Revised: 1/12/2021

Summarizing Data II

Michael C. Hackett

Assistant Professor, Computer Science

Community College of Philadelphia

Contingency Tables

- A **contingency table** is a table that summarizes data for two categorical variables.

- Contingency tables show the frequency of combinations between the two variables.
  - For example, the table below shows there were 3496 observations in the data set that had an application type of “individual” and homeownership type of “rent”
  - Another example, there were 183 observations in the data set that had an application type of “joint” and homeownership type of “own”

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

Contingency Tables

```
> library(readr)
> loans <- read_csv("loans.csv")
Parsed with column specification:
cols(
  .default = col_double(),
  emp_title = col_character(),
  state = col_character(),
  homeownership = col_character(),
  verified_income = col_character(),
  verification_income_joint = col_character(),
  loan_purpose = col_character(),
  application_type = col_character(),
  grade = col_character(),
  sub_grade = col_character(),
  issue_month = col_character(),
  loan_status = col_character(),
  initial_listing_status = col_character(),
  disbursement_method = col_character()
)
See spec(...) for full column specifications.
> table(loans$application_type, loans$homeownership)

             MORTGAGE  OWN RENT
  individual     3839 1170 3496
  joint           950  183  362
> addmargins(table(loans$application_type, loans$homeownership))

             MORTGAGE   OWN  RENT   Sum
  individual     3839  1170  3496  8505
  joint           950   183   362  1495
  Sum            4789  1353  3858 10000
> |
```

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

[table function]
[addmargins function]

Contingency Tables

- A contingency table can also be used to summarize one categorical variable.

```
> table(loans$homeownership)

MORTGAGE       OWN      RENT
    4789      1353      3858
> addmargins(table(loans$homeownership))

MORTGAGE       OWN      RENT       Sum
    4789      1353      3858     10000
```

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

Contingency Tables

- Sometimes, it is useful for contingency tables display proportions instead of frequencies.

```
> t<-table(loans$application_type, loans$homeownership)
> t

             MORTGAGE  OWN RENT
  individual     3839 1170 3496
  joint           950  183  362
> prop.table(t)

             MORTGAGE     OWN    RENT
  individual   0.3839 0.1170 0.3496
  joint        0.0950 0.0183 0.0362
```

Frequencies

Proportions

prop.table function

Contingency Tables

- Row Proportions:
  - 63.5% of observations with an application type of “joint” have a homeownership type of “mortgage”.
  - 12.2% of observations with an application type of “joint” have a homeownership type of “own”.
  - 24.2% of observations with an application type of “joint” have a homeownership type of “rent”.

```
> t<-table(loans$application_type, loans$homeownership)
> t

            MORTGAGE  OWN RENT
individual      3839 1170 3496
joint            950  183  362
> prop.table(t, margin=1)

            MORTGAGE       OWN      RENT
individual 0.4513815 0.1375661 0.4110523
joint      0.6354515 0.1224080 0.2421405
```

Contingency Tables

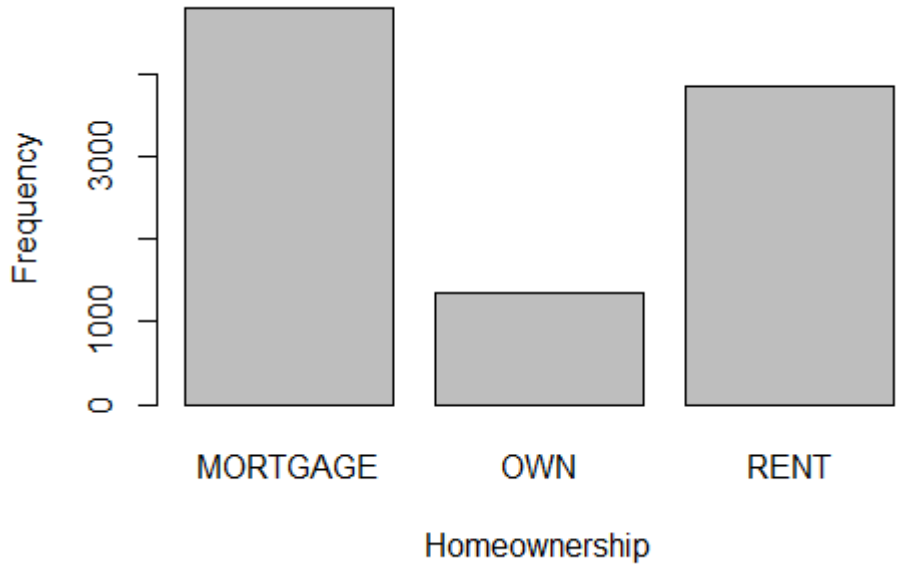
- Column Proportions:
  - 86.5% of observations with a homeownership type of “own” have an application type of “individual”.
  - 13.5% of observations with a homeownership type of “own” have an application type of “joint”.

```
> t<-table(loans$application_type, loans$homeownership)
> t

            MORTGAGE  OWN RENT
  individual     3839 1170 3496
  joint           950  183  362
> prop.table(t, margin=2)

             MORTGAGE       OWN      RENT
  individual 0.8016287 0.8647450 0.9061690
  joint      0.1983713 0.1352550 0.0938310
```

Bar plots

- Bar plots are used to visualize the distribution of categorical variables.

```
> t<-table(loans$homeownership)
> barplot(t, xlab="Homeownership", ylab="Frequency")
```

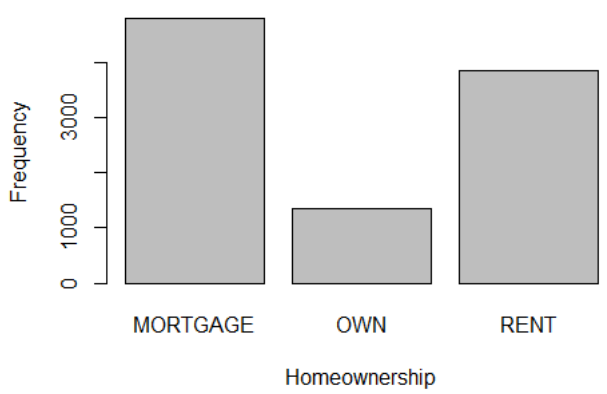


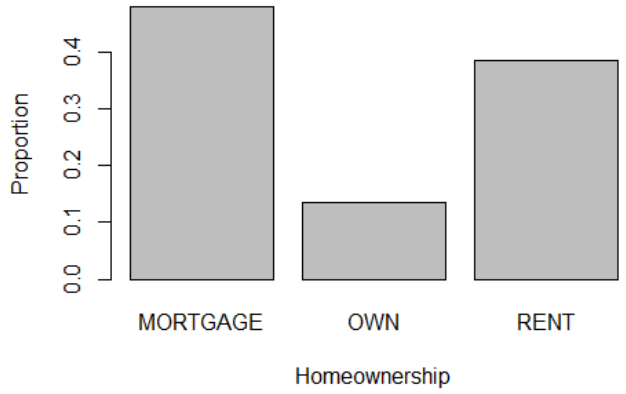
barplot function

Bar plots

- Can be used to display frequencies or proportions.
  - When displaying proportions, it is sometimes called a *relative frequency bar chart*.

```
> t<-table(loans$homeownership)
> barplot(t, xlab="Homeownership", ylab="Frequency")
> barplot(prop.table(t), xlab="Homeownership", ylab="Proportion")
```



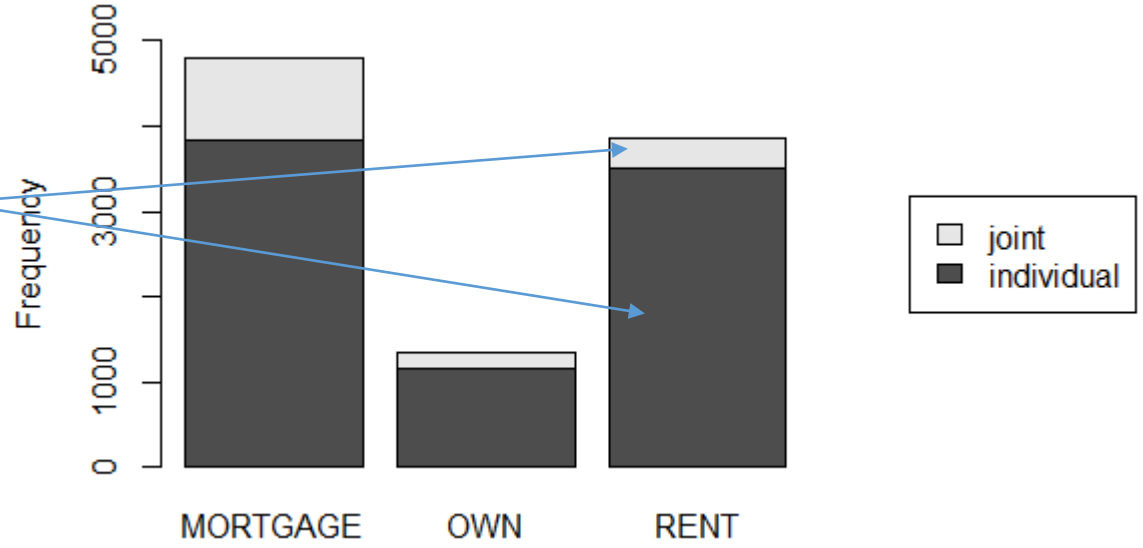


Bar plots

- Bar plots can be used to visualize contingency tables.
  - Below is a **stacked bar plot** showing frequency

```
> t<-table(loans$application_type, loans$homeownership)
> par(mar = c(5,4,4,10))
> barplot(t, ylab="Proportion", ylim=c(0,5000),legend.text=rownames(t), args.leg
end = list(x='right', inset=c(-0.50,0),xpd=TRUE))
```

	MORTGAGE	OWN	RENT
individual	3839	1170	3496
joint	950	183	362



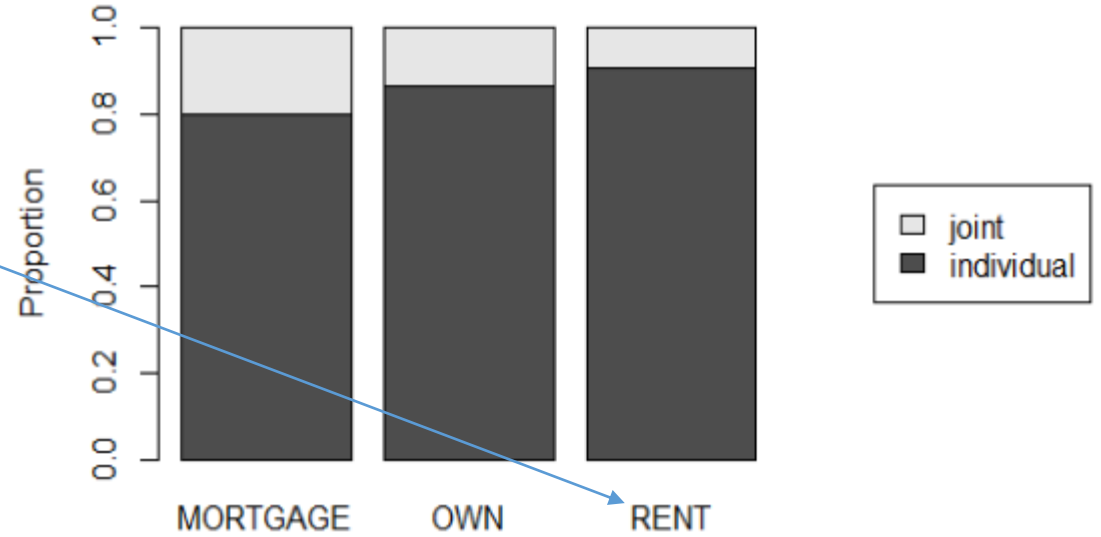
[par function]

Bar plots

- This stacked bar plot shows the proportion by column

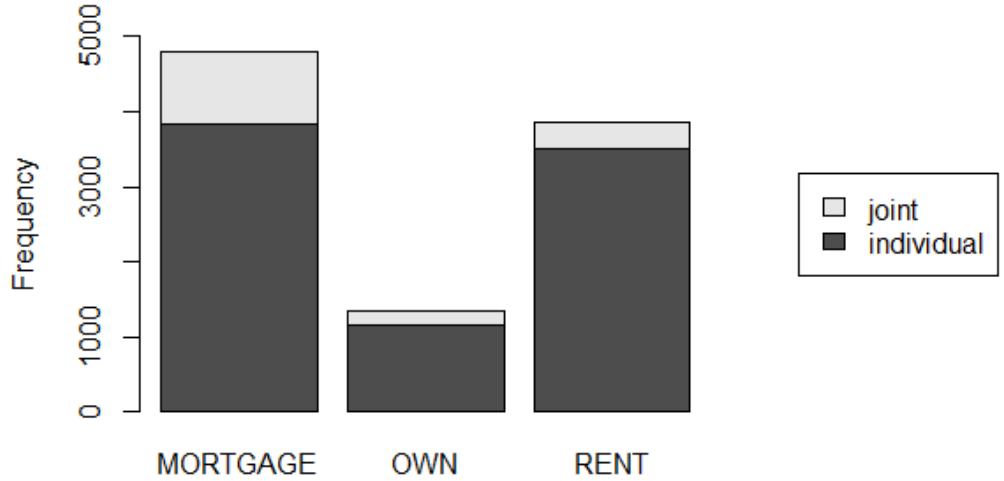
```
> t<-table(loans$application_type, loans$homeownership)
> pt<-prop.table(t, margin=2)
> barplot(pt, ylab="Proportion", ylim=c(0,1),legend.text=rownames(t), args.legen
d = list(x='right', inset=c(-0.50,0),xpd=TRUE))
```

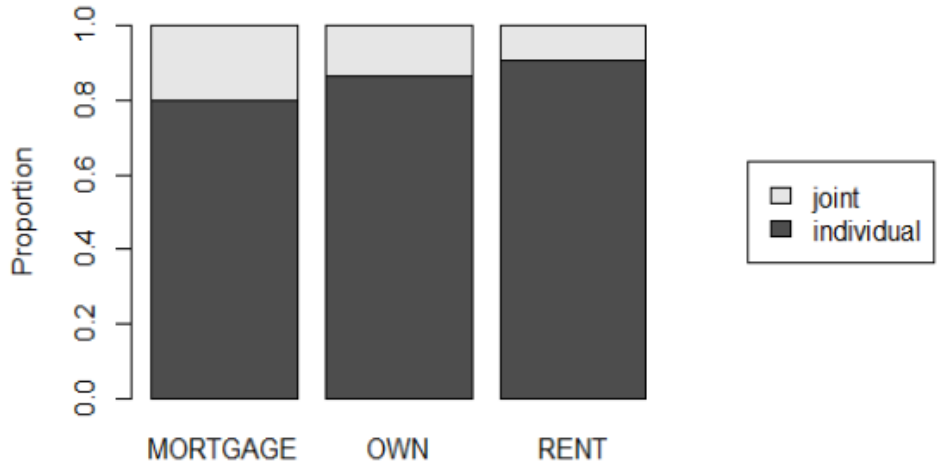
```
                MORTGAGE       OWN      RENT
individual 0.8016287 0.8647450 0.9061690
joint      0.1983713 0.1352550 0.0938310
```



Bar plots

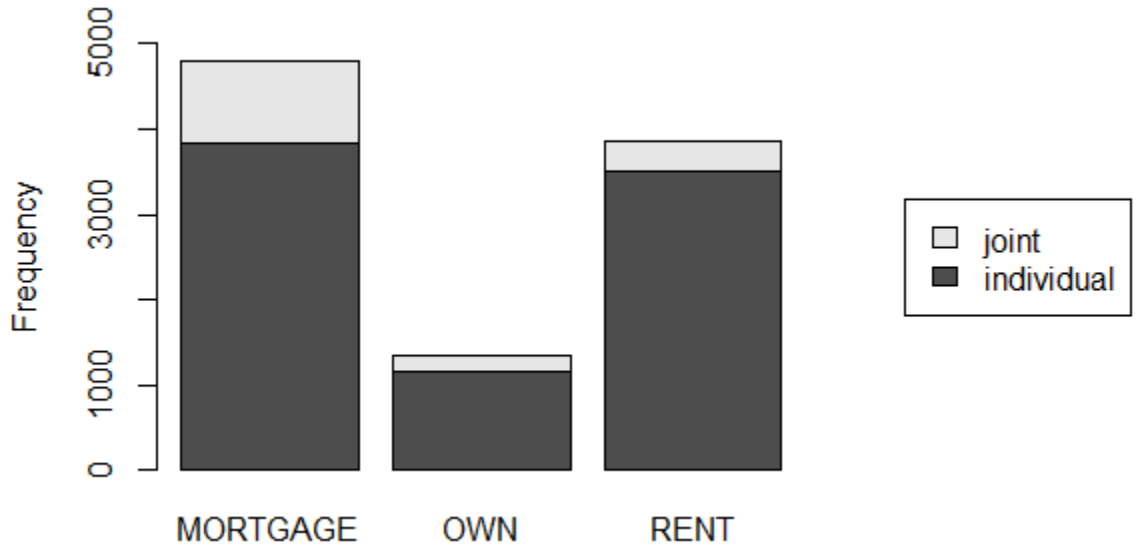
- Both plots below used the same data, but they each tell a different story about the data





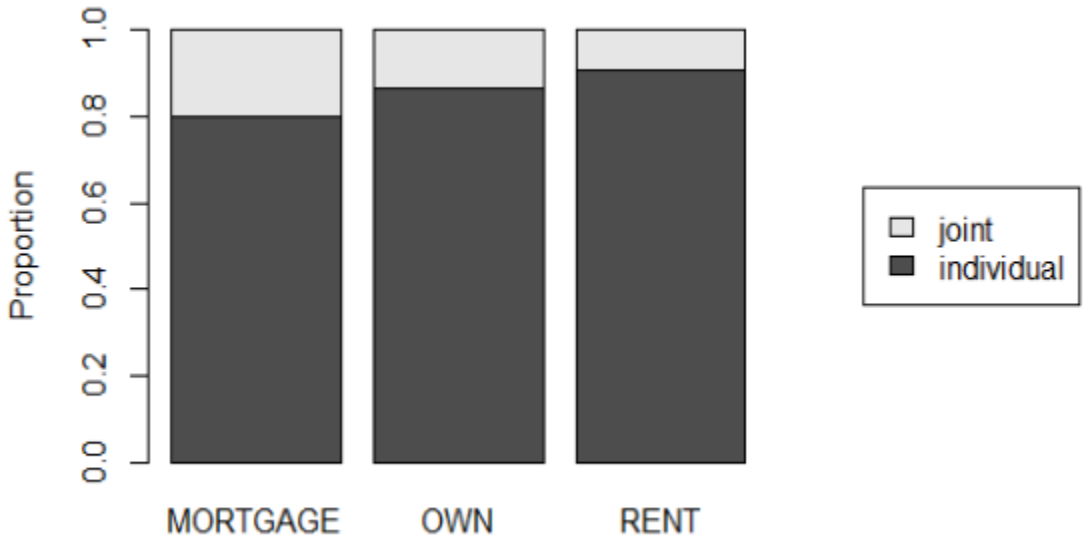
Bar plots

- This stacked bar plot clearly shows there are far fewer owners than renters or those with a mortgage



Bar plots

- This stacked bar plot clearly shows the proportion of joint applicants is roughly the same as individual applicants, regardless of homeowner status

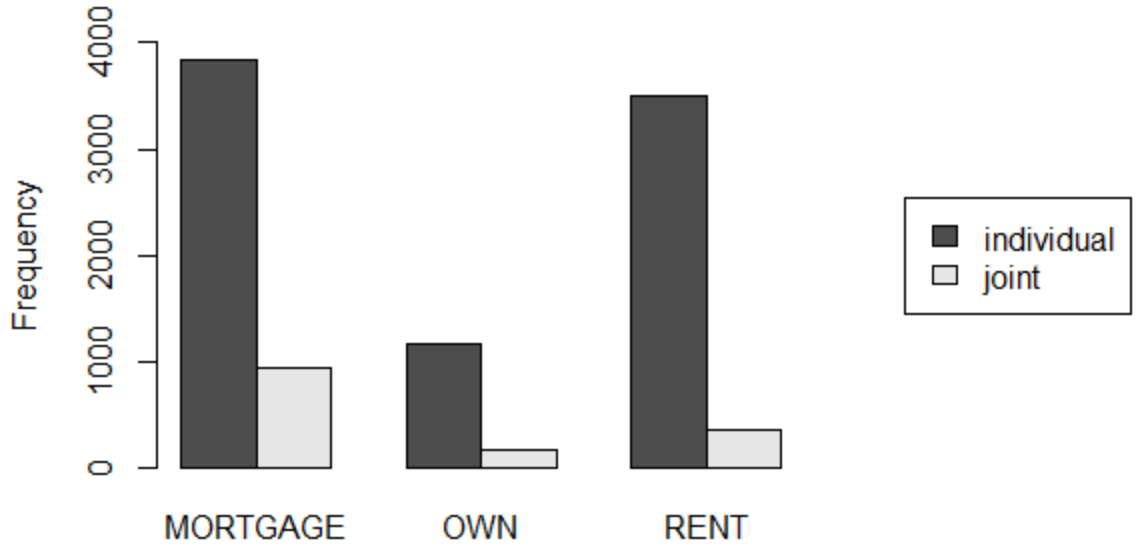


Bar plots

- **Side-by-side bar plots** show the bars next to each other instead of stacked

```
> t<-table(loans$application_type, loans$homeownership)
> pt<-prop.table(t, margin=2)
> barplot(t, ylab="Frequency", ylim=c(0,4000),legend.text=rownames(t), args.lege
nd = list(x='right', inset=c(-0.50,0),xpd=TRUE), beside=TRUE)
```

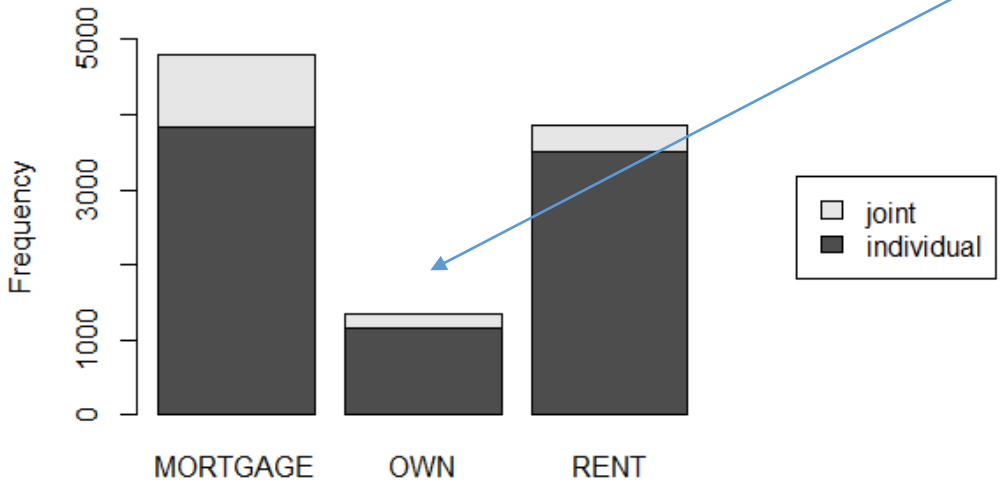
```
             MORTGAGE  OWN RENT
individual       3839 1170 3496
joint             950  183  362
```

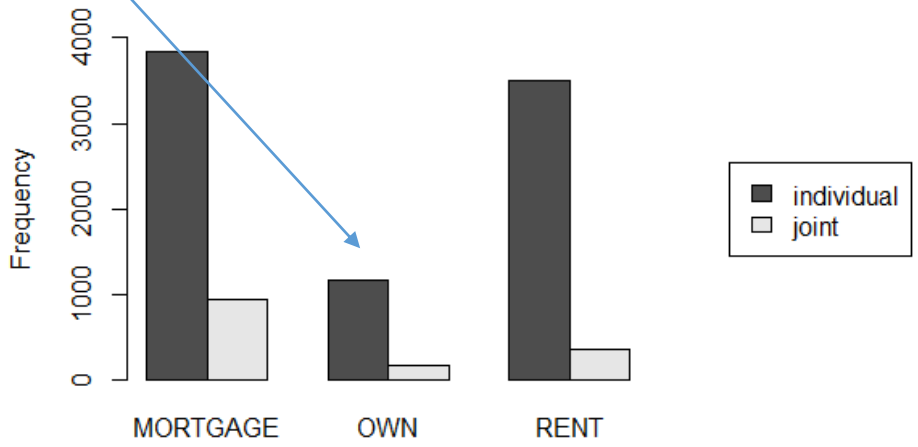


Bar plots

- Stacked bar plots best emphasize the total count of each column
- Side-by-side bar plots best emphasize the row data of each column

	MORTGAGE	OWN	RENT	Sum
individual	3839	1170	3496	8505
joint	950	183	362	1495
Sum	4789	1353	3858	10000

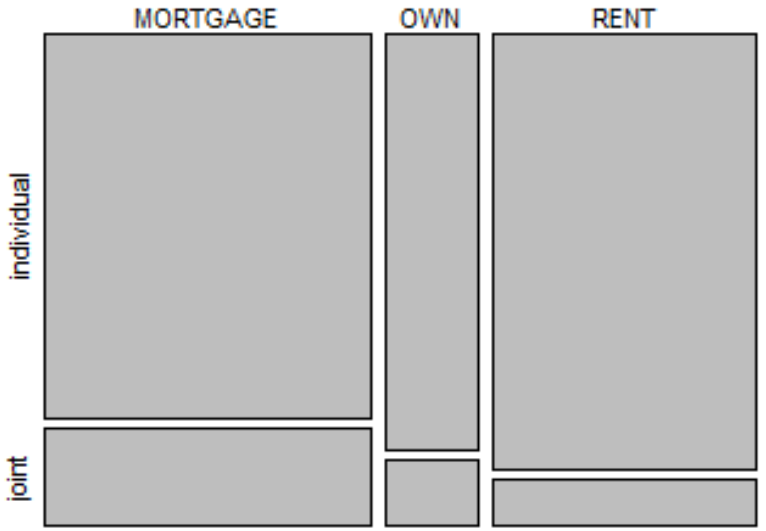




Mosaic plots

- Mosaic plots are another visualization for contingency tables.
  - Shows the relative group sizes of the variables

```
> t<-table(loans$homeownership, loans$application_type)
> mosaicplot(t, main="")
```



mosaicplot function

Mosaic plots

- Very similar to stacked bar plots, except widths now have meaning.
  - Width in the mosaic plot represents the frequency