Guidelines for Good Graphs

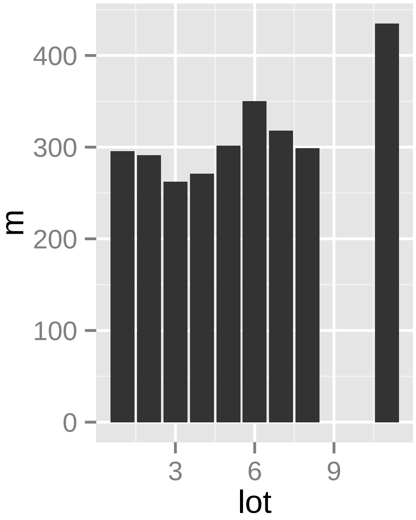
Te Rutherford

November 3, 2014

One of the goals of data visualization is to use graphics to present the information and draw the attention of the audience. Good data visualization is more or less like an art. There are some general principles you can follow to make visually-appealing graphs, but there is no right or wrong answer as long as the graphs achieve their communicative goals. In short, graphs must be information and beautiful. Here are some basic guidelines.

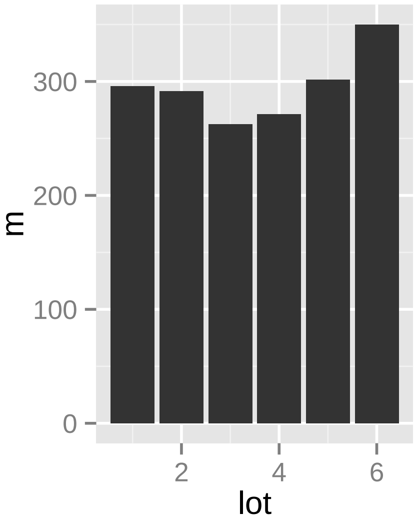
Don't show what you are not going to talk about.

```r
library(plyr)
library(ggplot2)
d = read.csv('http://www.typ-stats.com/datasets/homes.csv')
lot_means = ddply(d, c('lot'), summarize, m = mean(price))
ggplot(mapping=aes(x=lot,y=m),data=lot_means) + geom_bar(stat='identity')
```



It looks a bit ugly because there are no homes in certain types of lot sizes. Plus, if you will only talk about lot size from 1 - 6, then you should omit the rest by using the command `subset`.

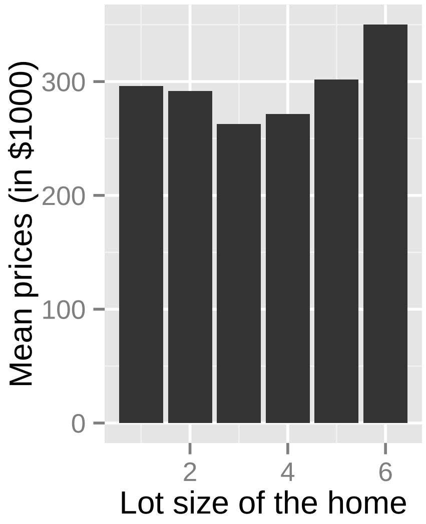
```r
small_lot_means = subset(lot_means, lot <= 6)
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity')
```



The X and Y axes must be labeled with description and unit.

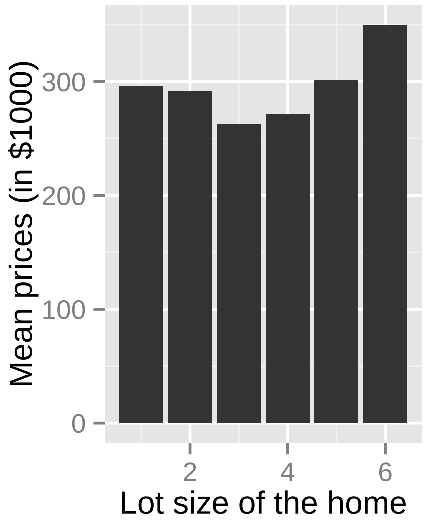
The audience should be able to read the graph and understand what is on each axis. What we have right now, we have m and lot, which we will have to explain later what they are. This is no good. We should label them using the commands `xlab` or `ylab`.

```r
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity') +
  xlab('Lot size of the home') +
  ylab('Mean prices (in $1000)')
```



Alternatively, you can use the command `scale_x_continuous` or `scale_y_continuous`.

```r
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity') +
  scale_x_continuous(name='Lot size of the home') +
  scale_y_continuous(name='Mean prices (in $1000)')
```

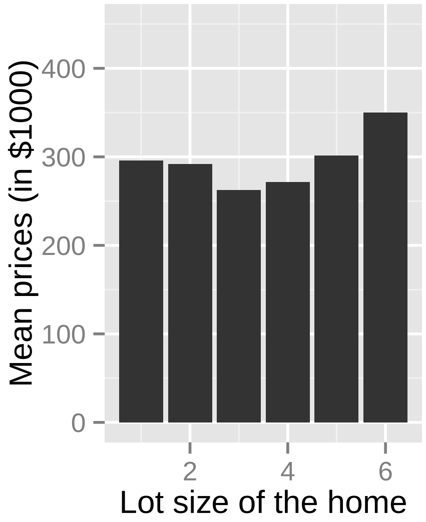


The results are the same. The axes now include the labels that are informative of what they represent. The price of the homes now includes the units, so we don't leave the audience guessing the units. Dollar? Euro?

The limits of the axes must be adjusted properly.

We note that the information of the graph is concentrated on the top because the bottom of each bar is all the same. We want to the highlight the height of each bar by putting it in the center. To do that, we have to adjust the limits of the axes by using the commands `xlim` or `ylim`.

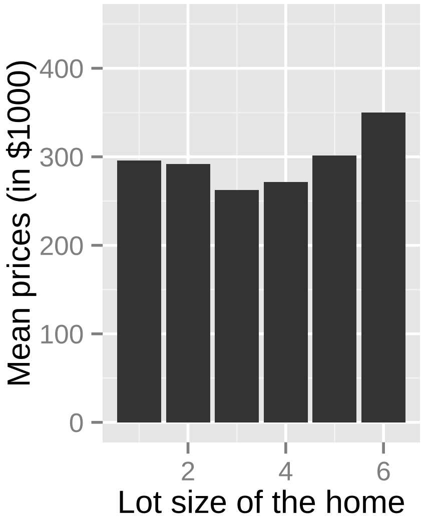
```
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity') +
  xlab('Lot size of the home') +
  ylab('Mean prices (in $1000)') +
  ylim(0, 450)
```



Now the heights of the bars are moved toward the center of the graph. That sounds fancy, but it just looks better.

Alternatively, you can use the command `scale_x_continuous` or `scale_y_continuous`.

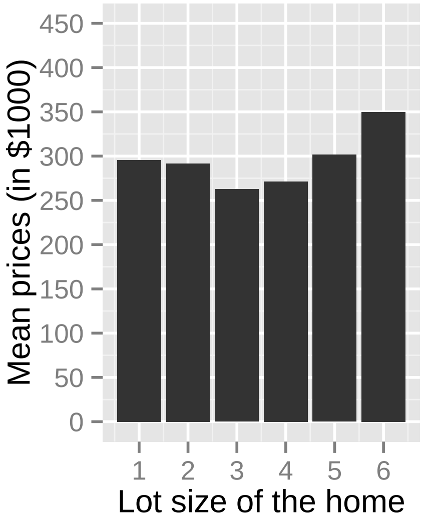
```
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity') +
  scale_x_continuous(name='Lot size of the home') +
  scale_y_continuous(name='Mean prices (in $1000)', limits=c(0,450))
```



Each tick mark must represent a significant increment.

The increment of $100,000 seems pretty steep. The difference between $100,000 and $200,000 is huge. We want each tick mark to represent the increment that is appropriate for the scale.

```
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity') +
  scale_x_continuous(name='Lot size of the home',
                     breaks=c(1,2,3,4,5,6)) +
  scale_y_continuous(name='Mean prices (in $1000)', limits=c(0,450),
                     breaks=c(0, 50, 100, 150, 200, 250, 300, 350, 400, 450))
```



The caption must be included.

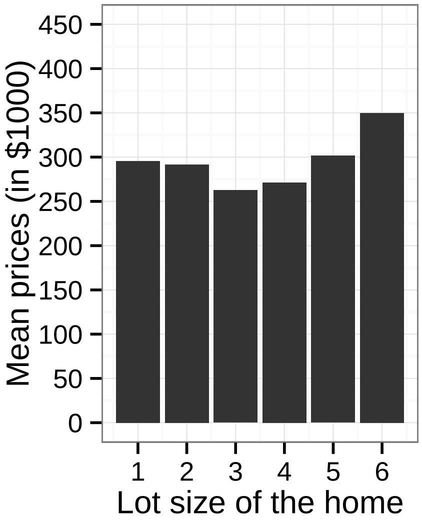
The caption should tell what the graph shows in words AND what it means (interpretation). For example, the caption might be

The size of the lots on which the homes are built does not directly influence the prices of the homes.

White background color is almost always the best

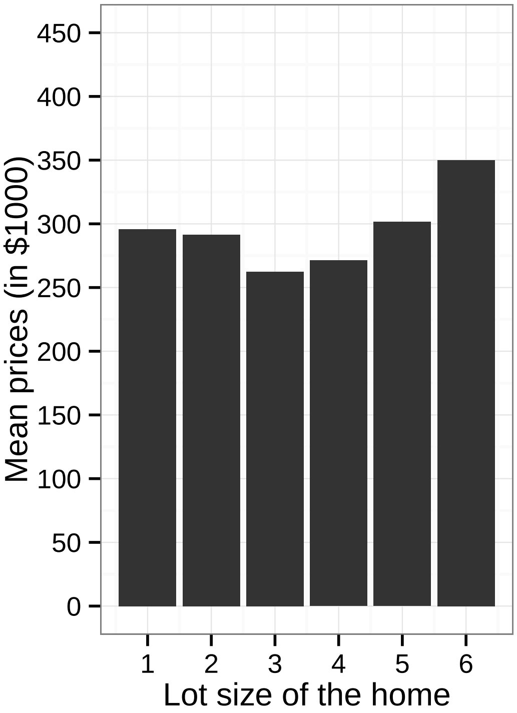
White background is usually the best because it does not distract the audience from the actual graph. But some people that grey background might be better contrast with the black bar plot. Again, this is your design aesthetic choice.

```r
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity') +
  scale_x_continuous(name='Lot size of the home',
                     breaks=c(1,2,3,4,5,6)) +
  scale_y_continuous(name='Mean prices (in $1000)', limits=c(0,450),
                     breaks=c(0, 50, 100, 150, 200, 250, 300, 350, 400, 450)) +
  theme_bw()
```

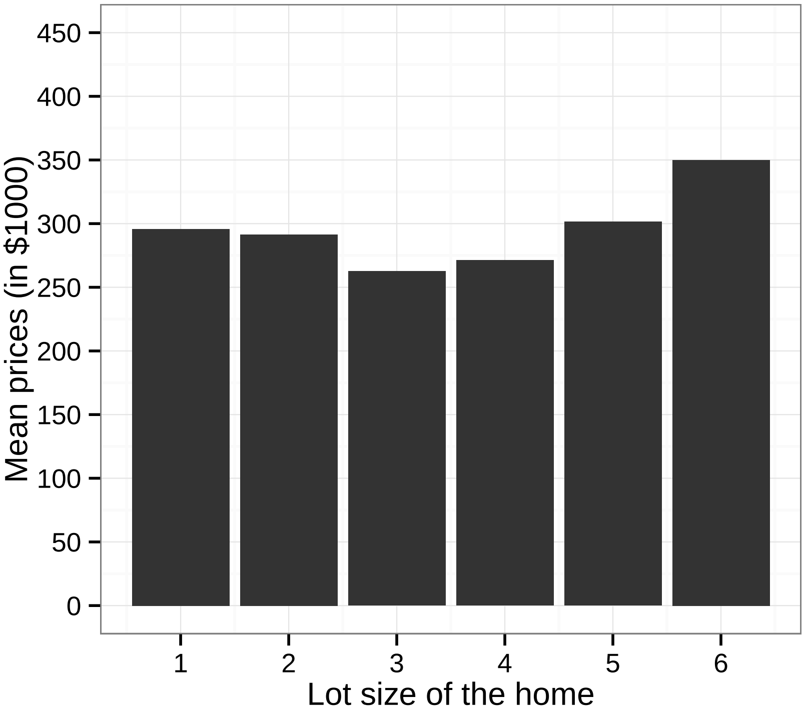


Adjust the width and height

This depends on what you want to put the graphs on. But adjust the width and the height until the graph looks good. Right now the graph looks a bit crowded. Let's change it up a bit.



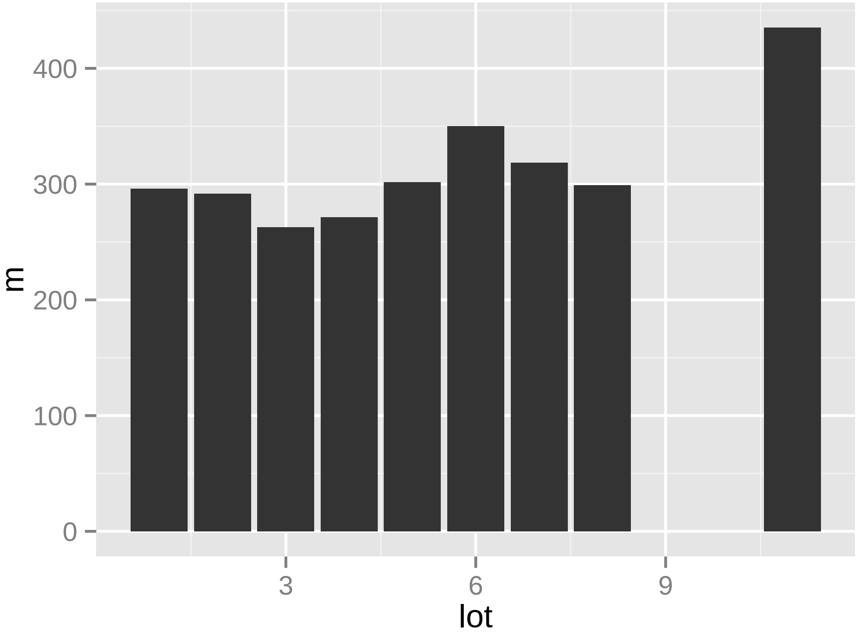
Make it a little bit wider.



Before and after

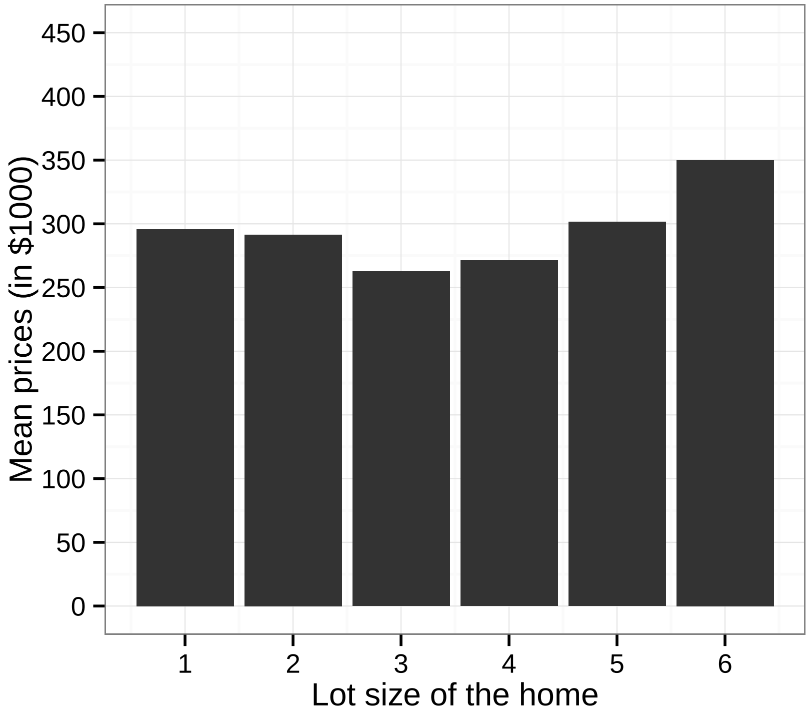
Before, it looks like this.

```
ggplot(mapping=aes(x=lot,y=m),data=lot_means) + geom_bar(stat='identity')
```



After, it looks like this.

```
ggplot(mapping=aes(x=lot,y=m),data=small_lot_means) + geom_bar(stat='identity') +
  scale_x_continuous(name='Lot size of the home',
                     breaks=c(1,2,3,4,5,6)) +
  scale_y_continuous(name='Mean prices (in $1000)', limits=c(0,450),
                     breaks=c(0, 50, 100, 150, 200, 250, 300, 350, 400, 450)) +
  theme_bw()
```


Mean prices (in $1000)
Lot size of the home