Describing Numerical and Categorical Data

STA 032: Gateway to data science Lecture 11

Jingwei Xiong

April 26, 2023

Reminders

- HW 2 due April 26 12pm.
- HW 3 due May 3 12pm.
    - Please start the homework as soon as possible.
- **Midterm 1** (Open book, take home, approximate 2 hours, time limit 5 hours)
    - **Due April 29 midnight, cover lecture 1-12**
    - Preparing guides: Finish Homework 1-3, be familiar to the lecture slides.
    - You can copy any your own codes in homework 1-3 to finish the open book exam.
- The PDF version of slides can be found on Canvas.
    - You can use search on it to find function examples.

Recap

- Relationships between numerical variables
  - Correlation coefficient
  - Line graph
- Describing categorical distributions
  - Bar plot
- Relationships between categorical data
  - Contingency tables

Today

- Relationships between categorical data
  - Stacked bar plot
  - Color pallettes
  - side by side bar plot
  - Customizing bar plots
- Relationships between numerical and categorical data
  - Fill and facet
  - Side-by-side boxplots
  - Other fancy plots
  - Numerical summaries in R

Data: Lending Club

- Lending Club is a platform that allows individuals to lend to other individuals

```
loans <- loans_full_schema %>%
  select(loan_amount, interest_rate, term, grade,
         state, annual_income, homeownership, debt_to_income,
         issue_month)
glimpse(loans)
```

```
Rows: 10,000
Columns: 9
$ loan_amount    <int> 28000, 5000, 2000, 21600, 23000, 5000, 24000, 20000, 2
$ interest_rate  <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6.72, 13.59, 11.99,
$ term           <dbl> 60, 36, 36, 36, 36, 36, 60, 60, 36, 36, 60, 60, 36, 60
$ grade          <fct> C, C, D, A, C, A, C, B, C, A, C, B, C, B, D, D, D, F,
$ state          <fct> NJ, HI, WI, PA, CA, KY, MI, AZ, NV, IL, IL, FL, SC, CO
$ annual_income  <dbl> 90000, 40000, 40000, 30000, 35000, 34000, 35000, 11000
$ homeownership  <fct> MORTGAGE, RENT, RENT, RENT, RENT, OWN, MORTGAGE, MORTG
$ debt_to_income <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6.46, 23.66, 16.19,
$ issue_month    <fct> Mar-2018, Feb-2018, Feb-2018, Jan-2018, Mar-2018, Jan-
```

Stacked bar plot

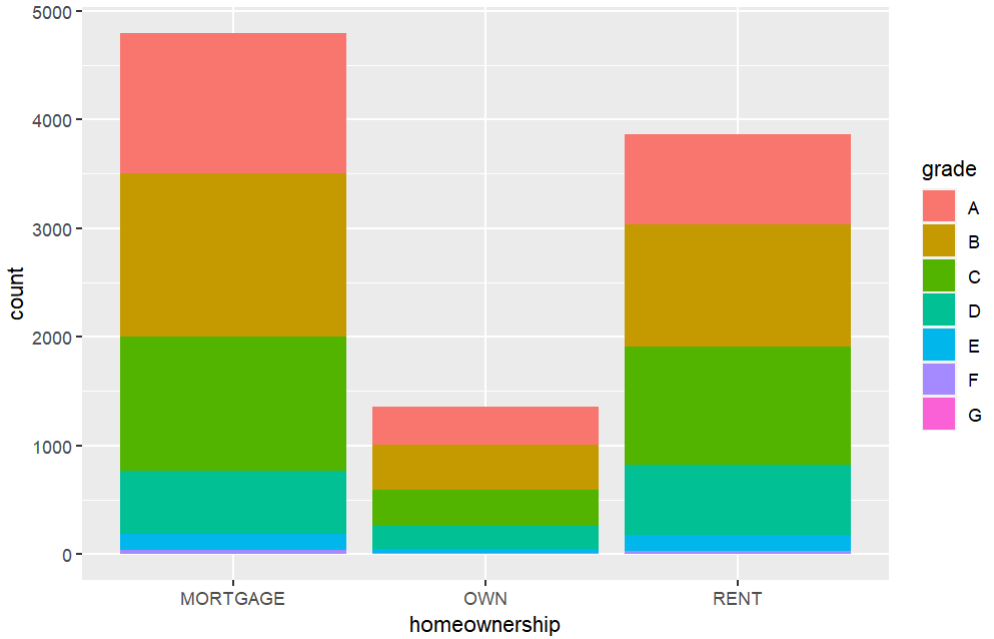
Overview | Plot | Change color | Plot 2

- A stacked bar plot looks at numeric values across two categorical variable
- Each bar in a standard bar plot is divided into stacked sub-bars, each one corresponding to a level of the second categorical variable.

Stacked bar plot

Overview | **Plot** | Change color | Plot 2

```
ggplot(loans, aes(x = homeownership,
                  fill = grade)) +
  geom_bar()
```



Stacked bar plot

Overview | Plot | Change color | Plot 2

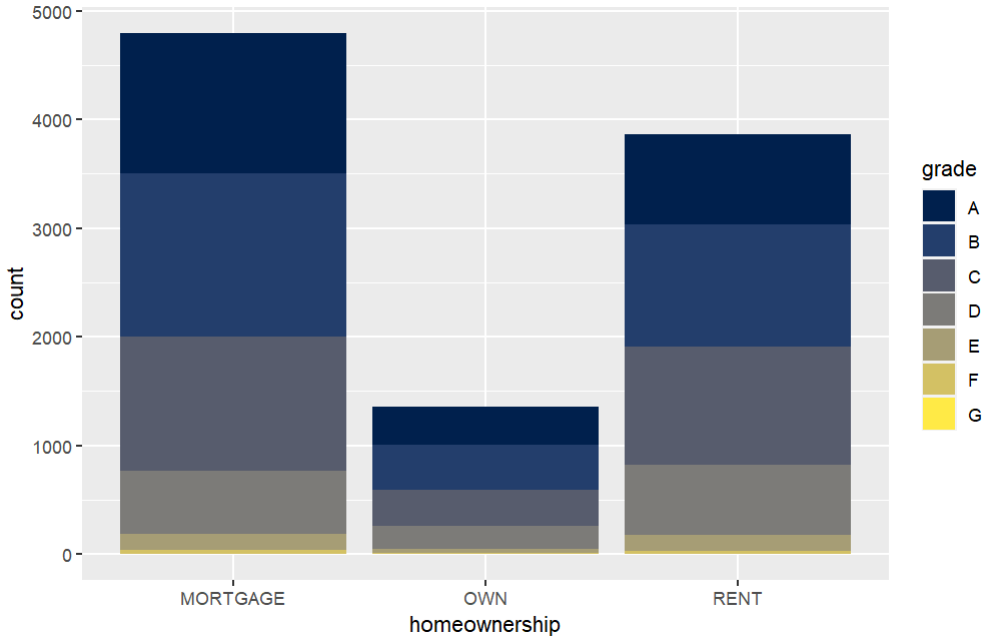
Change the color with `scale_fill_viridis_d(option = "D")` or `scale_colour_viridis_d(option = "D")`.

- option: A character string indicating the color map option to use. Eight options are available:
  - "magma" (or "A")
  - "inferno" (or "B")
  - "plasma" (or "C")
  - "viridis" (or "D")
  - "cividis" (or "E")
  - "rocket" (or "F")
  - "mako" (or "G")
  - "turbo" (or "H")

Stacked bar plot

Overview | Plot | Change color | **Plot 2**

```
ggplot(loans, aes(x = homeownership,  fill = grade)) +
  geom_bar() +
  scale_fill_viridis_d(option = "E")
```



Stacked bar plot: ordinal variable fill

ordinal variable | Example

Turning `grade` into an ordered variable makes `ggplot` use the `viridis` scale by default.

The `ggplot` will also try to order the variable in plot.

```
str(loans$grade)
```

```
 Factor w/ 8 levels "","A","B","C",..: 4 4 5 2 4 2 4 3 4 2 ...
```

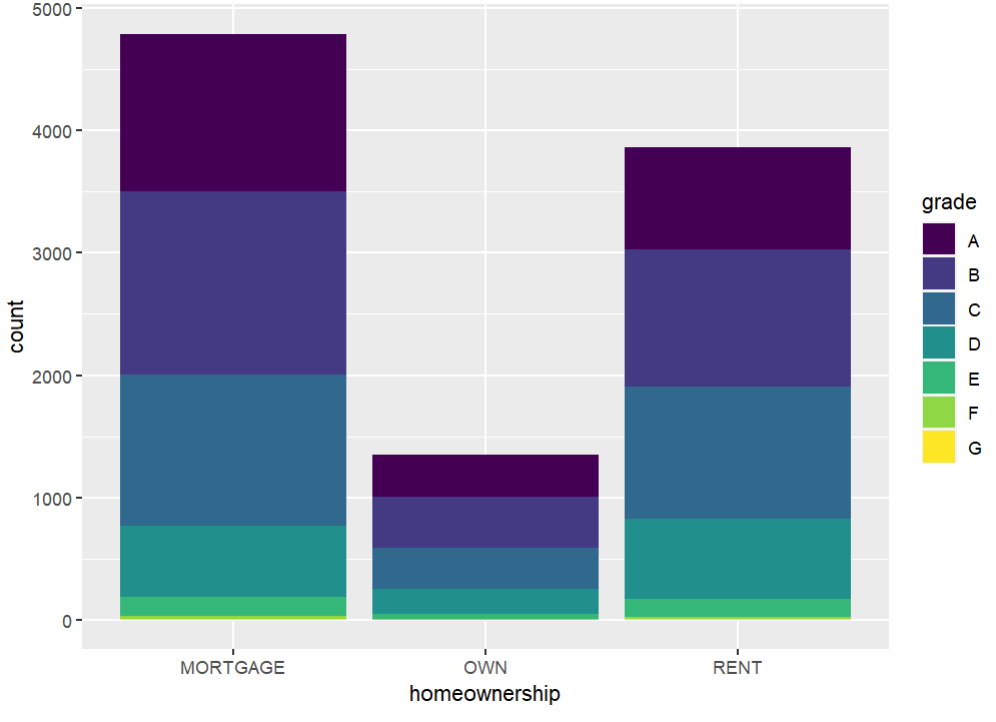
```
loans <- loans %>%
  mutate(grade = factor(grade, ordered = TRUE))
str(loans$grade)
```

```
 Ord.factor w/ 7 levels "A"<"B"<"C"<"D"<..: 3 3 4 1 3 1 3 2 3 1 ...
```

Stacked bar plot: ordinal variable fill

ordinal variable | Example

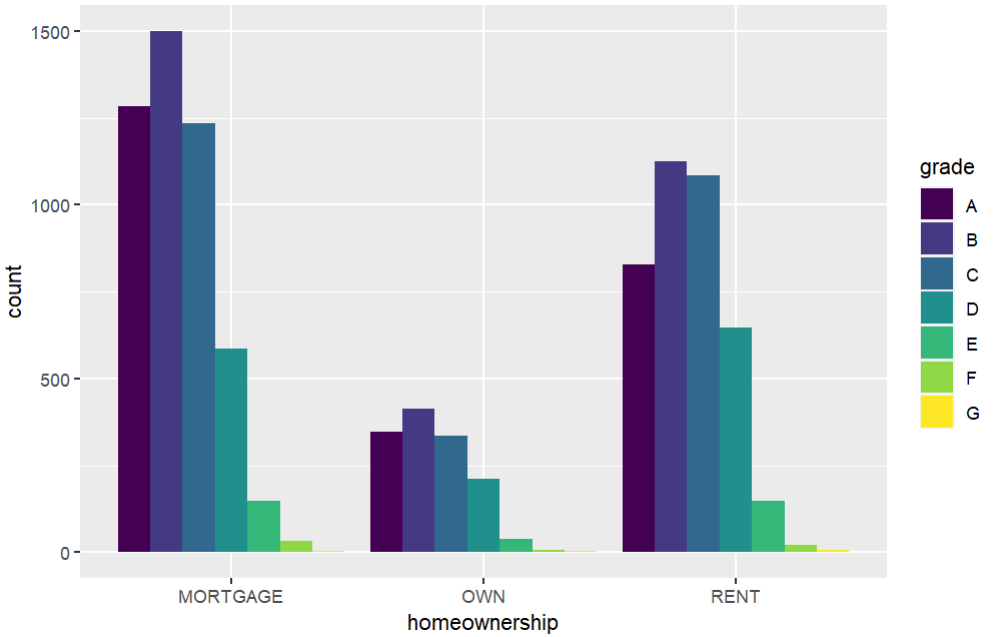
```
ggplot(loans, aes(x = homeownership,
                  fill = grade)) +
  geom_bar()
```



Side by side bar plot

Side by side | Swap the categories

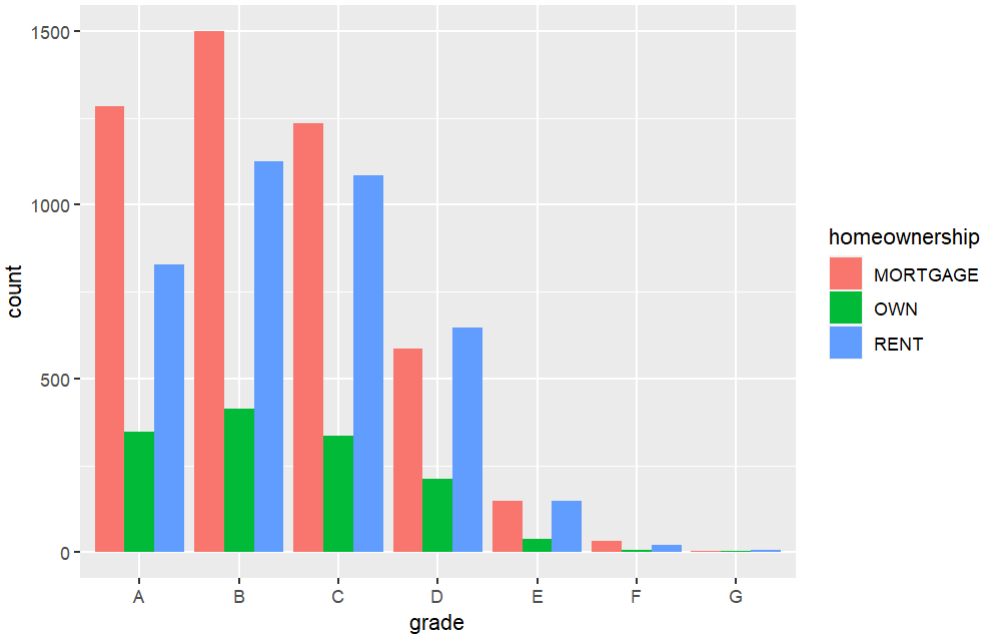
```
ggplot(loans, aes(x = homeownership,
                  fill = grade)) +
  geom_bar(position = 'dodge')
```



Side by side bar plot

Side by side | Swap the categories

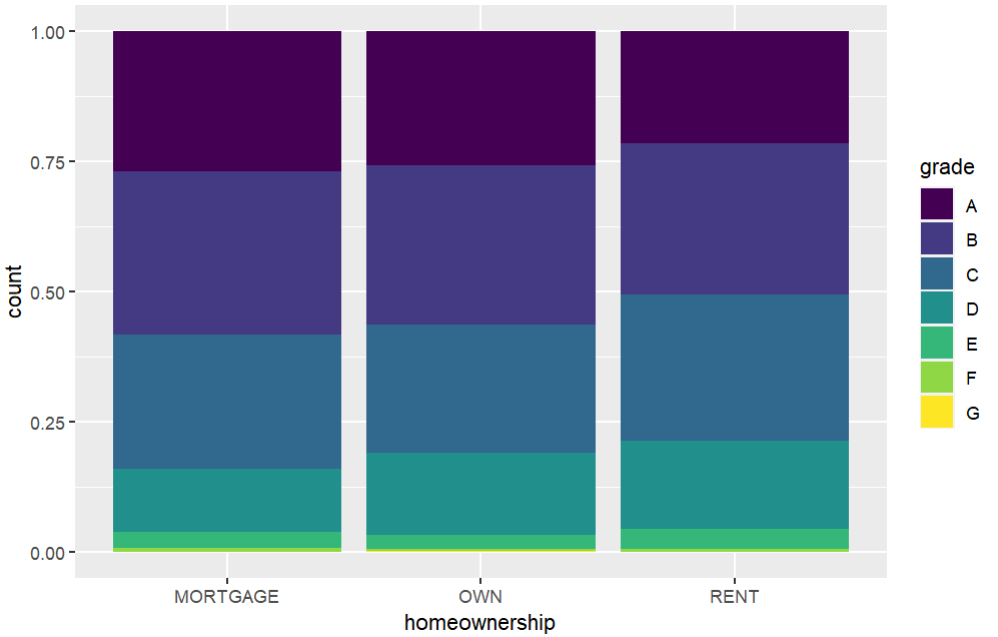
```
ggplot(loans, aes(x = grade,
                  fill = homeownership)) +
  geom_bar(position = 'dodge')
```



Stacked bar plot

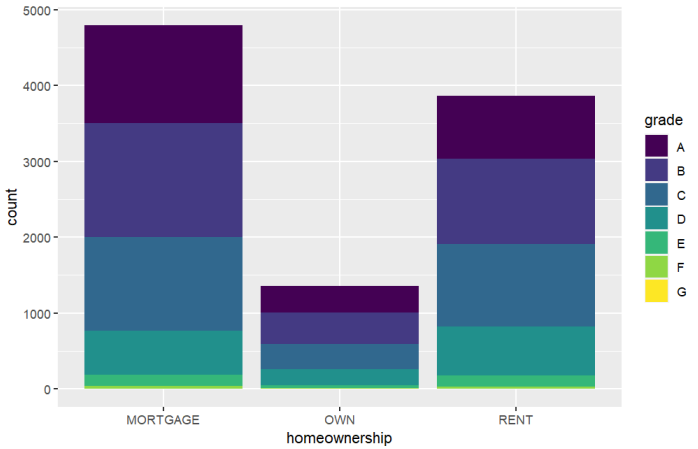
Adding `position = "fill"` argument changes visualization to proportions, and standardizes the height of columns into proportions.

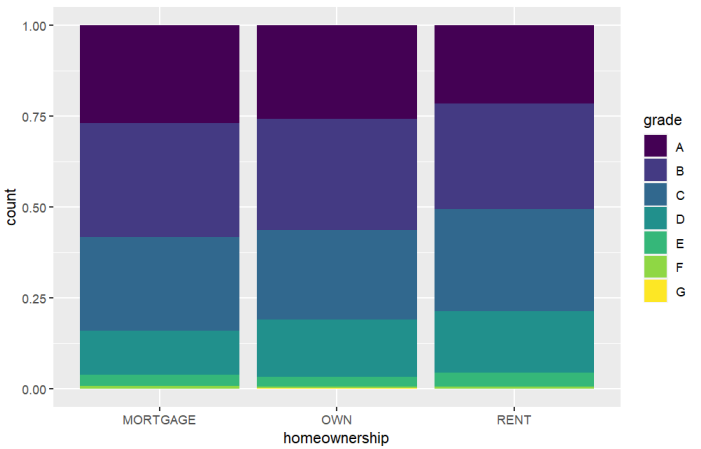
```
ggplot(loans, aes(x = homeownership, fill = grade)) +
  geom_bar(position = "fill")
```



Stacked bar plot: counts vs. proportions

Which bar plot is a more useful representation for visualizing the relationship between homeownership and grade?

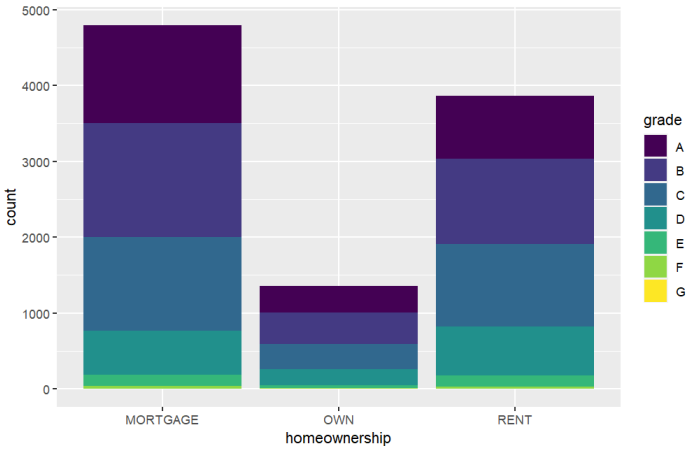


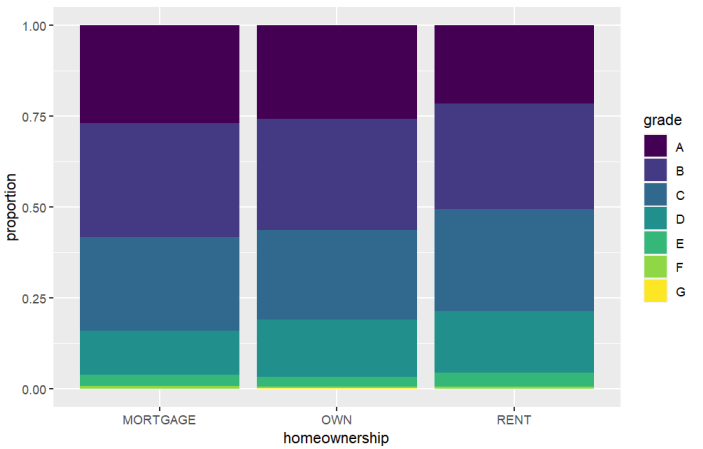


If there were no relationship between homeownership and grade, we would expect to see the bars to be similar lengths across the homeownership status (columns).

Stacked bar plot: counts vs. proportions

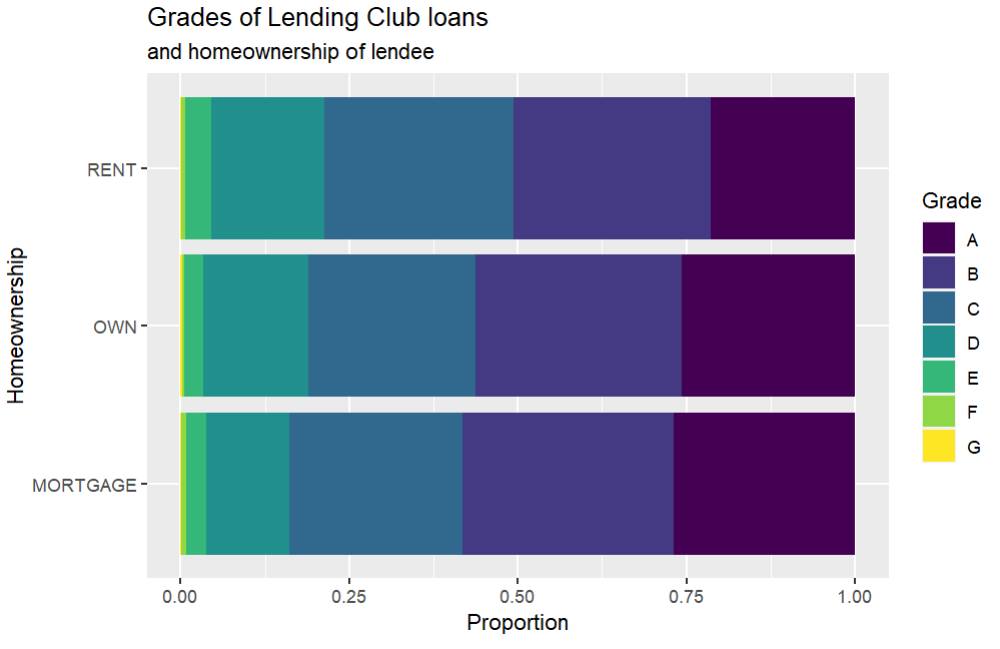
Is there a relationship between homeownership and grade?





Customizing bar plots

Plot | Code | Change color



Customizing bar plots

Plot | Code | Change color

```
ggplot(loans, aes(y = homeownership,
                  fill = grade)) +
  geom_bar(position = "fill") +
  labs(
    x = "Proportion",
    y = "Homeownership",
    fill = "Grade",
    title = "Grades of Lending Club loans",
    subtitle = "and homeownership of lendee"
  ) +
  scale_fill_viridis_d(option = "D")
```

Customizing bar plots

Plot Code Change color

Change the color with `scale_fill_viridis_d(option = "D")` or `scale_colour_viridis_d(option = "D")`.

- option: A character string indicating the color map option to use. Eight options are available:
  - "magma" (or "A")
  - "inferno" (or "B")
  - "plasma" (or "C")
  - "viridis" (or "D")
  - "cividis" (or "E")
  - "rocket" (or "F")
  - "mako" (or "G")
  - "turbo" (or "H")

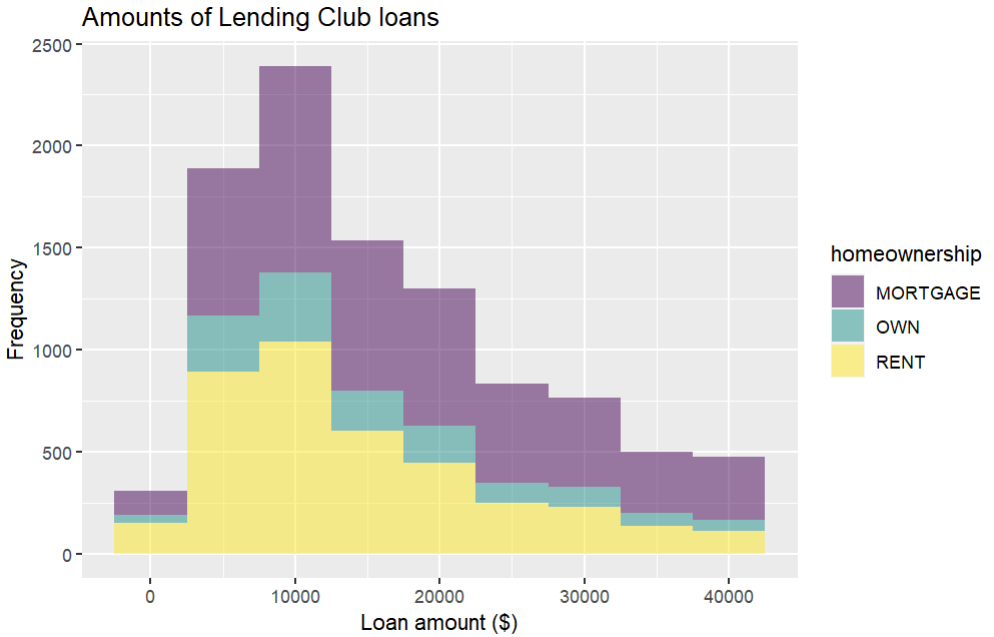
Relationships between numerical and categorical data

- We saw histograms, boxplots, and density plots earlier, for describing a single numerical variable
- To look at relationships between these numerical data and a categorical variable, we can:
  - Fill and facet histograms and density plots
  - Use side-by-side boxplots
  - Violin plots
  - Ridge plots
- Numerical summaries
  - `group_by()` and output table

Fill a histogram with a categorical variable

Is there a relationship between loan amount and home-ownership status?

Plot | Code



Fill a histogram with a categorical variable

Is there a relationship between loan amount and home-ownership status?

Plot | Code

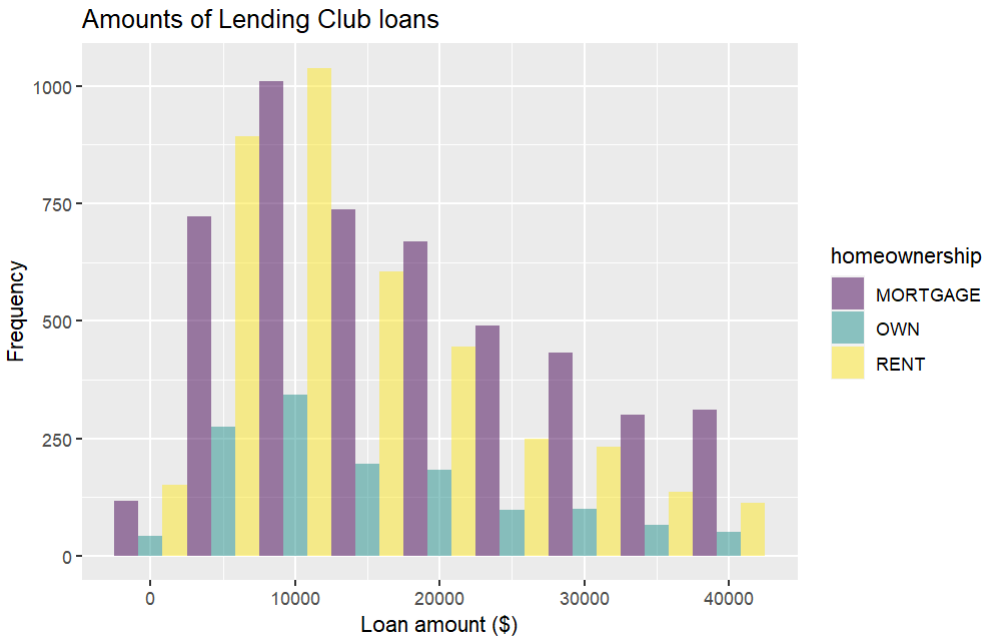
```
ggplot(loans, aes(x = loan_amount,
                  fill = homeownership)) +
  geom_histogram(binwidth = 5000,
                 alpha = 0.5) +
  labs(
    x = "Loan amount ($)",
    y = "Frequency",
    title = "Amounts of Lending Club loans"
  ) +
  scale_fill_viridis_d()
```

Fill a histogram with a categorical variable

Is there a relationship between loan amount and home-ownership status?

- Need `position = "dodge"` argument if we don't want histogram bars to be stacked on top of one another

Plot | Code



Fill a histogram with a categorical variable

Is there a relationship between loan amount and home-ownership status?

- Need `position = "dodge"` argument if we don't want histogram bars to be stacked on top of one another

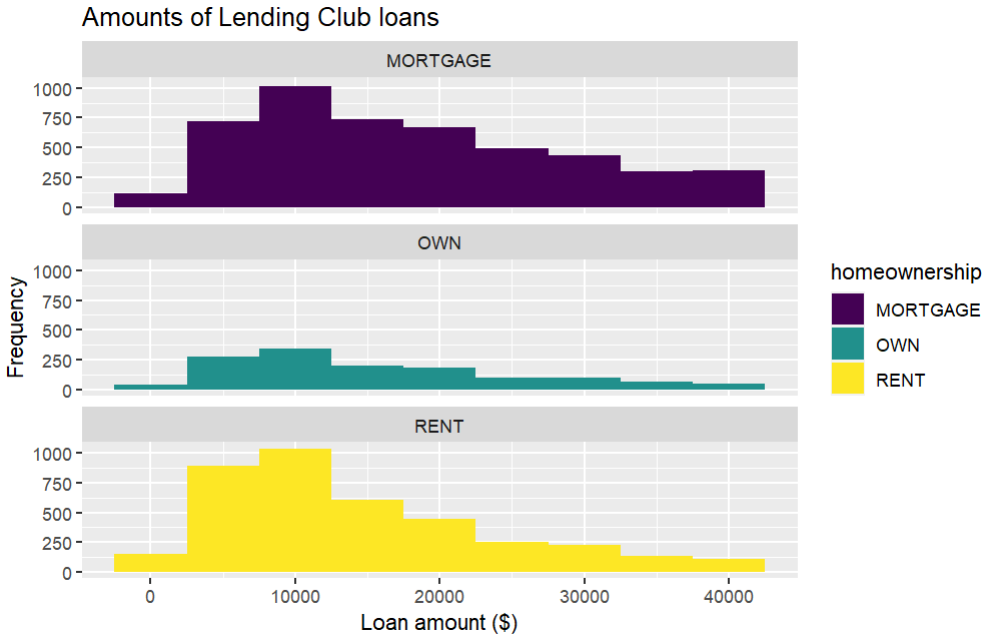
Plot | Code

```
ggplot(loans, aes(x = loan_amount,
                  fill = homeownership)) +
  geom_histogram(position = "dodge",
                 binwidth = 5000,
                 alpha = 0.5) +
  labs(
    x = "Loan amount ($)",
    y = "Frequency",
    title = "Amounts of Lending Club loans"
  ) +
  scale_fill_viridis_d()
```

Facet a histogram with a categorical variable

Is there a relationship between loan amount and home-ownership status?

Plot | Code



Facet a histogram with a categorical variable

Is there a relationship between loan amount and home-ownership status?

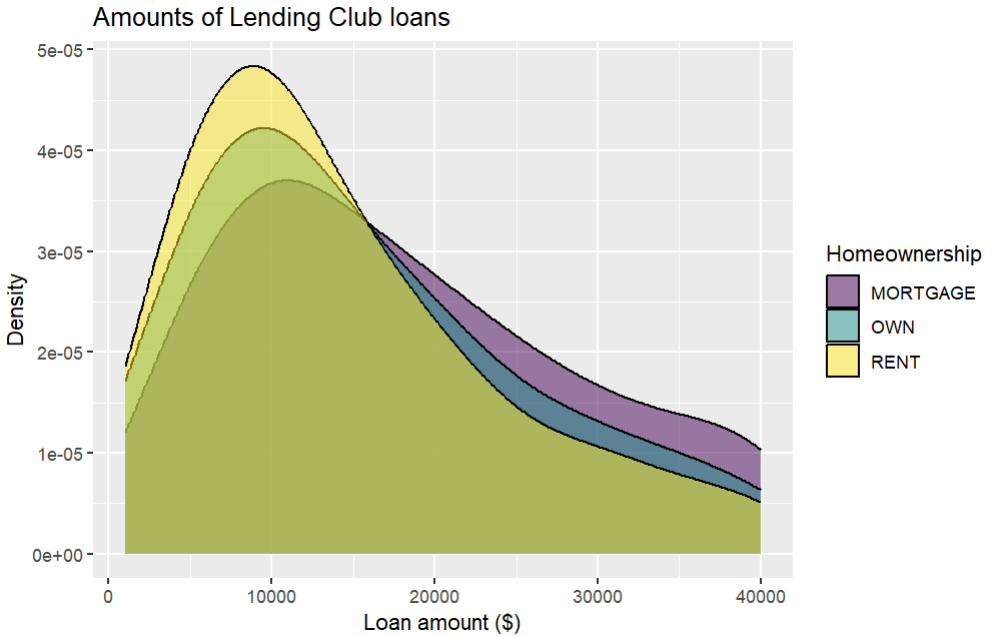
Plot | Code

```
ggplot(loans, aes(x = loan_amount, fill = homeownership)) +
  geom_histogram(binwidth = 5000) +
  labs(
    x = "Loan amount ($)",
    y = "Frequency",
    title = "Amounts of Lending Club loans"
  ) +
  facet_wrap(~ homeownership, nrow = 3) +
  scale_fill_viridis_d()
```

Filling density plots with a categorical variable

Is there a relationship between loan amount and home-ownership status?

Plot | Code



Filling density plots with a categorical variable

Is there a relationship between loan amount and home-ownership status?

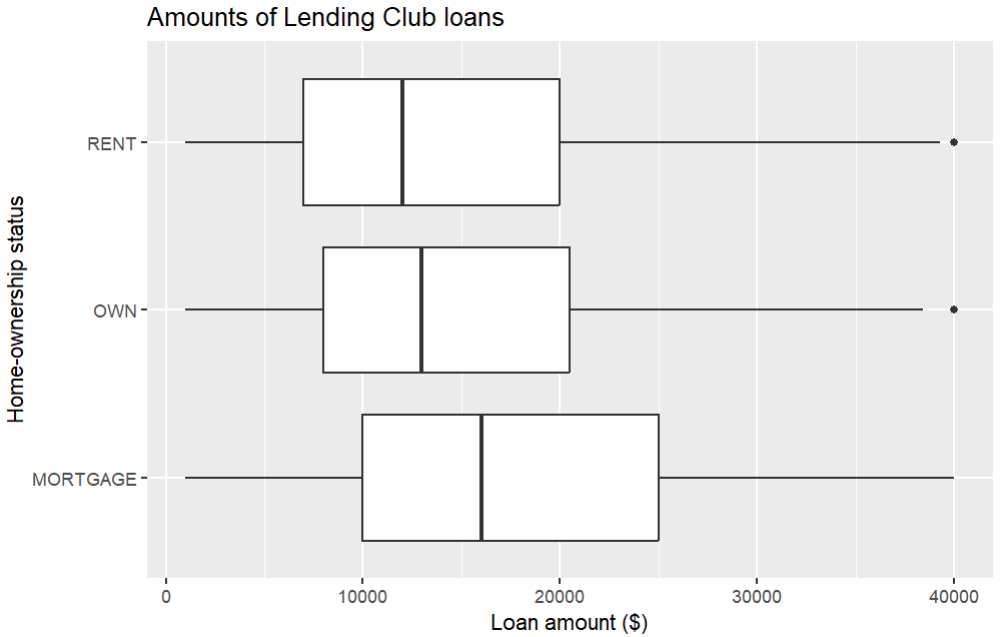
Plot | Code

```
ggplot(loans, aes(x = loan_amount,
                  fill = homeownership)) +
  geom_density(adjust = 2,  # adjust for smooth parameter
               alpha = 0.5) +
  labs(
    x = "Loan amount ($)",
    y = "Density",
    title = "Amounts of Lending Club loans",
    fill = "Homeownership"
  ) +
  scale_fill_viridis_d()
```

Side-by-side boxplots

Is there a relationship between loan amount and home-ownership status?

Plot | Code



Side-by-side boxplots

Is there a relationship between loan amount and home-ownership status?

Plot | Code

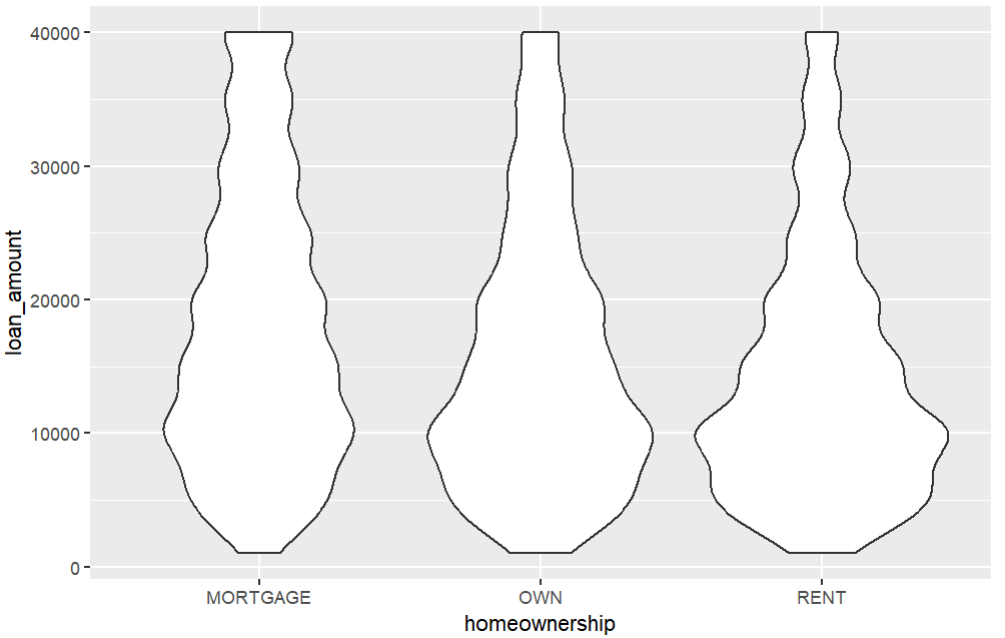
```
ggplot(loans, aes(x = loan_amount,
                  y = homeownership)) +
  geom_boxplot() +
  labs(
    x = "Loan amount ($)",
    y = "Home-ownership status",
    title = "Amounts of Lending Club loans"
  )
```

Violin plots

Is there a relationship between loan amount and home-ownership status?

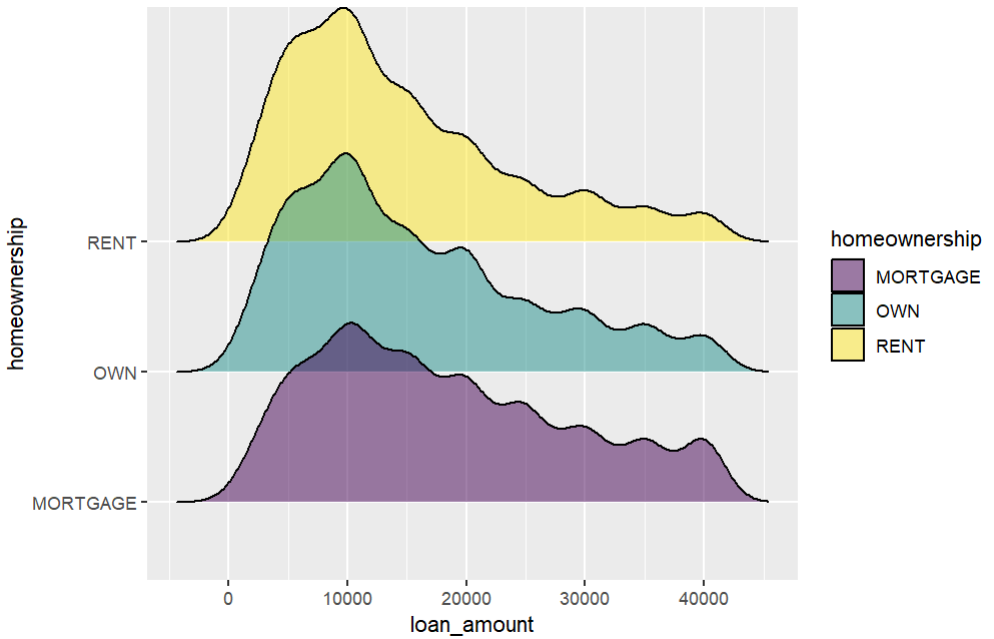
- A violin plot is a hybrid of a boxplot and a density plot

```
ggplot(loans, aes(x = homeownership, y = loan_amount)) +
  geom_violin()
```



Ridge plots

```
library(ggridges)
ggplot(loans, aes(x = loan_amount, y = homeownership,
                  fill = homeownership)) +
  geom_density_ridges(alpha = 0.5) +
  scale_fill_viridis_d()
```

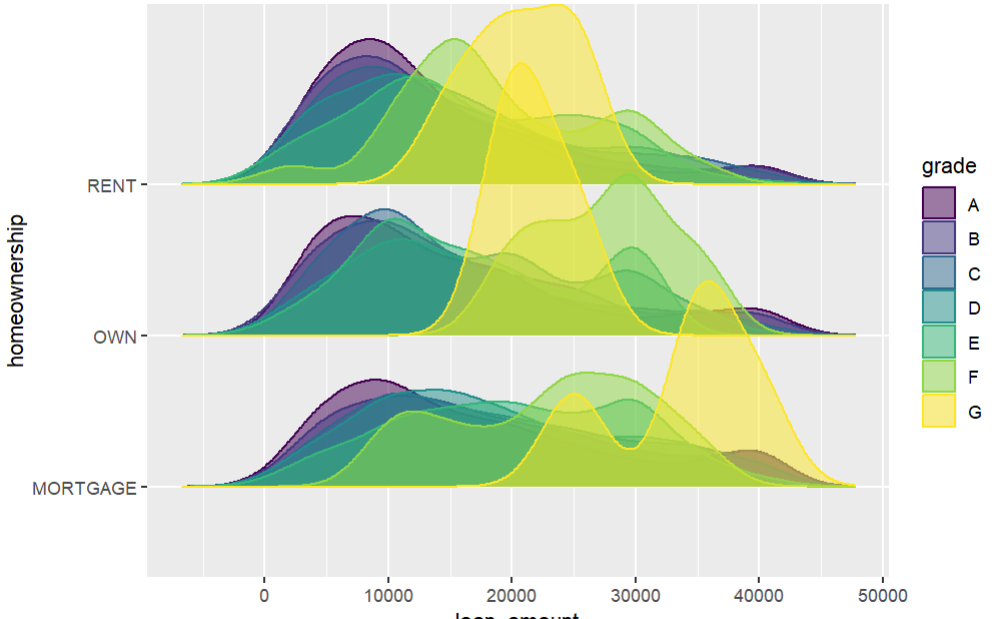


Ridge plots

Ridge plots can also be used to investigate more complicated relationships, such as those between a categorical and numerical variable, conditional on another categorical variable

Here, we consider the relationship between loan grade and loan amount, conditional on each level of home ownership

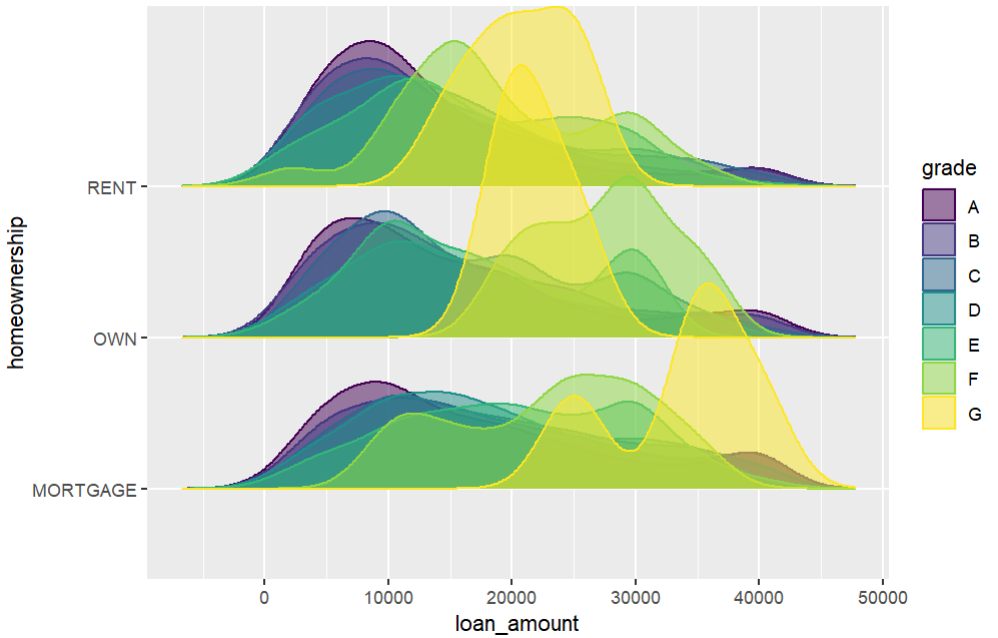
```
ggplot(loans, aes(x = loan_amount, y = homeownership, fill = grade, c
  geom_density_ridges(alpha = 0.5)
```



Ridge plots

Here, we consider the relationship between loan grade and loan amount, conditional on each level of home ownership

```
ggplot(loans, aes(x = loan_amount, y = homeownership, fill = grade, c
  geom_density_ridges(alpha = 0.5)
```



Interestingly, those who had mortgages tend to have a higher proportion of grade G loans that have higher loan amounts.

Numerical summaries in R, grouping by a categorical variable

Question: `homeownership` is a factor variable with three levels, `MORTGAGE`, `OWN` and `RENT`. How do we calculate the mean loan amount for each type of home ownership status?

```
loans %>%
  group_by(homeownership) %>%
  summarize(meanLoan = mean(loan_amount))
```

```
# A tibble: 3 × 2
  homeownership meanLoan
  <fct>            <dbl>
1 MORTGAGE        18129.
2 OWN             15684.
3 RENT            14406.
```

But how to make the result output as a table, not in R output format?

knitr::kable function for output

We can use the `knitr::kable` function.

```
t1 = loans %>%
  group_by(homeownership) %>%
  summarize(meanLoan = mean(loan_amount))
knitr::kable(t1, caption = "Summary table 1", digits=1,
             align = "l")
```

Table: Summary table 1

homeownership	meanLoan
MORTGAGE	18129.0
OWN	15683.9
RENT	14406.2

What descriptive statistics to use? Which plots to produce?

- Fancier does not always mean better; a pretty plot can look great but tell us nothing.
- Complex does not always mean better; you may be lost in too much information.
- Think about what question you are trying to answer and pick the figure that best suits the purpose.
- Hadley Wickham on exploratory data analysis: "EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind." (Chapter 7, R for Data Science)
- Extra examples: Chapter 11 Data visualization principles

Other fancy plots example:

- Pie chart
- The hourly heatmap
- Bubble plot
- Choropleth Map

Summary

- Relationships between categorical data
  - Contingency tables
  - Stacked bar plot
- Relationships between numerical and categorical data
  - Fill and facet
  - Side-by-side boxplots
  - Other fancy plots
  - Numerical summaries in R

Reading

- Chapter 10 Data visualization in practice
- Chapter 11 Data visualization principles