

## 4.1 Contingency tables and bar plots

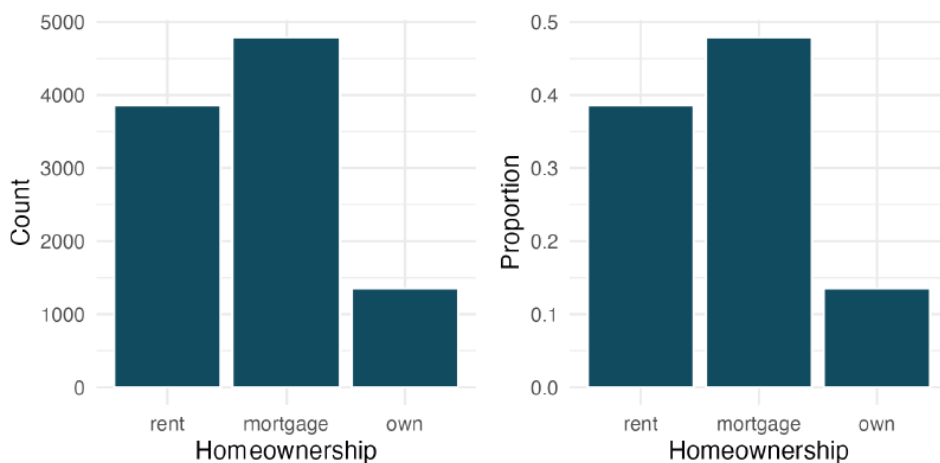
### Single Categorical Variable

Table 4.1 below summarizes two variables: `application_type` and `homeownership`. A table that summarizes data for two categorical variables in this way is called a contingency table. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 3496 corresponds to the number of loans in the dataset where the borrower rents their home and the application type was by an individual.

Table 4.1: A contingency table for application type and homeownership.

| application_<br>type | homeownership |          |      | Total |
|----------------------|---------------|----------|------|-------|
|                      | rent          | mortgage | own  |       |
| joint                | 362           | 950      | 183  | 1495  |
| individual           | 3496          | 3839     | 1170 | 8505  |
| Total                | 3858          | 4789     | 1353 | 10000 |

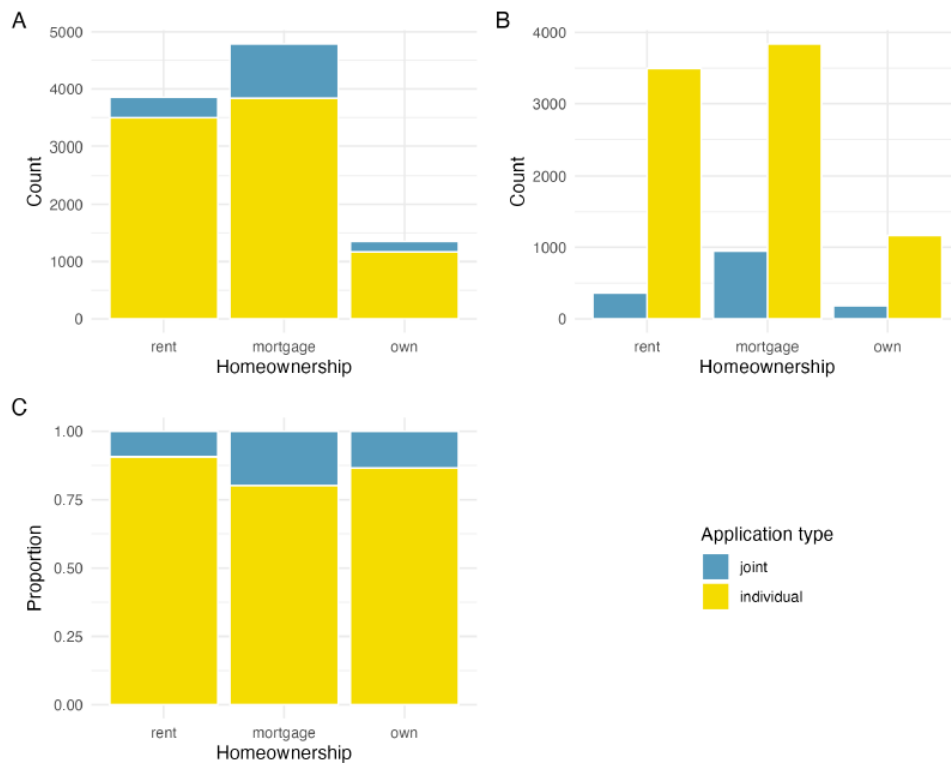
Bar charts for the frequencies and proportions for the single variable of homeownership by different types.



### Two Categorical Variables

Now create bar charts using the two variables, `homeownership` and `applicant type`.

- Shows a **“stacked”** bar graph of applicant type
- Shows a **“dodge”** bar graph, where applicant type is shown as separate bars
- Shows a **“fill”** bar graph, where applicant type fills each bar proportionally to add up to 1.00.



### 4.3 Row and column proportions

Table 4.3: A contingency table with row proportions for the application type and homeownership variables.

| application_<br>type | homeownership |          |       | Total |
|----------------------|---------------|----------|-------|-------|
|                      | rent          | mortgage | own   |       |
| joint                | 0.242         | 0.635    | 0.122 | 1     |
| individual           | 0.411         | 0.451    | 0.138 | 1     |

Table 4.4: A contingency table with column proportions for the application type and homeownership variables.

| application_type | homeownership |          |       |
|------------------|---------------|----------|-------|
|                  | rent          | mortgage | own   |
| joint            | 0.094         | 0.198    | 0.135 |
| individual       | 0.906         | 0.802    | 0.865 |
| Total            | 1.000         | 1.000    | 1.000 |

Row and column proportions can also be thought of as **conditional proportions** as they tell about the proportion of observations in a given level of a categorical variable conditional on the level of another categorical variable.

## Two-way Tables for Two Categorical Variables and Conditional Proportions

Often, we want to find a relationship, or an association between 2 categorical variables. We can use a two-way table to show all values and then compare proportions.

### Example: Smoking and Pregnancy Rate

|              | Smoker | Non-smoker | Total |
|--------------|--------|------------|-------|
| Pregnant     | 38     | 206        | 244   |
| Not pregnant | 97     | 337        | 434   |
| Total        | 135    | 543        | 678   |

*Be careful with what you designate as the denominator of the proportion.*

- What proportion **of this sample** were pregnant while smoking?
- What proportion **of this sample** were pregnant while not smoking?
- What proportion **of those who smoked** did NOT get pregnant?
- What proportion **of those who did not get pregnant** were non-smokers?

## Bar Charts and Waffle Charts (NEVER use pie charts if you want to show proportions)

Pie charts have entirely fallen out of favor in the data science world. People are not nearly as good at perceiving angle differences as they are at length differences:

Homeownership

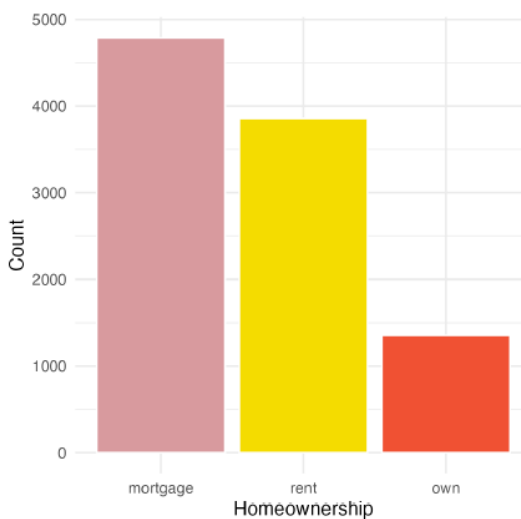
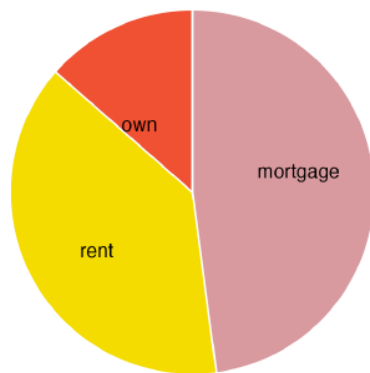


Figure 4.5: A pie chart and bar plot of homeownership.

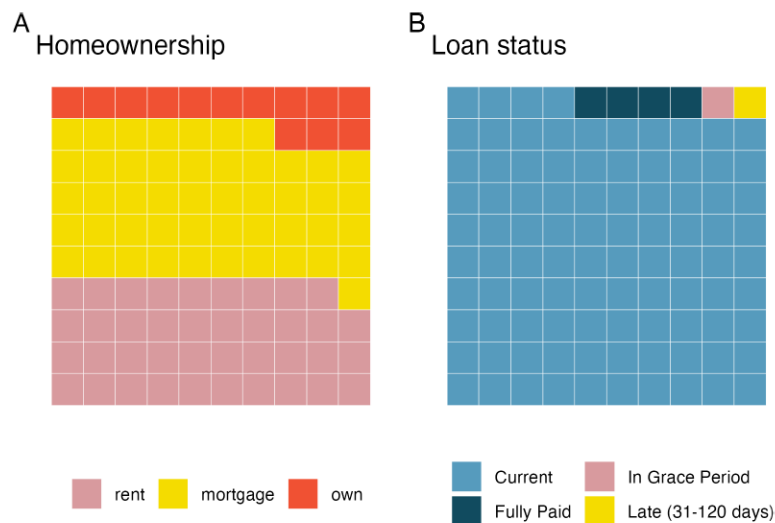


Figure 4.7: Plot A: Waffle chart of homeownership, with levels rent, mortgage, and own. Plot B: Waffle chart of loan status, with levels current, fully paid, in grade period, and late.

## 4.6 Comparing numerical data across groups

Revisit the **county** dataset and compare the **median household income** for counties that gained population from 2010 to 2017 versus counties that had no gain. While we might like to make a causal connection between income and population growth, remember that these are observational data and so such an interpretation would be, at best, half-baked.

We have data on 3142 counties in the United States. We are missing 2017 population data from 3 of them, and of the remaining 3139 counties, in 1541 the population increased from 2010 to 2017 and in the remaining 1598 the population decreased. Table 4.6 shows a sample of 5 observations from each group.

Table 4.6: The median household income from a random sample of 5 counties with population gain between 2010 to 2017 and another random sample of 5 counties with no population gain.

| State      | County              | Population change (%) | Gain / No gain | Median household income |
|------------|---------------------|-----------------------|----------------|-------------------------|
| Colorado   | Custer County       | 14.28                 | gain           | 41330                   |
| Georgia    | Murray County       | 1.35                  | gain           | 41617                   |
| Georgia    | Pickens County      | 7.41                  | gain           | 61542                   |
| Texas      | Wharton County      | 2.12                  | gain           | 50145                   |
| Washington | Grays Harbor County | 2.30                  | gain           | 45483                   |
| Alabama    | Conecuh County      | -3.40                 | no gain        | 30434                   |
| Illinois   | McDonough County    | -4.32                 | no gain        | 42911                   |
| Iowa       | Iowa County         | -1.08                 | no gain        | 58077                   |
| Michigan   | Genesee County      | -1.95                 | no gain        | 45231                   |
| Wyoming    | Campbell County     | -3.76                 | no gain        | 80178                   |

- Show the **median household income** data as a **histrogram**, filled by **change in population (gain or no gain)**.
- Show the **median household income** data as **side-by-side boxplots**, one is the **gain** and the other is **no gain**.
- Show the **median household income** data as **ridge plots** one is the **gain** and the other is **no gain**.

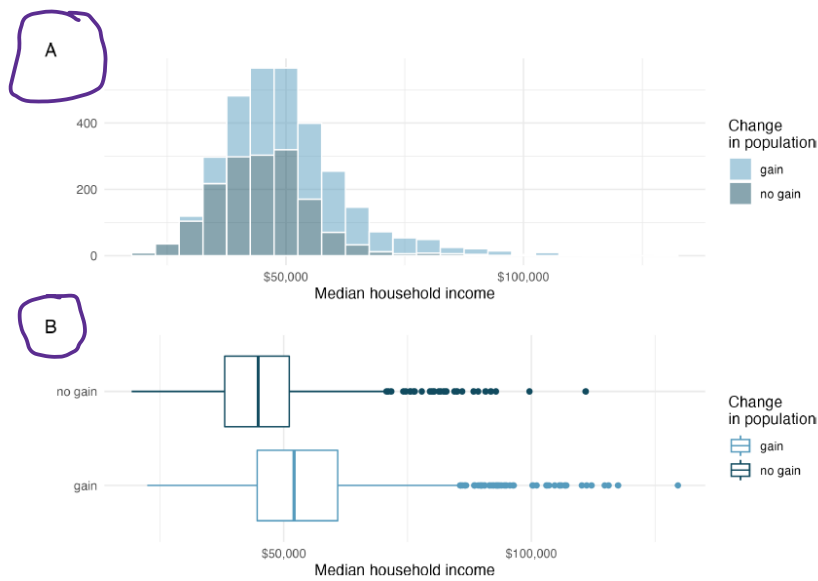


Figure 4.8: Histograms (Plot A) and side by-side box plots (Plot B) for median household income, where counties are split by whether there was a population gain or not.

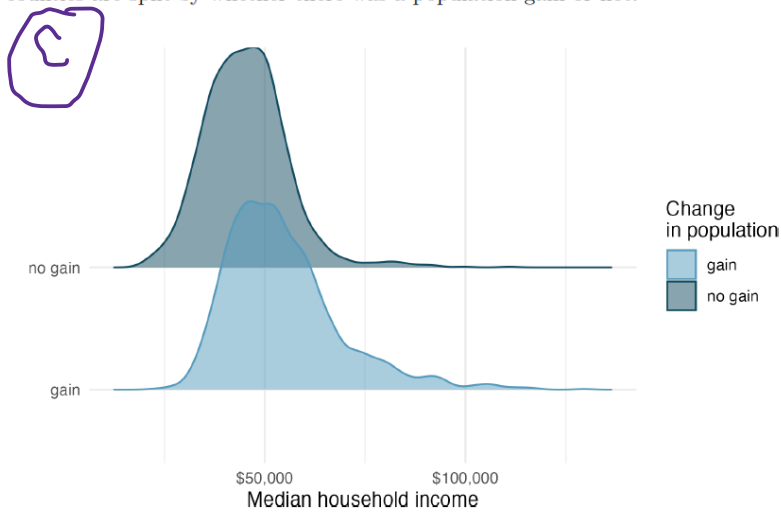


Figure 4.9: Ridge plot for median household income, where counties are split by whether there was a population gain or not.

## Comparing data across more than 2 groups



## Homework Chapter 4

1. Review section 4.7 (the chapter review)
2. Suggested problems from textbook section 4.8 exercises: 1, 3, and 5
3. Suggested tutorials to learn code: [Tutorial 1 – Lessons 1, 2, and 3](#)
  - [Visualizing categorical data](#)
  - [Visualizing numerical data](#)
  - [Summarizing with statistics](#)