Data visualization

Visualising categorical data

Matilde Trevisani

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** (or **nominal**) based on whether or not the levels have a natural ordering.

Data

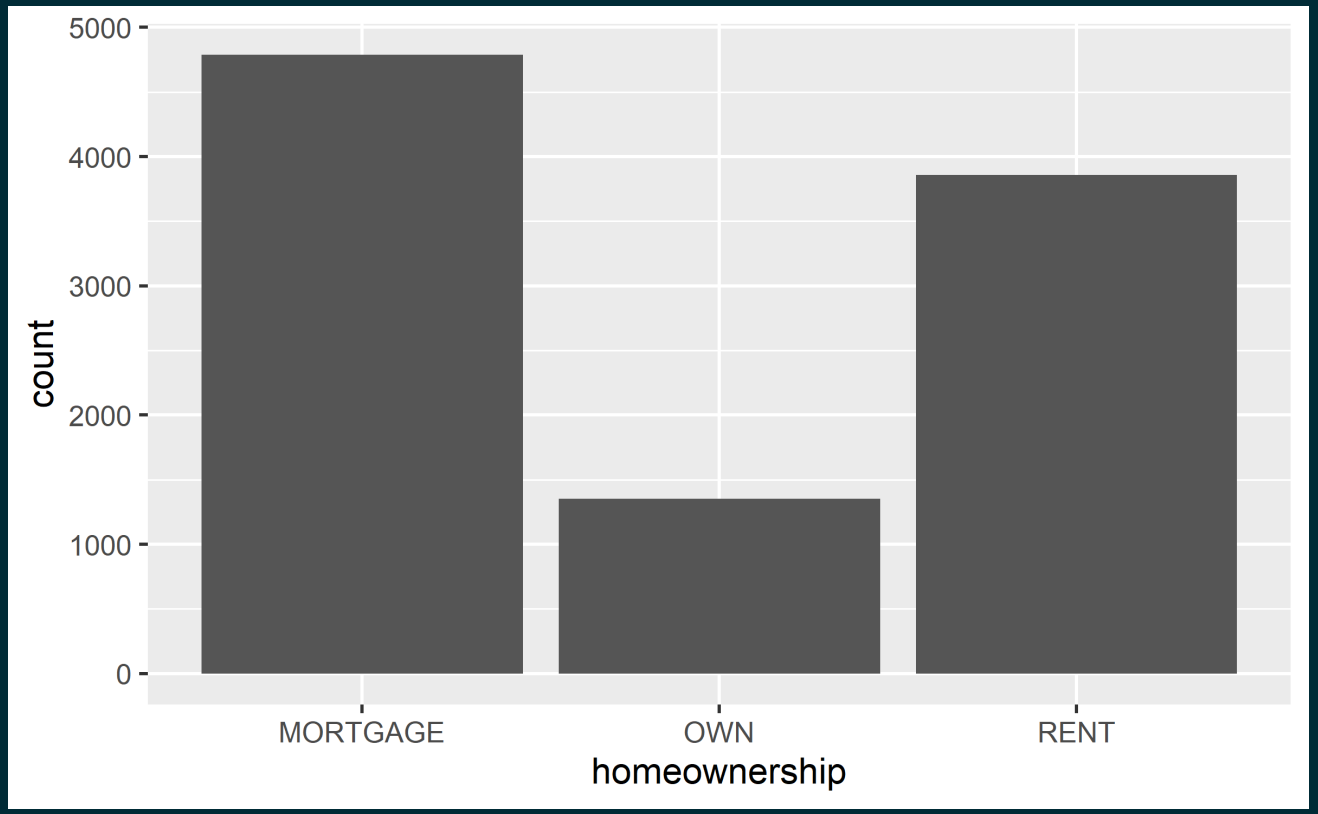
```r
library(openintro)
loans <- loans_full_schema %>%
  select(loan_amount, interest_rate, term, grade,
         state, annual_income, homeownership, debt_to_income)
glimpse(loans)
```

```
## 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

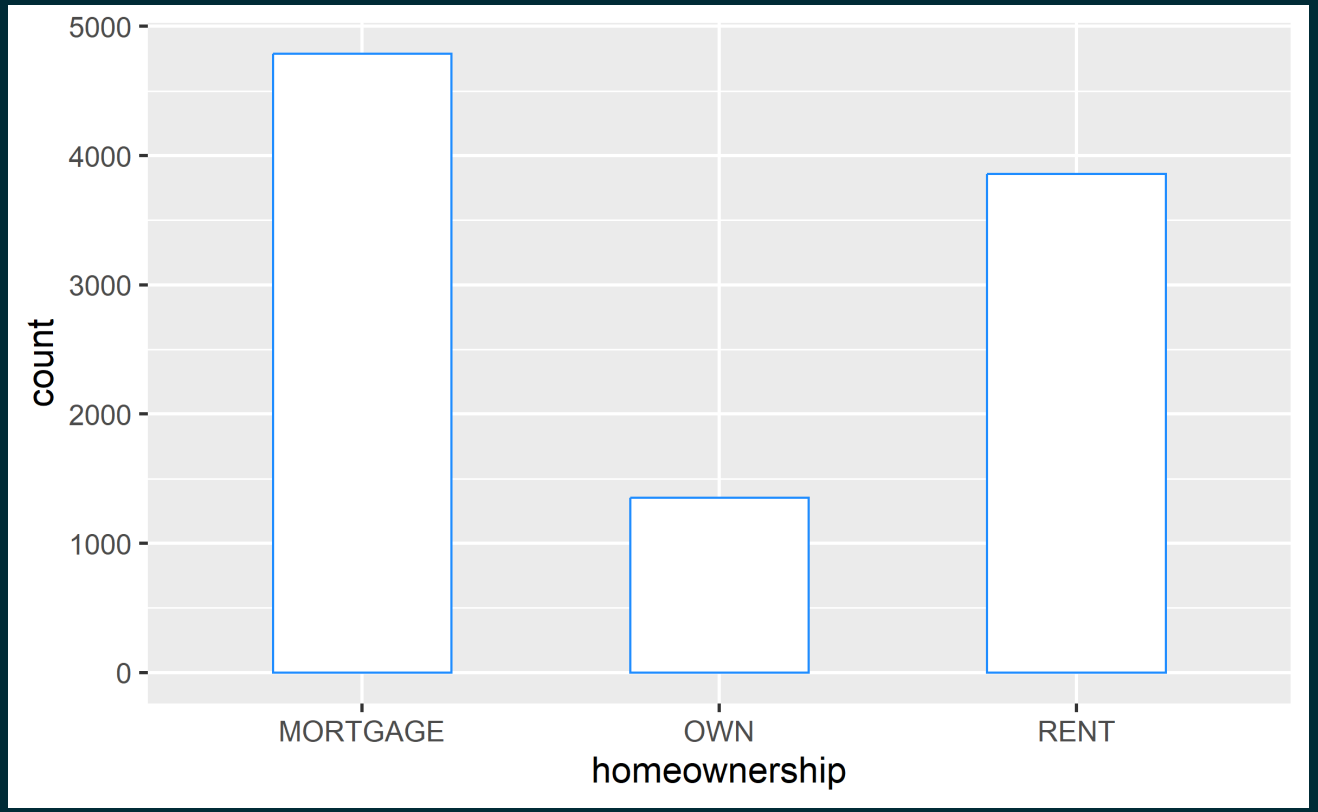
Bar plot

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



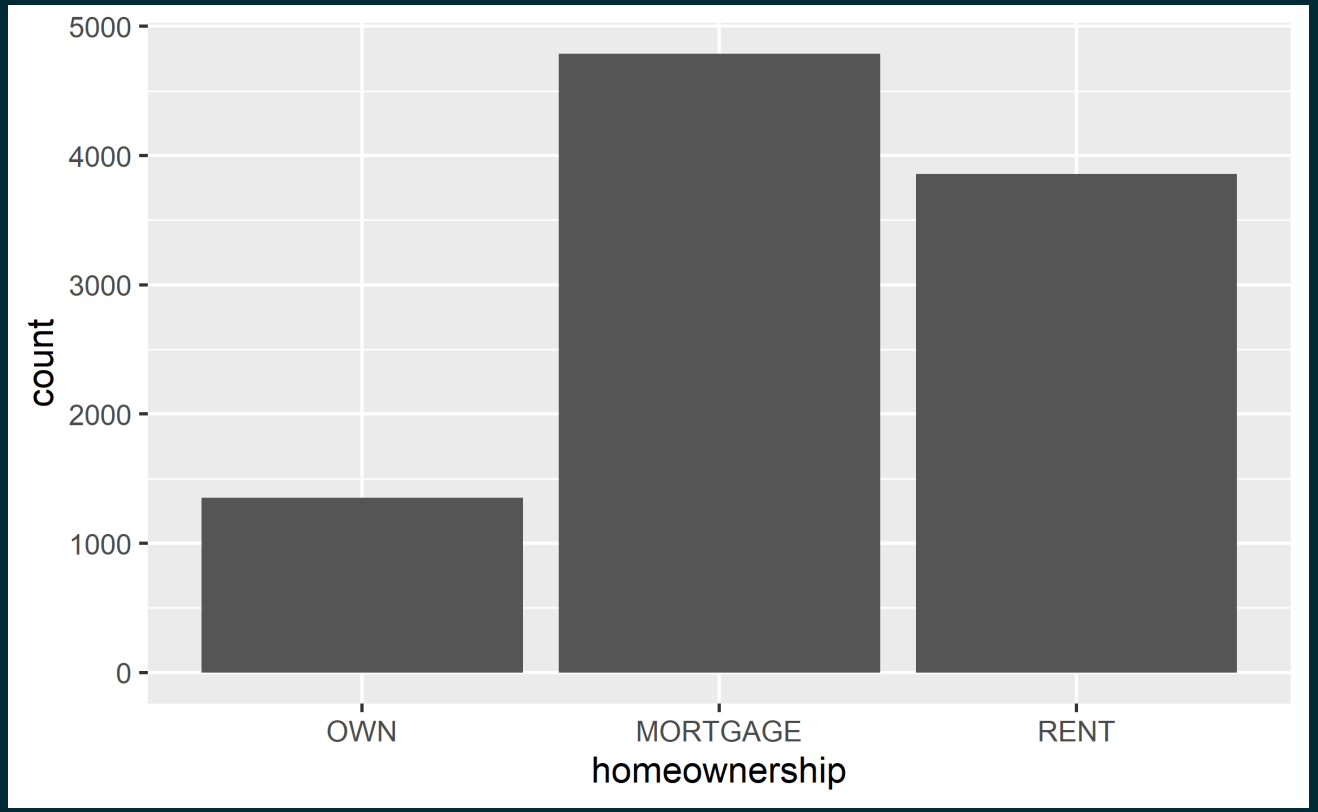
Setting a bar plot

```
ggplot(loans, aes(x = homeownership)) +
  geom_bar(width=0.5, col="dodgerblue", fill="white")
```



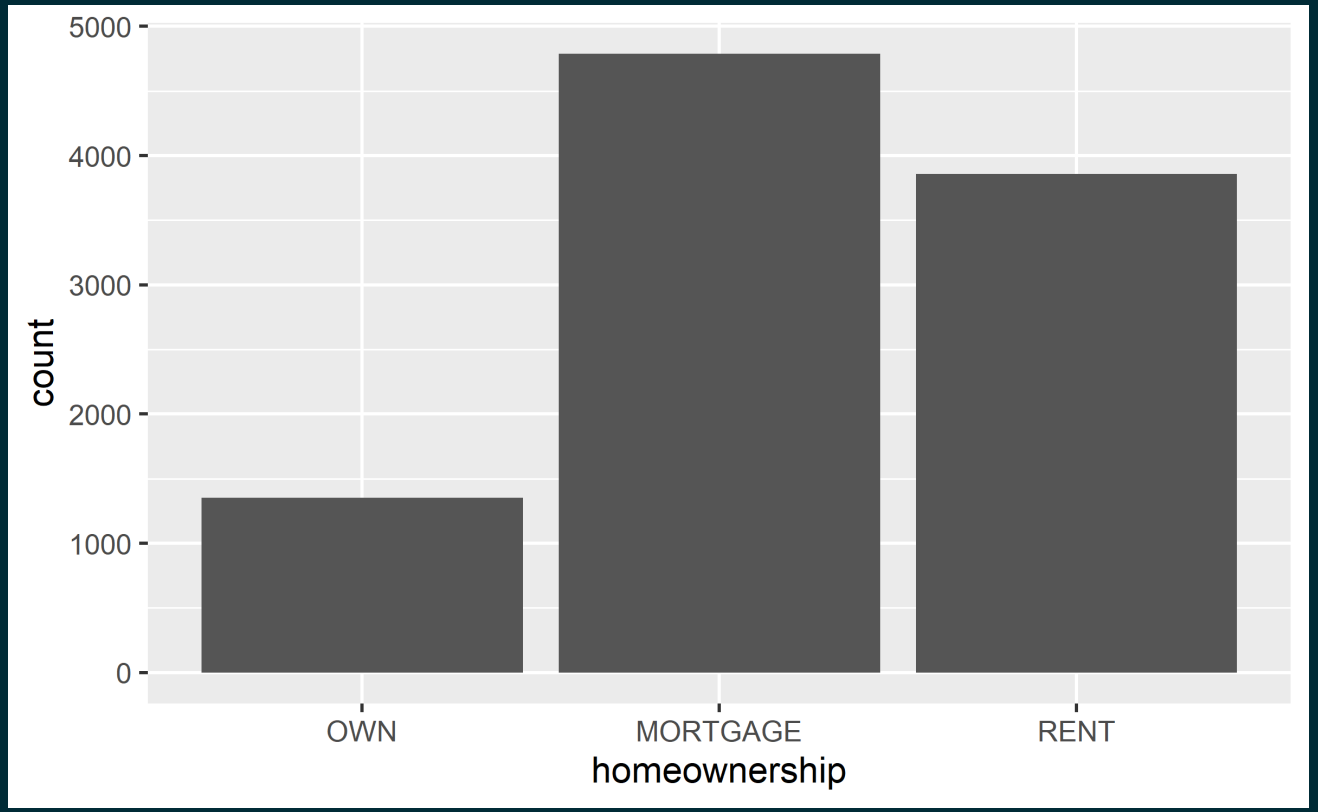
Changing the order of the items

```
ggplot(loans, aes(x = factor(homeownership, levels=c("OWN", "MORTGAGE", "RENT")))) +
  geom_bar() + labs(x="homeownership")
```



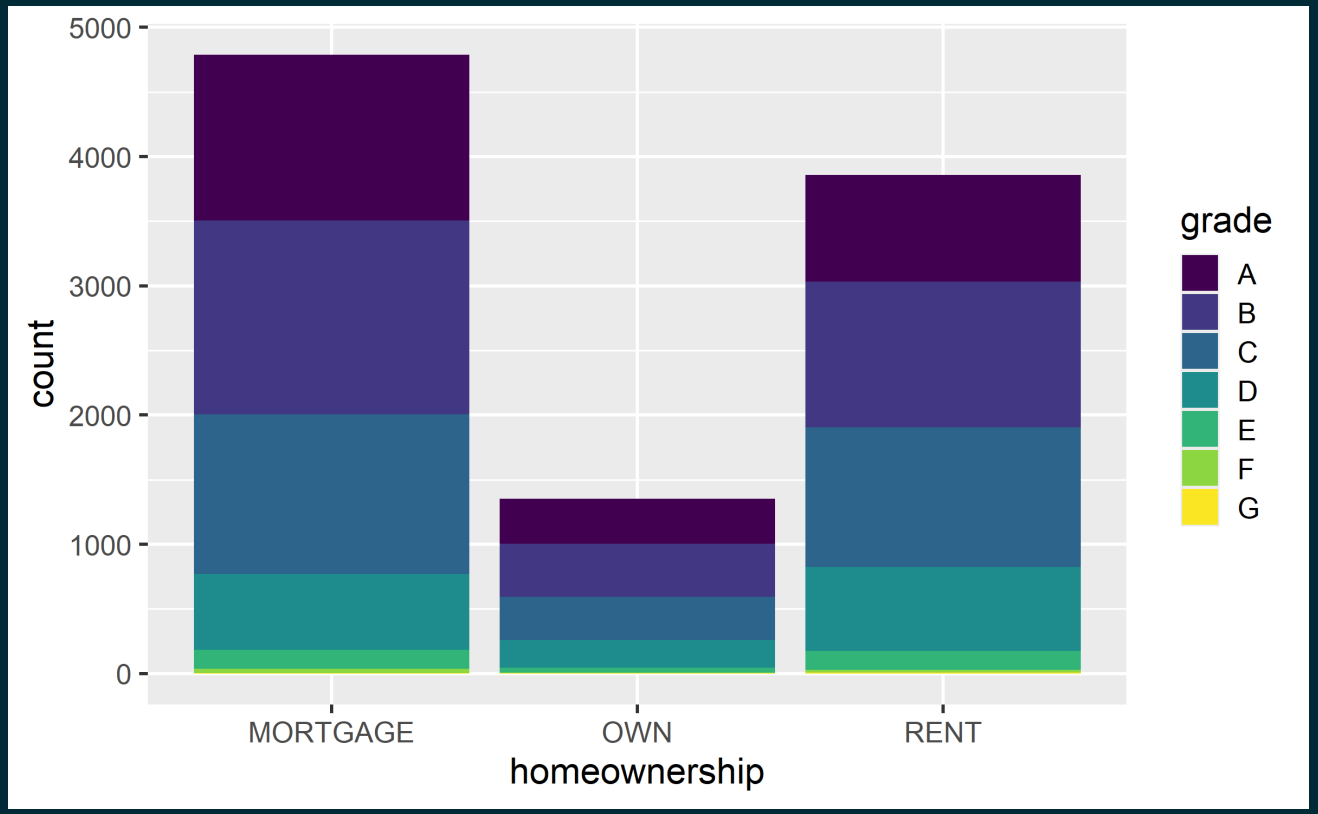
Changing the order of the items (2nd way)

```
ggplot(loans, aes(x = homeownership)) +
  geom_bar() + scale_x_discrete(limits=c("OWN", "MORTGAGE", "RENT"))
```



Component (aka stacked or segmented) bar plot

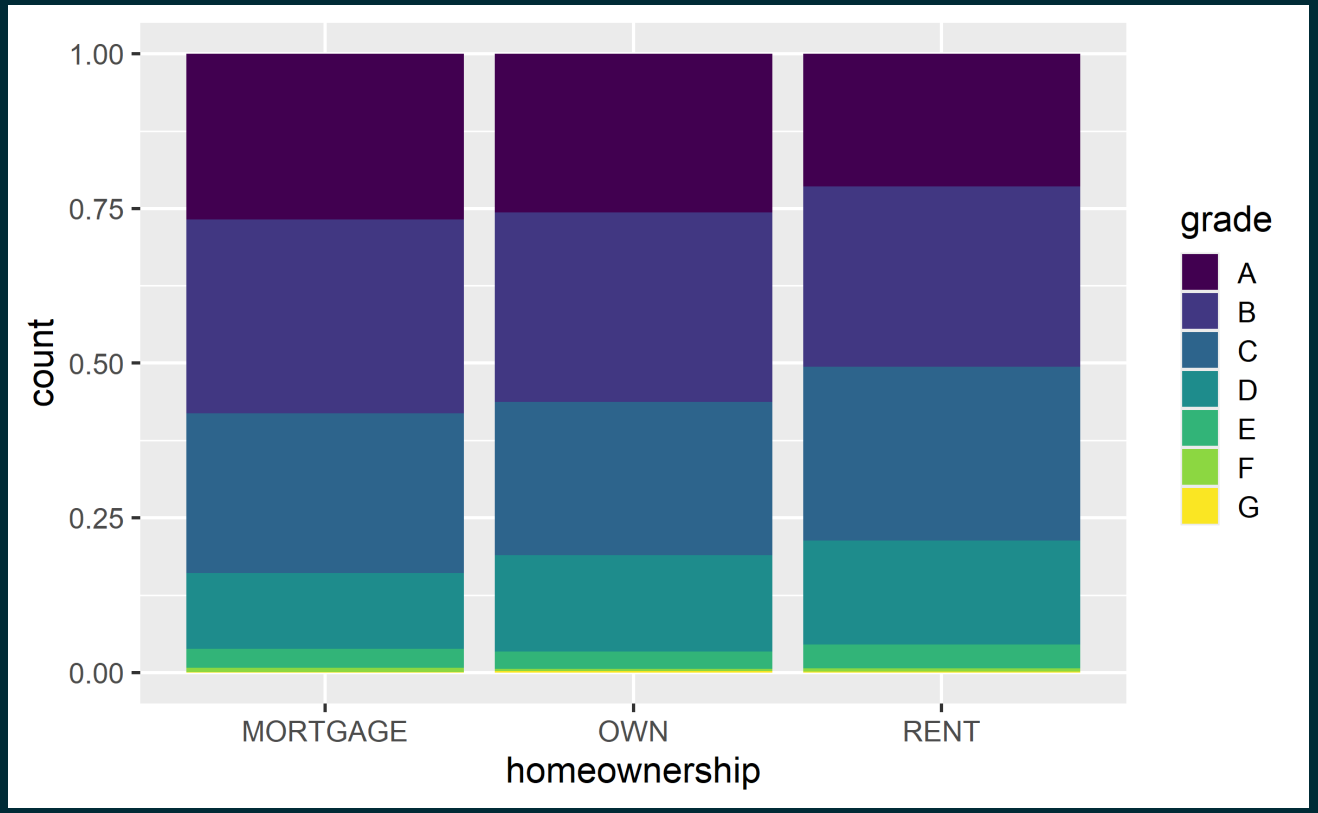
```
ggplot(loans, aes(x = homeownership,
                  fill = grade)) +
  geom_bar()
```



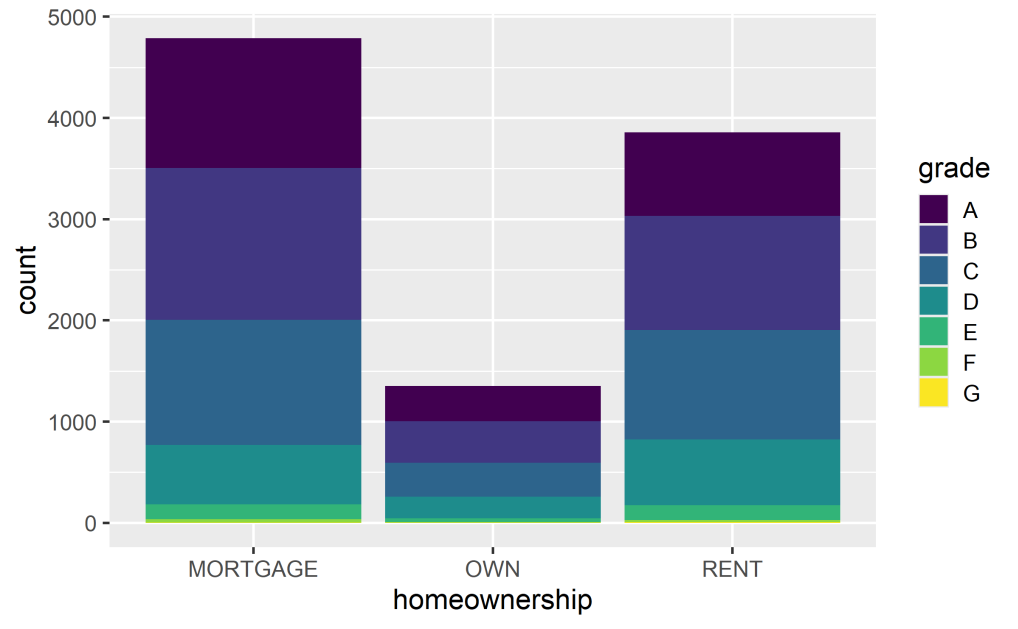
We stacked bars end-to-end.

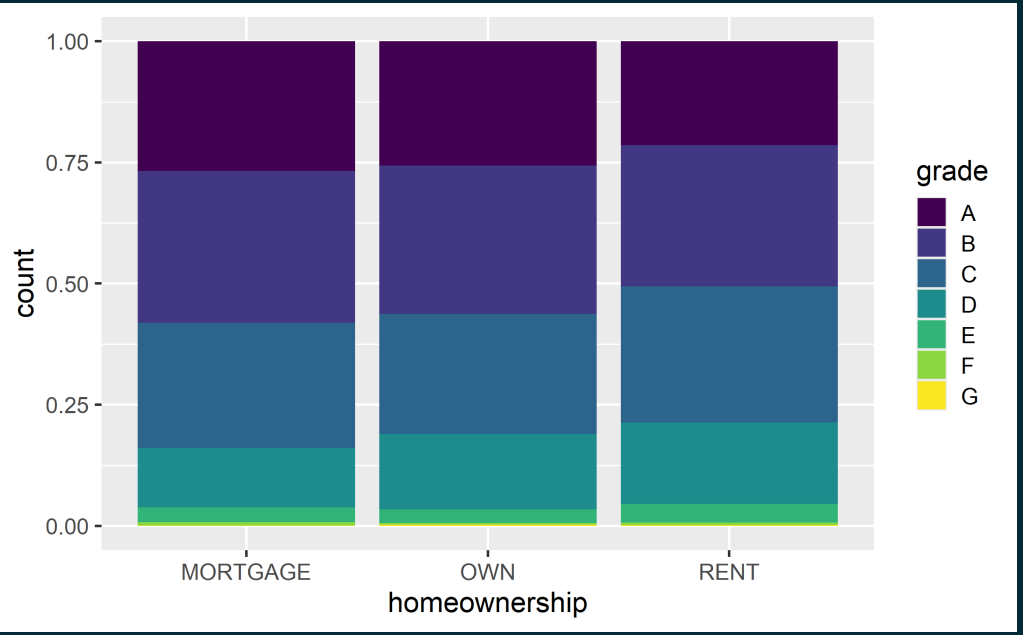
Component bar plot: conditional distribution

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



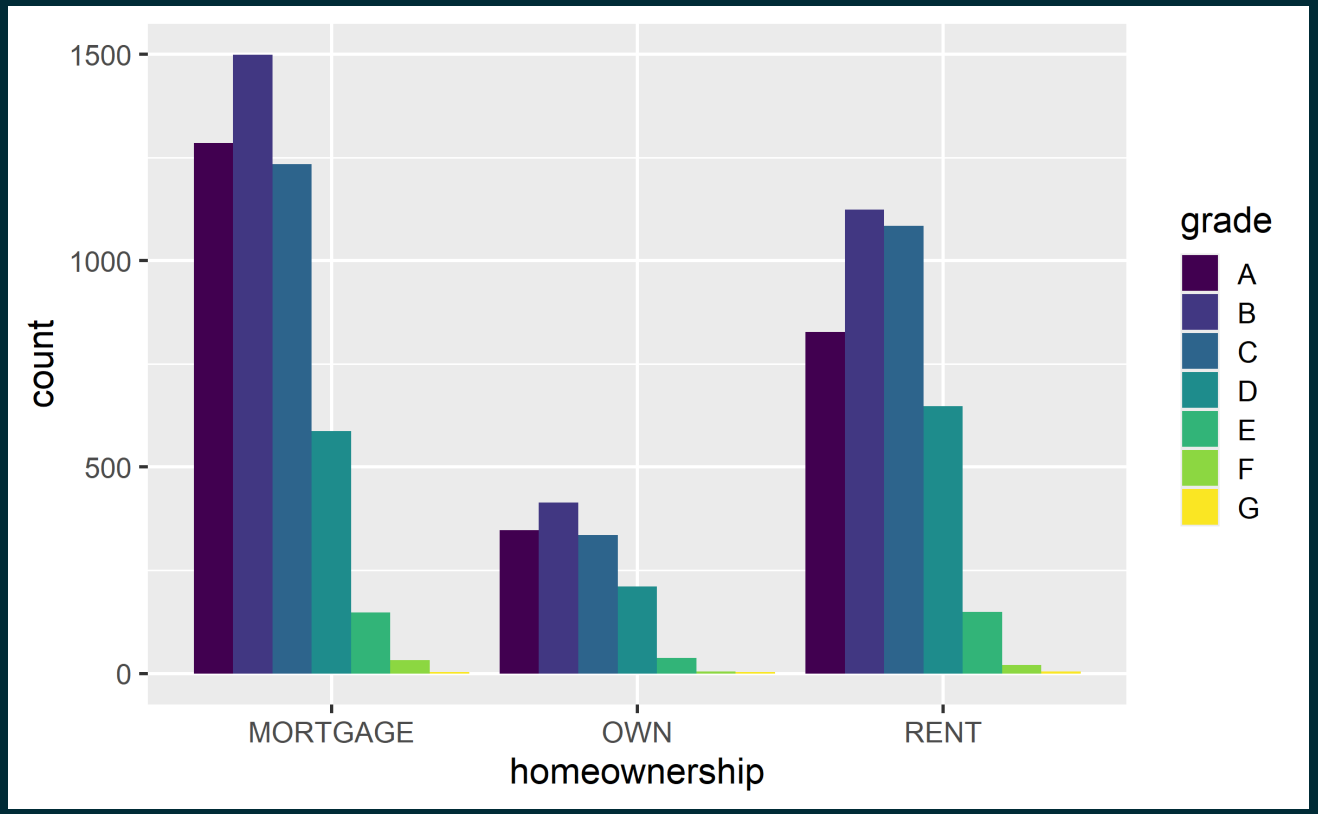
Which bar plot is a more useful representation for visualizing the relationship between homeownership and grade?





Clustered (aka grouped or multiseries) bar plot

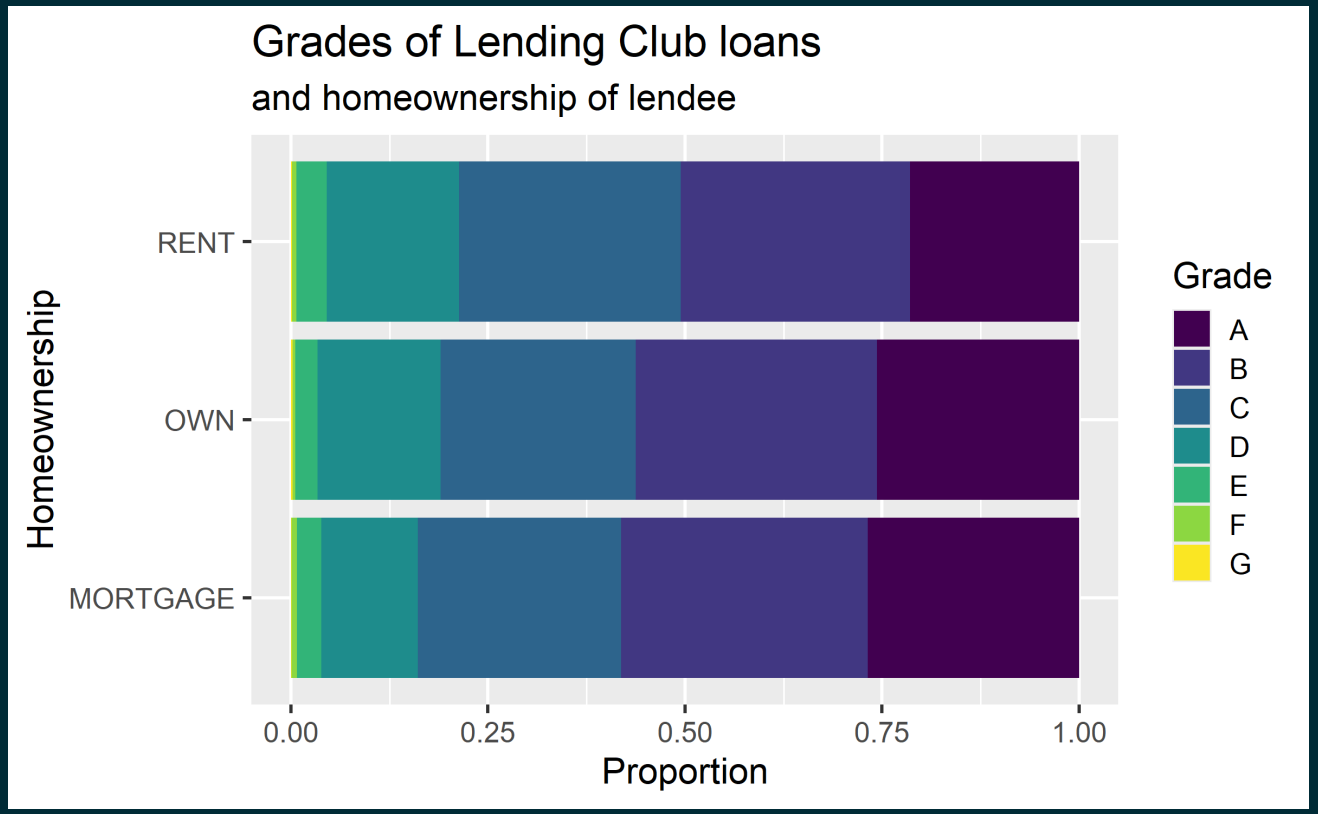
```
ggplot(loans, aes(x = homeownership,
                  fill = grade)) +
  geom_bar(position = "dodge")
```



We stacked bars side-by-side.

Customizing bar plots

Plot Code



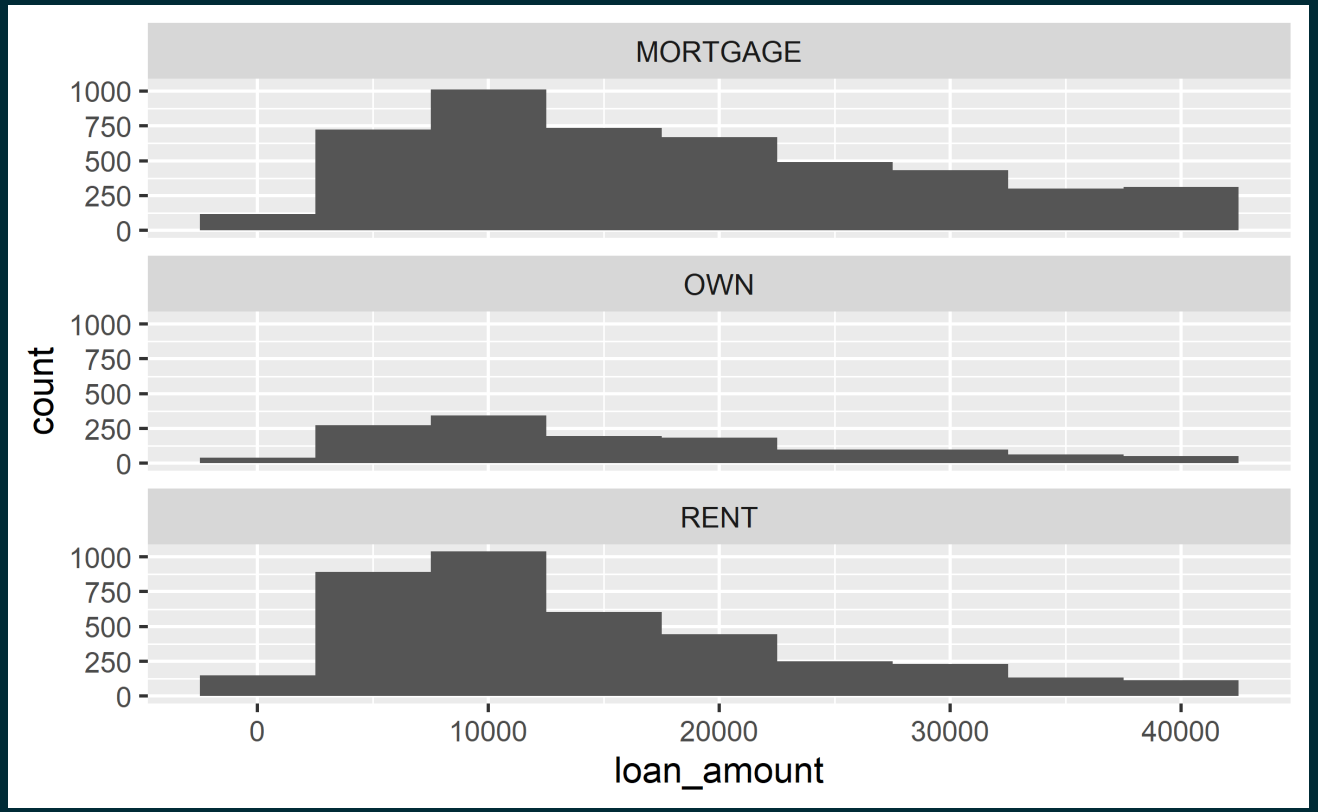
Relationships between numerical and categorical variables

Already talked about...

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

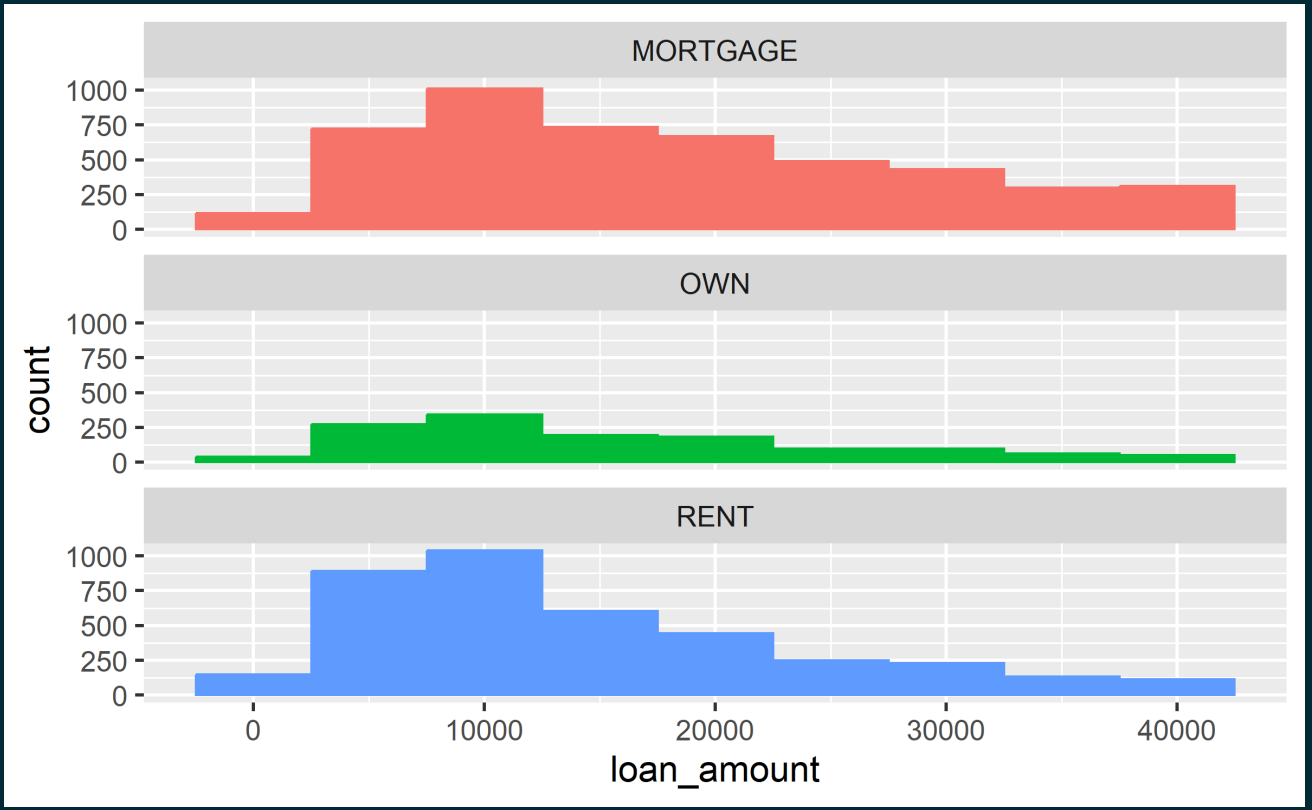
Faceting histograms

```
ggplot(loans, aes(x = loan_amount)) +
  geom_histogram(binwidth=5000) +
  facet_wrap(~ homeownership, nrow = 3)
```



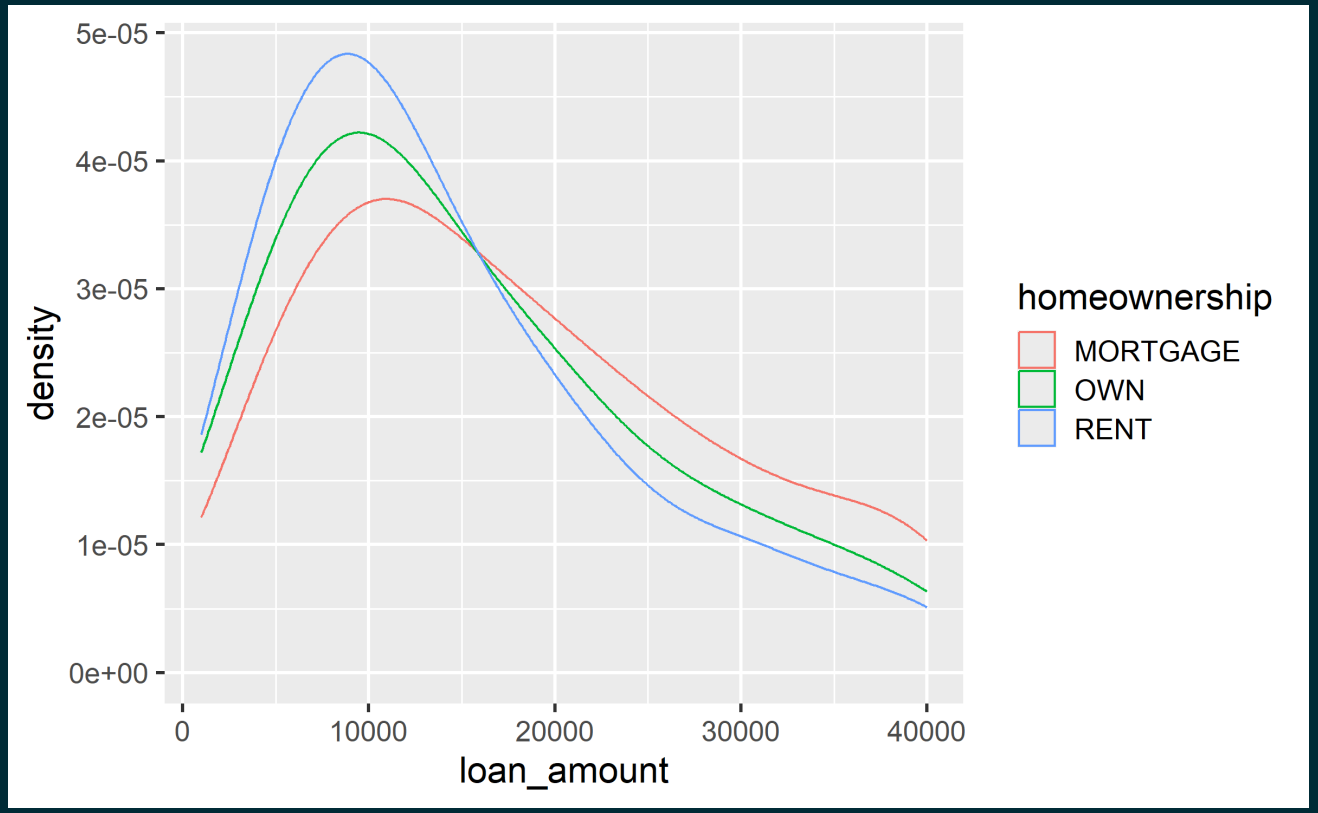
We can add colouring/filling (even if not necessary)

```
ggplot(loans, aes(x = loan_amount,
                  col = homeownership, fill = homeownership)) +
  geom_histogram(binwidth=5000) +
  facet_wrap(~ homeownership, nrow = 3) +
  guides(col = "none", fill = "none")
```



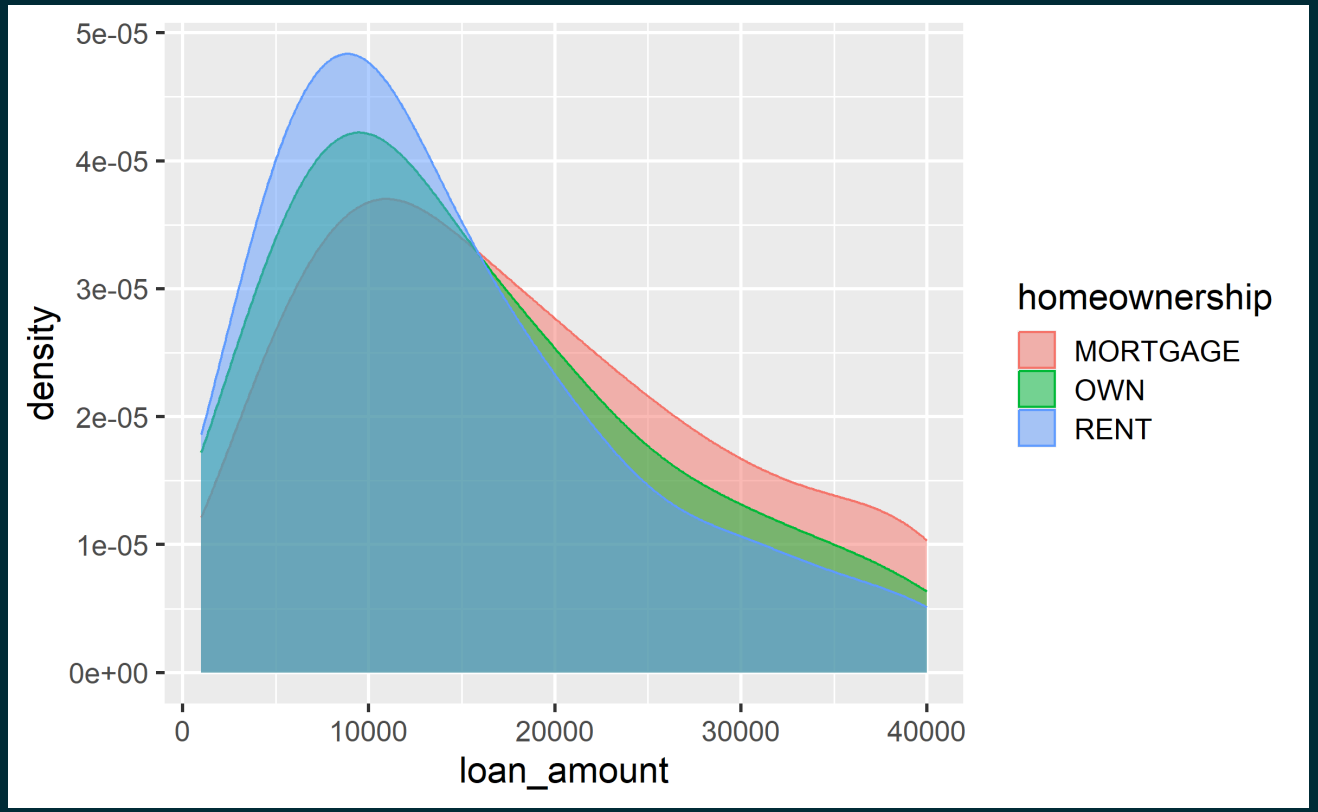
Colouring density plots

```
ggplot(loans, aes(x = loan_amount,
                  col = homeownership)) +
  geom_density(adjust=2, alpha=0.5)
```



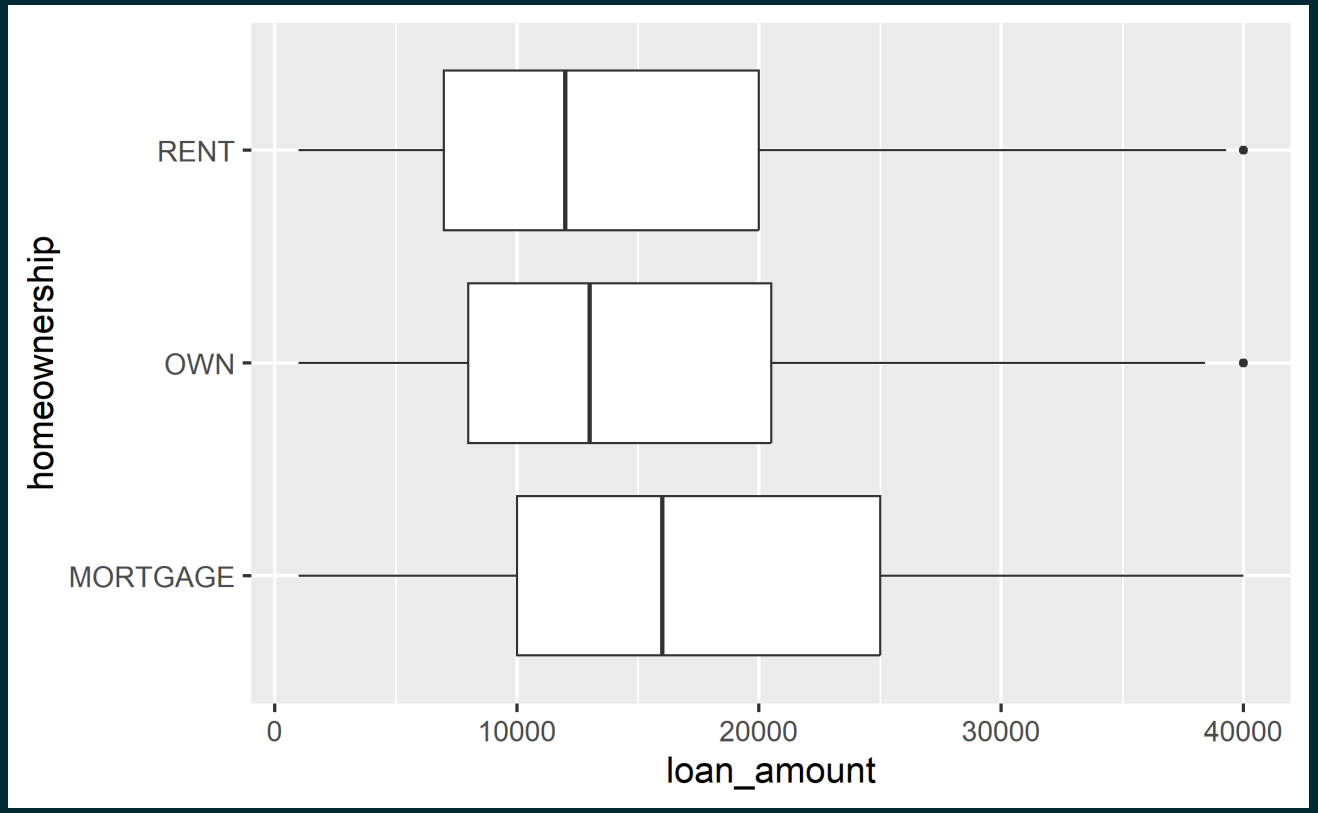
Colouring and / or filling

```
ggplot(loans, aes(x = loan_amount,
                  col = homeownership, fill = homeownership)) +
  geom_density(adjust=2, alpha=0.5)
```



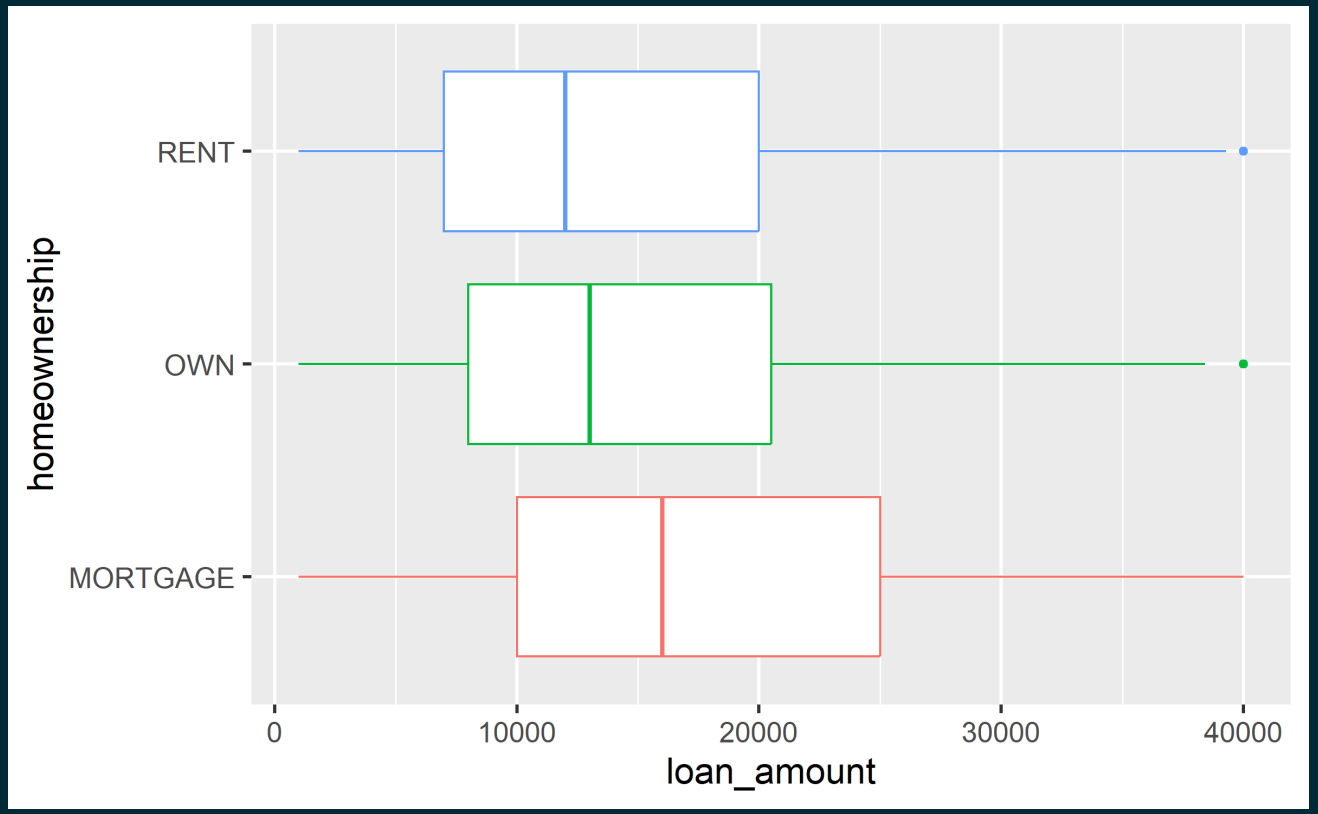
Side-by-side box plots

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



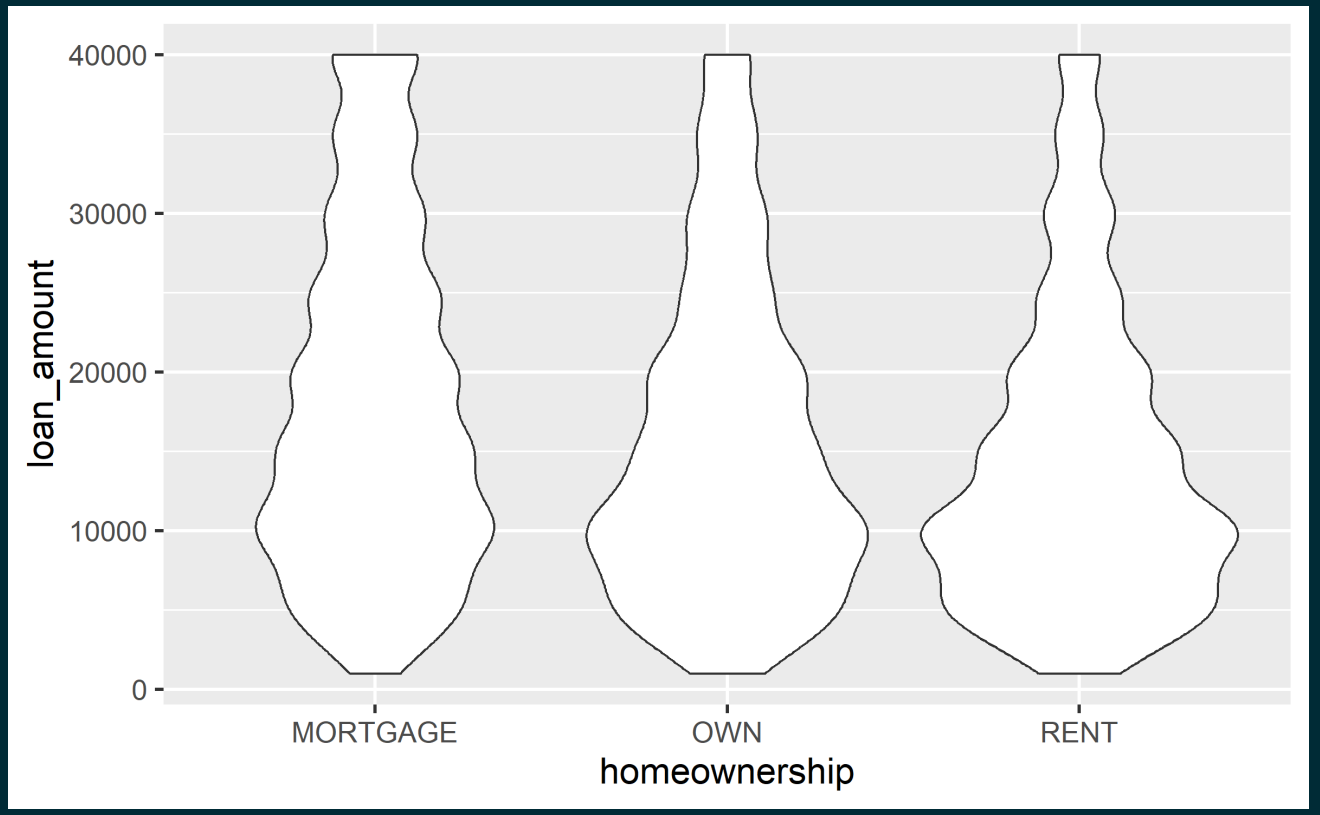
Again, we can add colouring/filling (not necessary)

```
ggplot(loans, aes(x = loan_amount,
                  y = homeownership,
                  col = homeownership)) +
  geom_boxplot() +
  guides(col = "none")
```



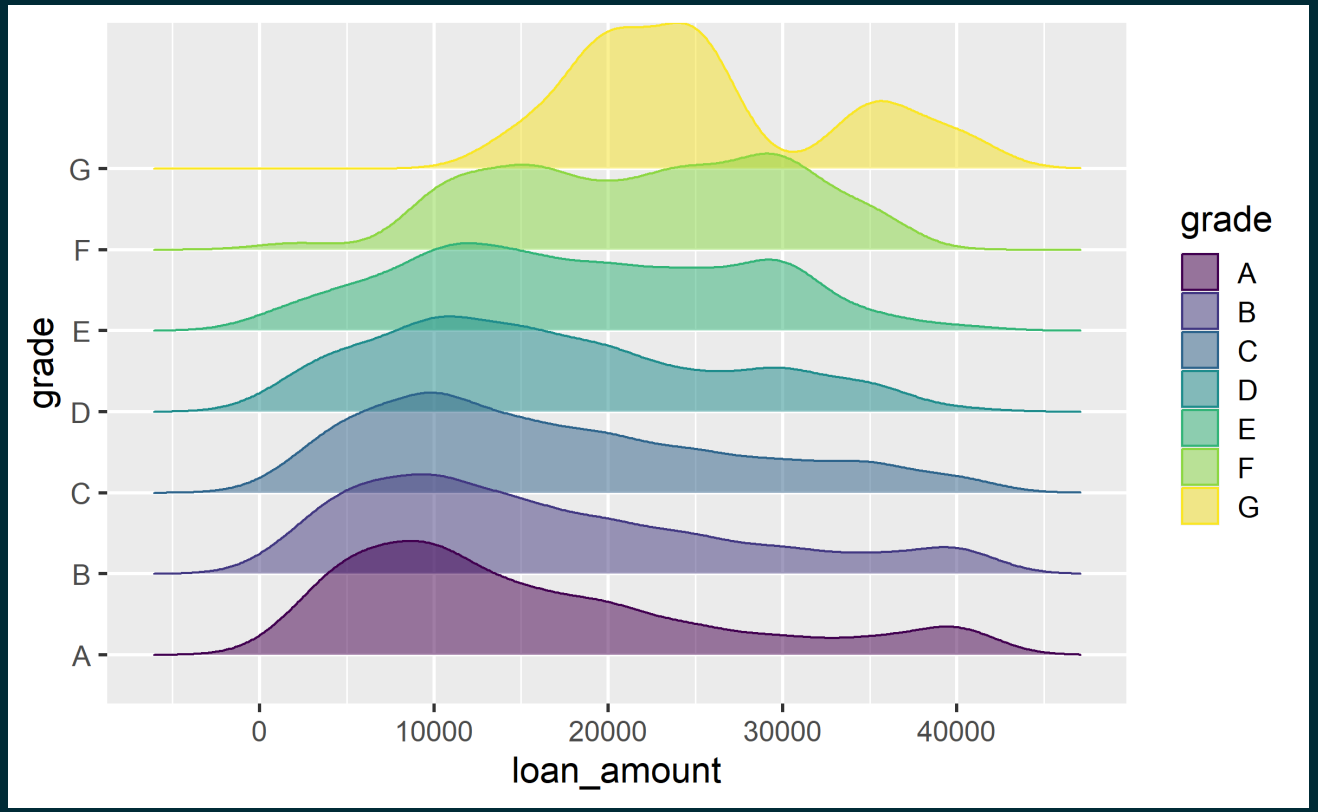
Violin plots

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



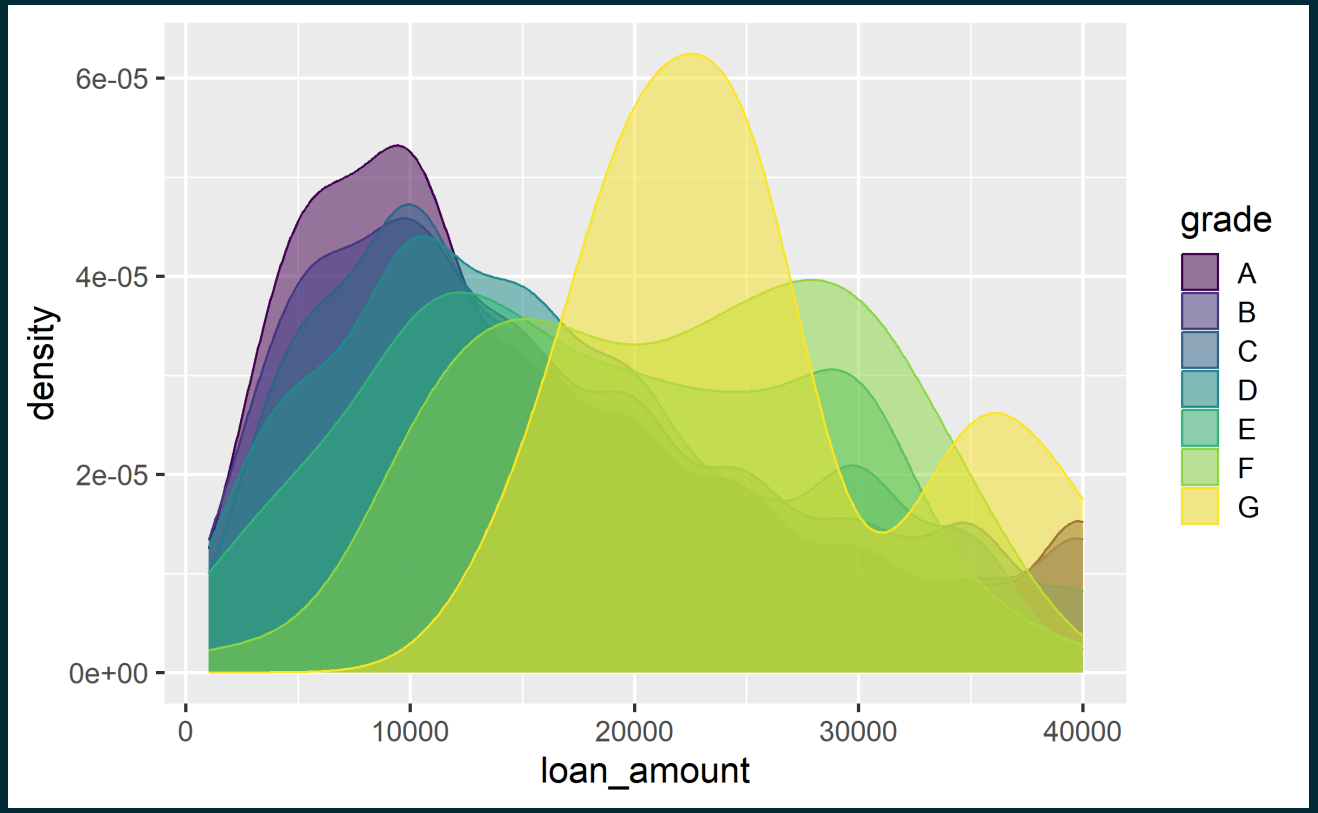
Ridge plots (when many categories)

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



Density plots are less readable

```
ggplot(loans, aes(x = loan_amount, fill = grade, color = grade)) +
  geom_density(alpha = 0.5)
```



Boxplots are the most used alternative (though much more synthetic)

```
ggplot(loans, aes(x = loan_amount,
                  y = grade,
                  col = grade, fill = grade)) +
  geom_boxplot() +
  guides(col = "none", fill = "none")
```