Lecture 4: Categorical variables

Michael Lopez, Skidmore College

Prelim: Awesome Viz's in the news

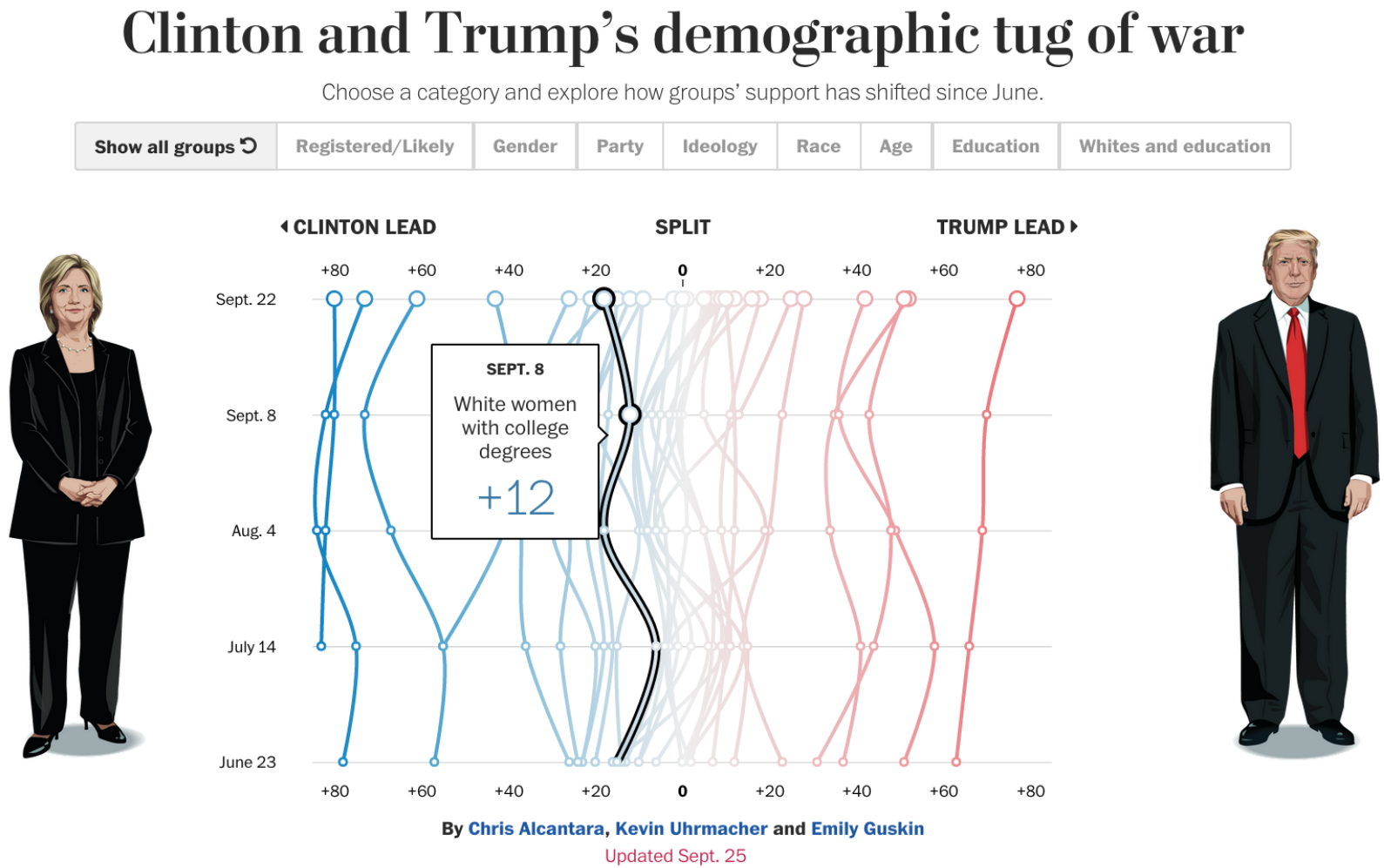


Figure 1:Tug of War (Washington Post)

Prelim: Awesome Viz's in the news

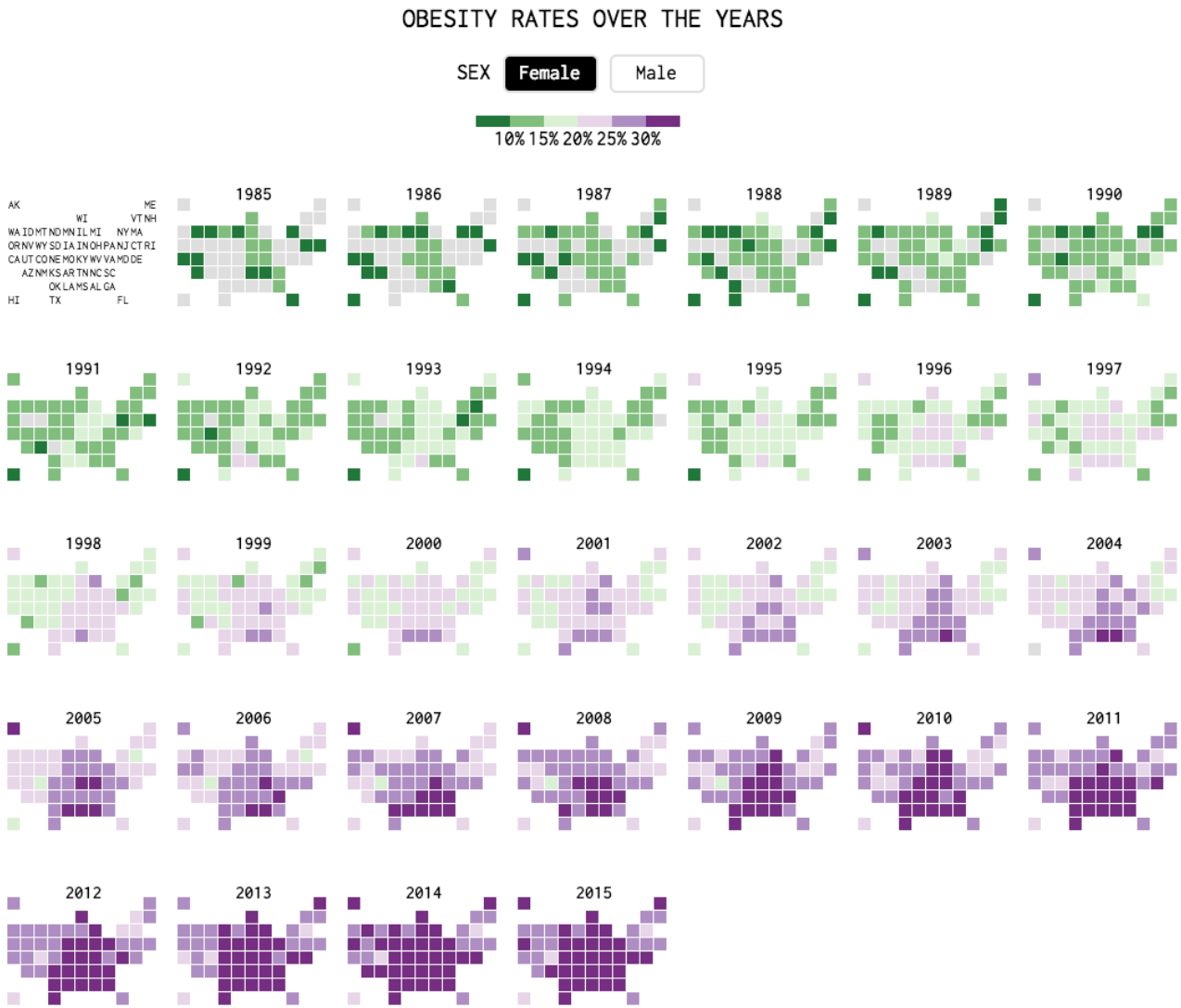


Figure 2:Obesity (Flowing Data)

Prelim: The 2016 election

Table 1:Voting preferences, sample of 2016 voters

	Clinton	Trump
Black	623	108
White	3190	4100
Other	1009	802

1. What does this data look like?
2. What questions can we answer?

Goals: charts for continuous data

1. Univariate: barchart
2. Bivariate: contingency tables, stacked barchart, mosaic plot
3. Tricks: axes limits and label details

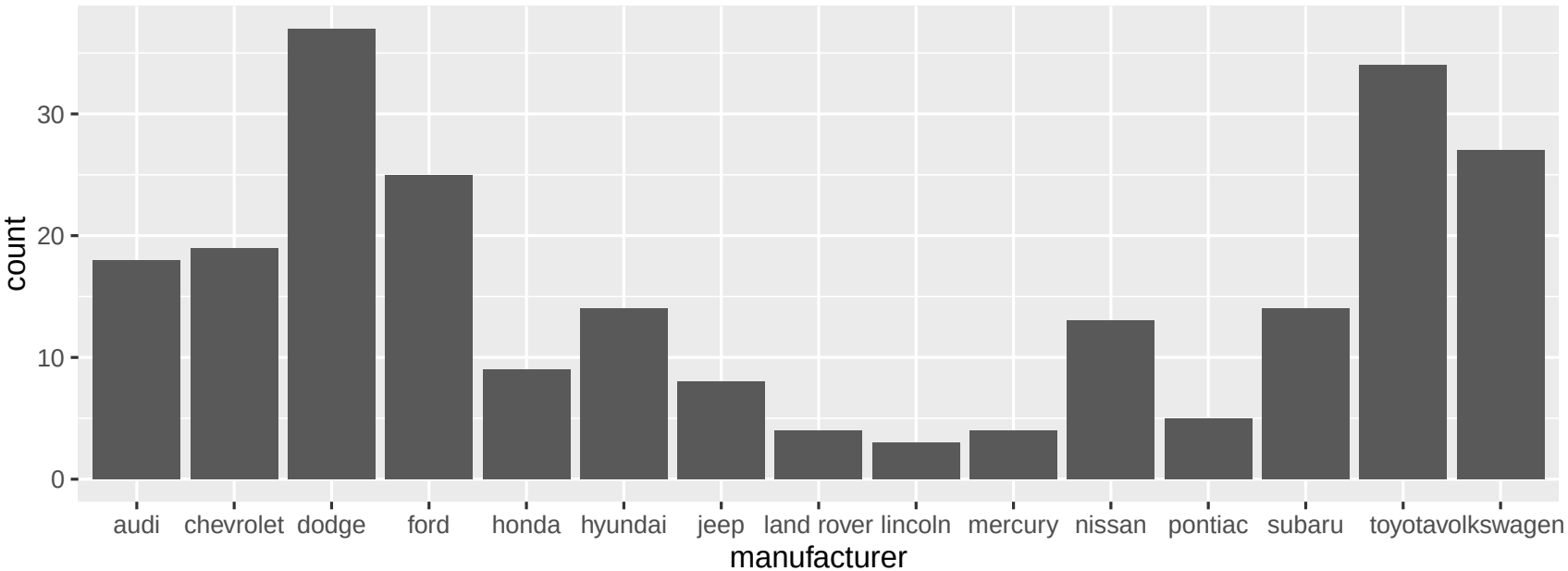
Data set

```
library(ggplot2); library(dplyr)
mpg %>% head(4)
```

```
## Source: local data frame [4 x 11]
##
##   manufacturer model displ  year   cyl      trans   drv   cty   hwy    fl
##          (chr) (chr) (dbl) (int) (int)      (chr) (chr) (int) (int) (chr)
## 1         audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
## 2         audi    a4   1.8  1999     4 manual(m5)     f    21    29     p
## 3         audi    a4   2.0  2008     4 manual(m6)     f    20    31     p
## 4         audi    a4   2.0  2008     4   auto(av)     f    21    30     p
## Variables not shown: class (chr)
```

Barplot

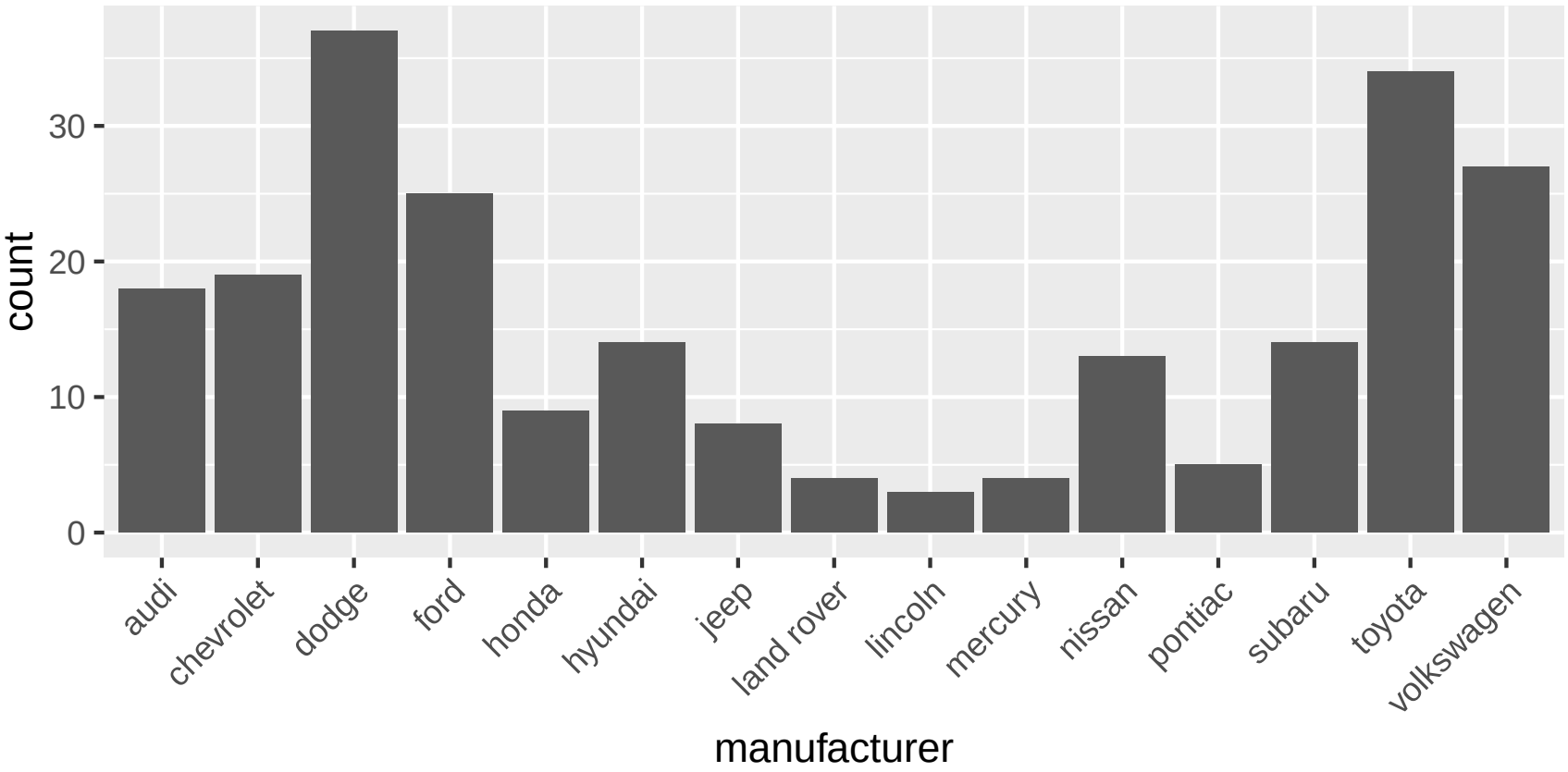
```
ggplot(data = mpg, aes(x = manufacturer)) +
  geom_bar()
```



Note: How to improve the axis?

Barplot

```
ggplot(data = mpg, aes(x = manufacturer)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



What does a barchart show?

A note on the y-axis

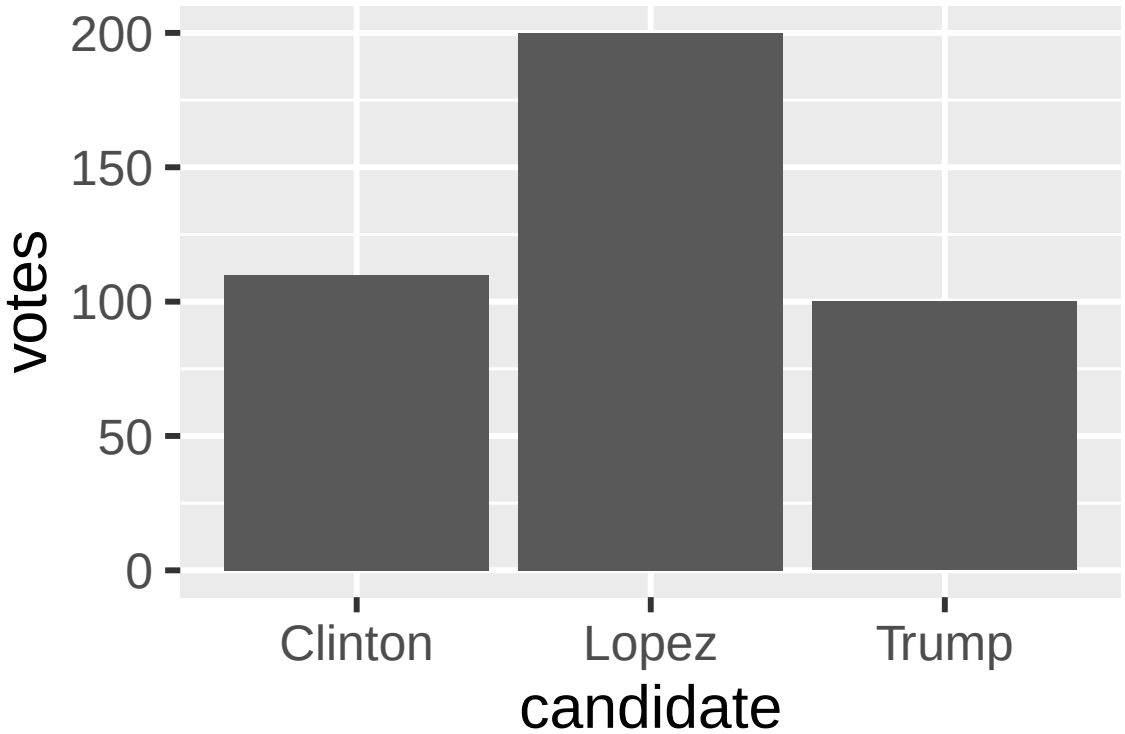
Barplot with summary data

```
count.responses <- data.frame(
  candidate = c("Lopez", "Trump", "Clinton"),
  votes = c(200, 100, 110)
)
count.responses
```

```
##   candidate votes
## 1     Lopez   200
## 2     Trump   100
## 3   Clinton   110
```

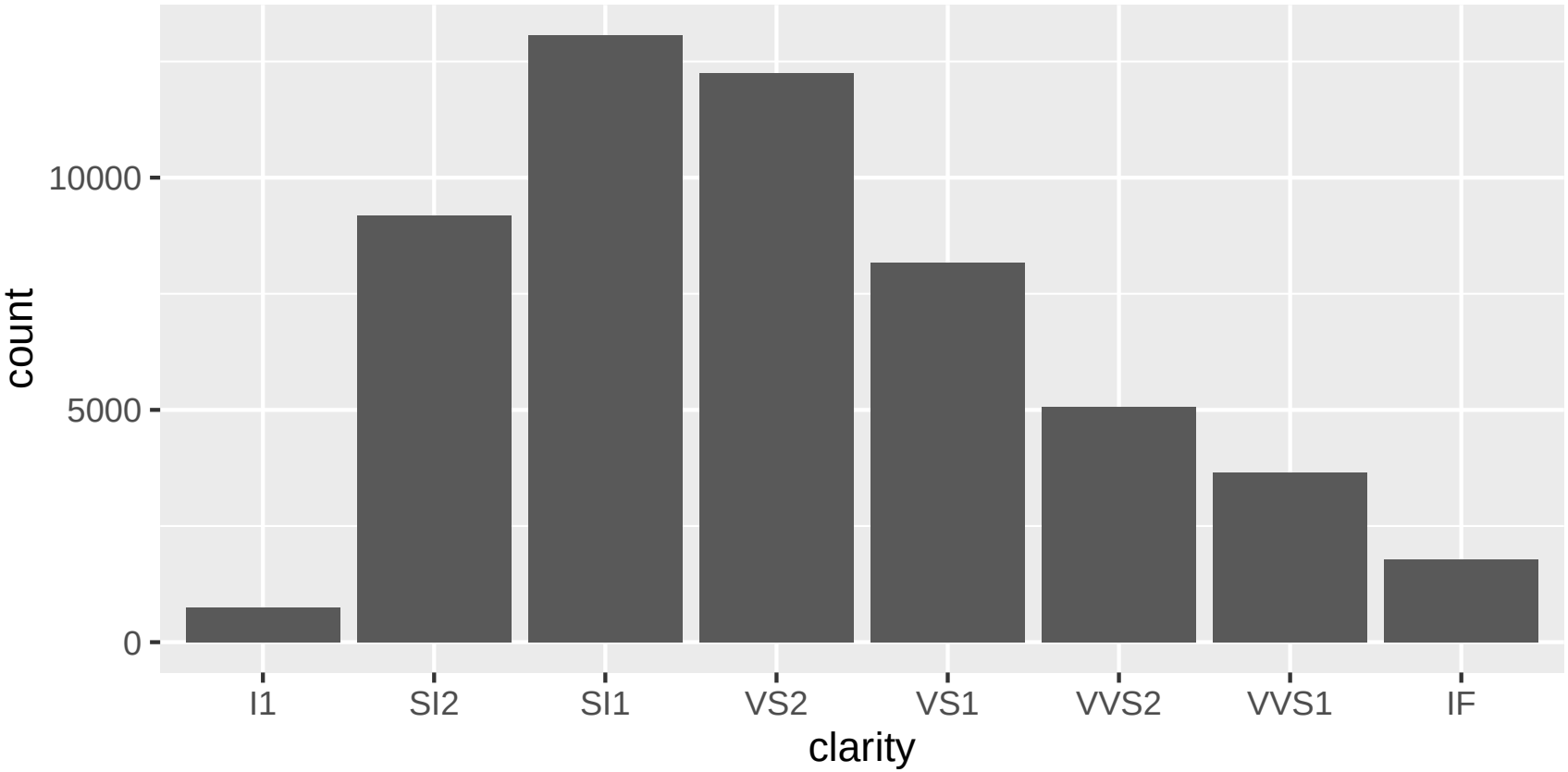
Barplot with summary data

```
ggplot(count.responses, aes(candidate, votes)) +
  geom_bar(stat = "identity")
```



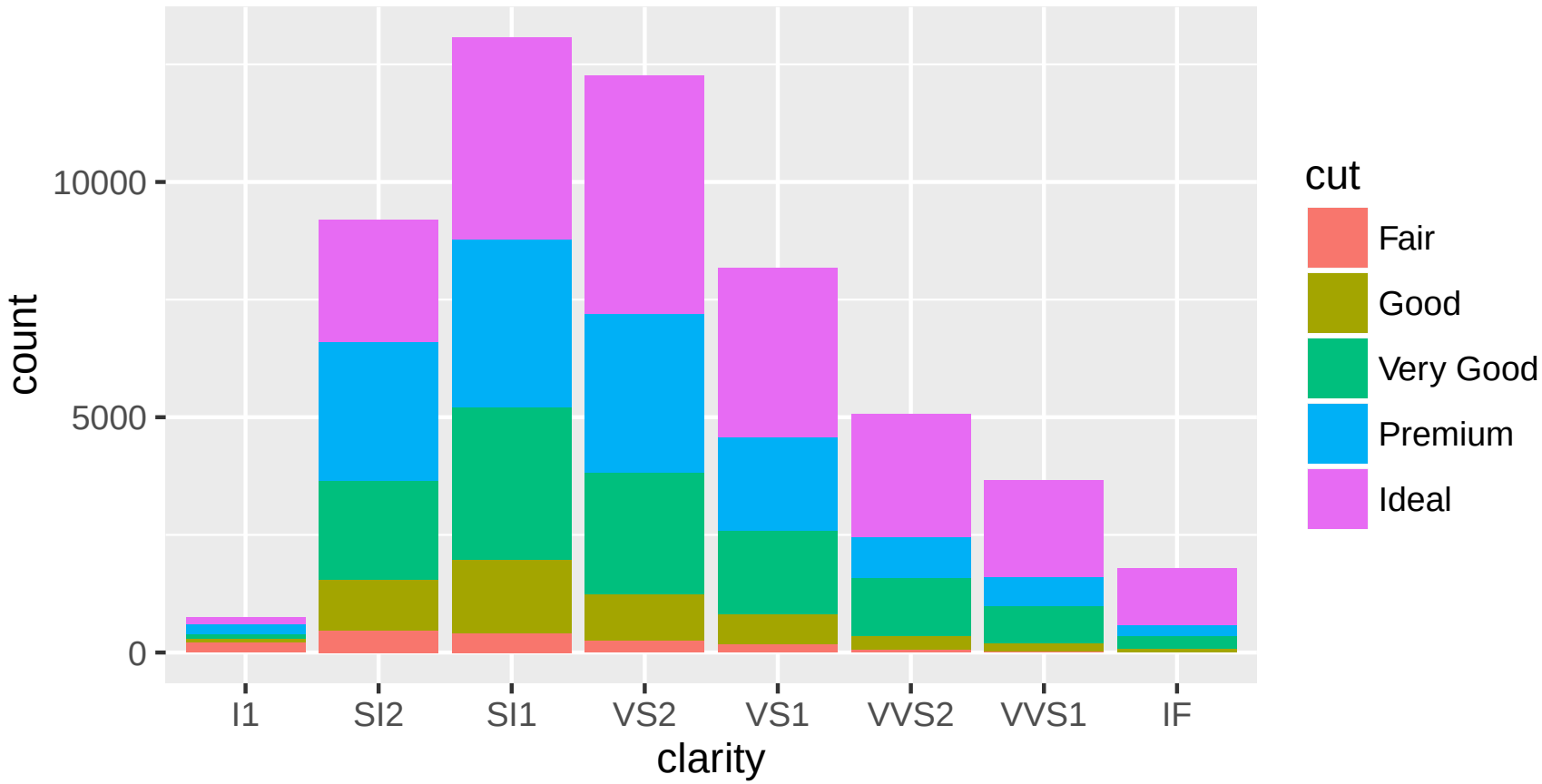
Barchart, diamonds dataset

```
ggplot(diamonds, aes(clarity)) +
  geom_bar()
```



Bivariate data: Stacked barcharts

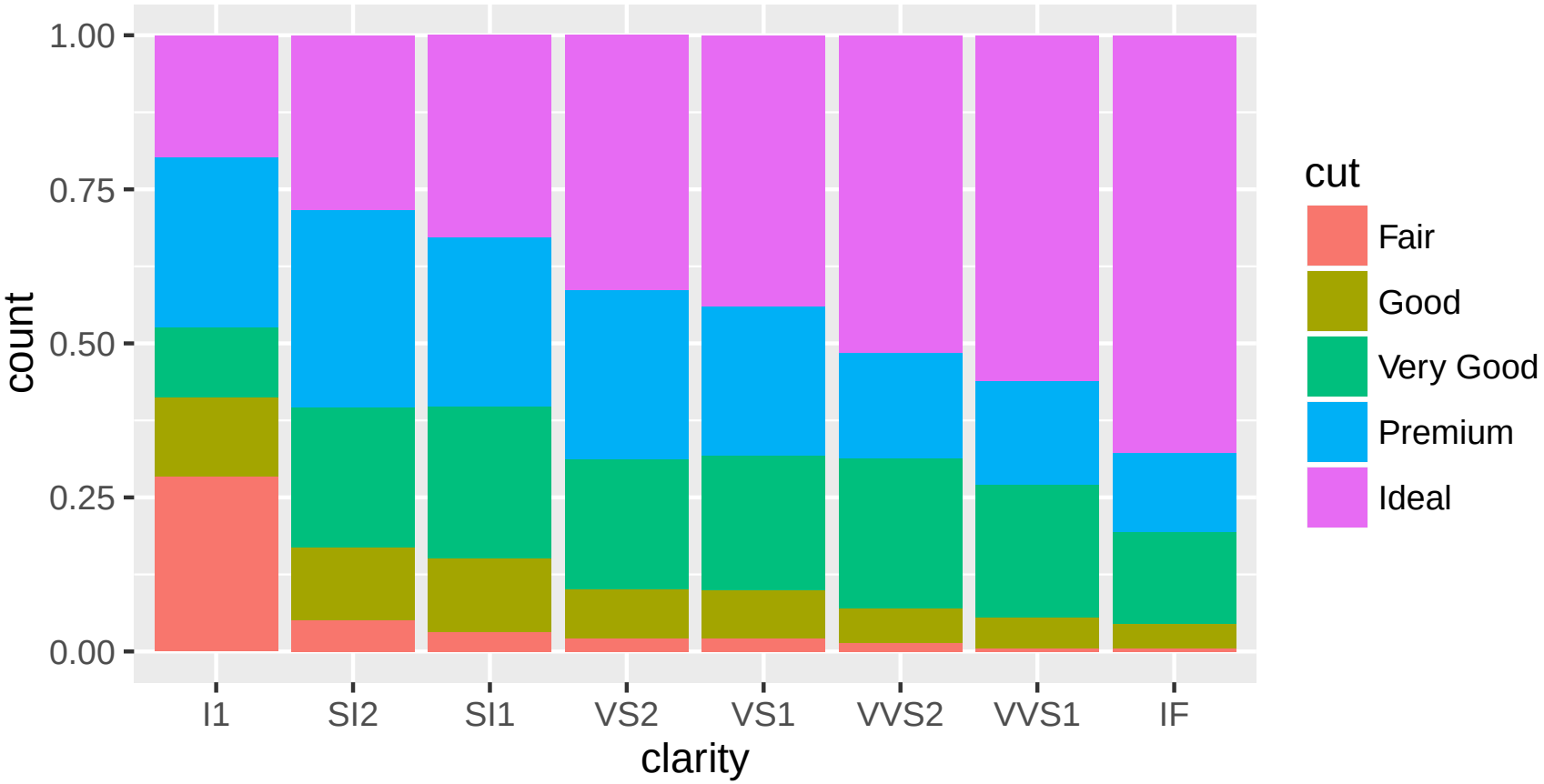
```
ggplot(diamonds, aes(clarity, fill=cut)) +
  geom_bar()
```



What is a stacked barchart?

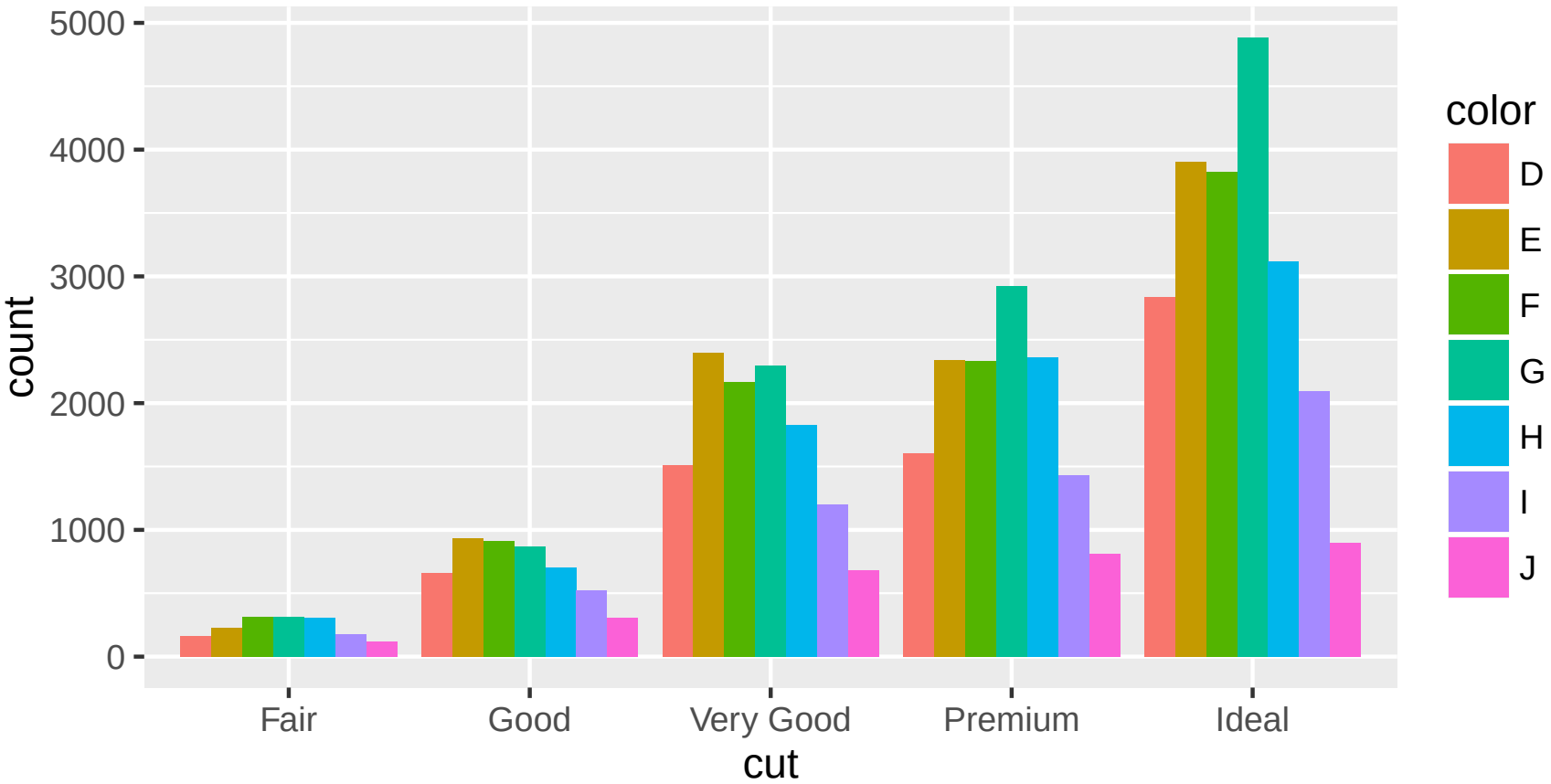
Bivariate data: Stacked barcharts with percents

```
ggplot(diamonds, aes(clarity, fill=cut)) +
  geom_bar(position = "fill")
```



Bivariate data: side-by-side bar charts

```
ggplot(diamonds, aes(cut, fill = color)) +
  geom_bar(position = "dodge")
```



What is a side-by-side barchart?

The contingency table

Mosaic plots

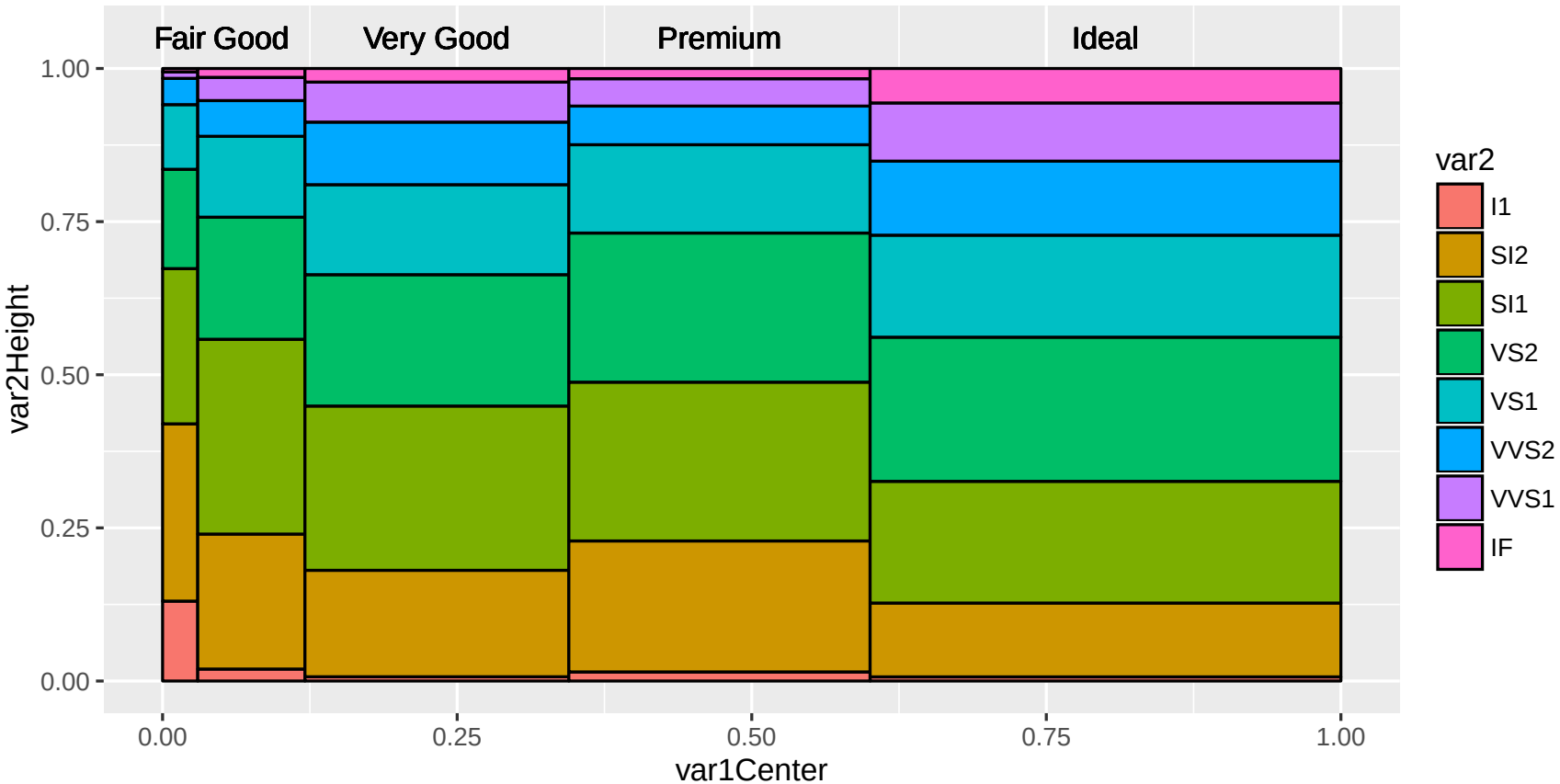
```
ggMMplot <- function(var1, var2){
  require(ggplot2)
  levVar1 <- length(levels(var1))
  levVar2 <- length(levels(var2))

  jointTable <- prop.table(table(var1, var2))
  plotData <- as.data.frame(jointTable)
  plotData$marginVar1 <- prop.table(table(var1))
  plotData$var2Height <- plotData$Freq / plotData$marginVar1
  plotData$var1Center <- c(0, cumsum(plotData$marginVar1)[1:levVar1 -1]) +
    plotData$marginVar1 / 2

  ggplot(plotData, aes(var1Center, var2Height)) +
    geom_bar(stat = "identity", aes(width = marginVar1, fill = var2), col = "Black") +
    geom_text(aes(label = as.character(var1), x = var1Center, y = 1.05))
  }
```

Mosaic plots

```
## Note: this is a specific function coded above.
## Not part of base ggplot
ggMMplot(diamonds$cut, diamonds$clarity)
```



Example: How to plot?

Table 2:Voting preferences, sample of 2016 voters

	Clinton	Trump
Black	623	108
White	3190	4100
Other	1009	802

Example plot 1

Example plot 2

Example plot 3