Exploring Data

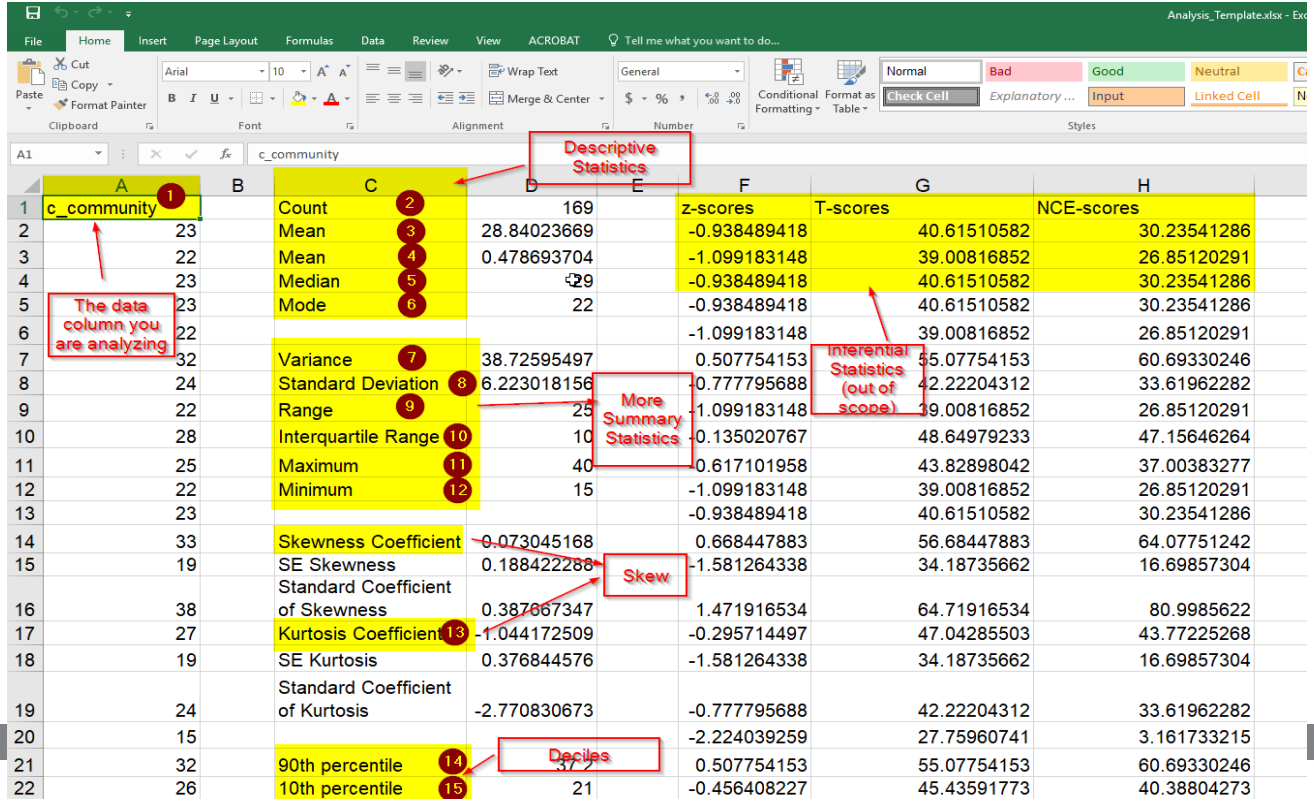
LEARNING VOYAGE

Table of Contents

- Using summary statistics to spot problems
- Spotting problems using graphics and visualization

Using summary statistics to spot problems

- With Analytics, you'll typically use the summary to take your first look at the data.
- https://github.com/fenago/ISM4415/tree/main/resources



Typical problems revealed by data summaries

- Missing Values
- Create formula in your spreadsheet to account for missing values (maybe C6)

Invalid values and outliers

*Define an outlier in your template as being 2X the standard deviation in either direction.

```
> summary(custdata$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  -8700   14600   35000   53500   67000
    Max.
 615000
```

Negative values for income could indicate bad data. They might also have a special meaning, like "amount of debt."

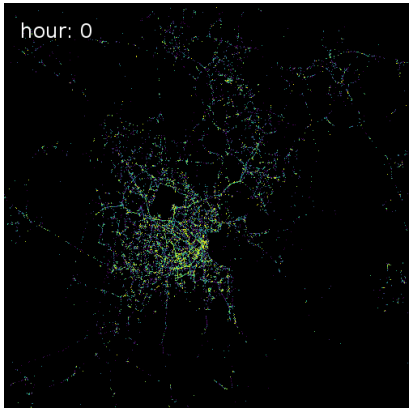
Either way, you should check how prevalent the issue is, and decide what to do: Do you drop the data with negative income? Do you convert negative values to zero?

```
> summary(custdata$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.
    0.0    38.0    50.0    51.7    64.0
    Max.
  146.7
```

Customers of age zero, or customers of an age greater than about 110 are outliers. They fall out of the range of expected customer values.

Outliers could be data input errors. They could be special sentinel values: zero might mean "age unknown" or "refuse to state." And some of your customers might be especially long-lived.

Data range



- Let's look at income again, in listing.
- Is the data range wide? Is it narrow?

```
> summary(custdata$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  -8700   14600   35000   53500   67000
   Max.
 615000
```

Income ranges from zero to over half a million dollars; a very wide range.

Units

- You might not notice the error during the modeling stage, but down the line someone will start inputting hourly wage data into the model and get back bad predictions in return.

```
> summary(Income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -8.7    14.6    35.0    53.5    67.0   615.0
```

The variable Income is defined as Income = custdata$income/1000. But suppose you didn't know that. Looking only at the summary, the values could plausibly be interpreted to mean either "hourly wage" or "yearly income in units of $1000."

Table of Contents

- Using summary statistics to spot problems
- **Spotting problems using graphics and visualization**

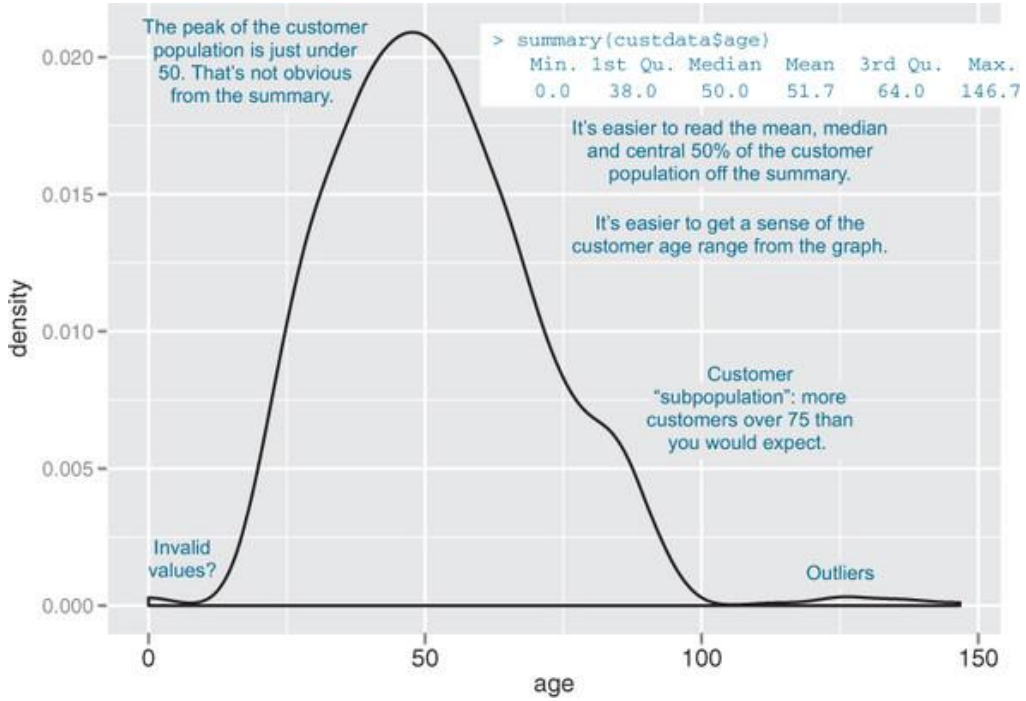
Spotting problems using graphics & visualization

We cannot expect a small number of numerical values [summary statistics] to consistently convey the wealth of information that exists in data. Numerical reduction methods do not retain the information in the data.

William Cleveland The Elements of Graphing Data

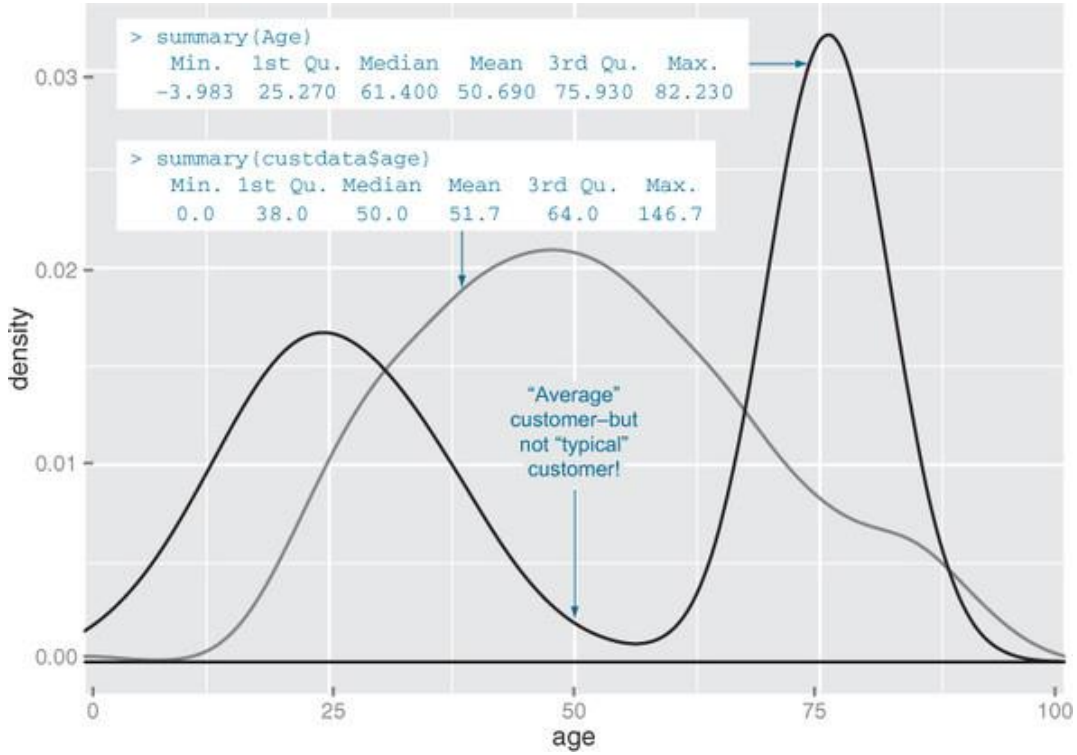
Spotting problems using graphics & visualization

- Some information is easier to read from a graph, and some from a summary.



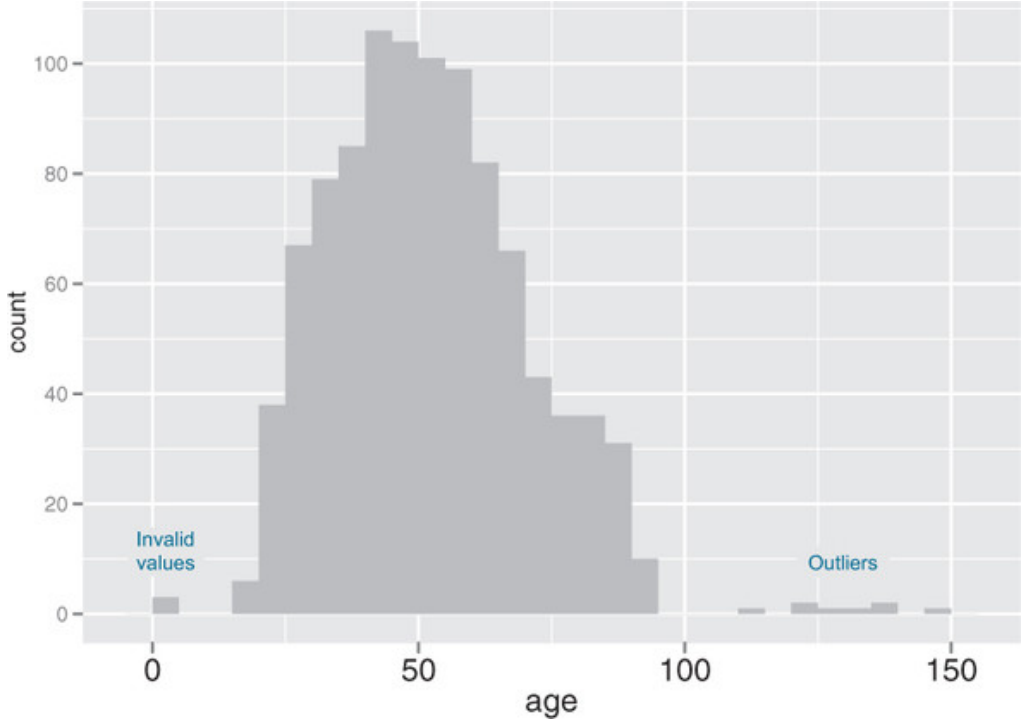
Visually checking distributions for a single variable

- A unimodal distribution (gray) can usually be modeled as coming from a single population of users.
- With a bimodal distribution (black), your data often comes from two populations of users.



Histograms

- A histogram tells you where your data is concentrated.
- It also visually highlights outliers and anomalies.



Histograms

- You create the histogram in figure in ggplot2 with the geom_histogram layer.

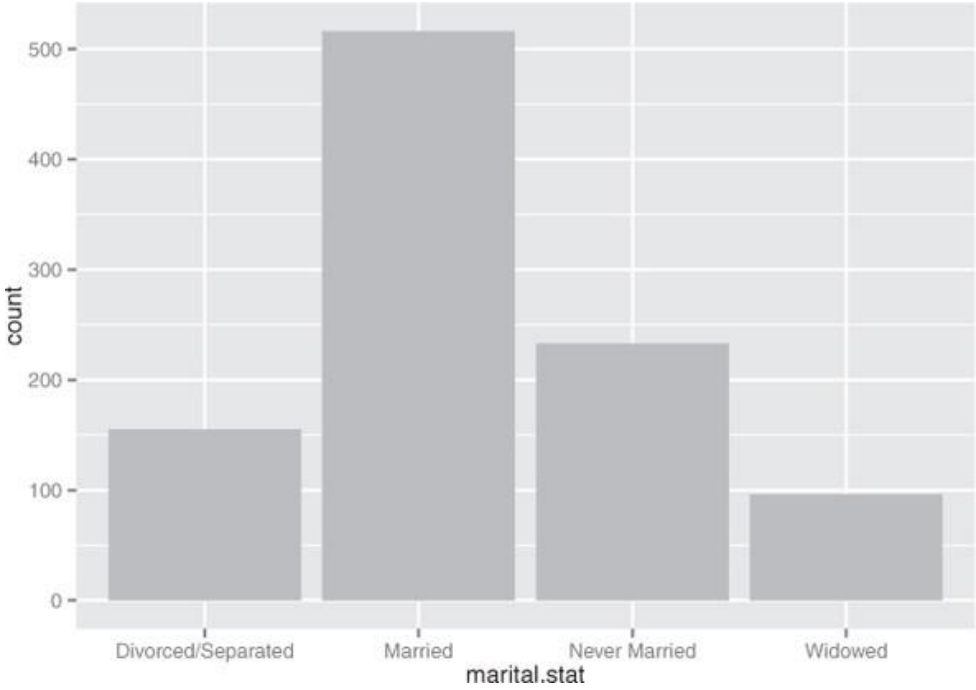
```
library(ggplot2)

ggplot(custdata) +
   geom_histogram(aes(x=age),
   binwidth=5, fill="gray")
```

Load the ggplot2 library, if you haven't already done so.

The binwidth parameter tells the geom_histogram call how to make bins of five-year intervals (default is datarange/30). The fill parameter specifies the color of the histogram bars (default: black).

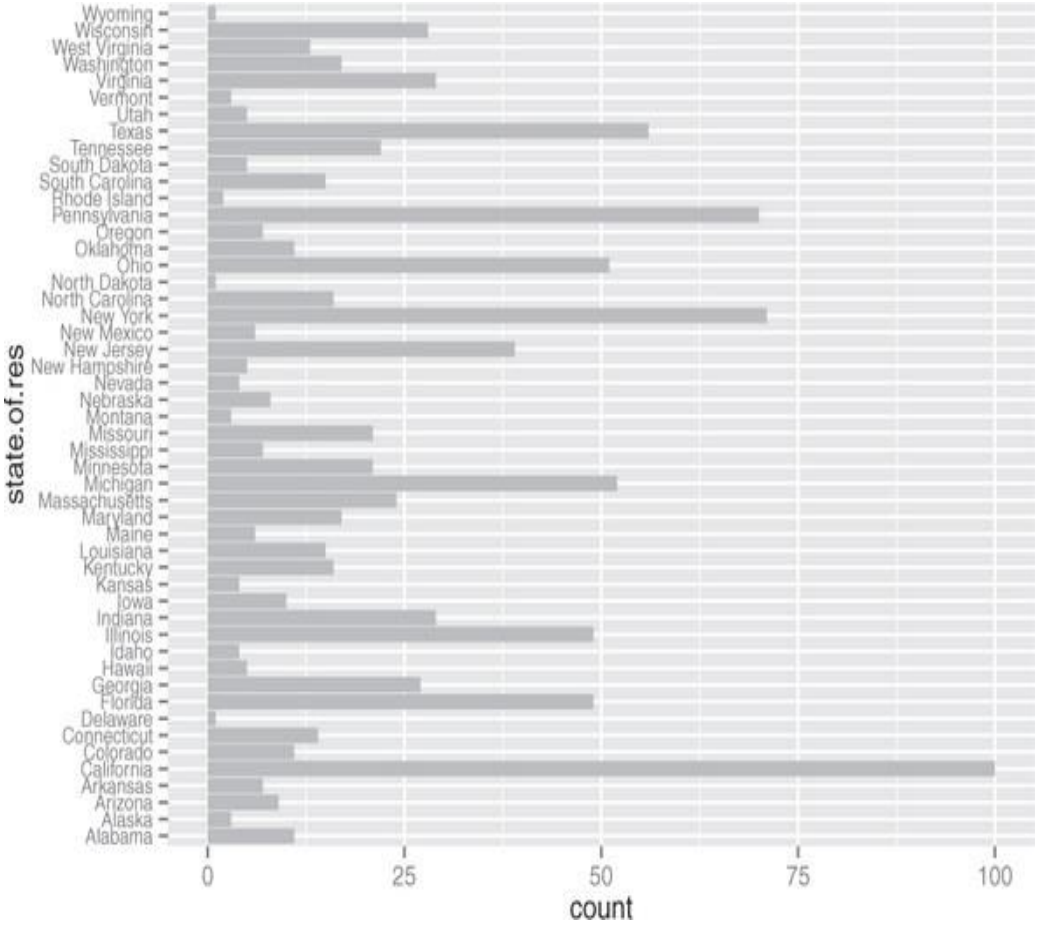
Bar charts

- Bar charts show the distribution of categorical variables.



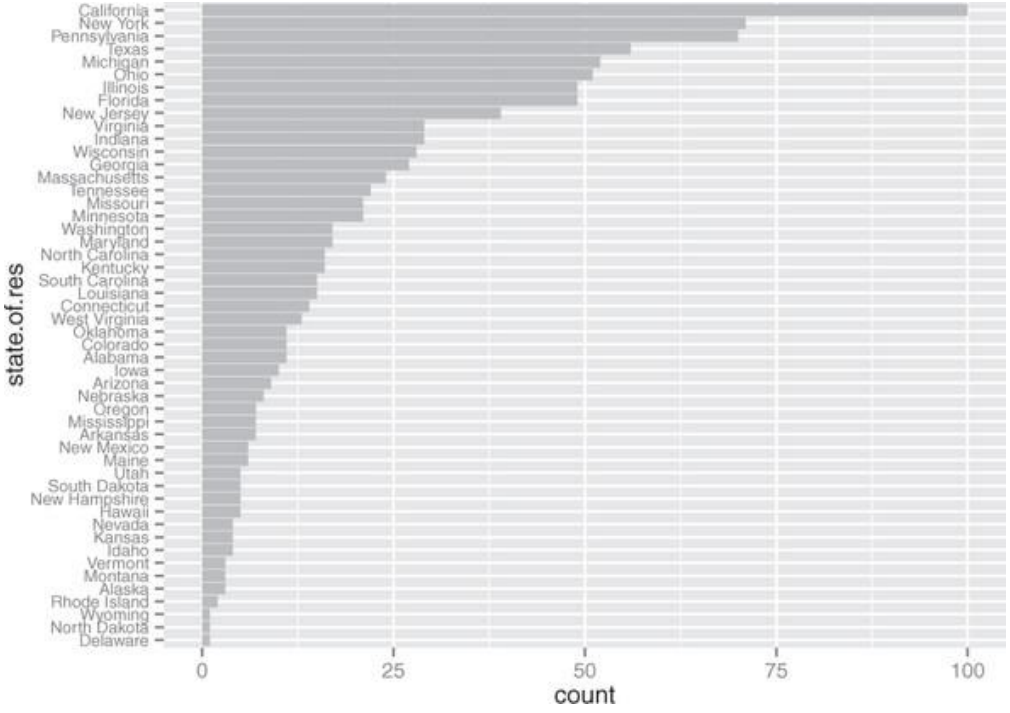
Bar charts

- A horizontal bar chart can be easier to read when there are several categories with long names.



Bar charts

- Sorting the bar chart by count makes it even easier to read.



Bar charts

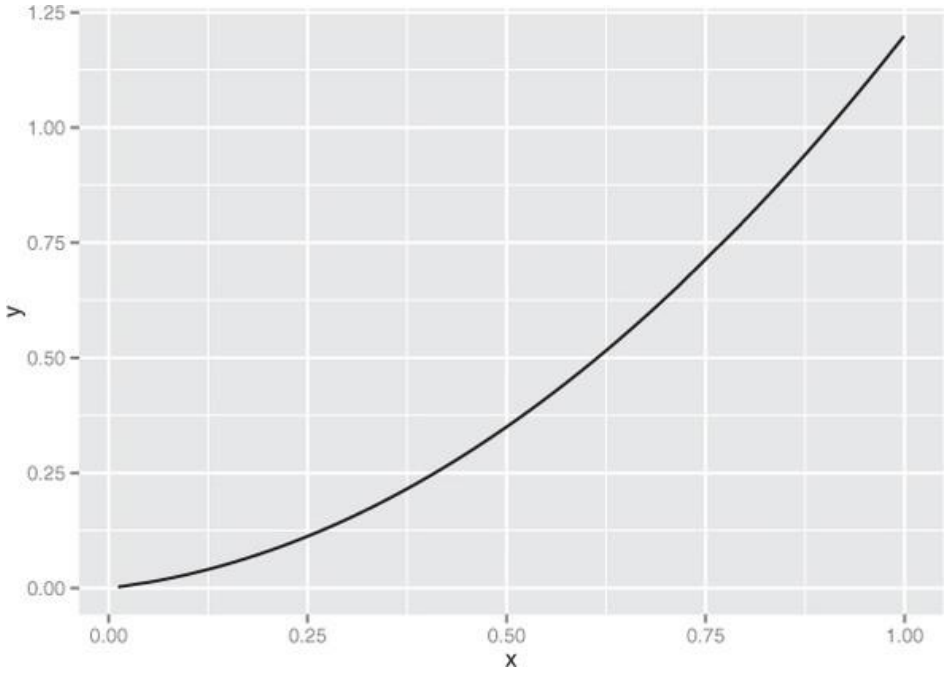
Graph type	Uses
Histogram or density plot	Examines data range Checks number of modes Checks if distribution is normal/lognormal Checks for anomalies and outliers
Bar chart	Compares relative or absolute frequencies of the values of a categorical variable

Visually checking relationships between two variables

In addition to examining variables in isolation, you'll often want to look at the relationship between two variables. For example, you might want to answer questions like these:

- Is there a relationship between the two inputs age and income in my data?
- What kind of relationship, and how strong?
- Is there a relationship between the input marital status and the output health insurance? How strong?

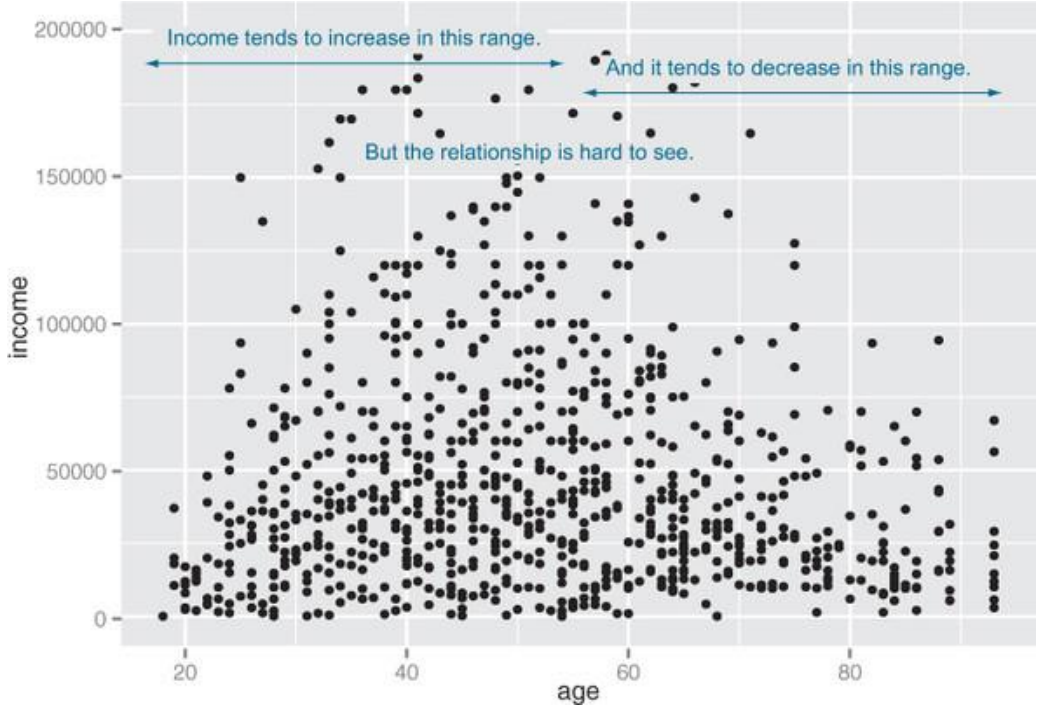
Example of a line plot



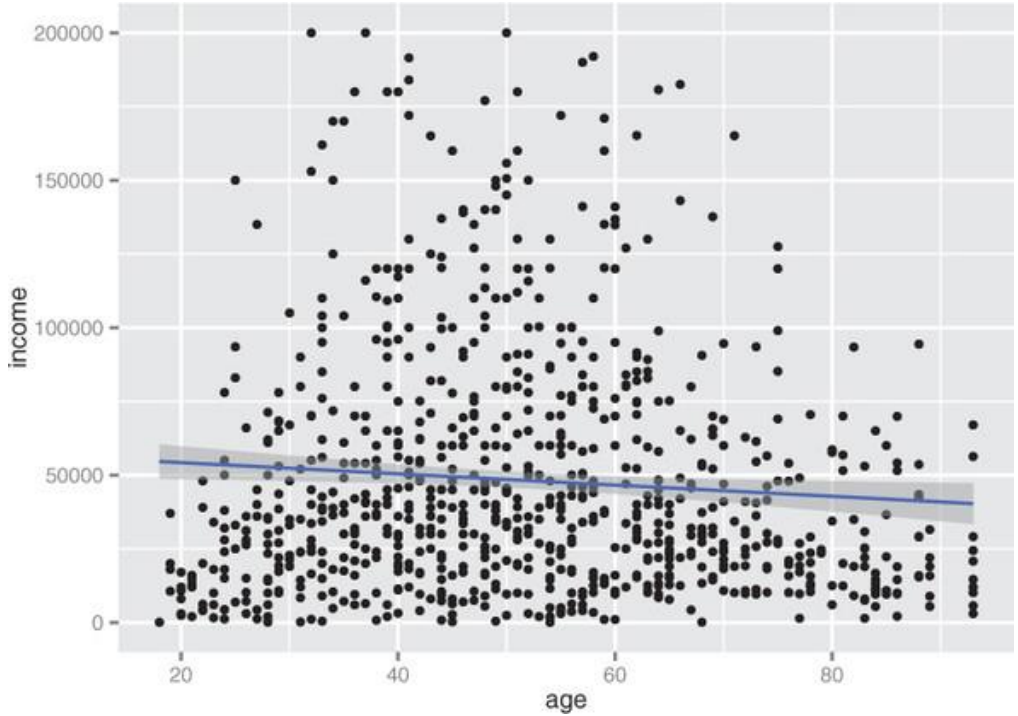
Scatter plots and smoothing curves

- The appropriate summary statistic is the correlation, which we compute on a safe subset of our data.

Scatter plots and smoothing curves

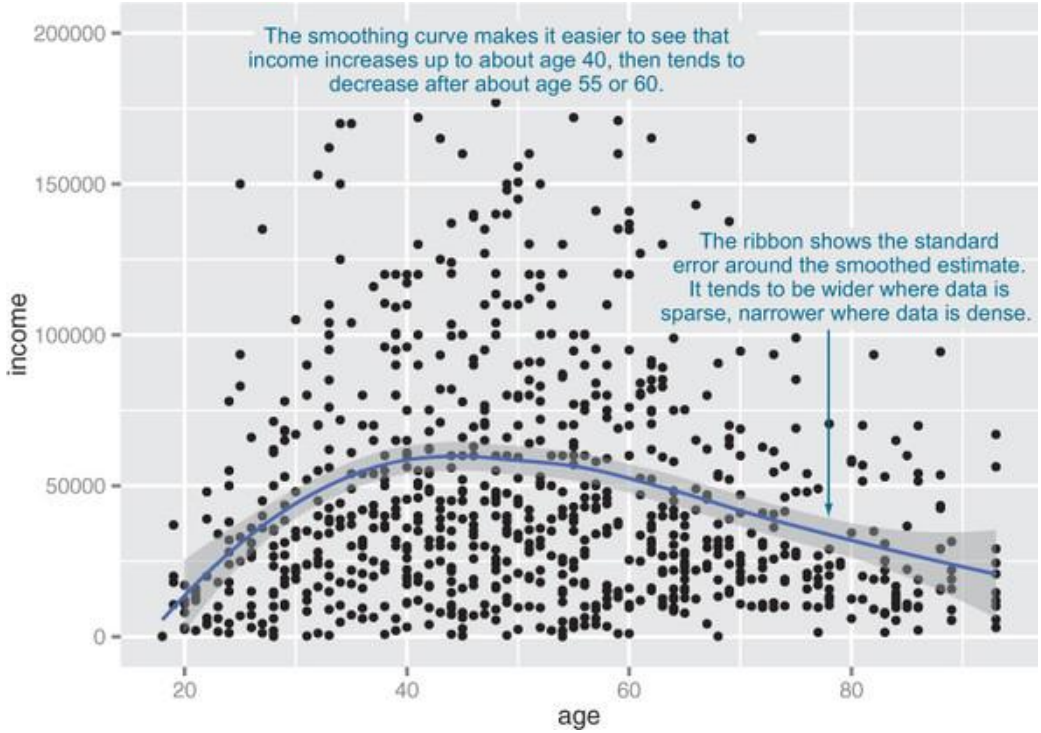


Scatter plots and smoothing curves



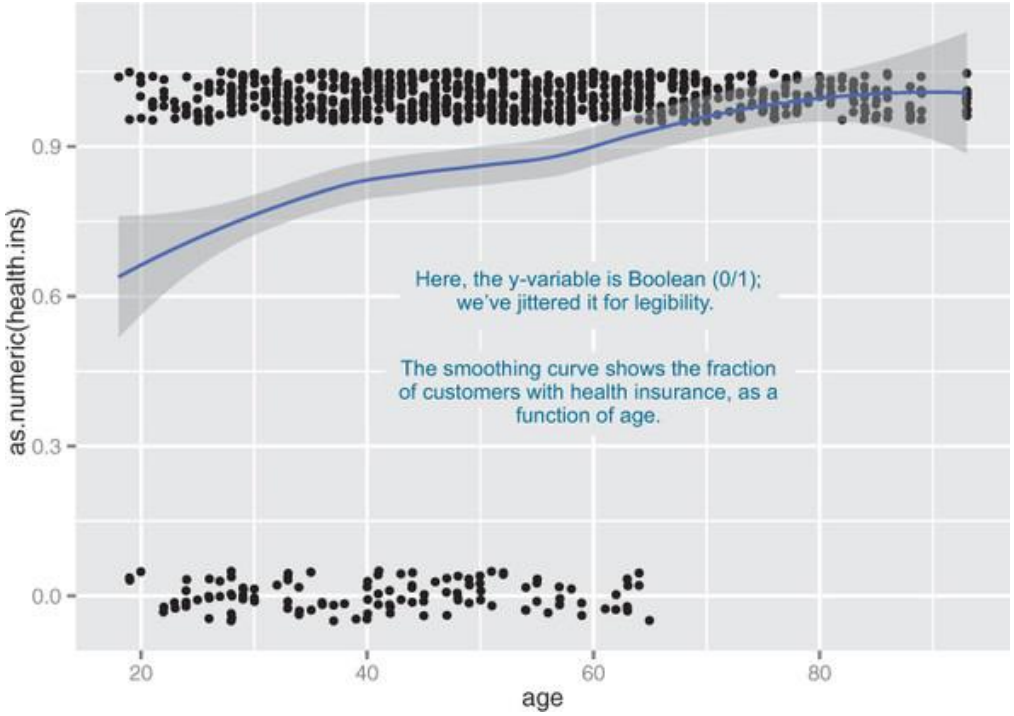
Visually checking relationships between two variables

- A scatter plot of income versus age, with a smoothing curve



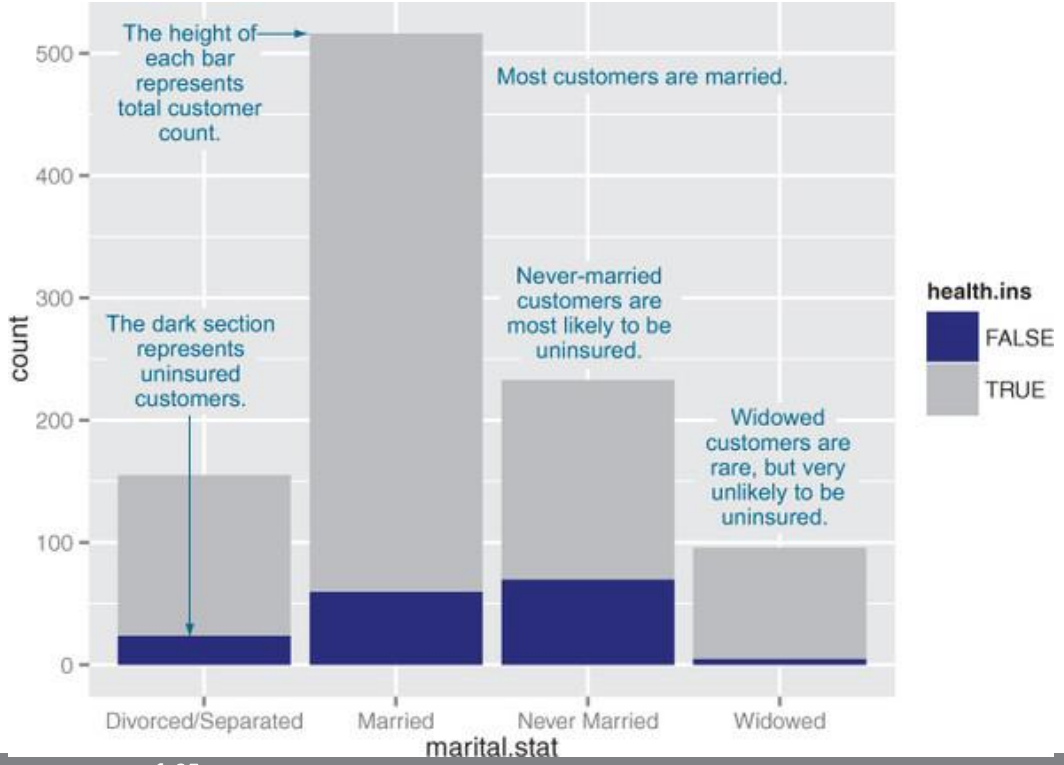
Visually checking relationships between two variables

- Distribution of customers with health insurance, as a function of age



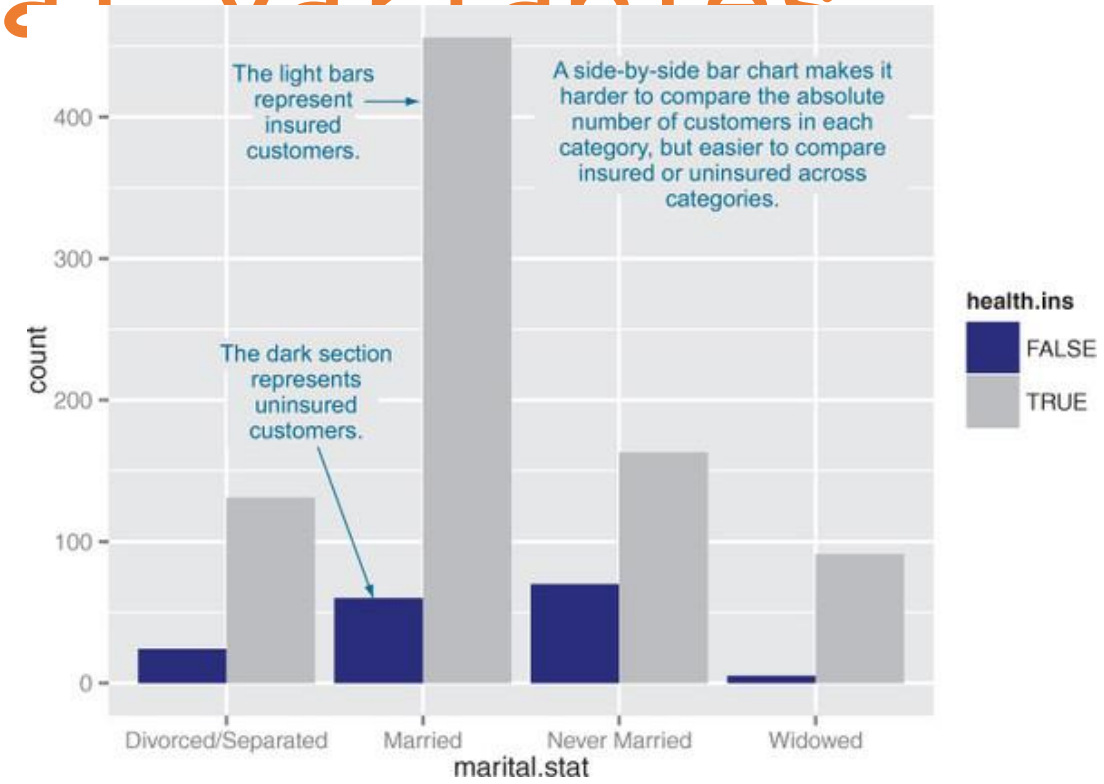
Bar charts for two categorical variables

- Health insurance versus marital status: stacked bar chart



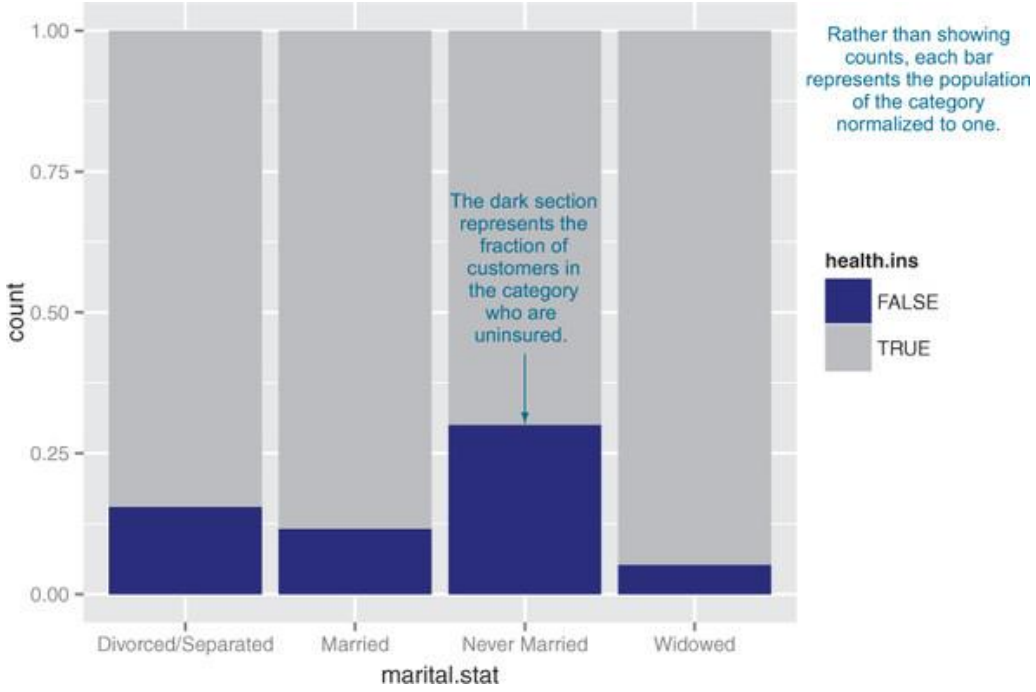
Bar charts for two categorical variables

- Health insurance versus marital status: side-by-side bar chart



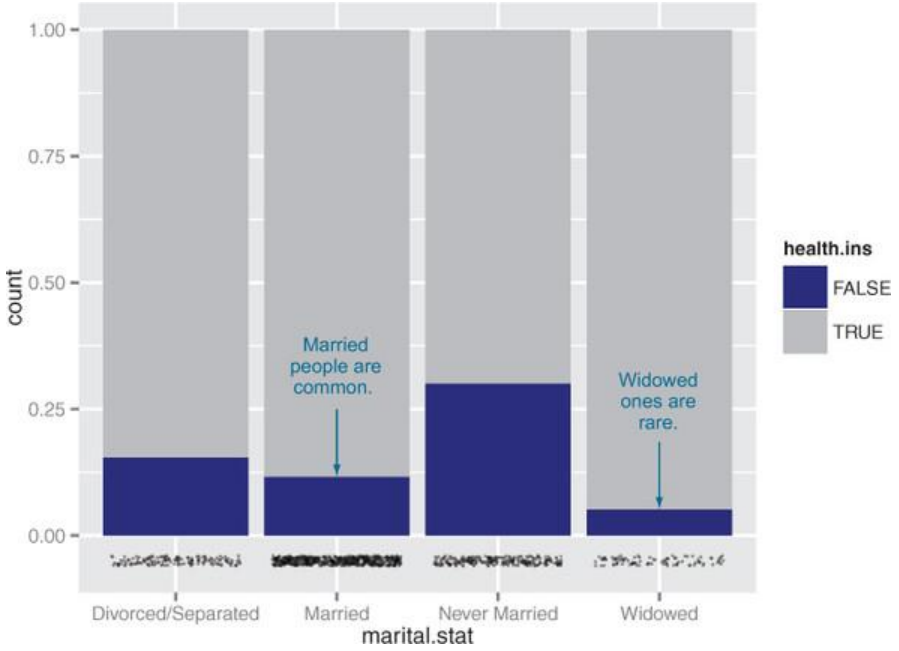
Bar charts for two categorical variables

- Health insurance versus marital status: filled bar chart



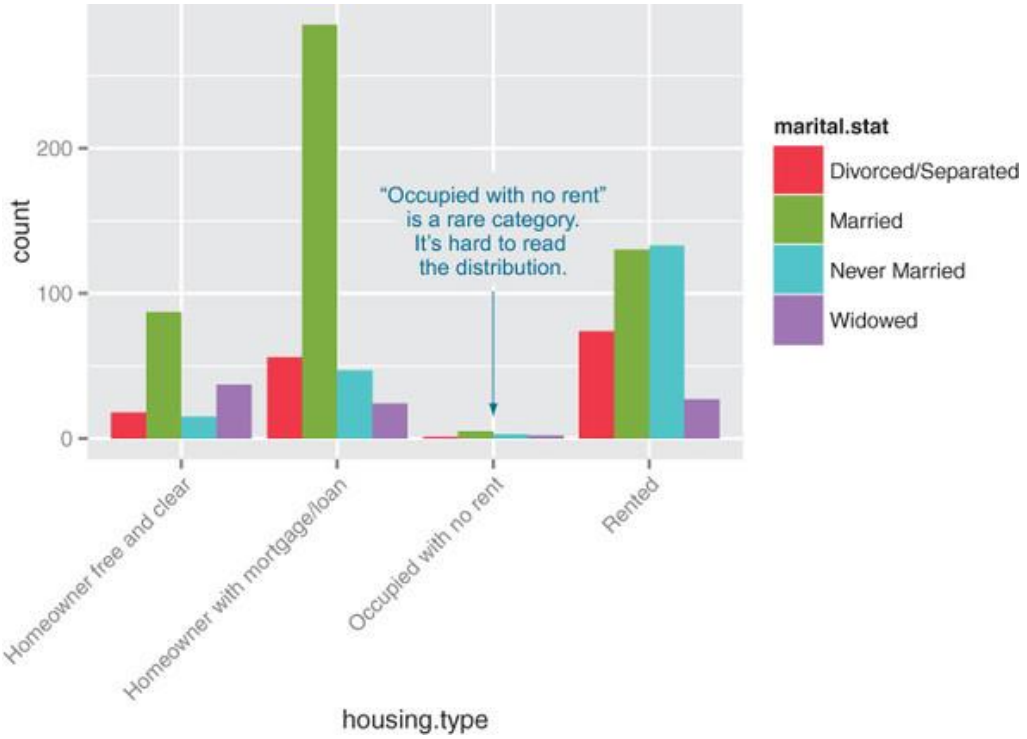
Bar charts for two categorical variables

Health insurance versus marital status: filled bar chart with rug



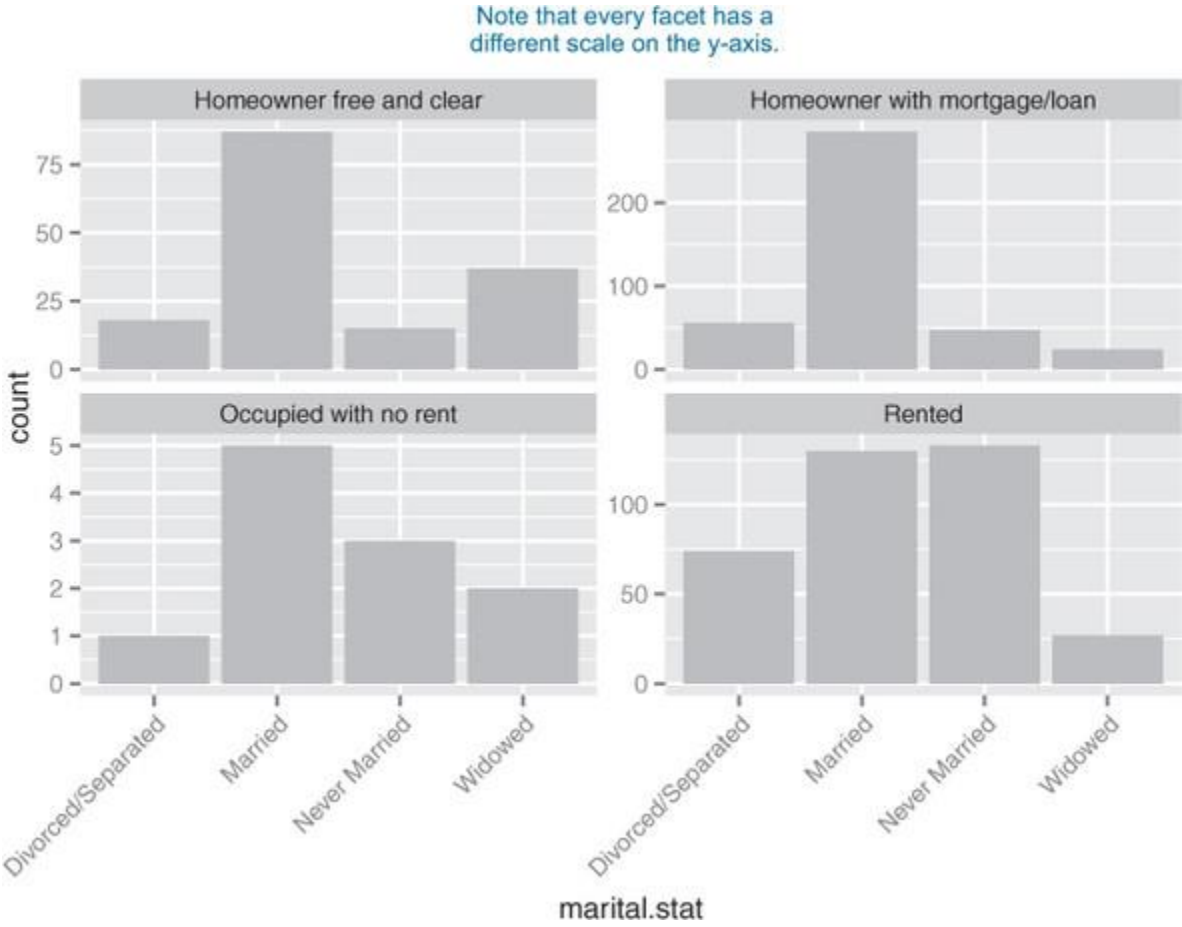
Bar charts for two categorical variables

- Distribution of marital status by housing type: side-by-side bar chart



Bar charts for two categorical variables

Distribution of marital status by housing type: faceted side-by-side bar chart



Summary

- At this point, you've gotten a feel for your data.
- You've explored it through summaries and visualizations; you now have a sense of the quality of your data, and of the relationships among your variables.
- You've caught and are ready to correct several kinds of data issues—although you'll likely run into more issues as you progress.

Complete Lab"