

Visualising categorical data

Data Science in a Box

datasciencebox.org



Recap



Variables

- **Numerical** variables can be classified as **continuous** or **discrete** based on whether or not the variable can take on an infinite number of values or only non-negative whole numbers, respectively.
- If the variable is **categorical**, we can determine if it is **ordinal** based on whether or not the levels have a natural ordering.



Data

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

```
## Rows: 10,000
## Columns: 8
## $ loan_amount    <int> 28000, 5000, 2000, 21600, 23000, 5000, 2...
## $ interest_rate  <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6.72, ...
## $ term           <dbl> 60, 36, 36, 36, 36, 36, 60, 60, 36, 36, ...
## $ grade          <ord> C, C, D, A, C, A, C, B, C, A, C, B, C, B...
## $ state          <fct> NJ, HI, WI, PA, CA, KY, MI, AZ, NV, IL, ...
## $ annual_income  <dbl> 90000, 40000, 40000, 30000, 35000, 34000...
## $ homeownership  <fct> MORTGAGE, RENT, RENT, RENT, RENT, OWN, M...
## $ debt_to_income <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6.46, ...
```

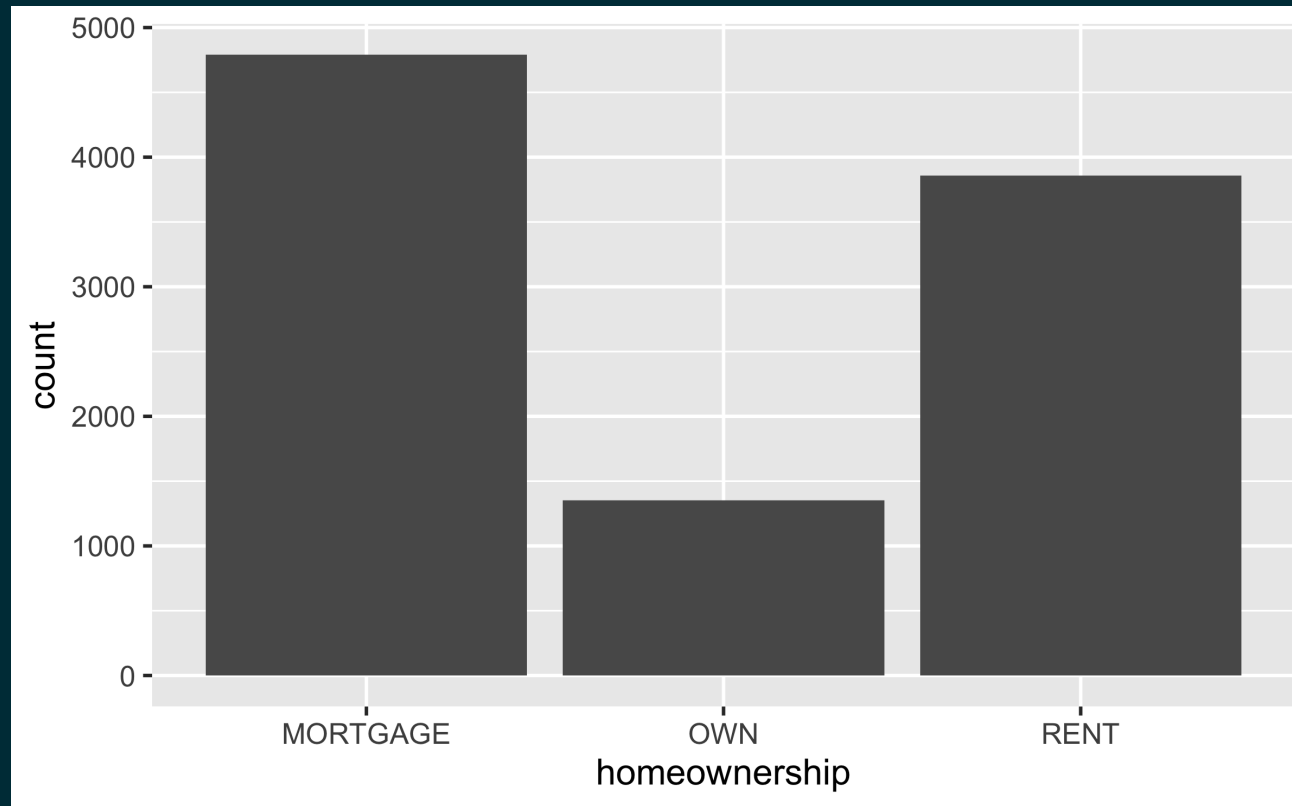


Bar plot



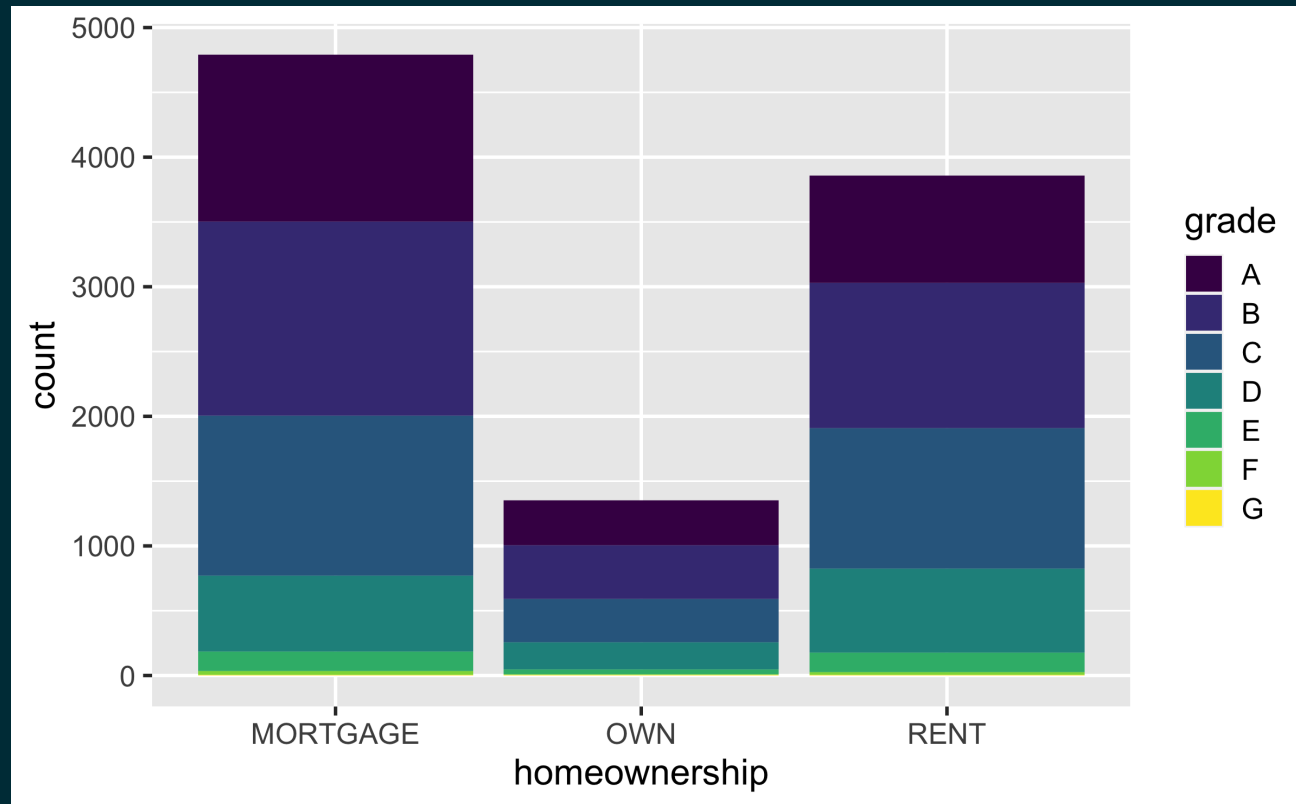
Bar plot

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



Segmented bar plot

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

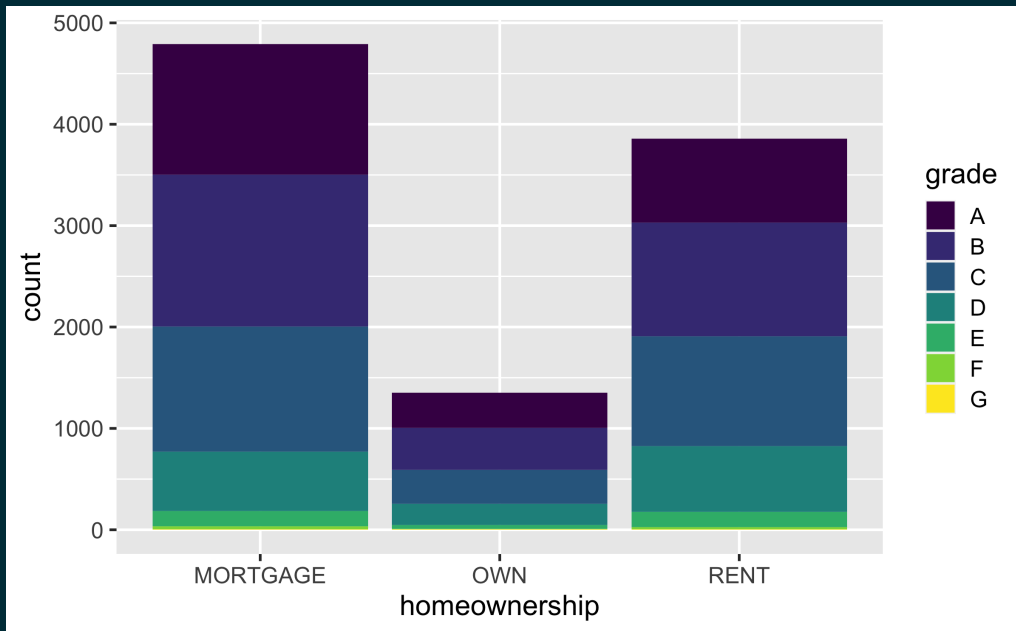


Segmented bar plot

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



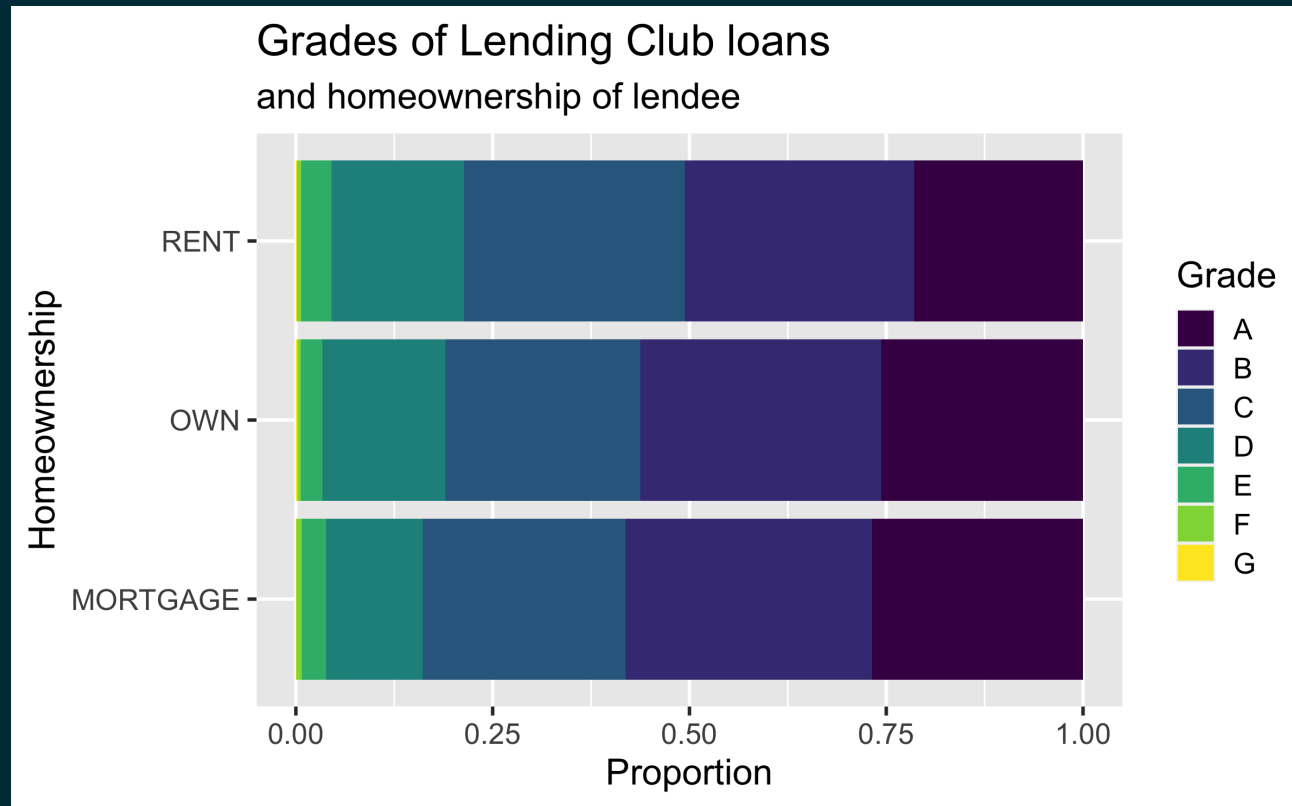
Which bar plot is a more useful representation for visualizing the relationship between homeownership and grade?



Customizing bar plots

Plot

Code



Customizing bar plots

Plot

Code

```
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 lendeer"  
  )
```



Relationships between numerical and categorical variables



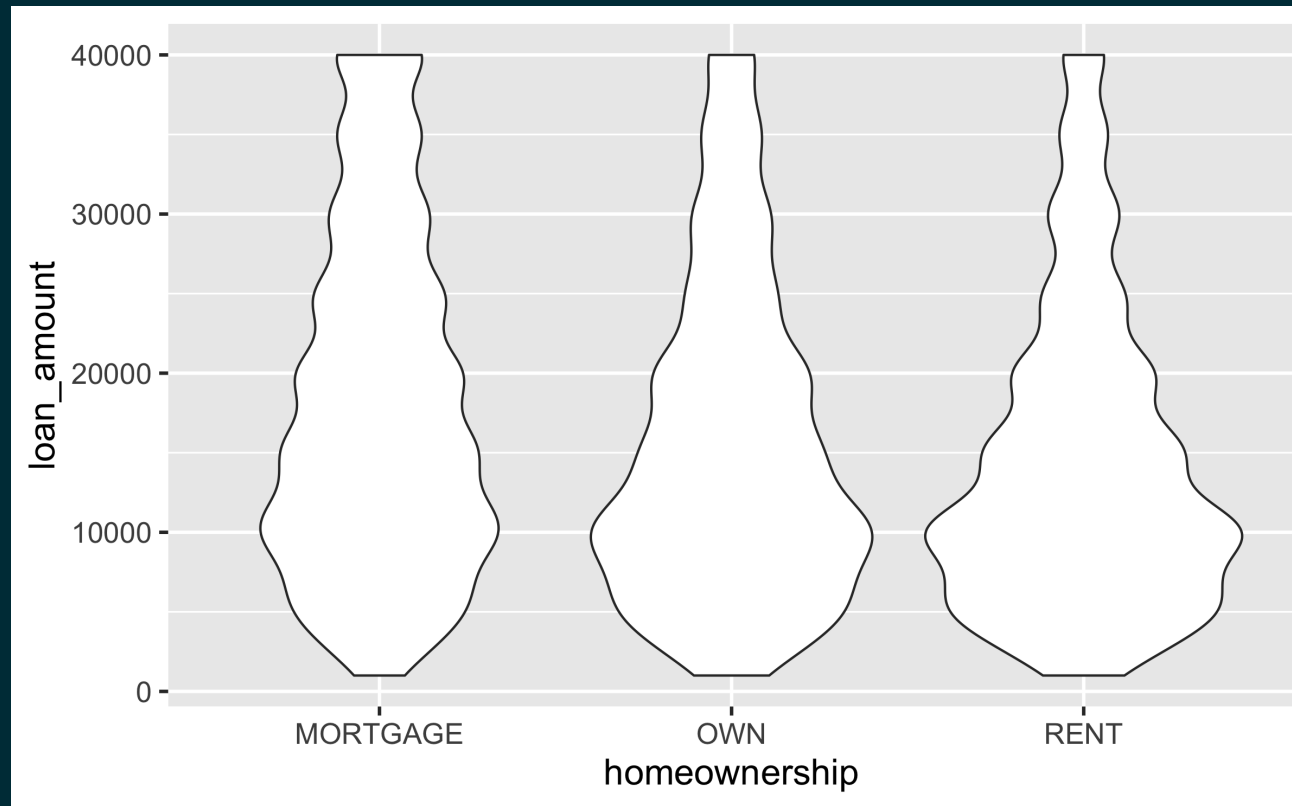
Already talked about...

- Colouring and faceting histograms and density plots
- Side-by-side box plots



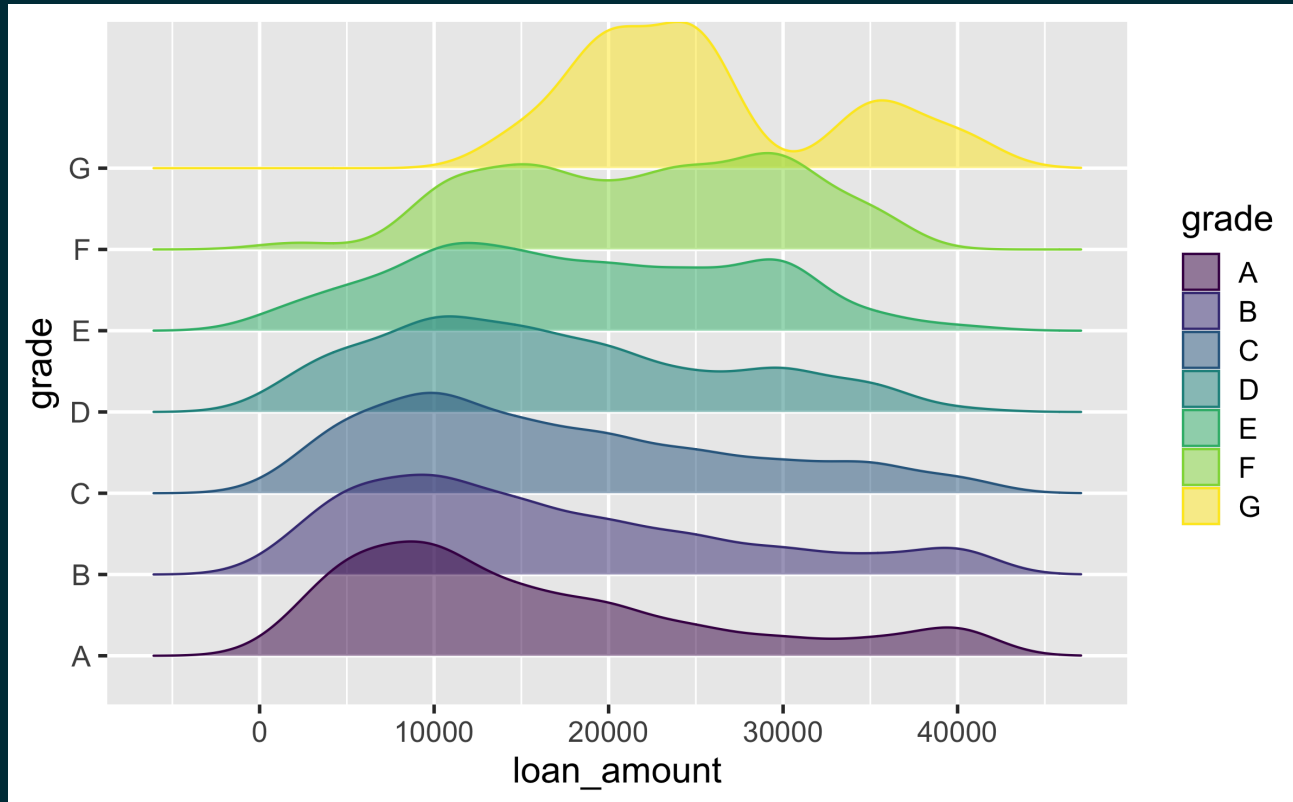
Violin plots

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



Ridge plots

```
library(ggribes)  
ggplot(loans, aes(x = loan_amount, y = grade, fill = grade, color = grade)) +  
  geom_density_ridges(alpha = 0.5)
```



My favorite ggridges plot

