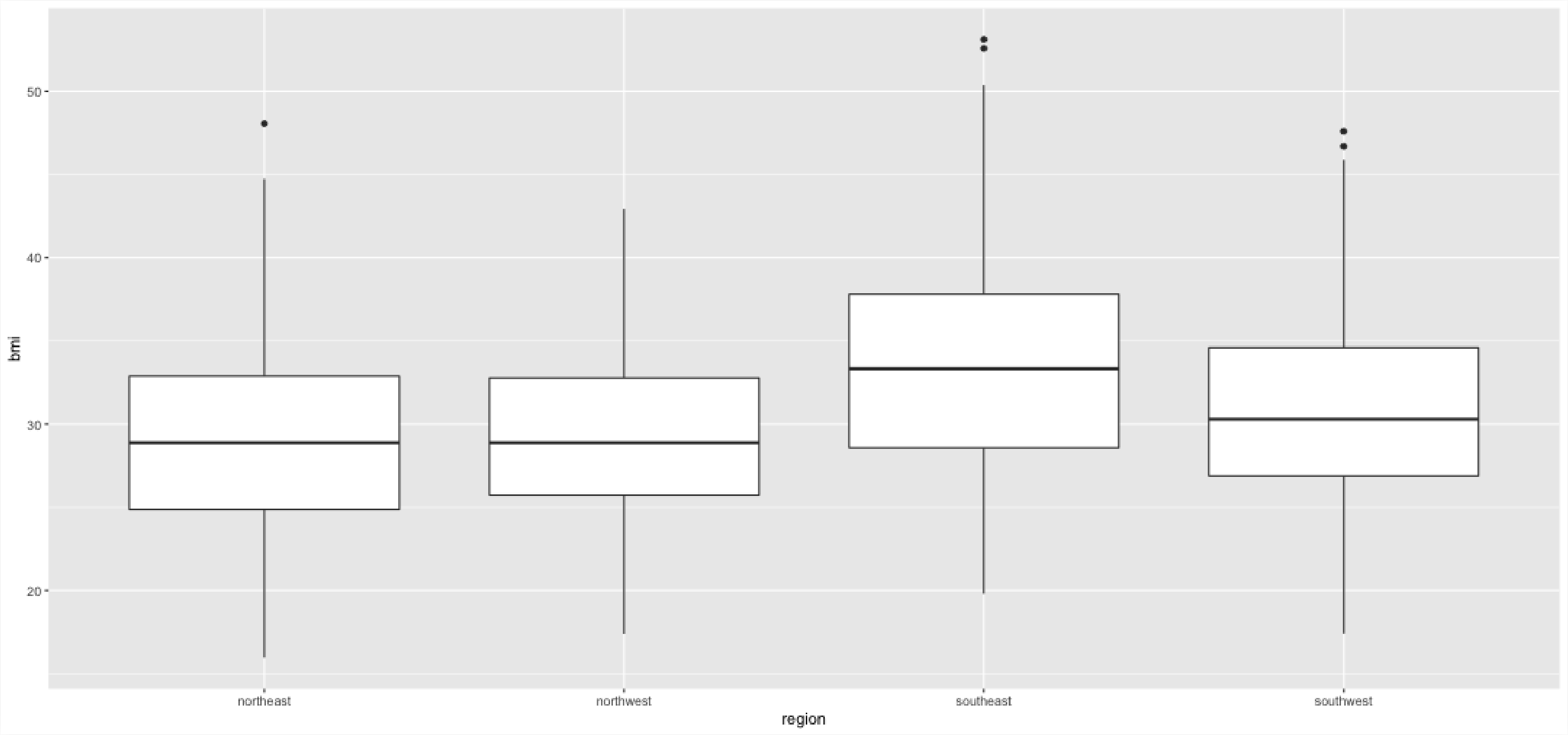
Num. vs Num.: Smooth Line Plot

```
ggplot(data=insurance,
       mapping=aes(x=bmi,
                   y=charges))+
  geom_point(size=3)+
  geom_smooth(se=F)
```



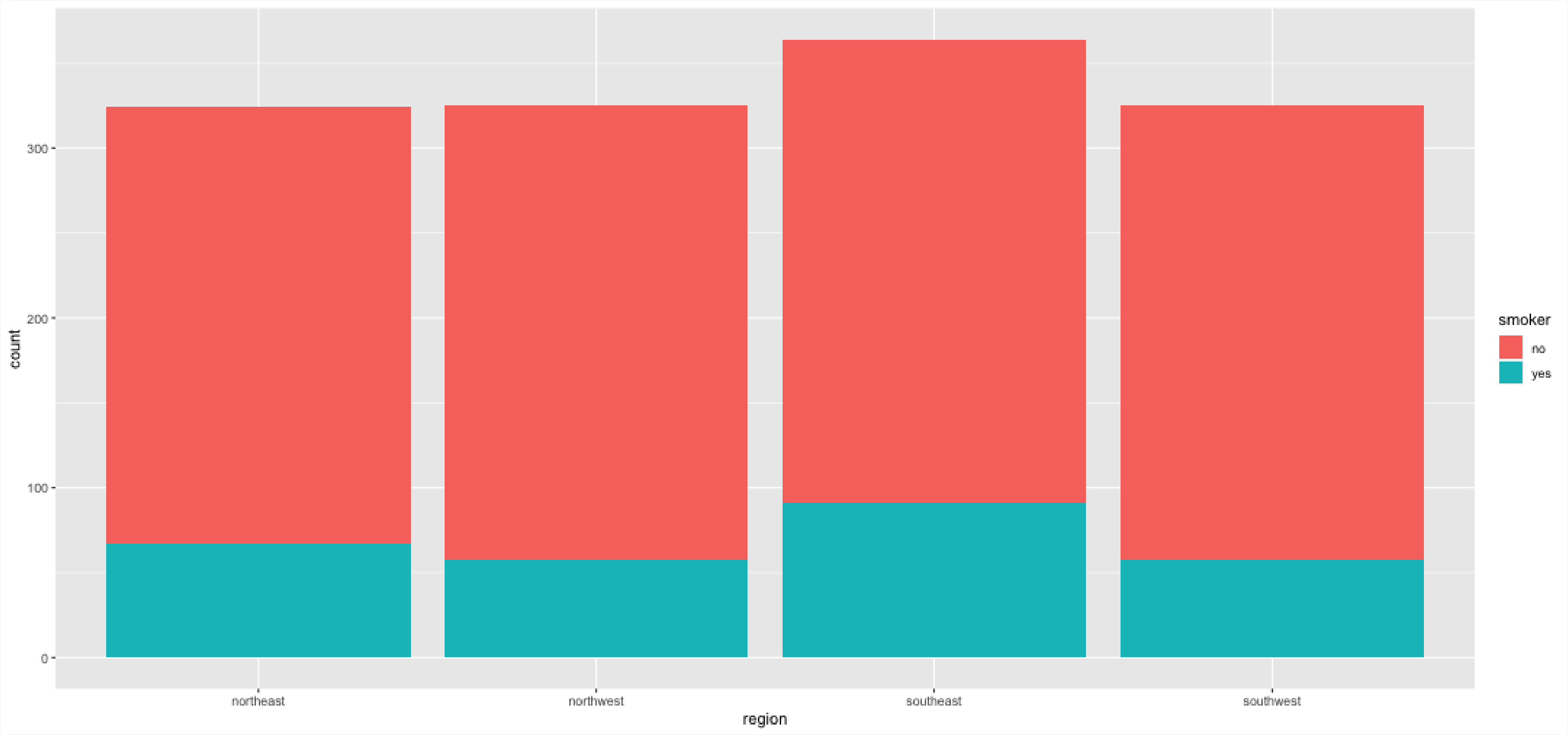
Num. vs Cat.: box (and whisker) plots

```
ggplot(data = insurance, mapping = aes(y =
bmi, x = region)) +
  geom_boxplot()
```



Cat. vs Cat.: Segmented bar plots (counts)

```
ggplot(data = insurance, mapping = aes(x =
region, fill = smoker)) +
  geom_bar()
```



Recap

- Visualizing our data can help lead to powerful insights between variable relationships
- ggplot() is a package in R that allows us to make plots
- There are many ways you can vizualize your data!