Considering Categorical Data

Colby Community College

Example 1

The following table summarizes two categorical variables from the full `loans` data set.

		homeownership			
		rent	*mortgage*	*own*	Total
app_type	*individual*	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Definition

A table that summarizes data for two categorical variables in this way is called a **contingency table**.

Definition

The **row totals** provide the total counts across each row.

The **column totals** provide the total counts down each column.

Note

You can also create a table that considers only a single variable.

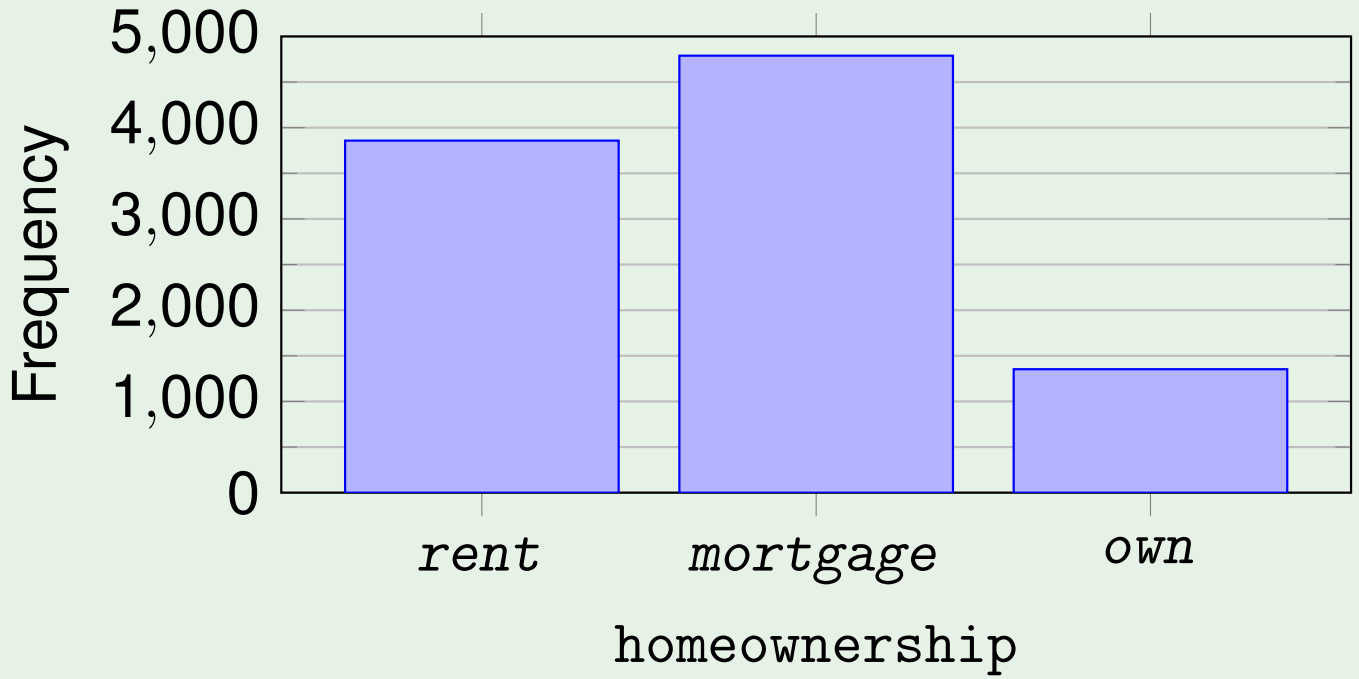
Example 2

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

Definition

A **bar plot** plots a bar for each variable outcome, the height is the frequency of the outcome.

Example 3



Note

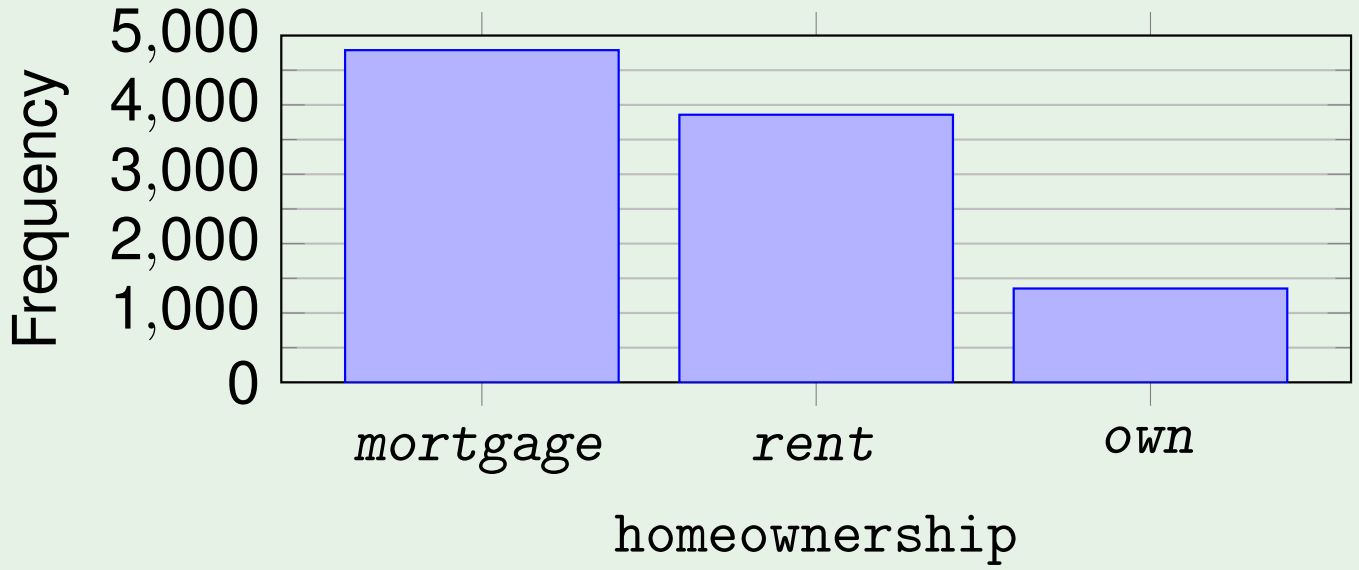
A histogram has no gaps between the bars, where as bar plot does.

Definition

A bar plot that has been sorted so the largest bar is on the left and the smallest bar on the right is called **Pareto chart**.

Example 4

Here is the Pareto chart for the bar plot in Example 3.



Note

Named after Vilfredo Pareto, a noted Italian economist.

Note

Instead of using frequencies, we could instead use proportions.

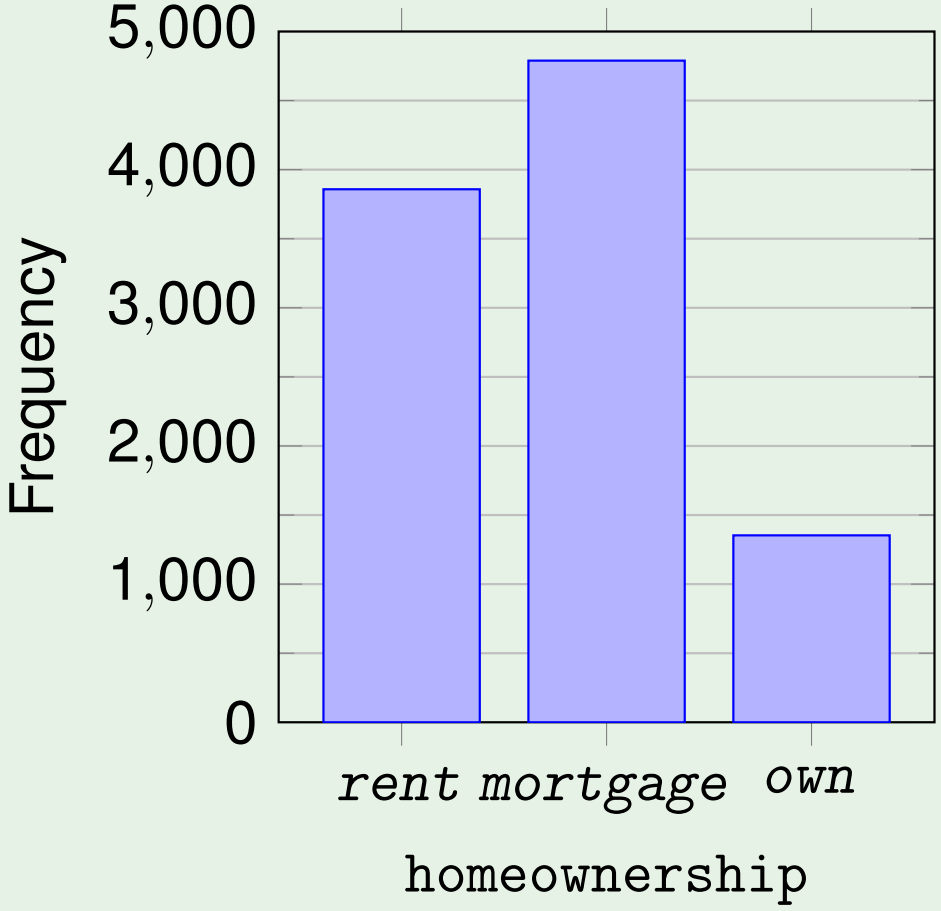
Example 5

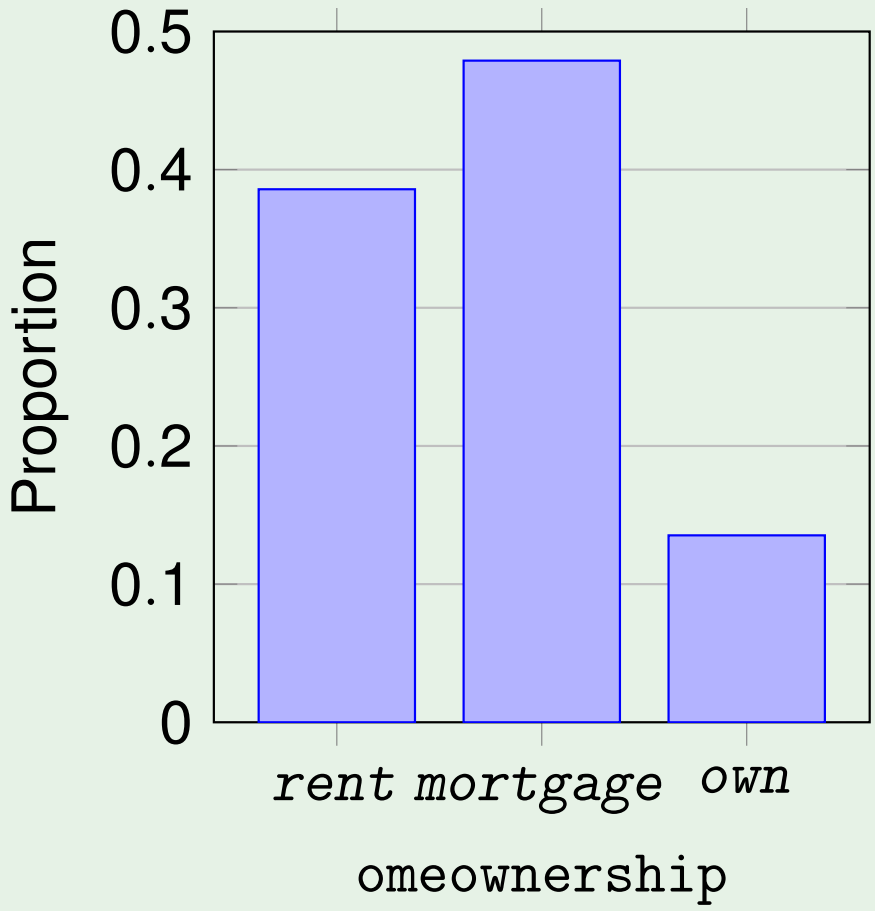
To find the proportion, divide each frequency by the total count.

`homeownership`	Frequency	Proportion
rent	3858	$\frac{3858}{10000} = 0.3858$
mortgage	4789	$\frac{4789}{10000} = 0.4789$
own	1353	$\frac{1353}{10000} = 0.1353$
Total	10000	1.0000

Example 6

Here are both the frequency and proportion for `homeownership`.





Example 7

Here, we use the **row proportions** for the contingency table from Example 1. Where we divide each count by their row total.

		homeownership			
		rent	*mortgage*	*own*	Total
app_type	*individual*	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

$\downarrow\downarrow\downarrow$

		homeownership			
		rent	*mortgage*	*own*	Total
app_type	*individual*	0.4111	0.4514	0.1376	1.0000
	joint	0.2421	0.6355	0.1224	1.0000
	Total	0.3858	0.4789	0.1353	1.0000

What does the number 0.4111 represent?

That 41.11% of those that applied as individuals are renters.

Example 8

Here, we use the **column proportions** for the contingency table from Example 1. Where we divide each count by their column total.

		homeownership			
		rent	*mortgage*	*own*	Total
app_type	*individual*	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

$\downarrow\downarrow\downarrow$

		homeownership			
		rent	*mortgage*	*own*	Total
app_type	*individual*	0.9062	0.8016	0.8647	0.8505
	joint	0.0946	0.1984	0.1353	0.1495
	Total	1.0000	1.0000	1.0000	1.0000

What does the number 0.9062 represent?

That 90.62% of renters applied as individuals.

Example 9

Let us look at the column proportions from Example 8 more closely.

		homeownership			
		rent	*mortgage*	*own*	Total
app_type	*individual*	0.9062	0.8016	0.8647	0.8505
	joint	0.0946	0.1984	0.1353	0.1495
	Total	1.0000	1.0000	1.0000	1.0000

Notice that 90.62% of renters applied as individuals, which is higher than for those with mortgages (80.16%) or those who own (86.47%).

Because these rates vary between the three levels of homeowndership (*rent*, *mortgage*, *own*), there is evidence that the app_type and homeownsership variables are associated.

Note

If we had considered row proportions instead, we would look down columns to see if the fraction of loans where the borrower rents, has a mortgage, or owns varied across the *individual* to *joint* types.

Example 10

Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, an email may be classified as spam or not spam.

One such characteristic is the email format, which indicates whether or not an email has any HTML content, such as bolded text.

	text	*HTML*	Total
spam	209	158	367
not spam	986	2568	2554
Total	1195	2726	3921

	text	*HTML*	Total
spam	0.175	0.058	0.094
not spam	0.825	0.942	0.651
Total	1.000	1.000	1.000

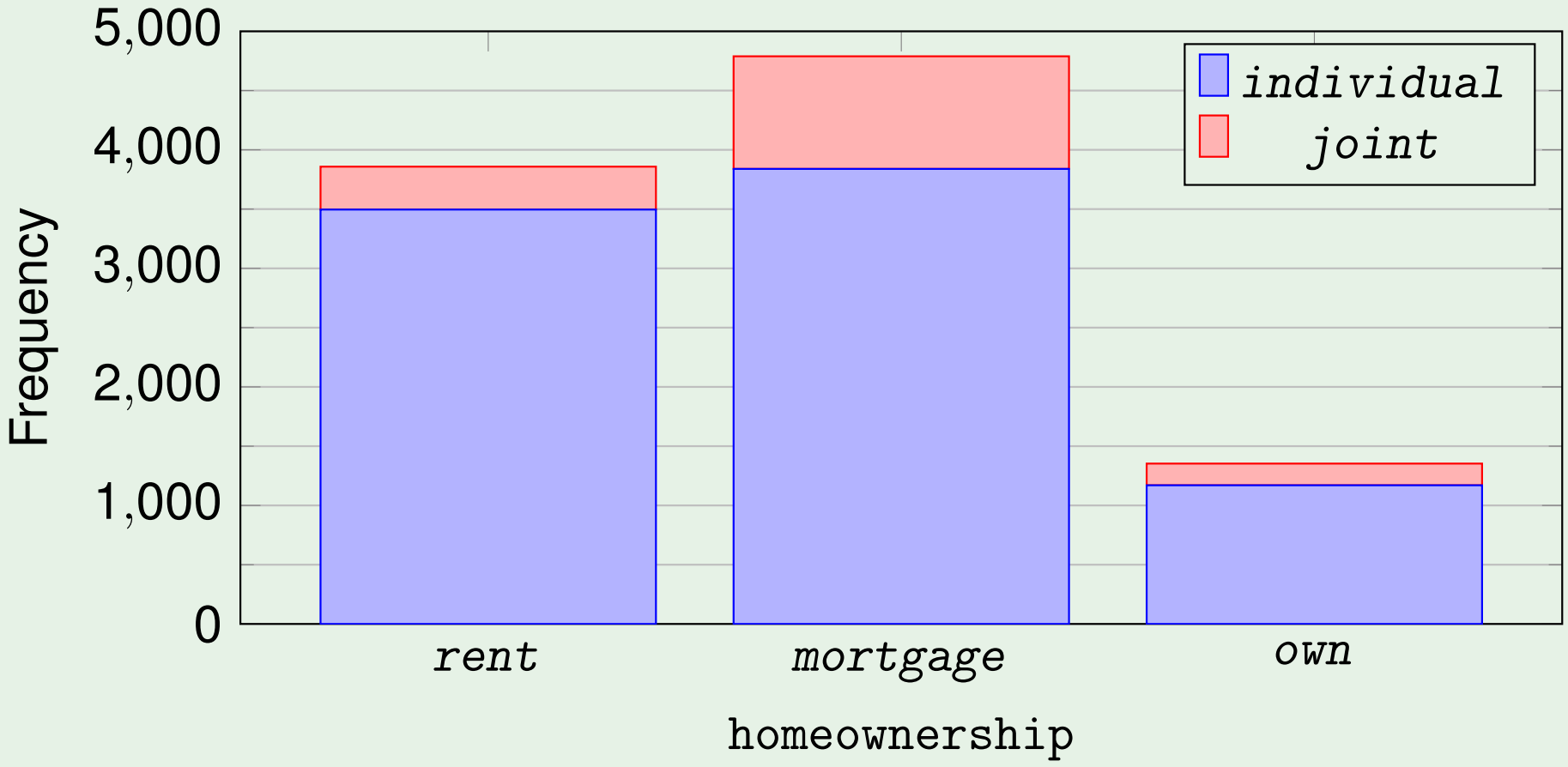
If we generate column proportions, we can see that a higher fraction of plain text emails are spam (17.5%) than compared to HTML emails (5.8%).

But, this information is not enough on its own to classify an email and *spam* or *not spam*, since more than 80% of plain text emails are not spam.

Definition

A **stacked bar plot** is a graphical display of contingency tables.

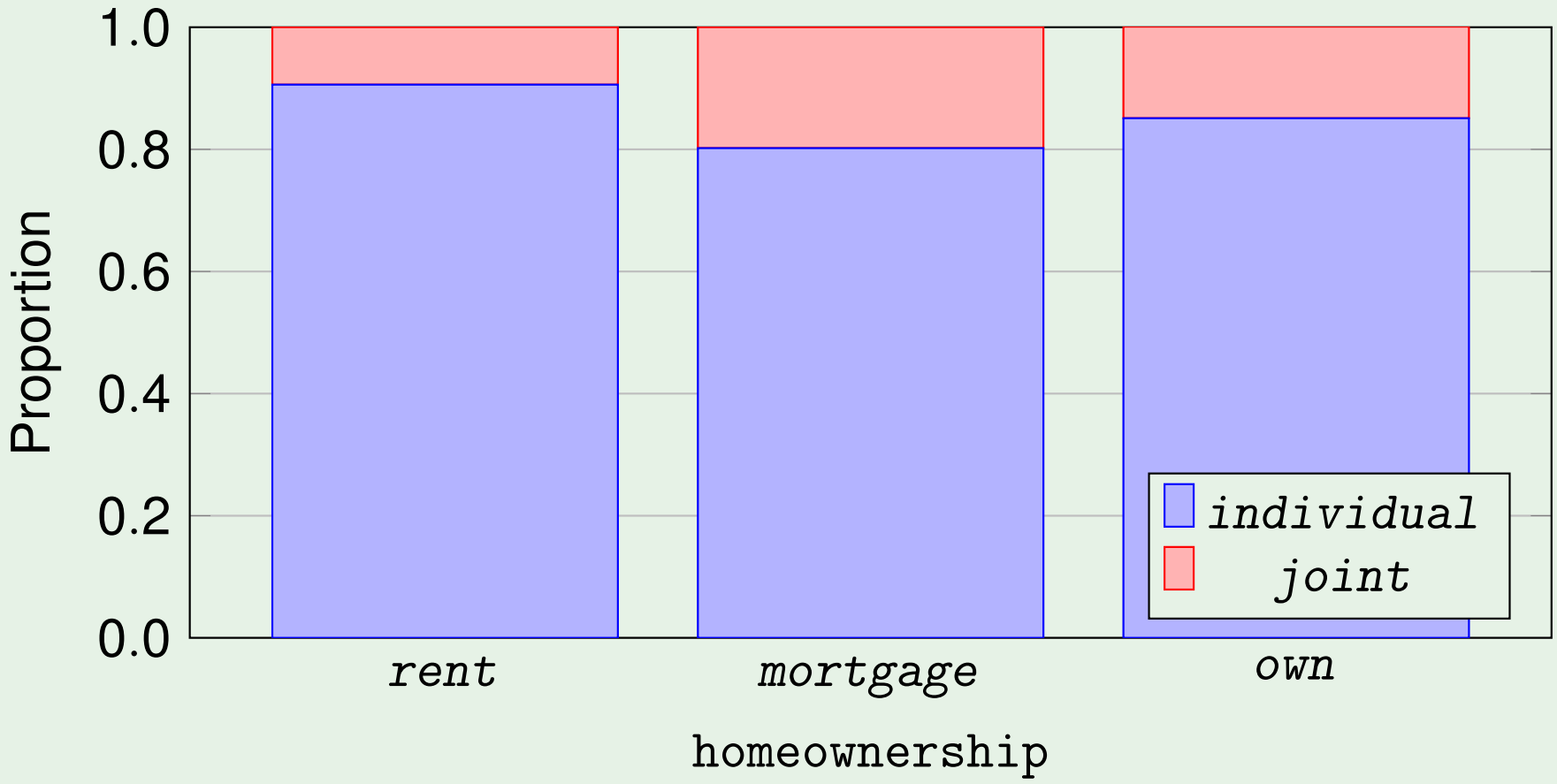
Example 11



Definition

A stacked bar plot generated from column proportions is called a **standardized stacked bar plot**.

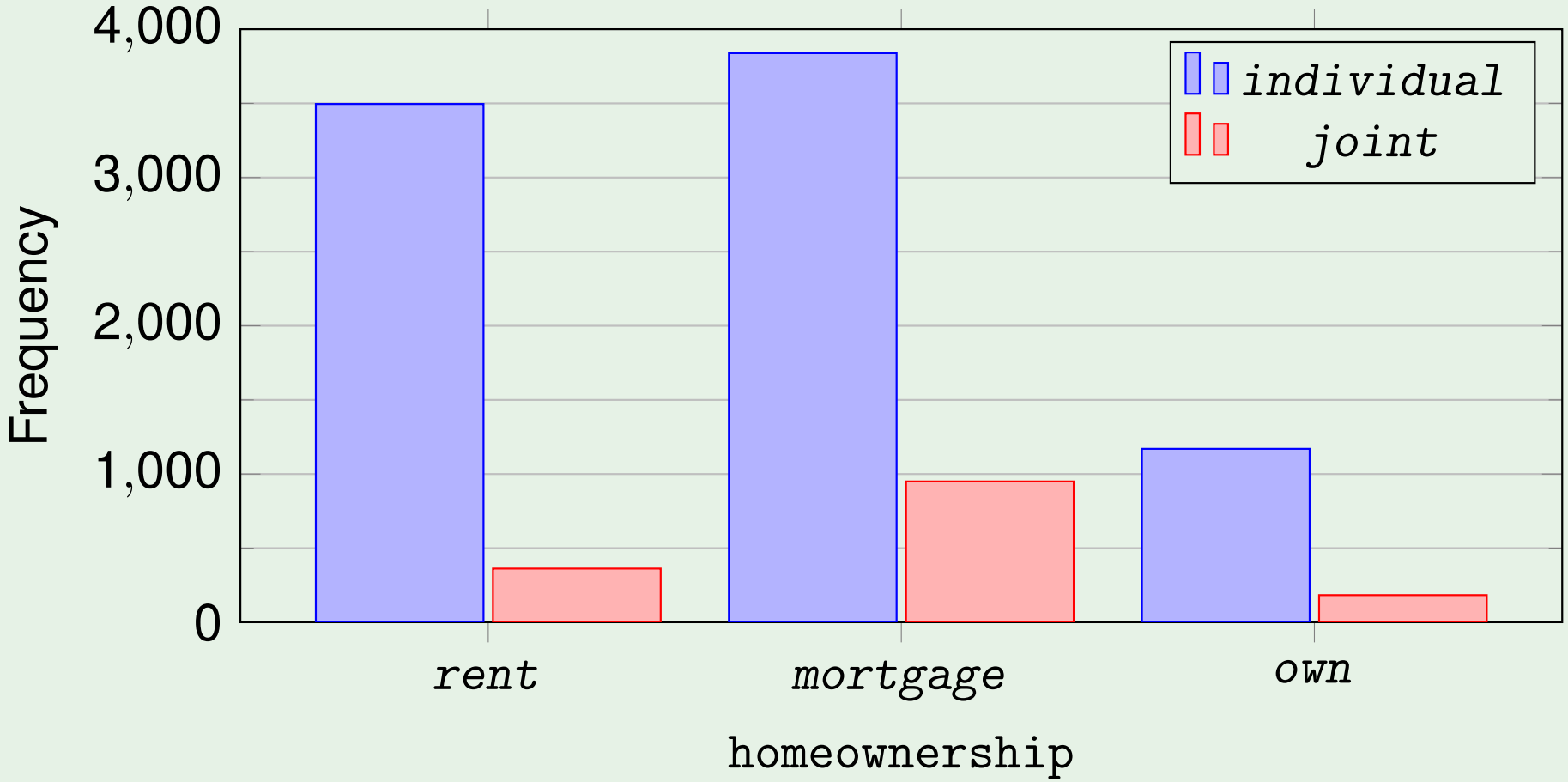
Example 12



Definition

A **side-by-side bar plot** is graphical display of contingency tables.

Example 13



Note

The stacked bar plot is most useful when it's reasonable to assign one variable as the explanatory variable and the other variable as the response.

Note

Side-by-side bar plots are more agnostic in their display about which variable, if any, represents the explanatory variable and which the response variable.

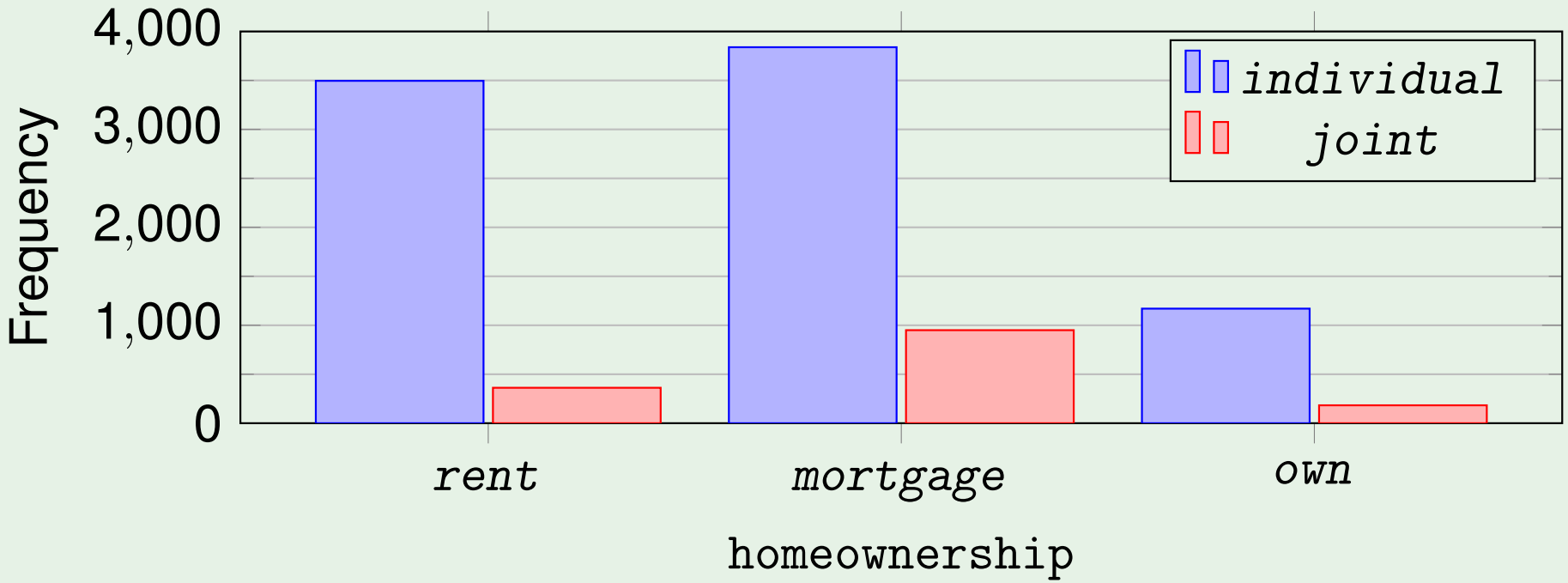
Note

The standardized stacked bar plot is useful when the primary variable in the stacked bar plot is relatively imbalanced.

Note

When a group of bars have very different sizes, relative to the other groups, it is difficult to discern if there is an association between the variables.

Example 14

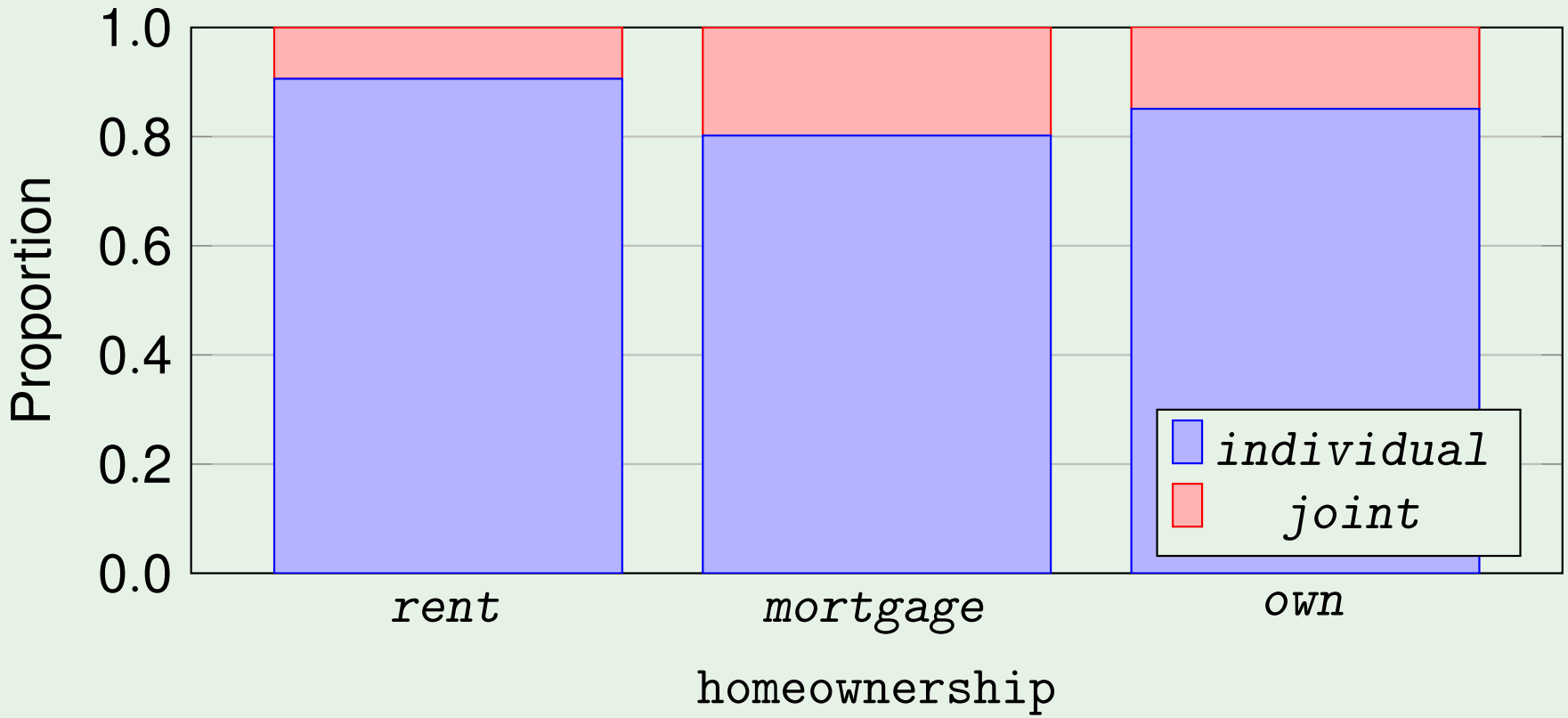


The bars for the *own* group are both much shorter than the other bars, so you can't tell from this plot if there is an association.

Note

A standardized stacked bar plot is useful for checking for association.

Example 15

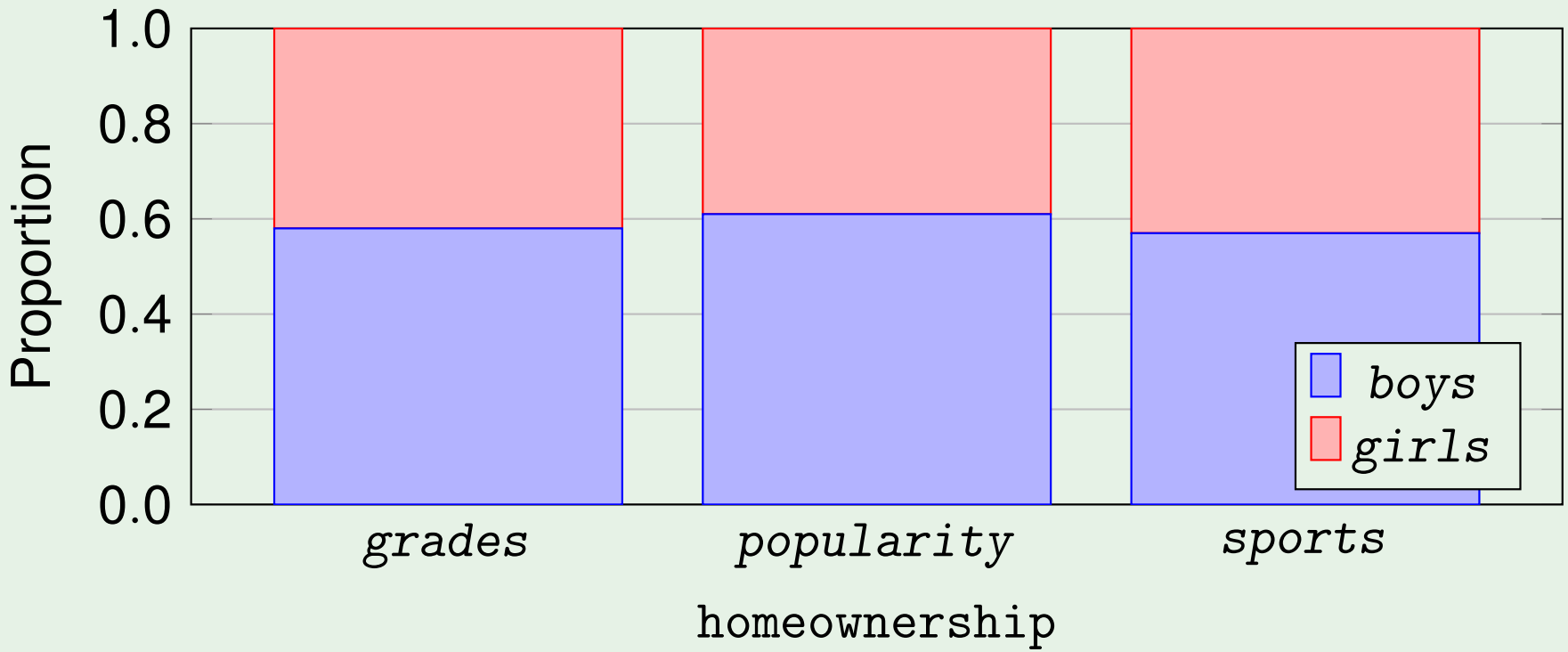


Since the proportions vary for each outcome of `homeownership`, there is an association.

Example 16

Suppose grade school students were surveyed and asked which of the following they though was the most important: grades, popularity, sports.

The standardized stacked bar plot represents the responses.



Since the proportions do not signifigantly vary for each outcome, the variables `gender` and `interest` are independent.

Note

To compare numerical and categorical variable, you can group the numbers based on the possible outcomes of the categorical variable, then draw a plot for each group.

Example 17

If we wish to compare the median income to whether a county gained population, we could build the table.

Median Income for U.S. Counties, in $1000s

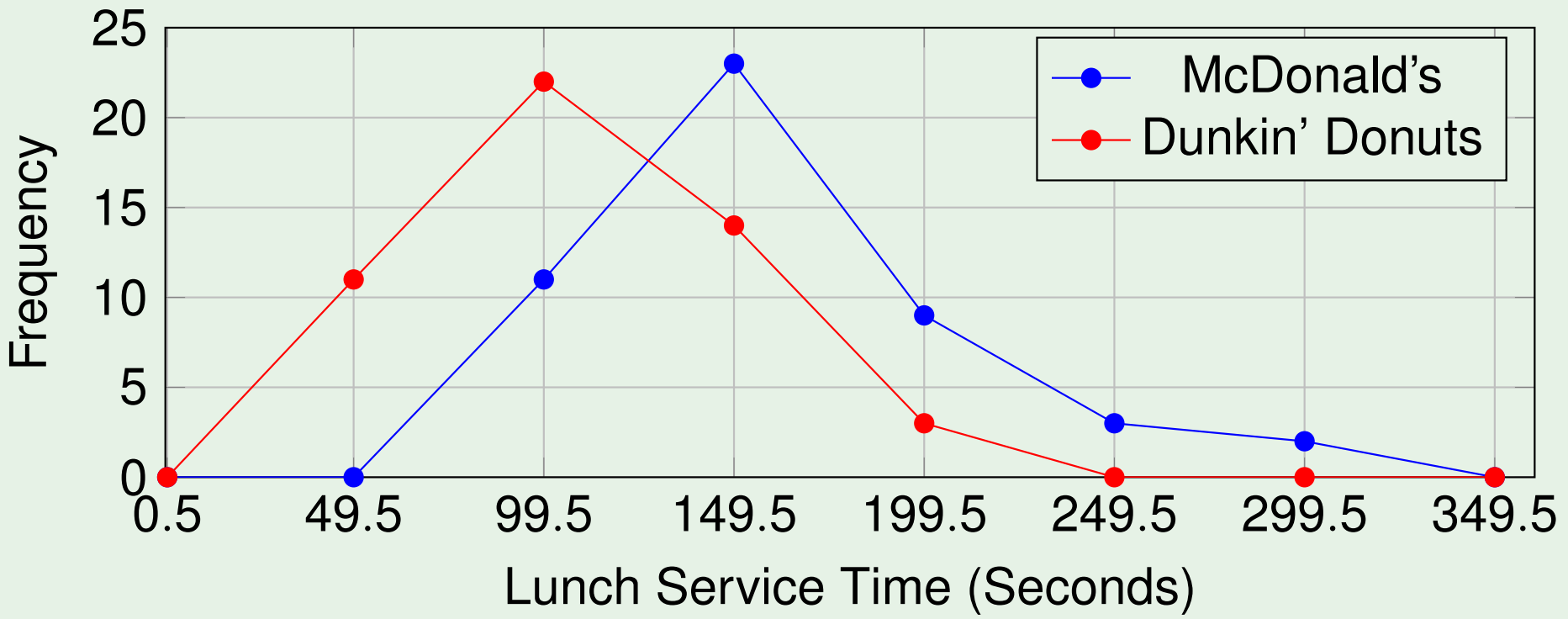
Population Gain						No Population Gain		
38.2	43.6	42.2	61.5	51.1	45.7	48.3	60.3	50.7
38.2	43.6	44.6	51.8	40.6	63.3	48.3	60.3	50.7
51.1	34.1	80.8	46.3	75.2	40.6	39.3	40.4	40.3
51.9	34.7	61.0	51.4	53.8	57.6	57.0	47.2	45.9
53.1	54.6	63.0	49.1	46.6	46.5	42.3	41.5	46.1
74.2	63.0	63.2	47.6	50.4	49.0	44.9	51.7	46.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Definition

A **frequency polygon** or **hollow histogram** allows us to compare numerical and categorical variables. It uses line segments connected to points located where the top of the bar would be.

Example 18

The plot shows the wait times for both McDonald's and Dunkin' Donuts.



Note

For small data sets (20 values or fewer), use a table instead of a graph.

Note

A graph of data should draw focus to the true nature of the data, not on other elements, such as eye-catching but distracting design features.

Note

Do not distort data. Construct a graph to reveal the true nature of the data.

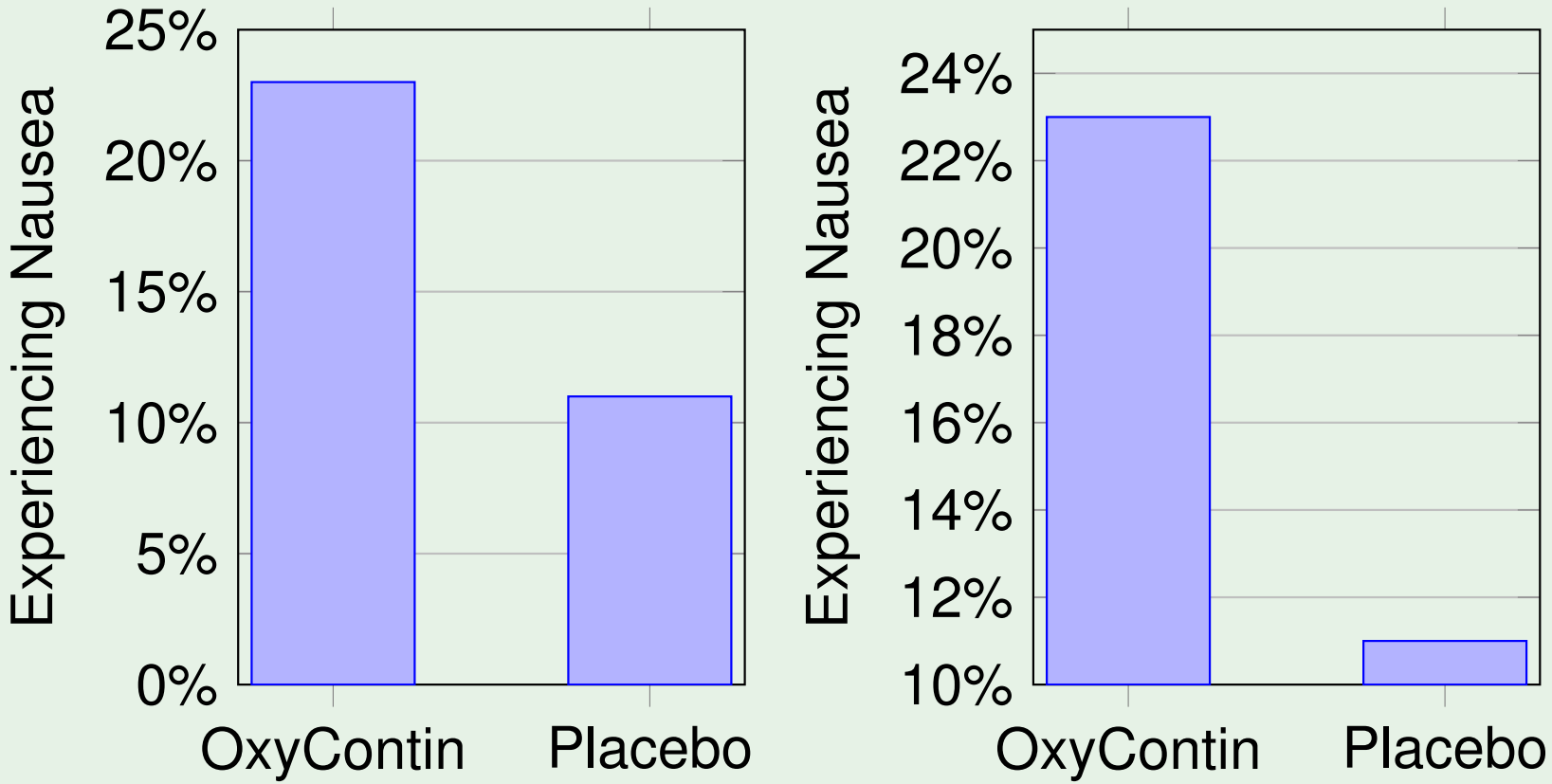
Note

Almost all of the ink in a graph should be used for data, not for other design elements.

Note

Always examine a graph carefully to see whether a vertical axis begins at some point other than zero so that differences are exaggerated.

Example 19



Note

When examining data depicted with a pictograph, beware when area or volume is used to depict amounts that are actually one-dimensional.

Example 20

The pictographs show data from the CDC.

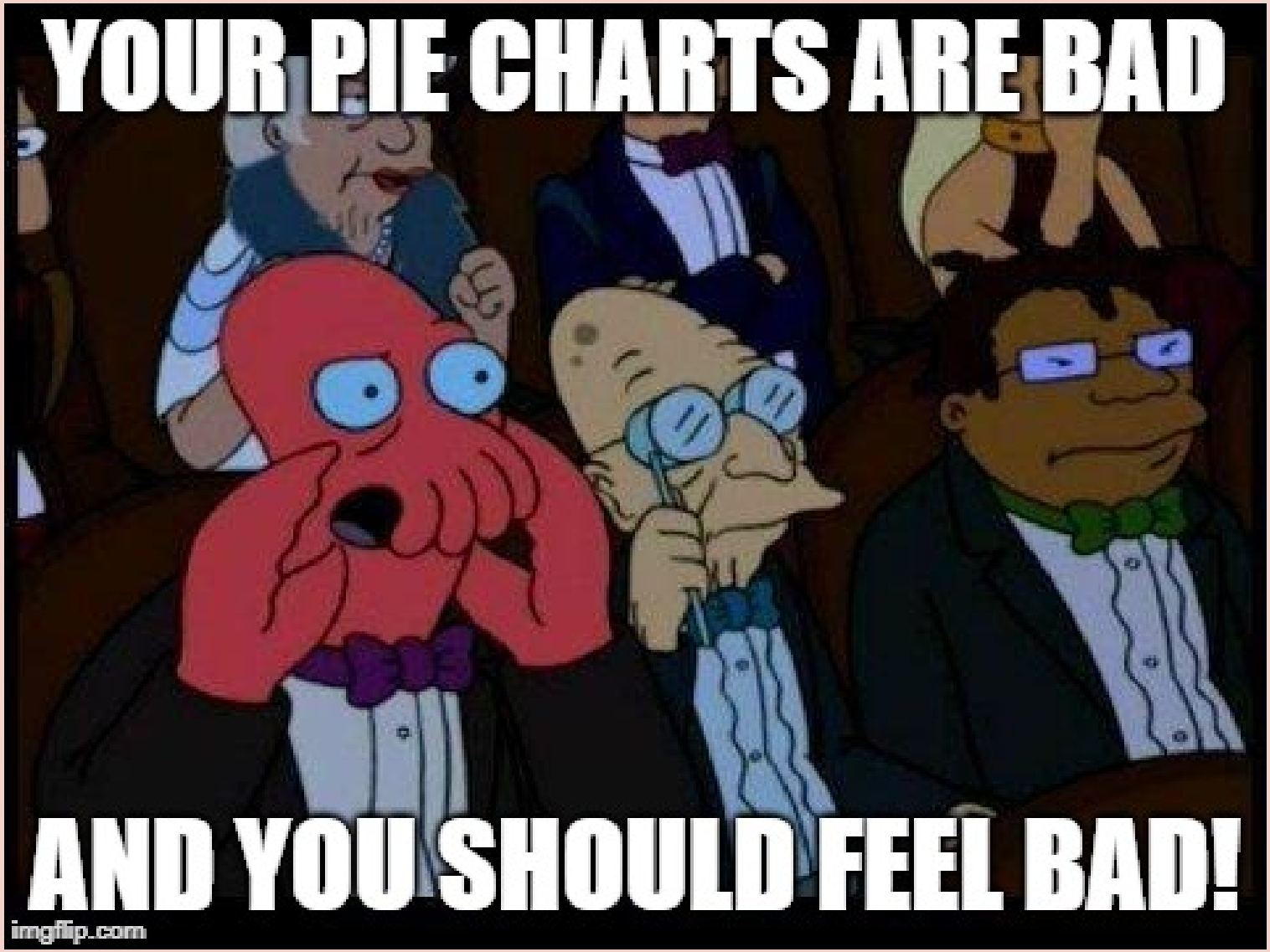
(a) 1970: 36% of U.S. adults smoked.

(b) 2013: 18% of U.S. adults smoked.

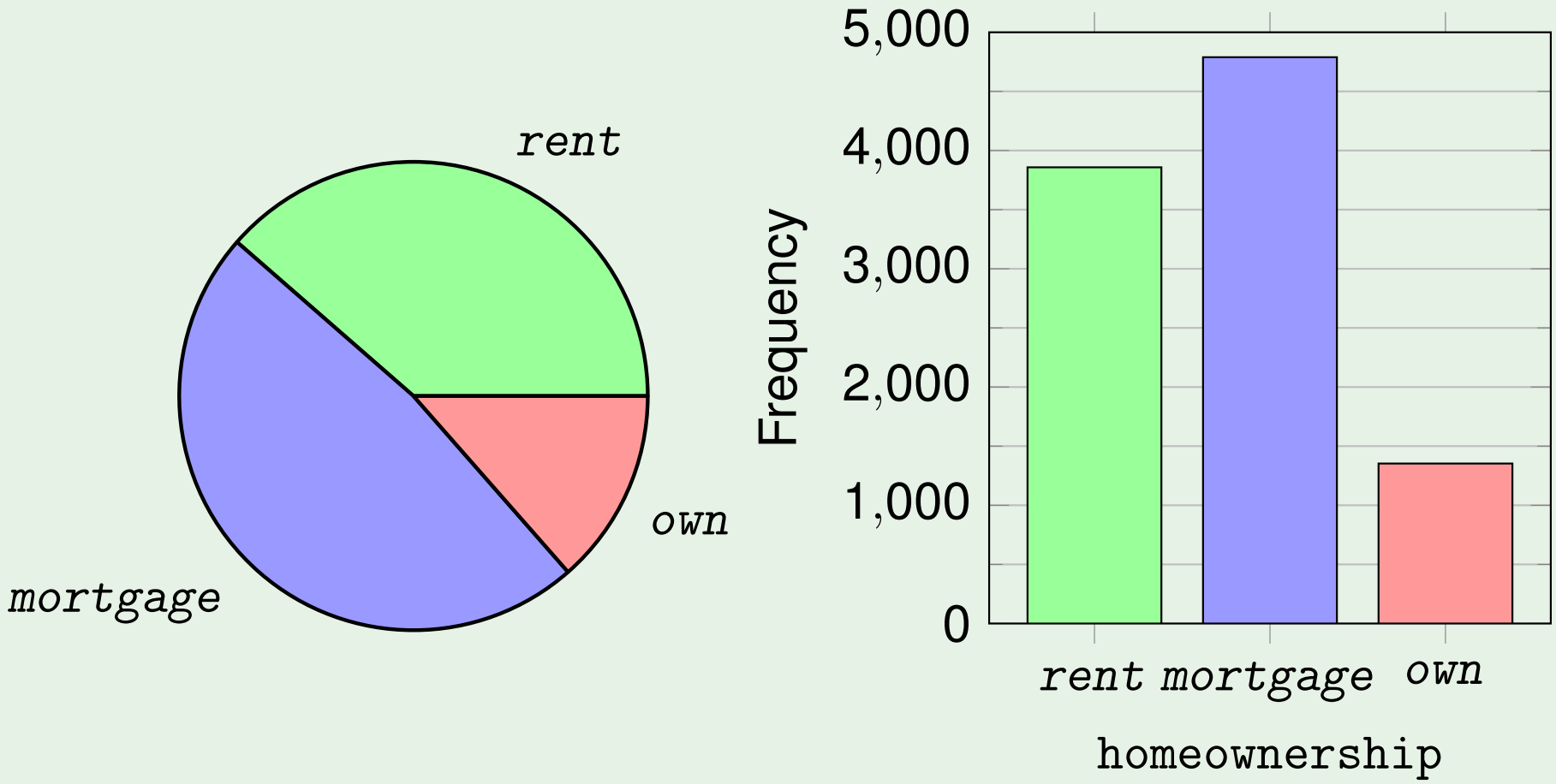
The larger cigarette is about twice as long as the smaller, which means it has four times the area of the smaller cigarette. While the data shows that 36% is only double 18%.

Definition

A **pie chart** is a worse version of the bar chart.



Example 21



The pie chart makes it harder to tell the relative sizes of each group, where the bar plot makes it much easier and gives the actual frequencies.