

Exploring Data





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>

The screenshot shows an Excel spreadsheet titled 'Analysis_Template.xlsx'. The data is organized into columns A through H. Column A contains the variable name 'c_community'. Column B contains numerical values. Column C lists various summary statistics, numbered 1 through 15. Columns D through H contain the corresponding numerical results for these statistics. Red boxes and arrows highlight specific sections of the data:

- Descriptive Statistics:** Rows 2-6, covering Count, Mean, Median, Mode, Variance, and Standard Deviation.
- More Summary Statistics:** Rows 7-12, covering Range, Interquartile Range, Maximum, Minimum, Skewness Coefficient, SE Skewness, Standard Coefficient of Skewness, Kurtosis Coefficient, SE Kurtosis, and Standard Coefficient of Kurtosis.
- Inferential Statistics (out of scope):** Rows 13-15, covering T-scores and NCE-scores.
- Skew:** Rows 16-18, covering Skewness Coefficient, SE Skewness, and Standard Coefficient of Skewness.
- Deciles:** Rows 19-20, covering 90th and 10th percentiles.
- The data column you are analyzing:** A red box points to column A, which contains the variable name 'c_community'.

	A	B	C	D	E	F	G	H
1	c_community		Count	169		z-scores	T-scores	NCE-scores
2	23		Mean	28.84023669		-0.938489418	40.61510582	30.23541286
3	22		Median	0.478693704		-1.099183148	39.00816852	26.85120291
4	23		Mode	22		-0.938489418	40.61510582	30.23541286
5	23					-0.938489418	40.61510582	30.23541286
6	22					-1.099183148	39.00816852	26.85120291
7	32		Variance	38.72595497		0.507754153	55.07754153	60.69330246
8	24		Standard Deviation	6.223018156		0.777795688	42.22204312	33.61962282
9	22		Range	25		-1.099183148	39.00816852	26.85120291
10	28		Interquartile Range	10		-0.135020767	48.64979233	47.15646264
11	25		Maximum	40		0.617101958	43.82898042	37.00383277
12	22		Minimum	15		-1.099183148	39.00816852	26.85120291
13	23					-0.938489418	40.61510582	30.23541286
14	33		Skewness Coefficient	-0.073045168		0.668447883	56.68447883	64.07751242
15	19		SE Skewness	0.188422288		-1.581264338	34.18735662	16.69857304
16	38		Standard Coefficient of Skewness	0.387667347		1.471916534	64.71916534	80.9985622
17	27		Kurtosis Coefficient	-1.044172509		-0.295714497	47.04285503	43.77225268
18	19		SE Kurtosis	0.376844576		-1.581264338	34.18735662	16.69857304
19	24		Standard Coefficient of Kurtosis	-2.770830673		-0.777795688	42.22204312	33.61962282
20	15					-2.224039259	27.75960741	3.161733215
21	32		90th percentile	37.2		0.507754153	55.07754153	60.69330246
22	26		10th percentile	21		-0.456408227	45.43591773	40.38804273

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."

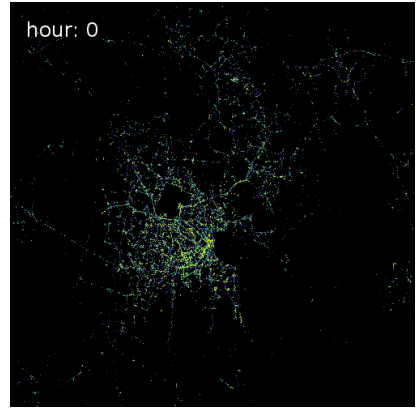
Either way, you should check how prevalent the issue is, and decide what to do: Do you drop the data with negative income? Do you convert negative values to zero?

```
> 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 $\text{Income} = \text{custdata\$Income} / 1000$. 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

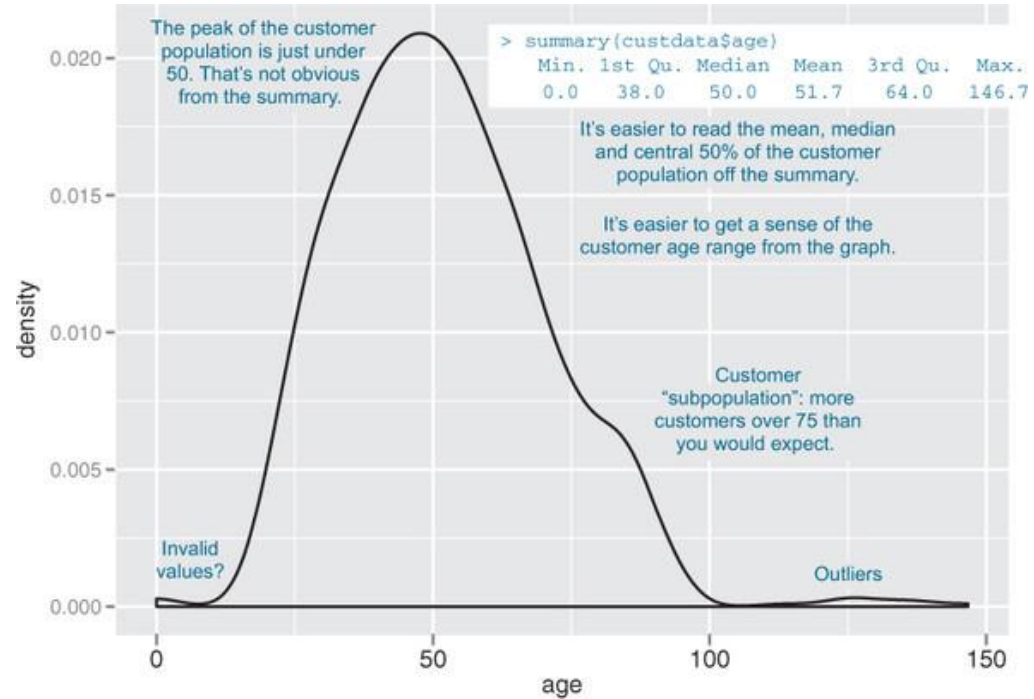
Spotting problems using graphics & visualization

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

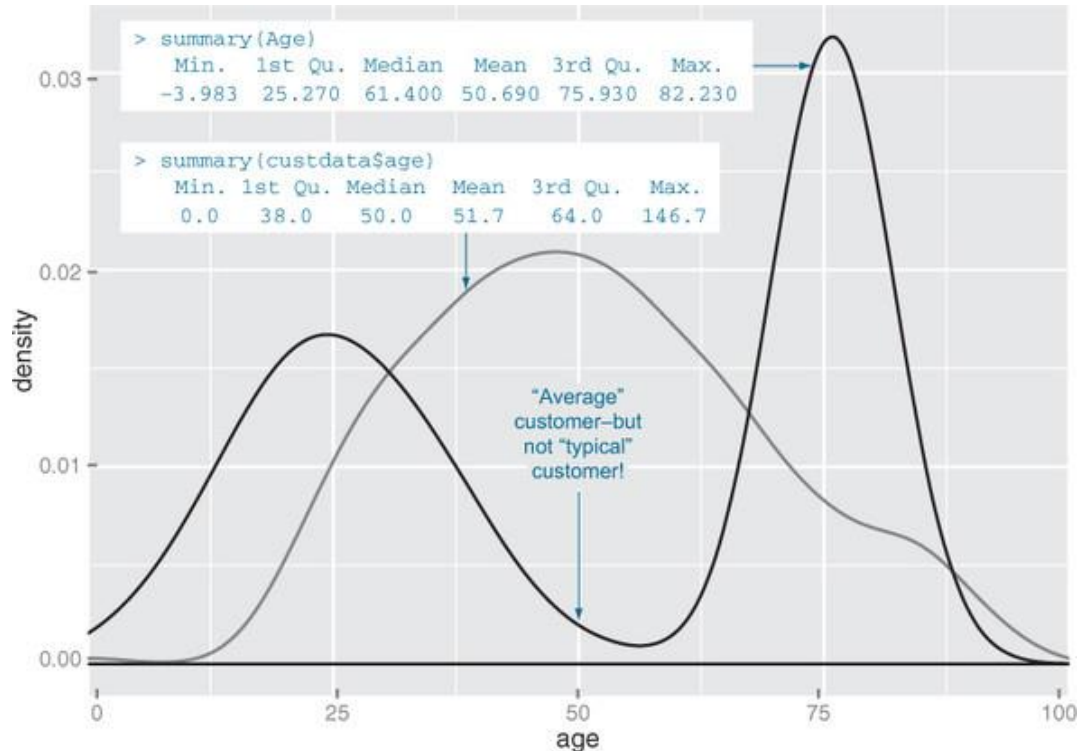
Spotting problems using graphics & visualization

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



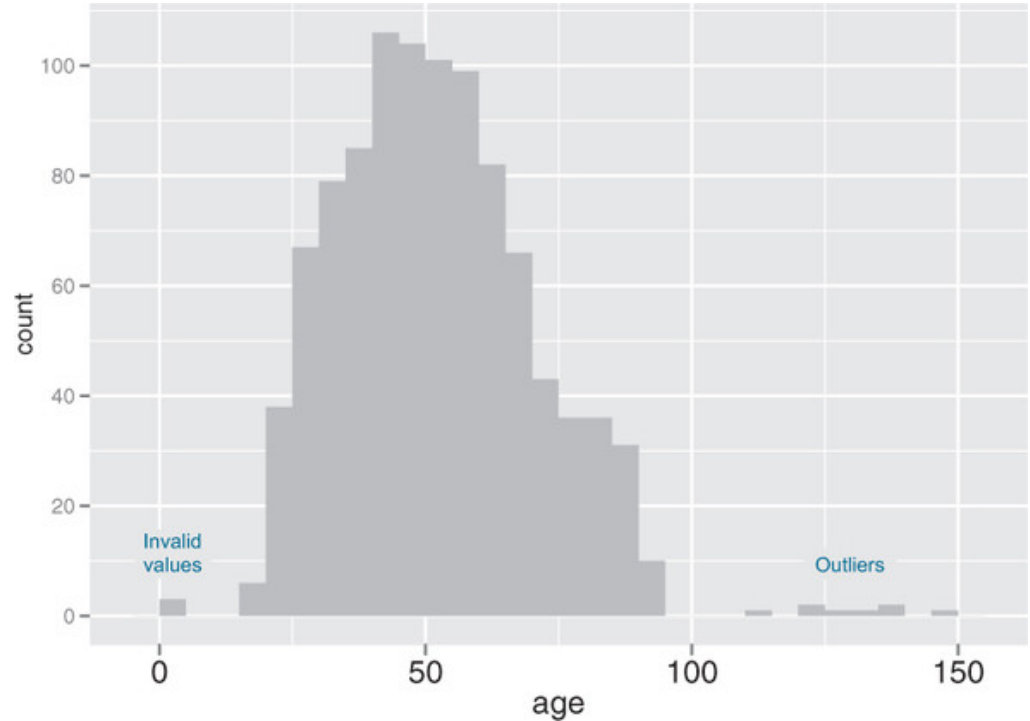
Visually checking distributions for a single variable

- 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)

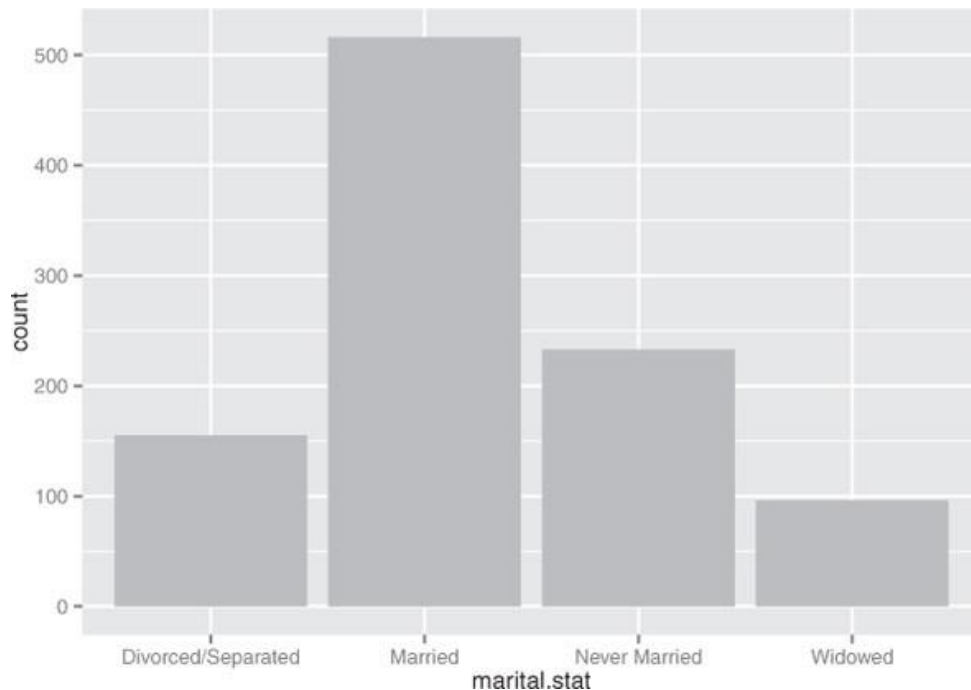
ggplot(custdata) +
  geom_histogram(aes(x=age),
    binwidth=5, fill="gray")
```

← 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).

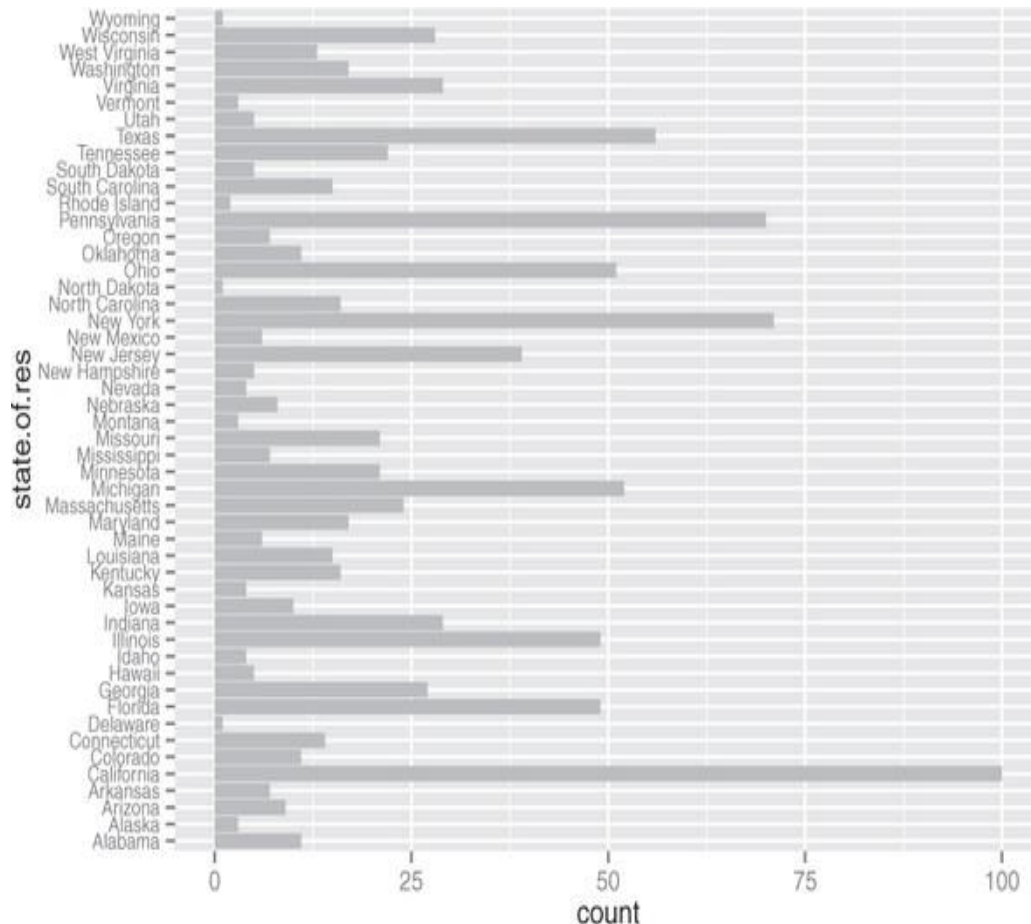
Bar charts

- Bar charts show the distribution of categorical variables.



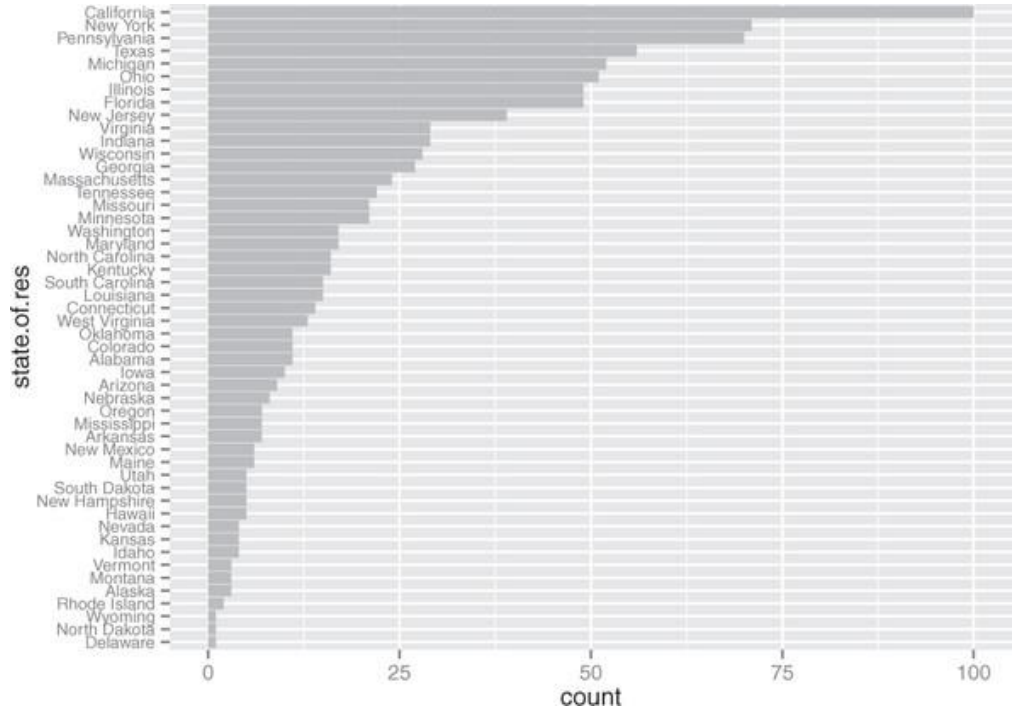
Bar charts

- 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

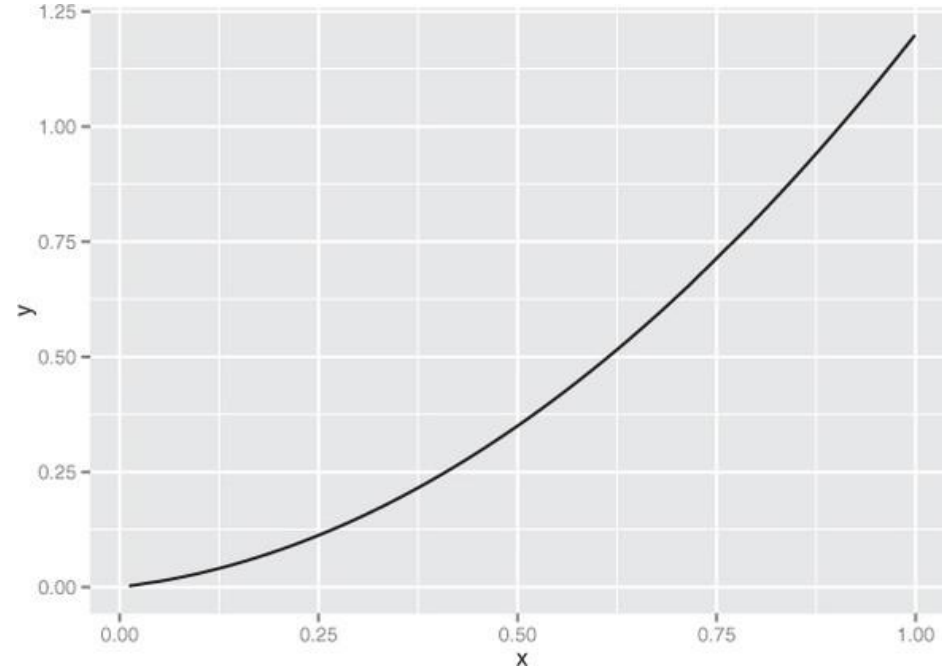
Graph type	Uses
Histogram or density plot	Examines data range Checks number of modes Checks if distribution is normal/lognormal Checks for anomalies and outliers
Bar chart	Compares relative or absolute frequencies of the values of a categorical variable

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?

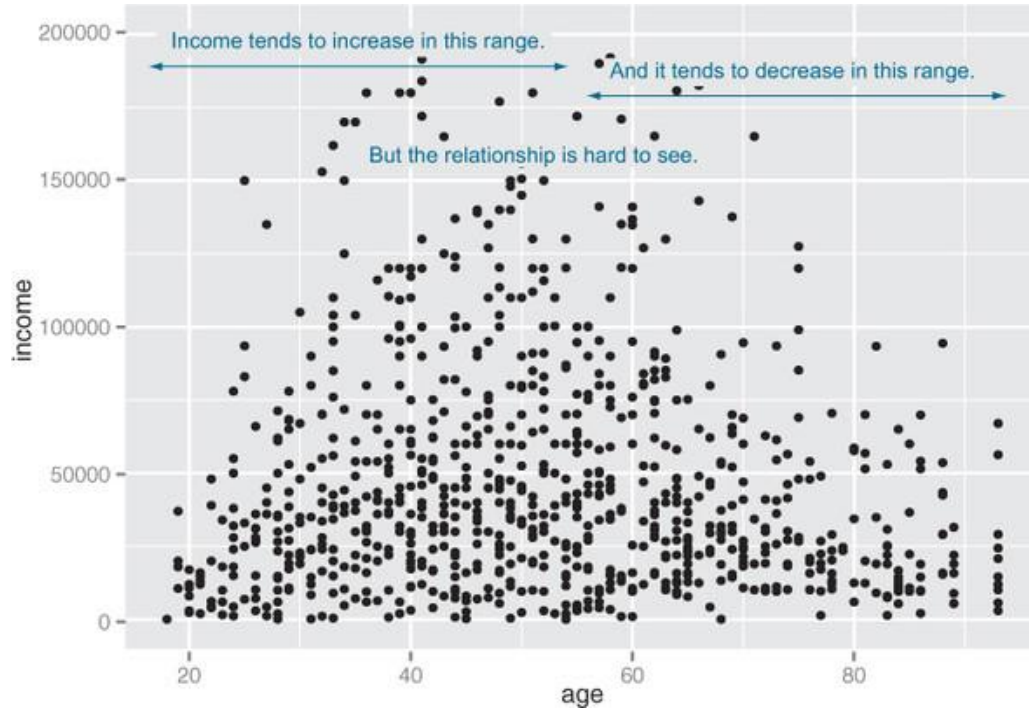
Example of a line plot



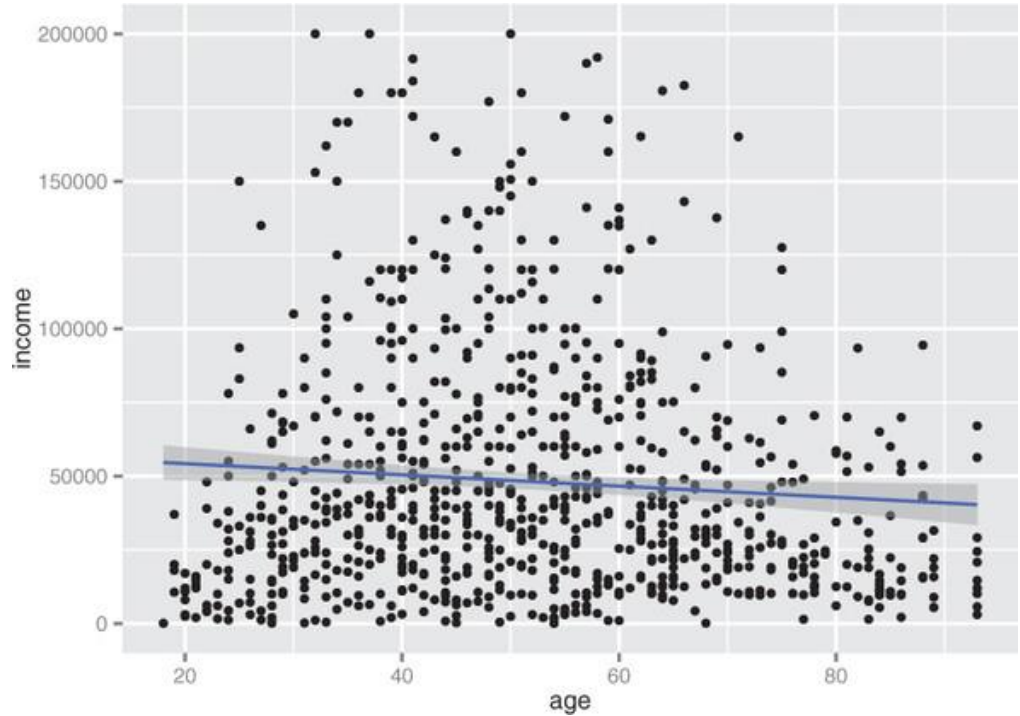
Scatter plots and smoothing curves

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

Scatter plots and smoothing curves

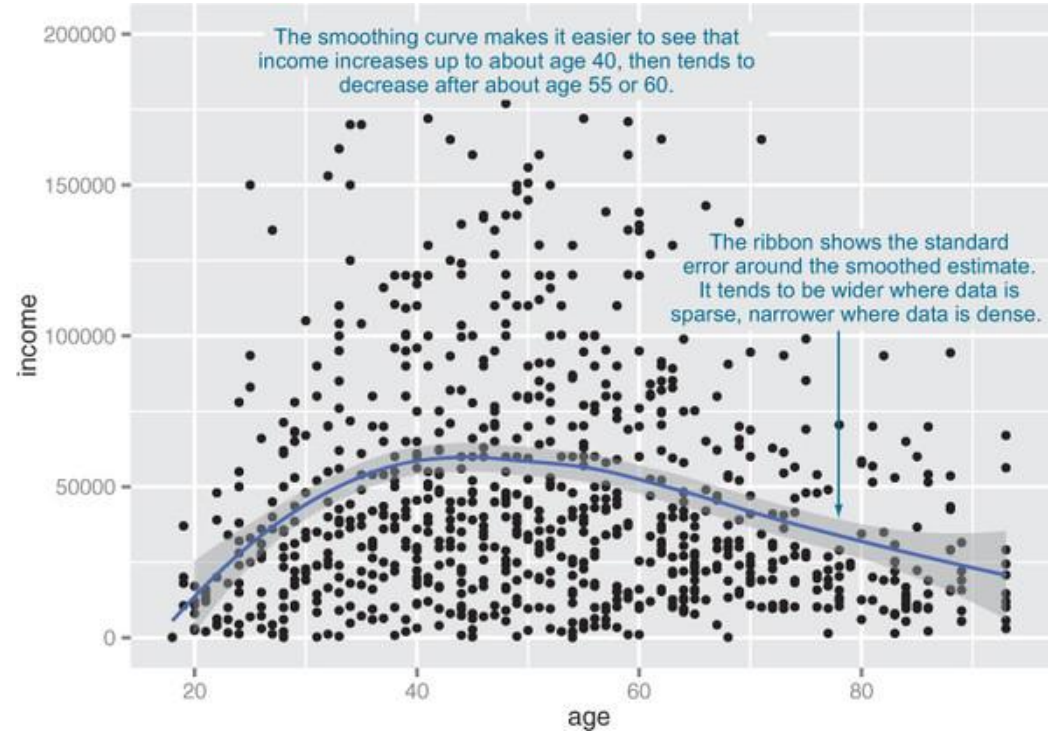


Scatter plots and smoothing curves



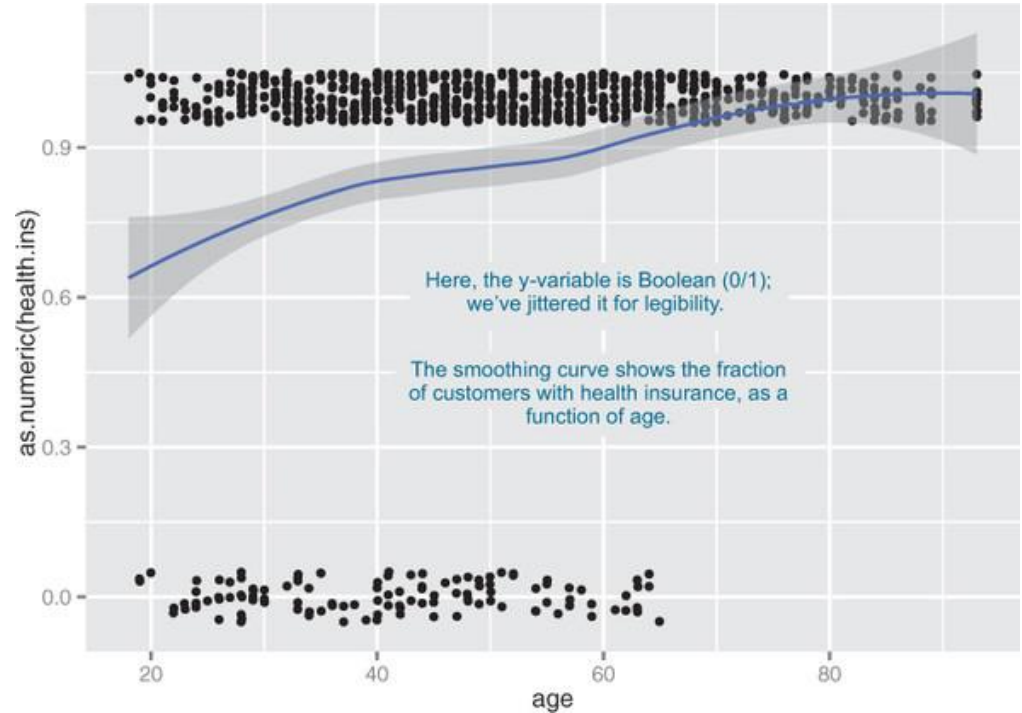
Visually checking relationships between two variables

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



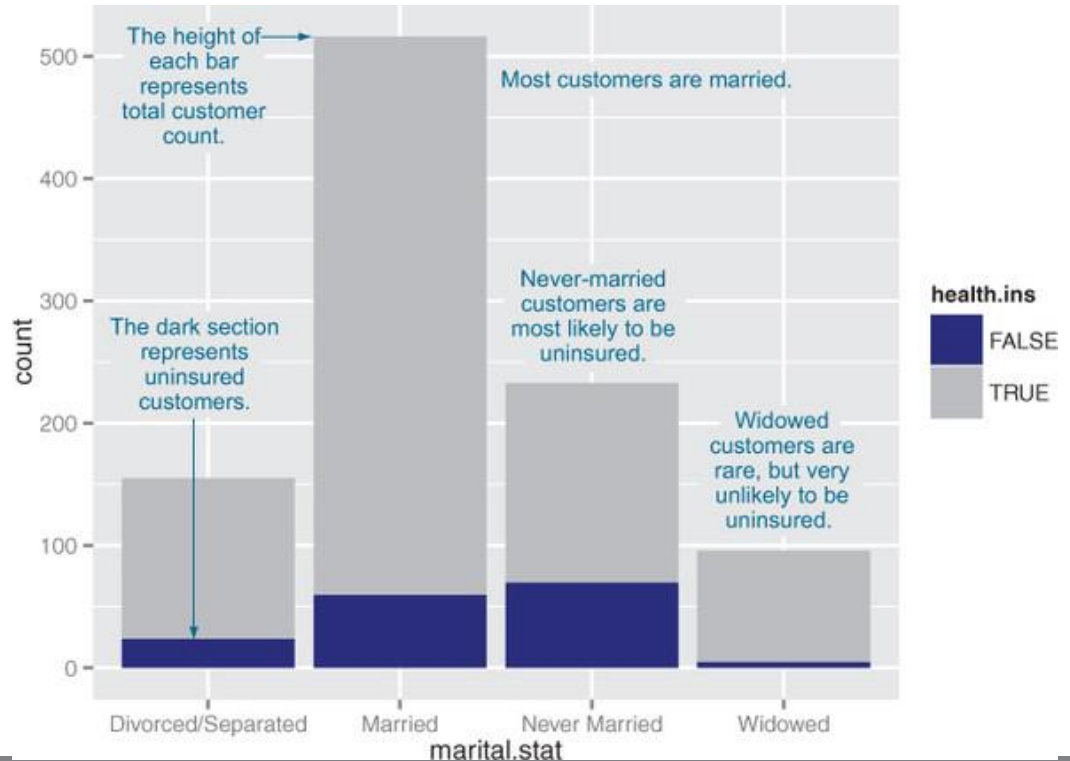
visually checking relationships between two variables

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



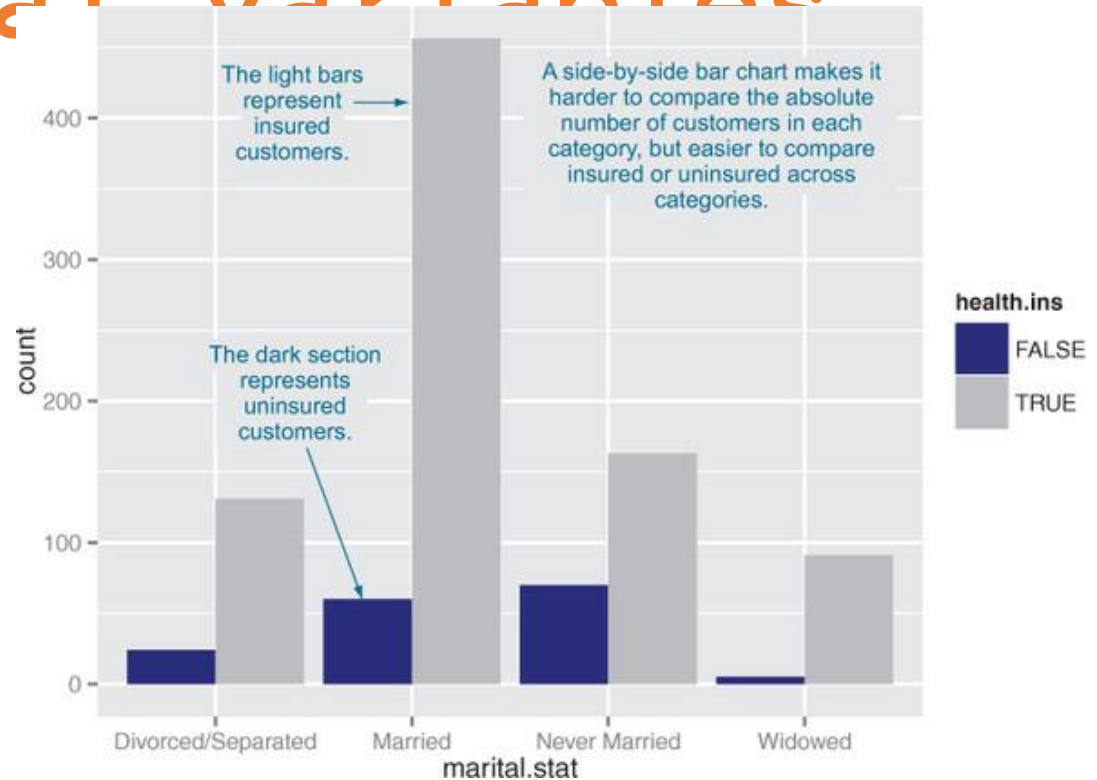
Bar charts for two categorical variables

- Health insurance versus marital status: stacked bar chart



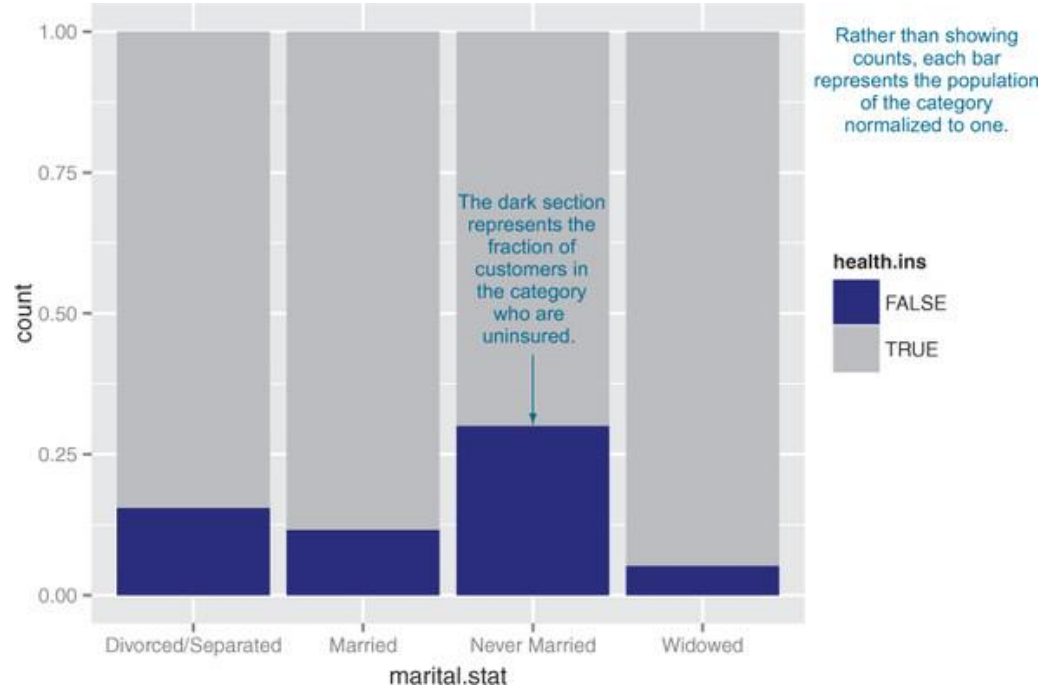
Bar charts for two categorical variables

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



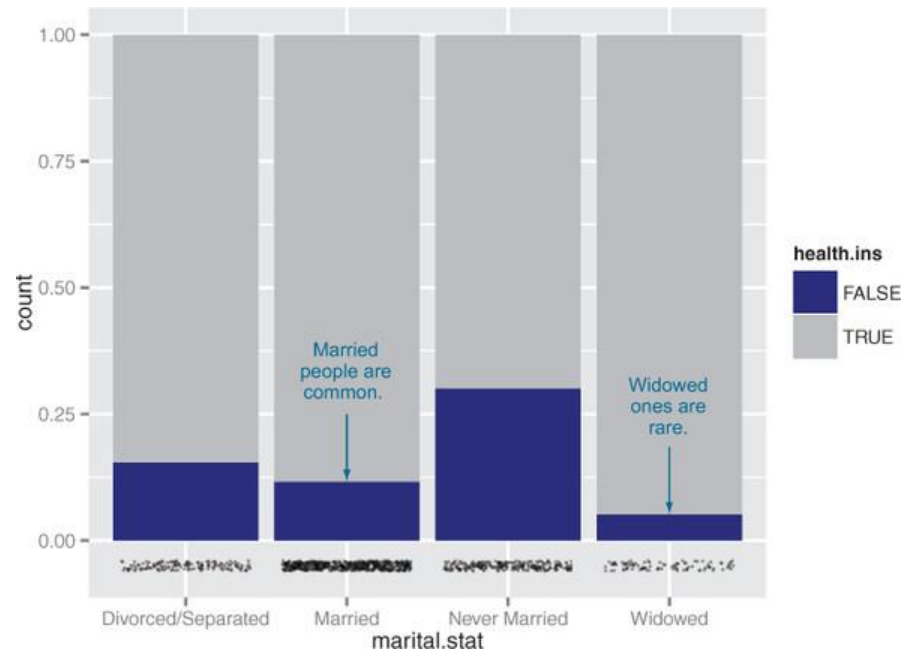
Bar charts for two categorical variables

- Health insurance versus marital status: filled bar chart



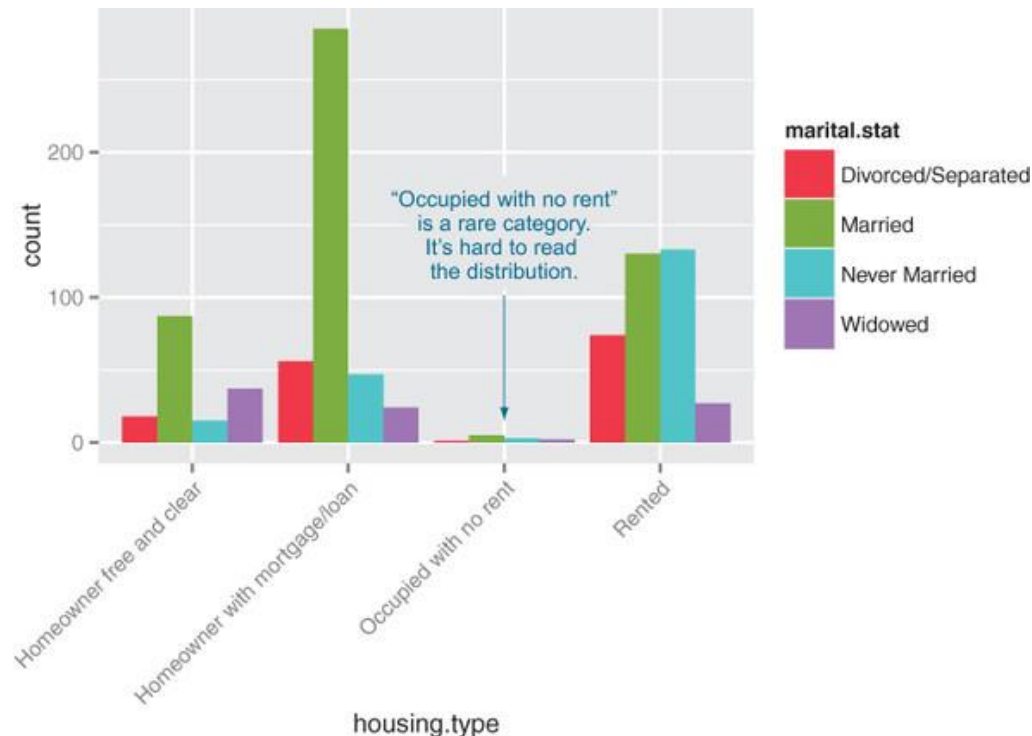
Bar charts for two categorical variables

Health insurance versus
marital status: filled
bar chart with rug



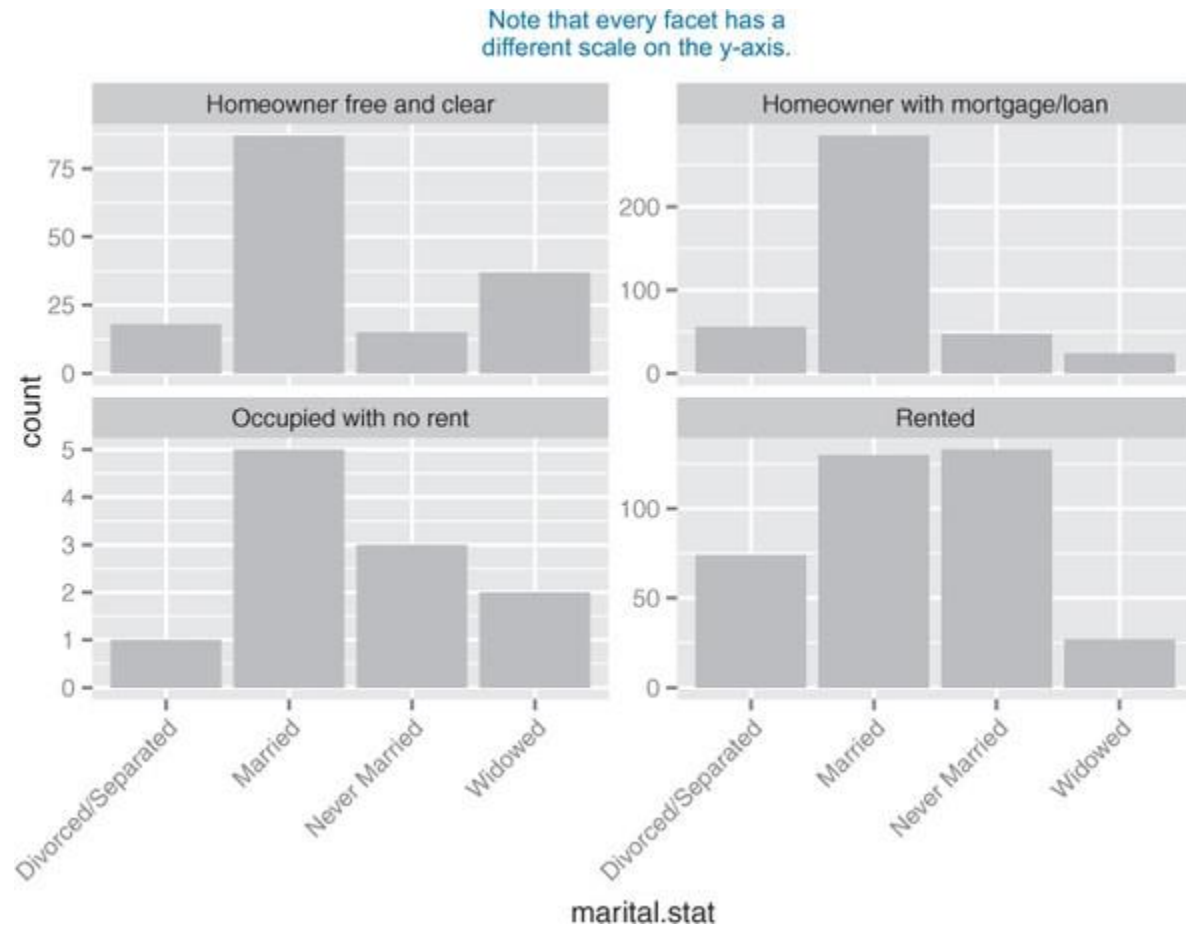
Bar charts for two categorical variables

- 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"