Exploratory data analysis

Exploring categorical data

Prof. Dr. Jan Kirenz
HdM Stuttgart

Loan data

Prof. Dr. Jan Kirenz

BORROW ▾ | INVEST ▾   ABOUT US   HELP   SIGN IN

Radius is now LendingClub.   **Visit our Banking website**   ✕

Personal Loans Up to $40,000

Check your rate. It won't impact your credit score.   Privacy & security
PROTECTION

$ | How much do you need?   What's the money for? ▾   **Check Your Rate**

✉ **Respond to Mail Offer**
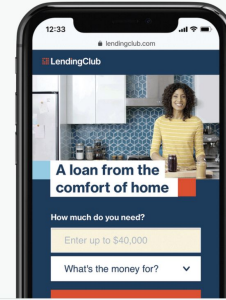
How LendingClub Works

Apply in Minutes
Get customized loan options based on what you tell us.

Choose a Loan Offer
Select the rate, term, and payment options you like best.

Get Funded
Once your loan is funded, we'll send the money straight to your bank account or pay your creditors directly.

Check Your Rate

12:33   🔒 lendingclub.com

LendingClub

A loan from the comfort of home

How much do you need?

Enter up to $40,000

What's the money for? ▾

Join Over 3 Million Members

★★★★★

Thank you so much for valuing me as a customer, and coming through for me and my family at a trying time in this world.

— **Roselyn**, a member from Texas

Read More Reviews

💲 **$60 Billion +**
Borrowed

👤 **3 Million +**
Members

b: Authentic Reviews
★★★★★
54,898 Reviews

We will explore these two variables

This data set represents thousands of loans made through the Lending Club platform, which is a platform that allows individuals to lend to other individuals.

- **home_ownership**: The ownership status of the applicant's residence.

- **application_type**: The type of application: either individual or joint.

Prof. Dr. Jan Kirenz

Head and tail of two categorical variables from the loans data

	homeownership	application_type
0	mortgage	individual
1	rent	individual
2	rent	individual
3	rent	individual
4	rent	joint

	homeownership	application_type
9995	rent	individual
9996	mortgage	individual
9997	mortgage	joint
9998	mortgage	individual
9999	rent	individual

Information about our two columns. What is the data type?

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 2 columns):
 #    Column               Non-Null Count   Dtype
---   ------               --------------   -----
 0    homeownership        10000 non-null   category
 1    application_type     10000 non-null   category
dtypes: category(2)
memory usage: 19.9 KB
```

What are the levels and frequencies?

```
homeownership: Index(['mortgage', 'own', 'rent'], dtype='object')

application_type: Index(['individual', 'joint'], dtype='object')
```

```
mortgage      4789
rent          3858
own           1353
Name: homeownership, dtype: int64

individual    8505
joint         1495
Name: application_type, dtype: int64
```

Contingency table (cross table)
- Part I -

Prof. Dr. Jan Kirenz

Columns of the table

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

Rows of the table

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

Take a look at **mortgage**

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

Contingency table

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

Bar plot

Homeownership is on the x-axes
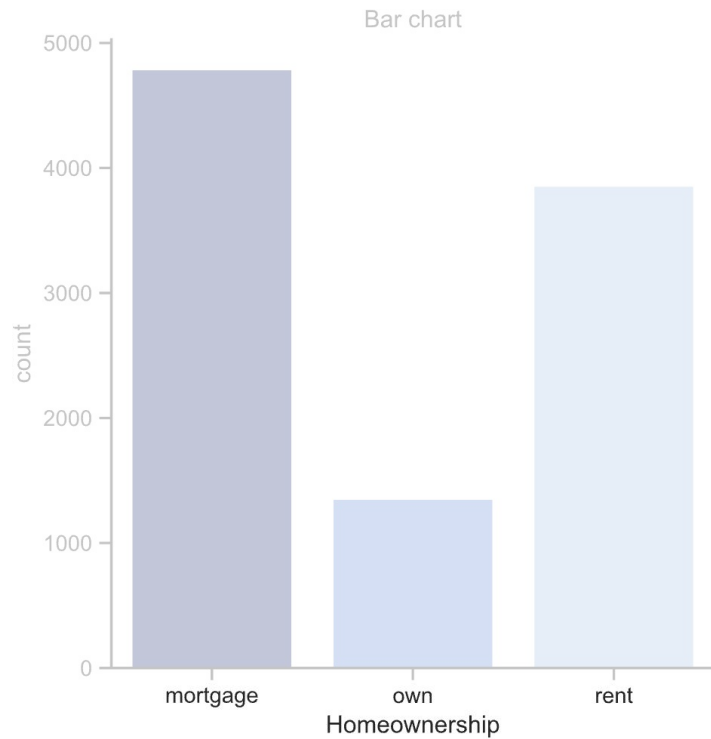
The frequencies (**count**) are on the y-axis
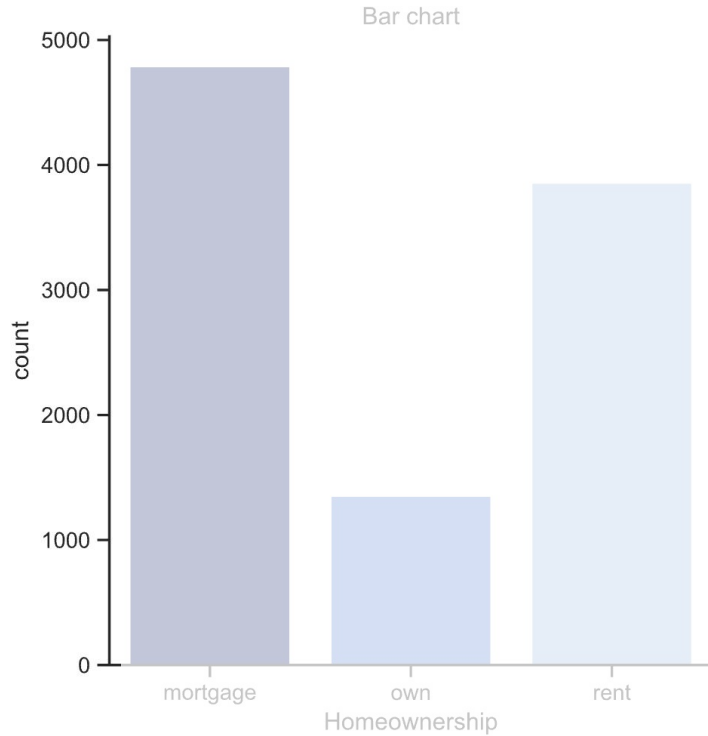
The number of applicants with mortgage

Prof. Dr. Jan Kirenz
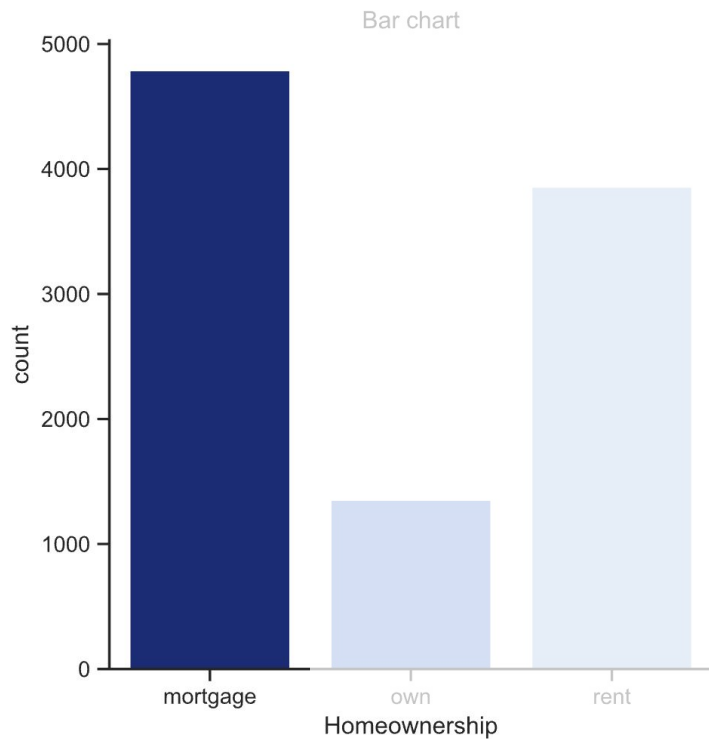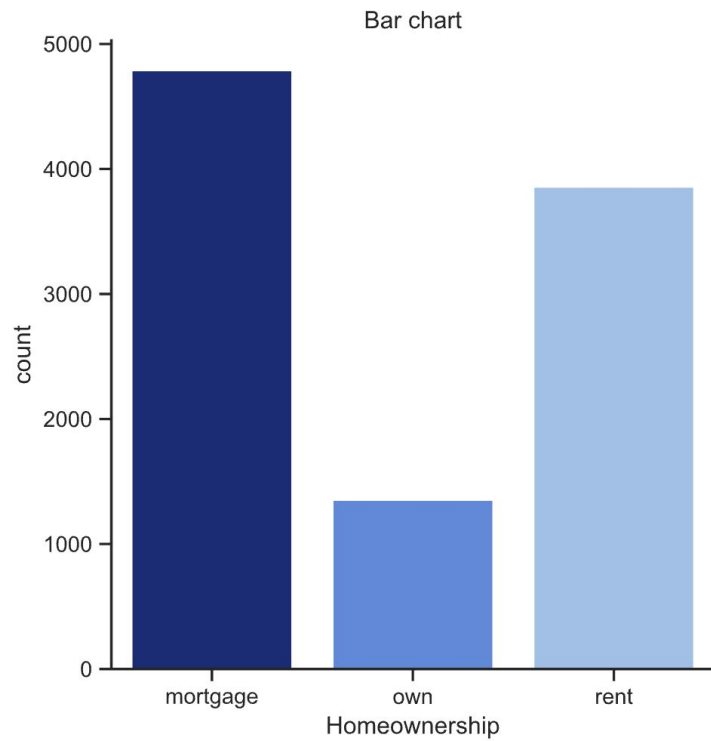
Bar chart

Visualizing two categorical variables

Dodged bar plot

Homeownership is on the x-axis

Prof. Dr. Jan Kirenz
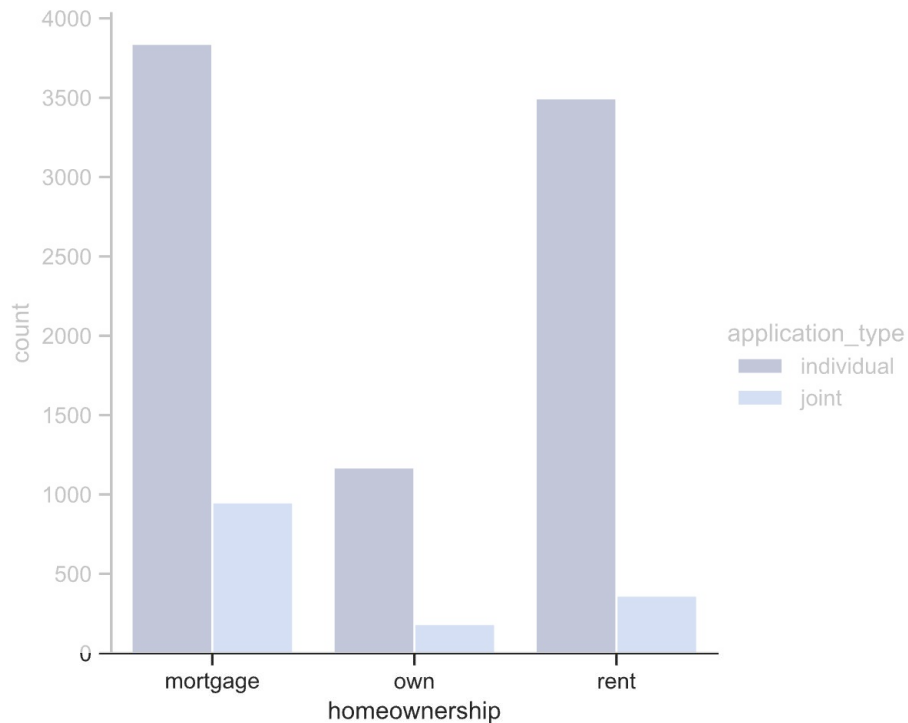
Frequencies (**count**) are on the y-axis

The two application types are next to each other

Prof. Dr. Jan Kirenz
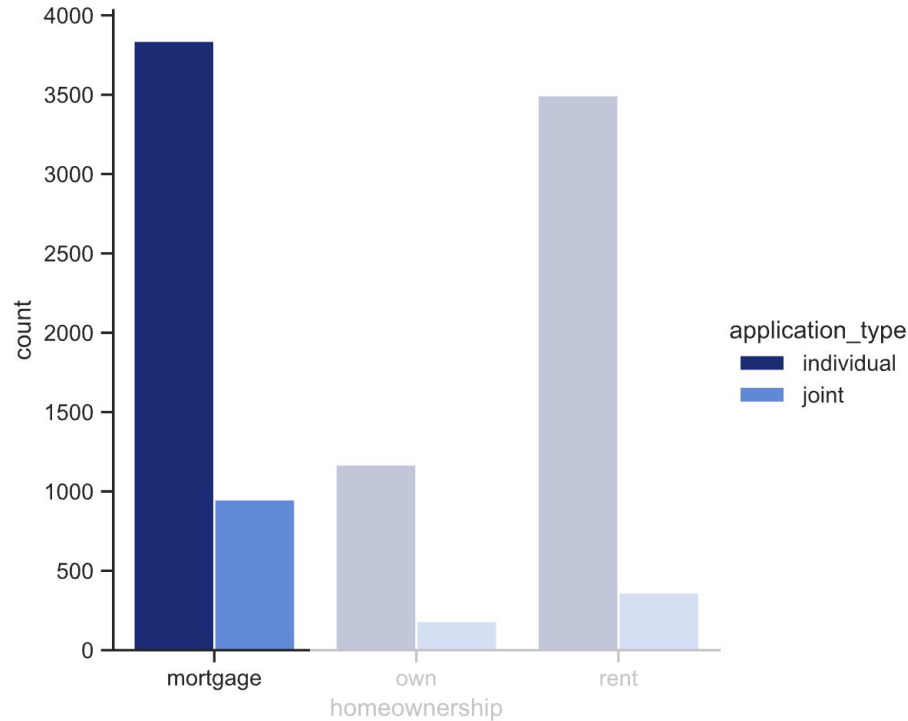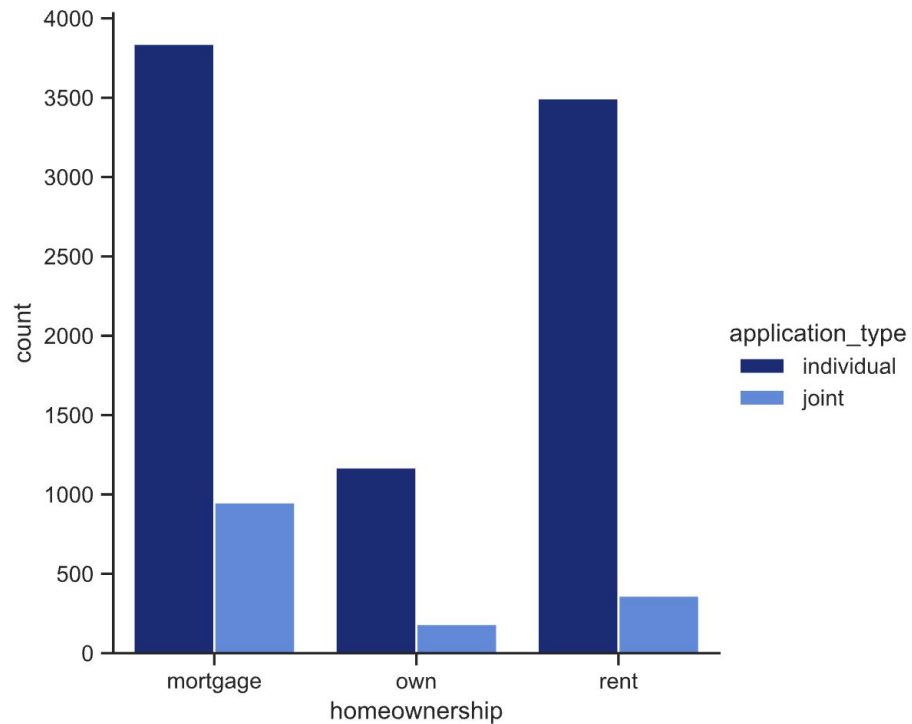
Dodged bar plot

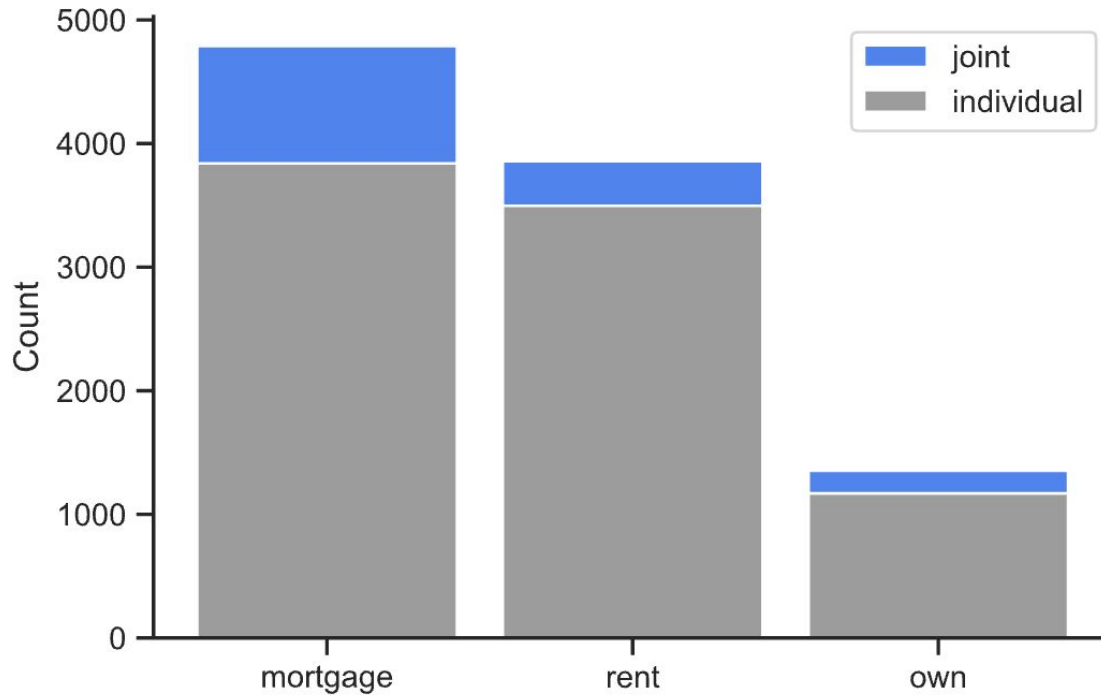# Stacked bar chart

Homeownership is on the x-axis

Frequencies (**count**) are on the y-axis

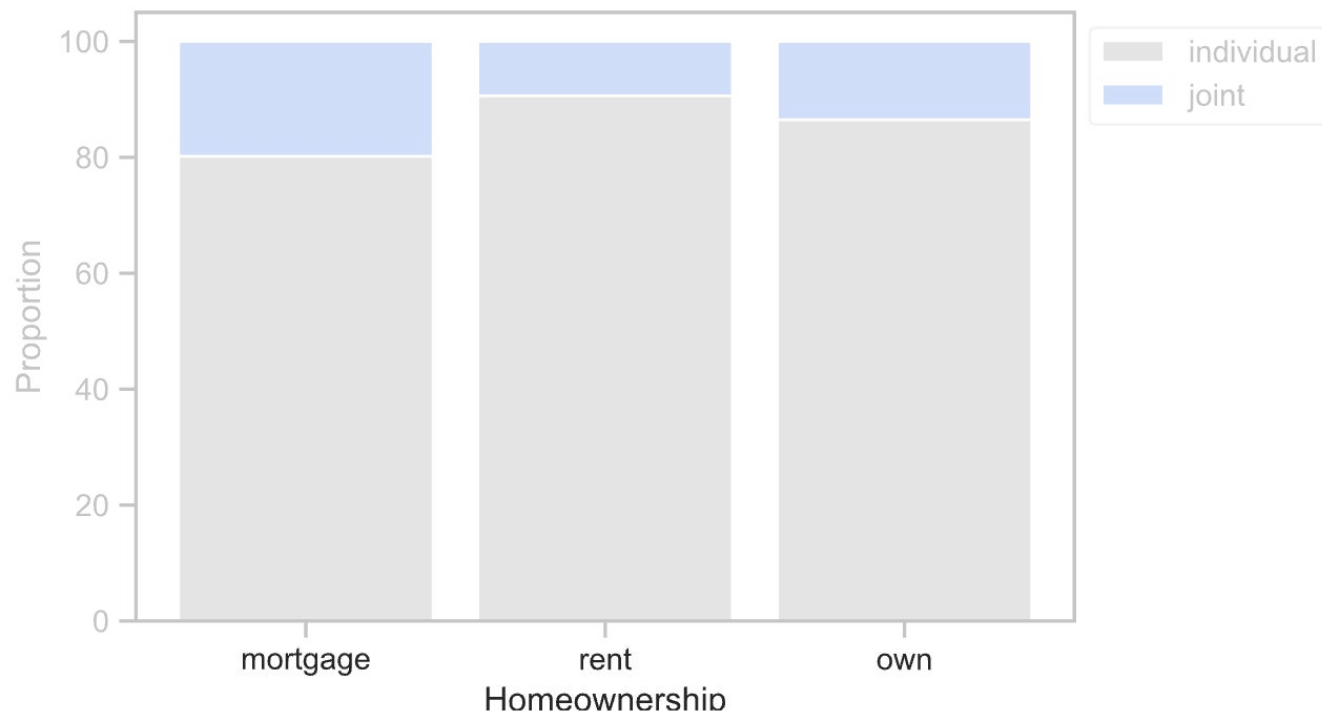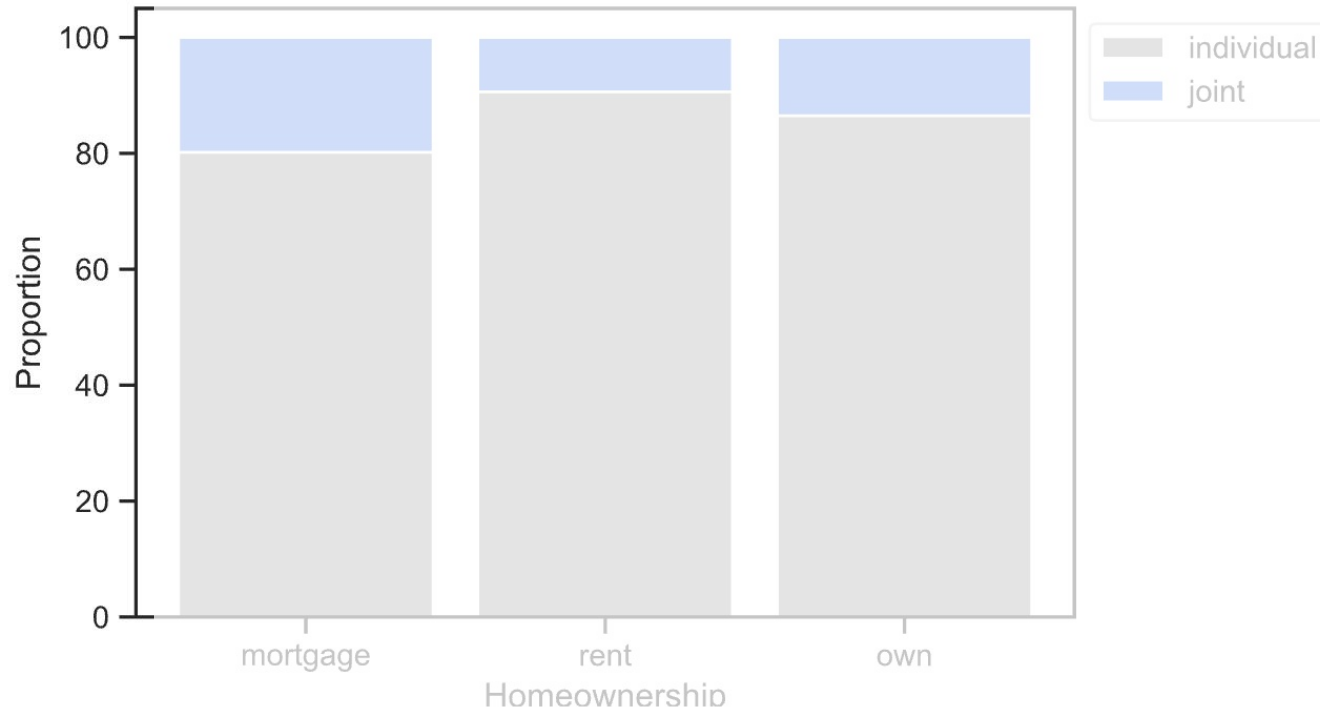# The two application types for mortgage are on top of each other
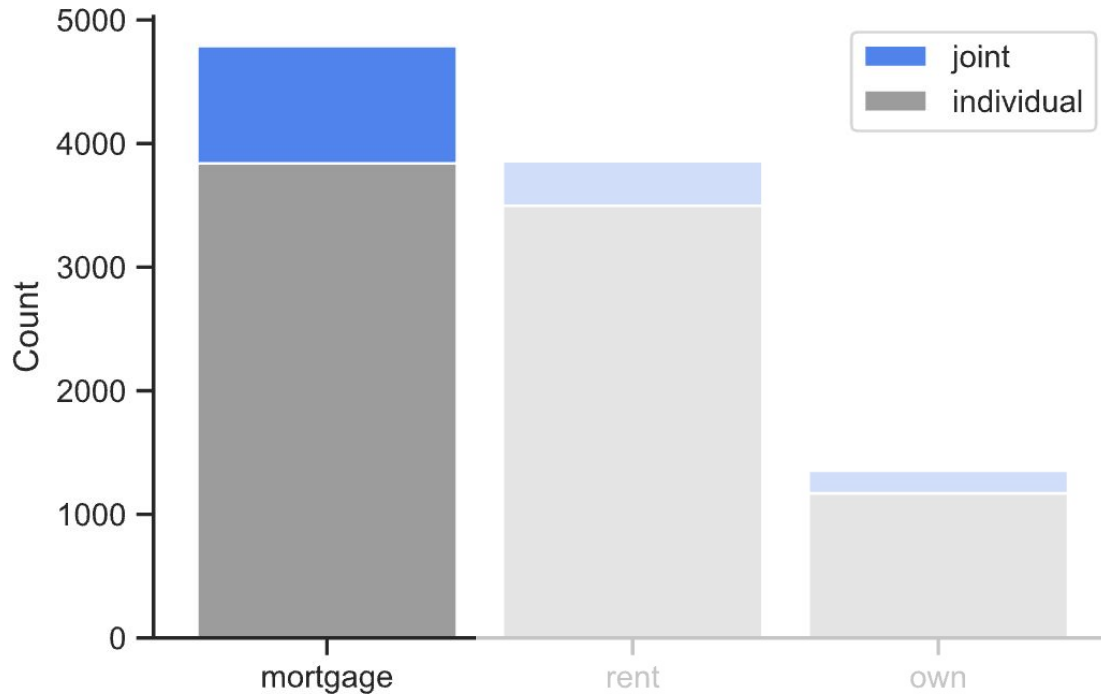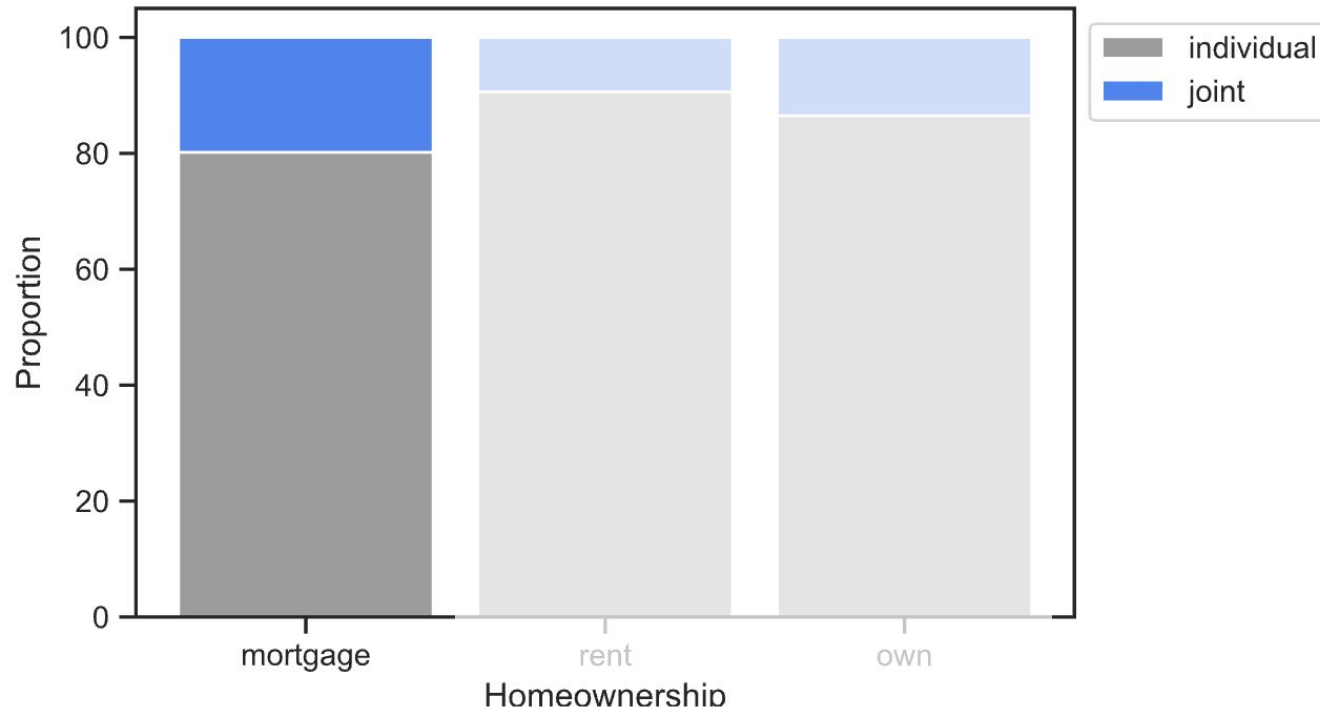
Stacked bar chart
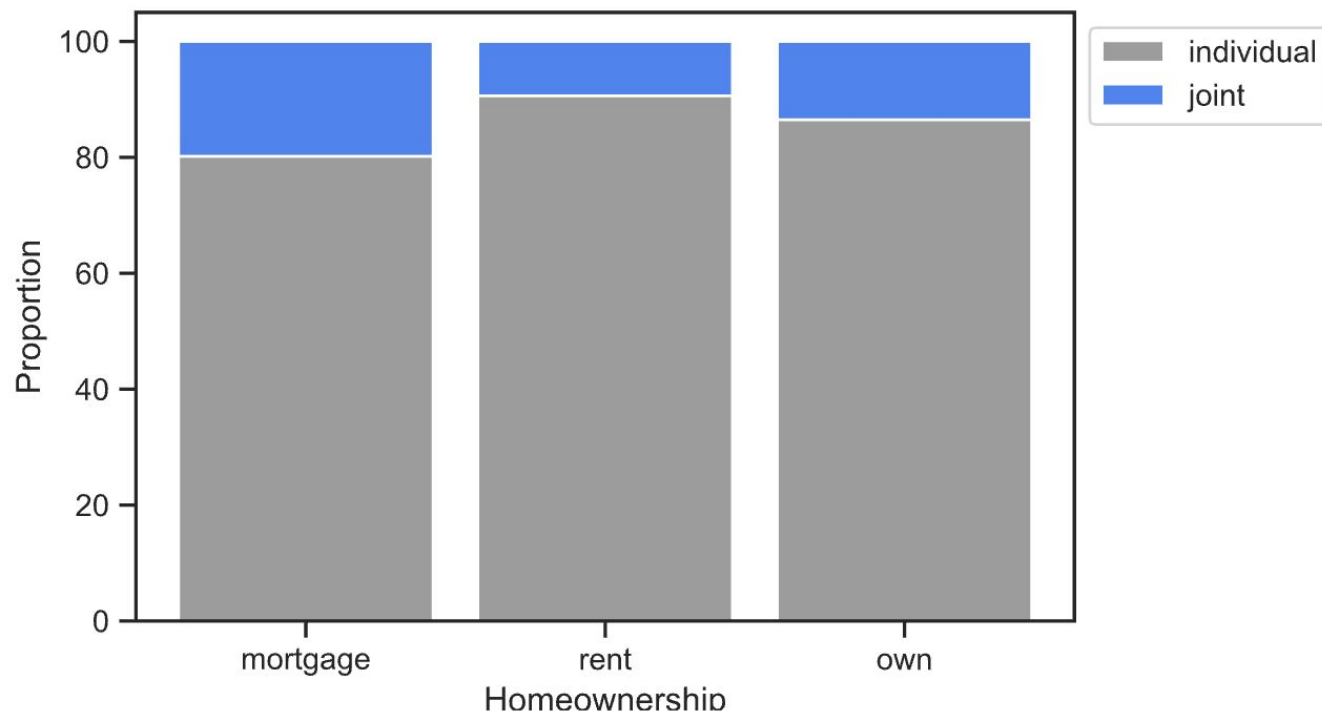
Standardized bar plot

Homeownership is on the x-axis

The **proportion** is on the y-axis (in percent)

Prof. Dr. Jan Kirenz

Proportions per application type for mortgage

Standardized bar chart

Mosaic plot

Example with only one variable (**homeownership**)

Moasic plot for homeownership

mortgage | rent | own
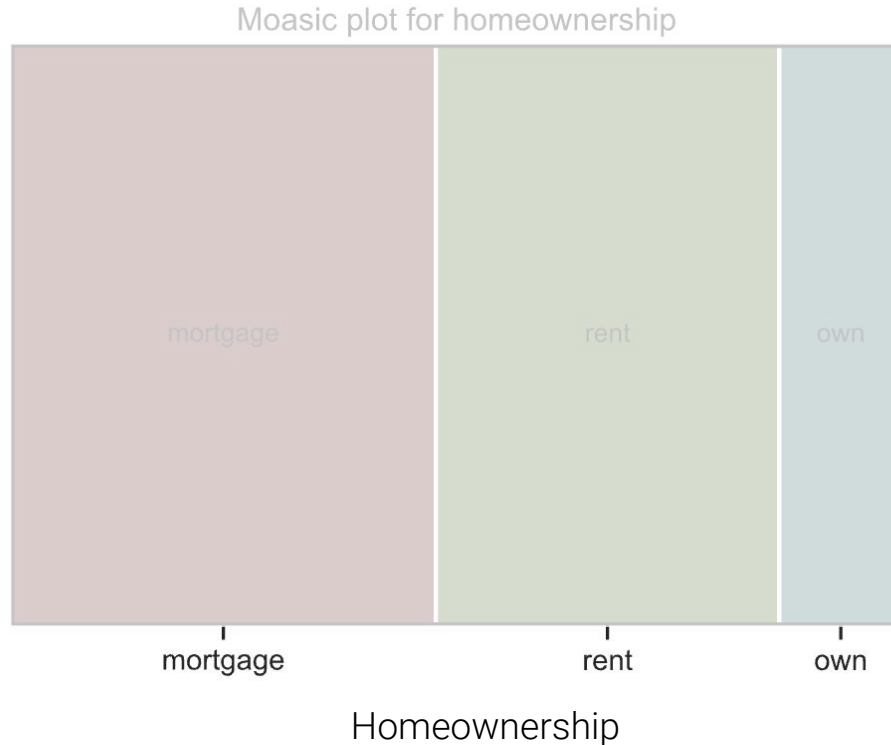
Homeownership

Resembles a standardized stacked bar plot

Moasic plot for homeownership

mortgage | rent | own

mortgage | rent | own
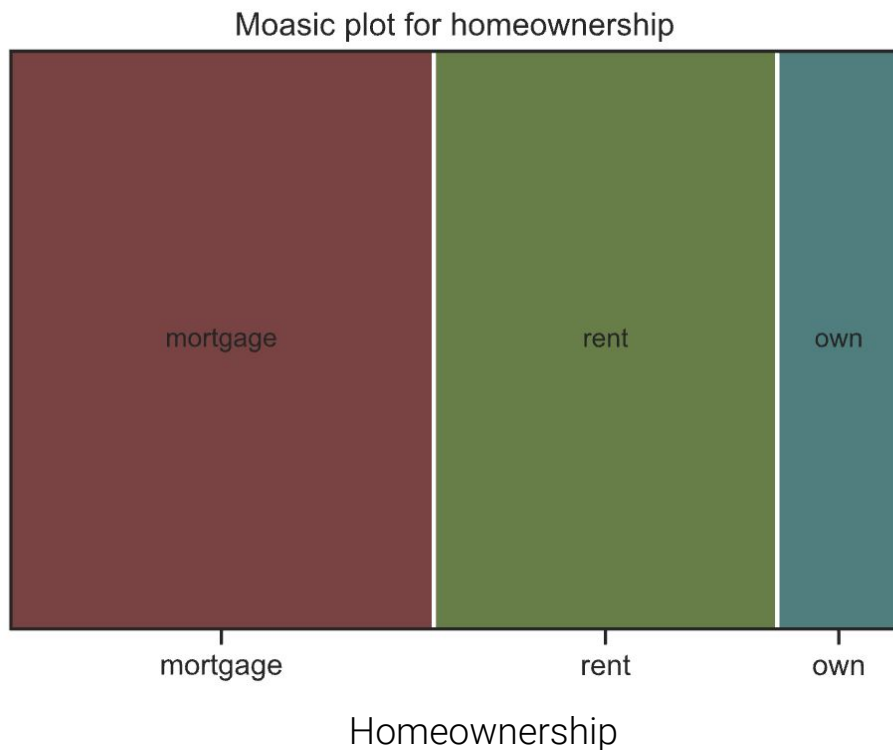
Homeownership

Column widths correspond to the proportion of loans in each of those categories

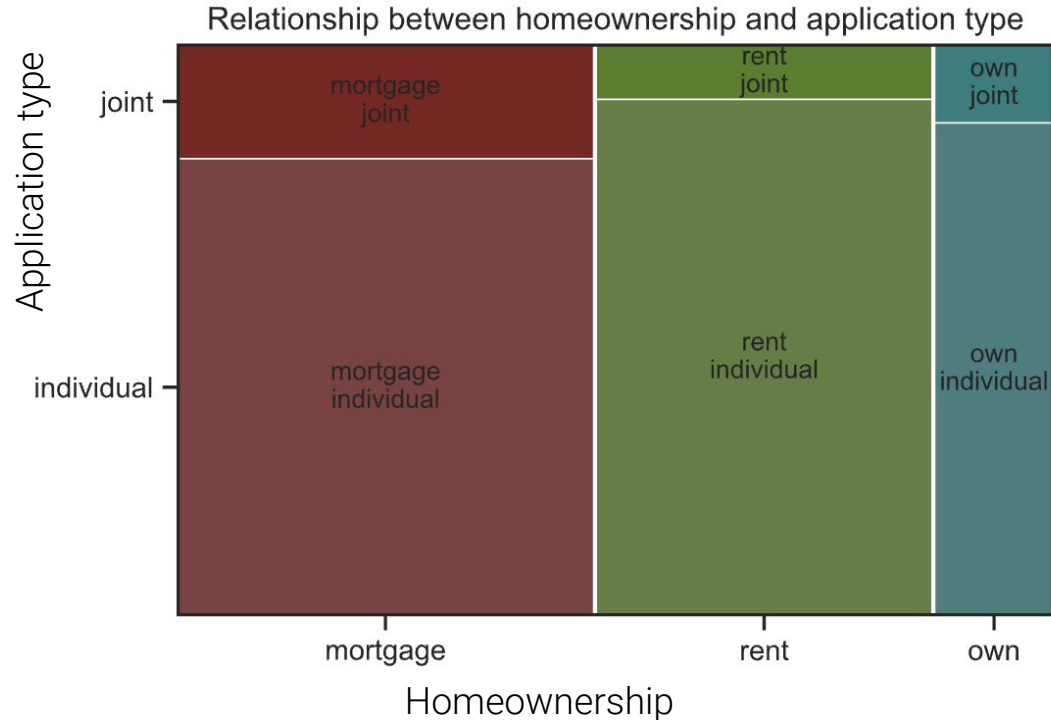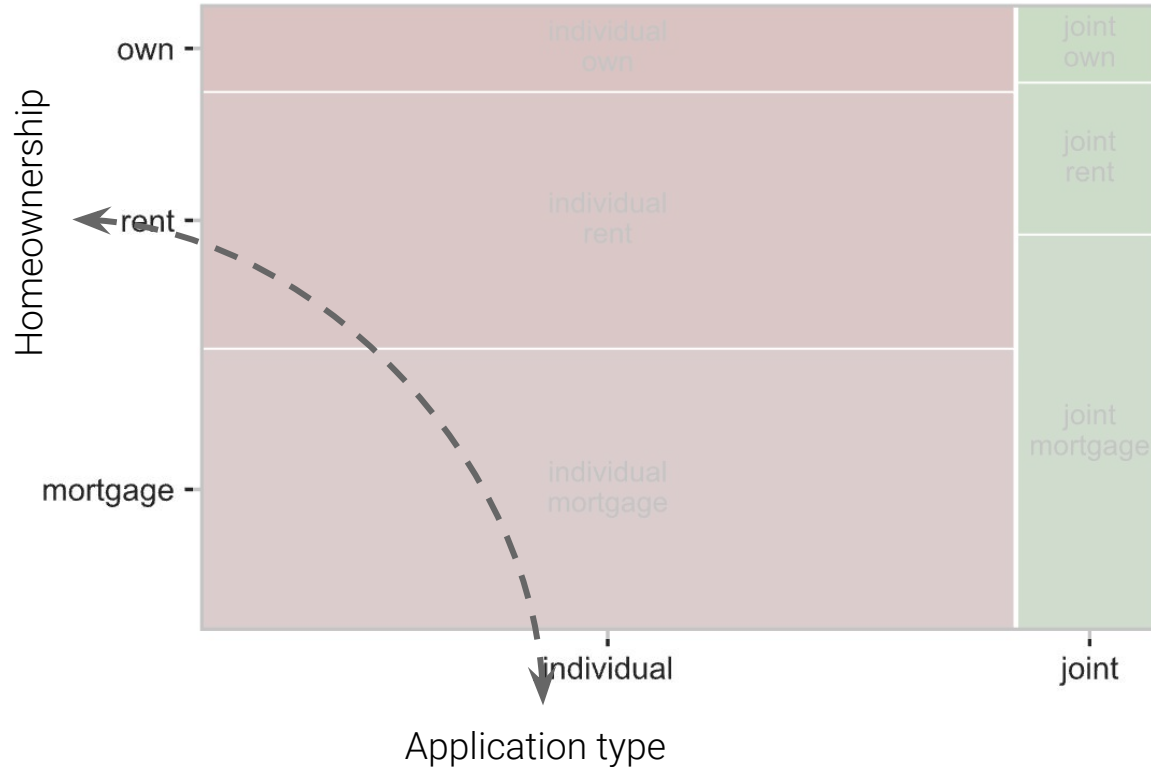# Relationship between **homeownership** and **application type**
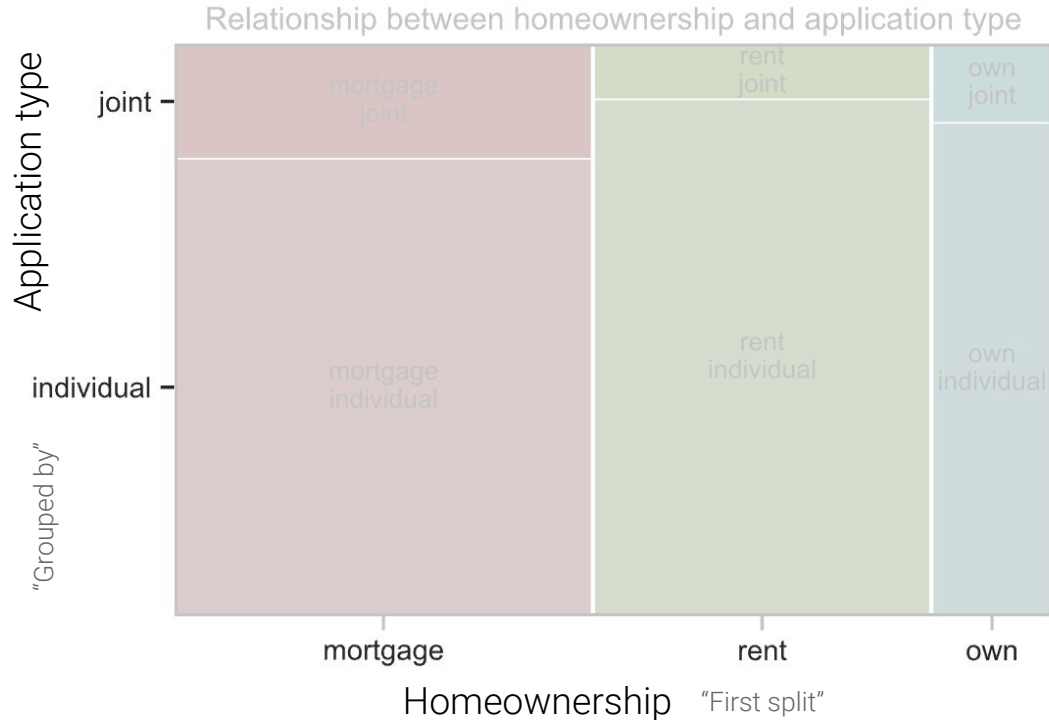
Relationship between homeownership and application type

Application type

"Grouped by"

- joint
- individual

mortgage joint

mortgage individual

rent joint

rent individual

own joint

own individual

mortgage   rent   own

Homeownership   "First split"

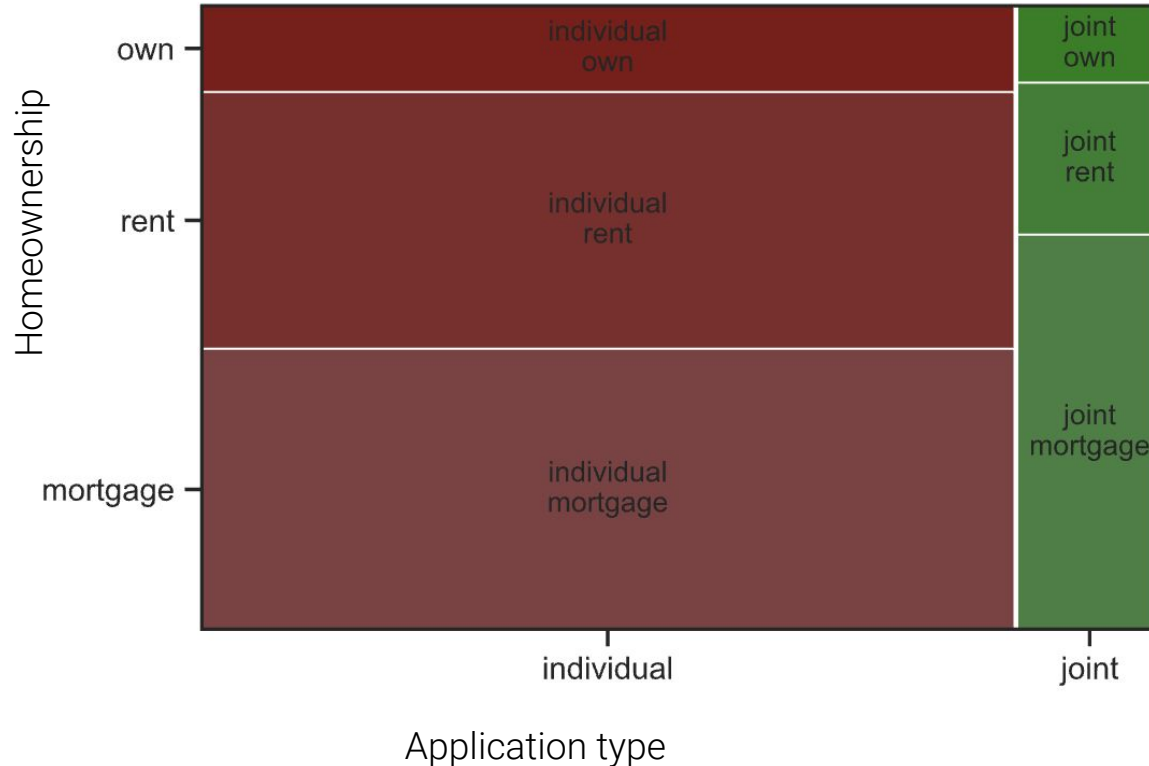# Relationship between **homeownership** and **application type**

New plot: grouped by homeownership

Homeownership

own

rent

mortgage

individual
own

joint
own

individual
rent

joint
rent

joint
rent

individual
mortgage

joint
mortgage

individual

joint

Application type

New plot: grouped by homeownership

Contingency table - Part II -

Prof. Dr. Jan Kirenz

Table with absolute values (focus on **mortgage**)

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

Inspection of application type "**individual**" (row)

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

→ **row**

Inspection of homeownership "**mortgage**" (column)

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

column

Contingency table
with proportions

Prof. Dr. Jan Kirenz

Table with **row** proportions

homeownership / application_type	mortgage	own	rent	Total
individual	0.451	0.138	0.411	1.0
joint	0.635	0.122	0.242	1.0

How many of the individual applicants **rent**?

homeownership	mortgage	own	rent
application_type			
individual	0.451	0.138	**0.411**
joint	0.635	0.122	0.242

Proportion of individual applicants who rent: 41.1%

Table with **column** proportions

homeownership	mortgage	own	rent
application_type			
individual	0.802	0.865	0.906
joint	0.198	0.135	0.094
Total	**1.0**	**1.0**	**1.0**

How many of the renters applied as **individuals**?

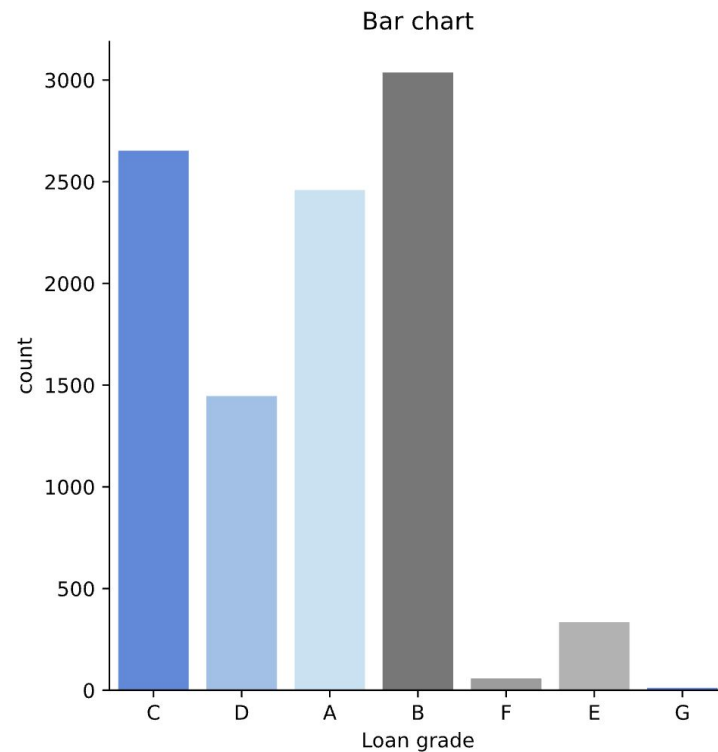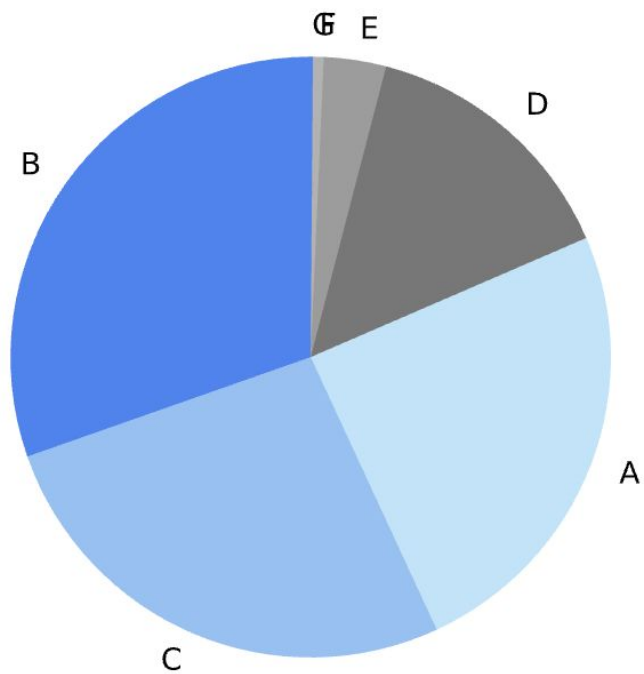| homeownership	mortgage	own	rent
application_type			
individual	0.802	0.865	**0.906**
joint	0.198	0.135	0.094

Proportion of renters who applied as individuals: 90.6%

Pie charts

Pie chart vs bar chart
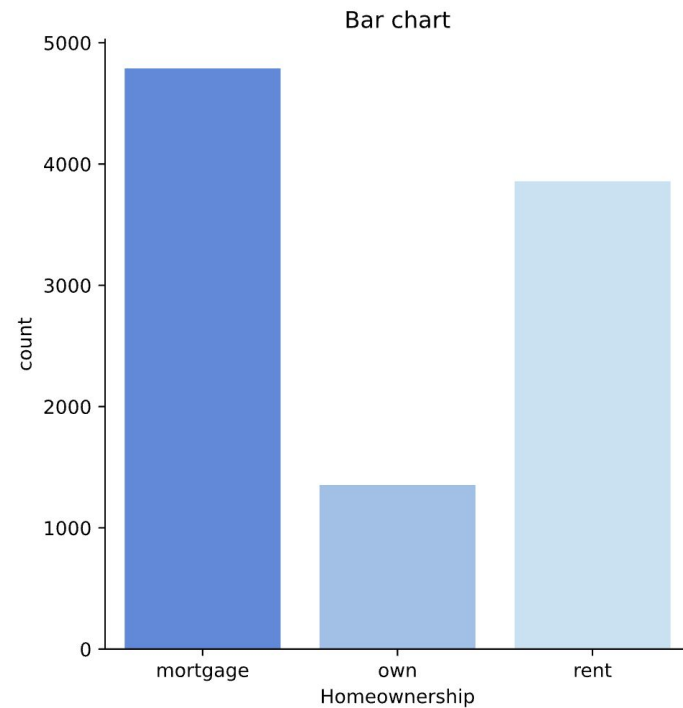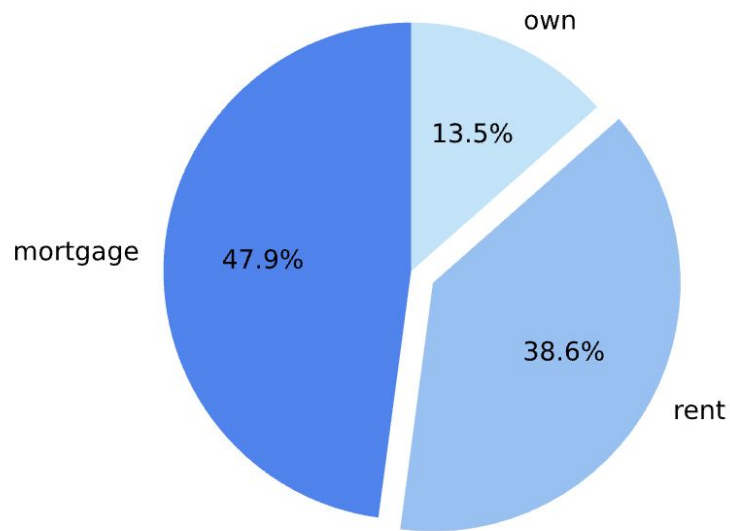
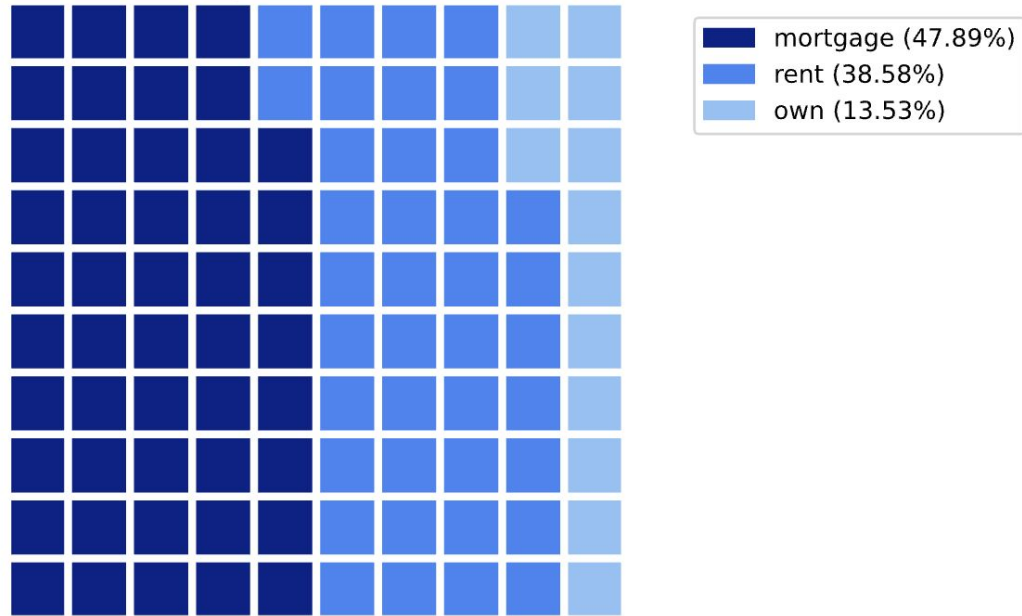# Pie chart vs bar chart

Waffle chart

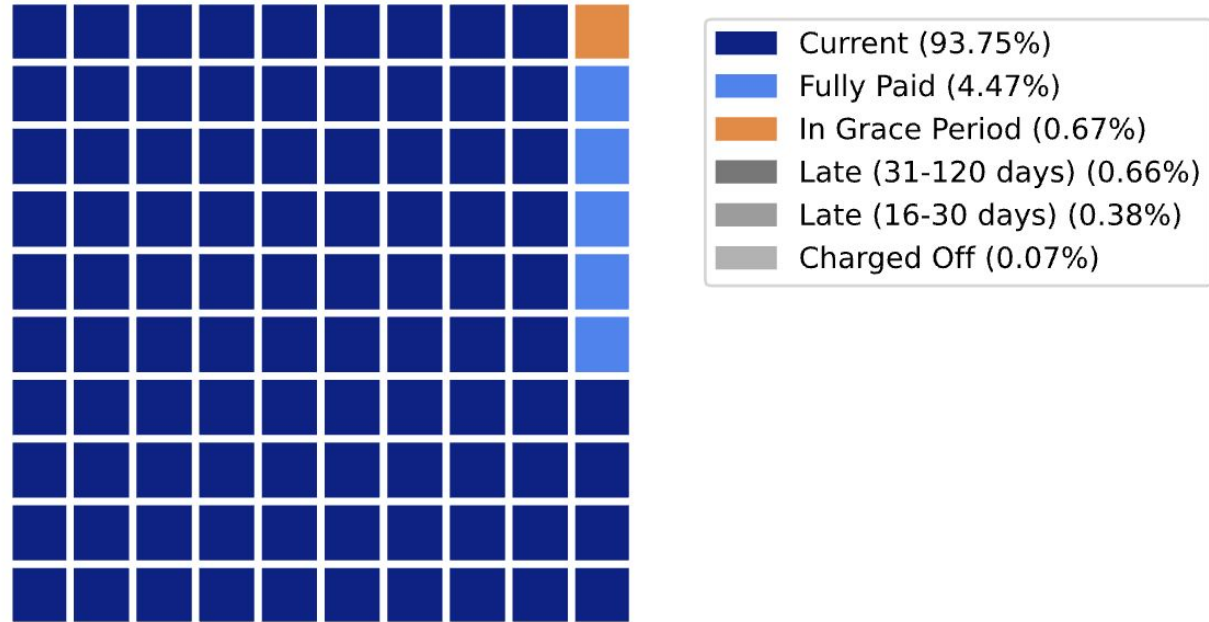# Waffle chart of homeownership, with levels rent, mortgage, and own

Homeownership

- mortgage (47.89%)
- rent (38.58%)
- own (13.53%)

Waffle chart of loan status, with levels current, fully paid, in grade period, and late

Loan status

Legend:
- Current (93.75%)
- Fully Paid (4.47%)
- In Grace Period (0.67%)
- Late (31-120 days) (0.66%)
- Late (16-30 days) (0.38%)
- Charged Off (0.07%)

Resources

The slides are based on the excellent book "Introduction to Modern Statistics" by Mine Çetinkaya-Rundel and Johanna Hardin.

The online version can be **accessed for free**:

https://openintro-ims.netlify.app/explore-categorical.html