R Notebook

Code

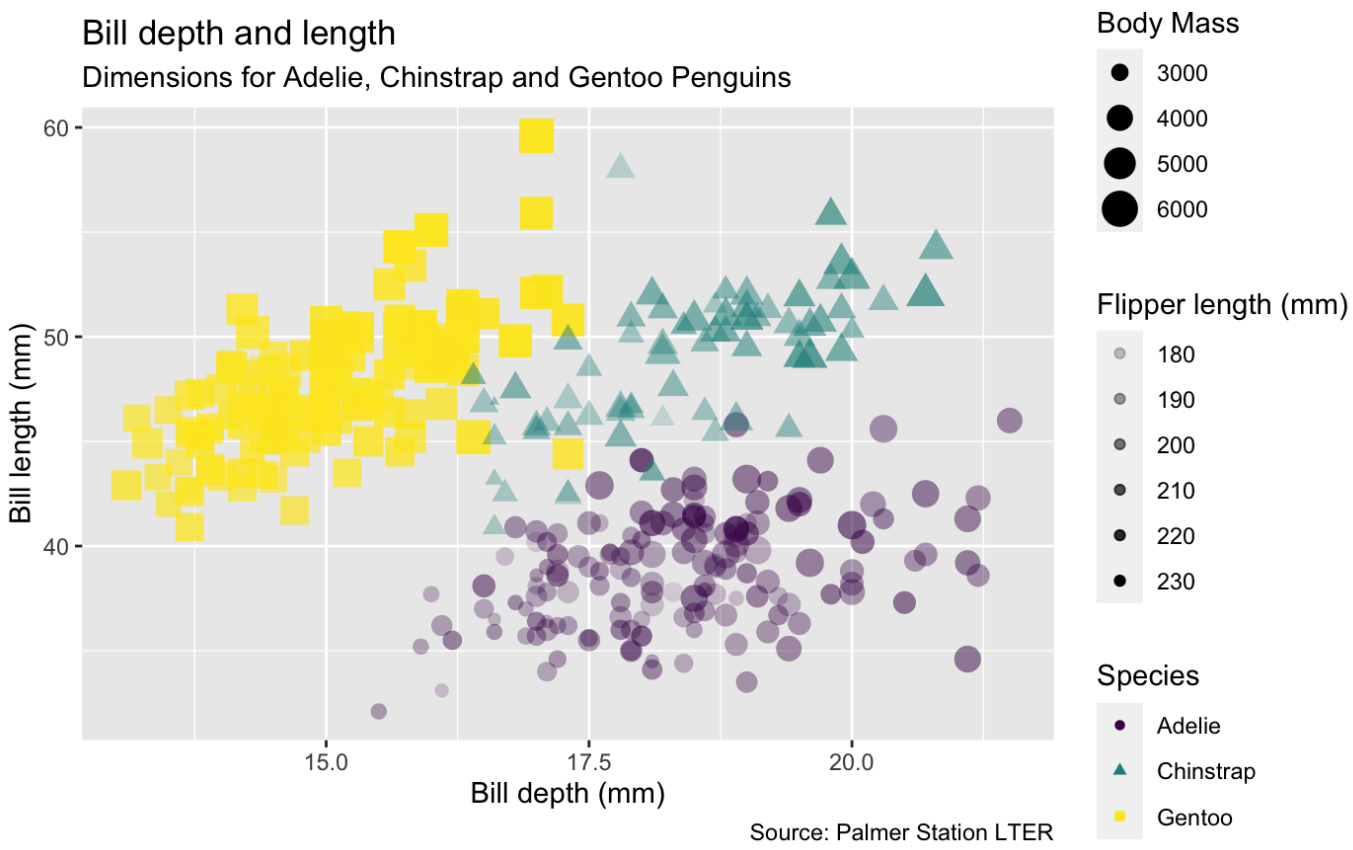
Hide

```
library(tidyverse)
library(palmerpenguins)
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie,…
$ island            <fct> Torgersen, Torgersen, Torgersen…
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.…
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.…
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 18…
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 365…
$ sex               <fct> male, female, female, NA, femal…
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2…
```

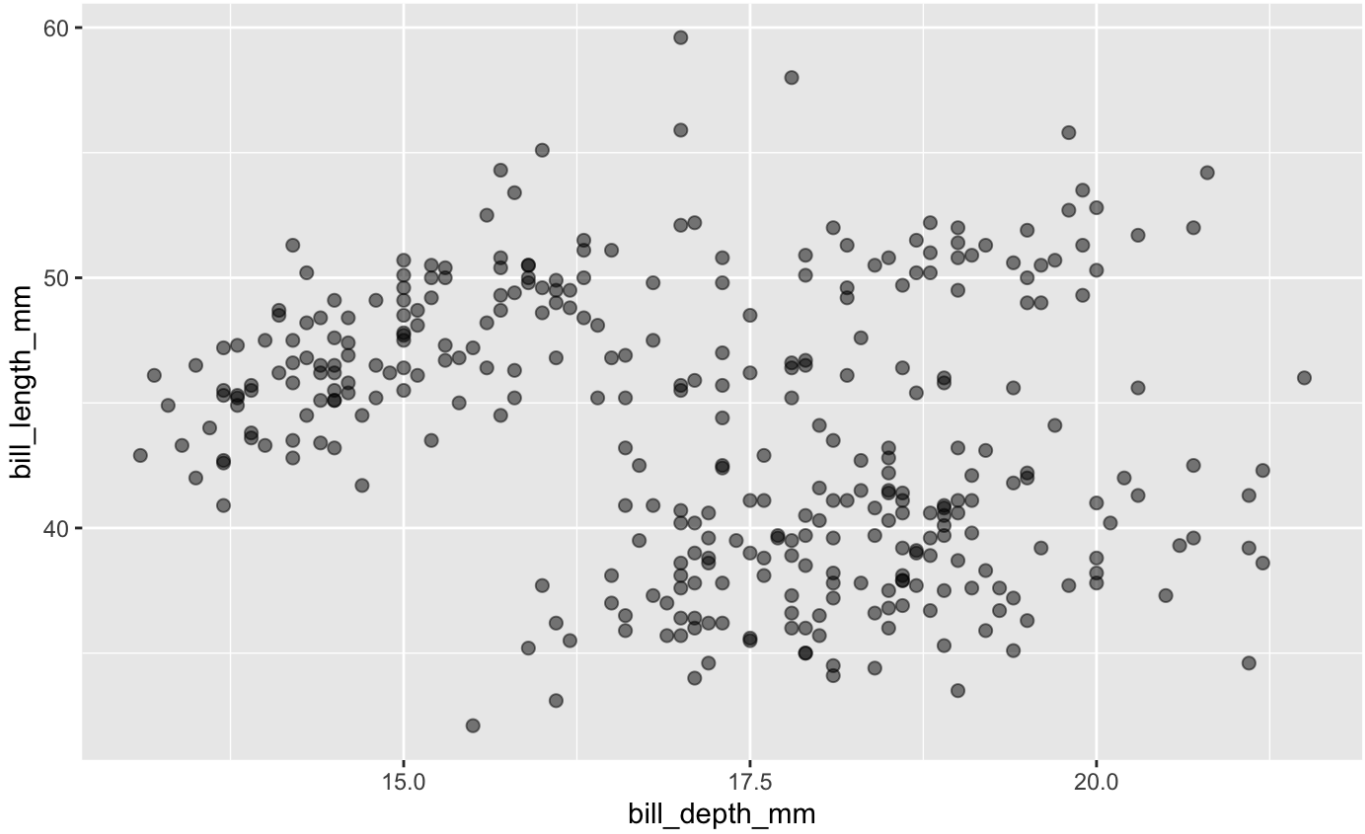
Hide

```
#Mapping: Determines the size, alpha, etc. of points based on the values of a variabl
e in the data --> goes into aes(): for eg. the higher the flipper length the higher t
he alpha.

ggplot(data = penguins) +
  aes(x = bill_depth_mm,
      y = bill_length_mm,
      colour = species,
      shape = species,
      size = body_mass_g,
      alpha = flipper_length_mm) +

  geom_point() +
  labs(title = "Bill depth and length",
       subtitle = "Dimensions for Adelie, Chinstrap and Gentoo Penguins",
       x = "Bill depth (mm)",
       y = "Bill length (mm)",
       colour = "Species",
       shape = "Species",
       size = "Body Mass",
       alpha = "Flipper length (mm)",
       caption = "Source: Palmer Station LTER") +
  scale_colour_viridis_d()
```



Hide

```
#Setting: Determines the size, alpha, etc. of points NOT based on the values of a var
iable in the data --> goes into geom_*()

ggplot (data = penguins) +
  aes( x = bill_depth_mm,
       y = bill_length_mm) +
  geom_point( size = 2,
              alpha = 0.5)
```



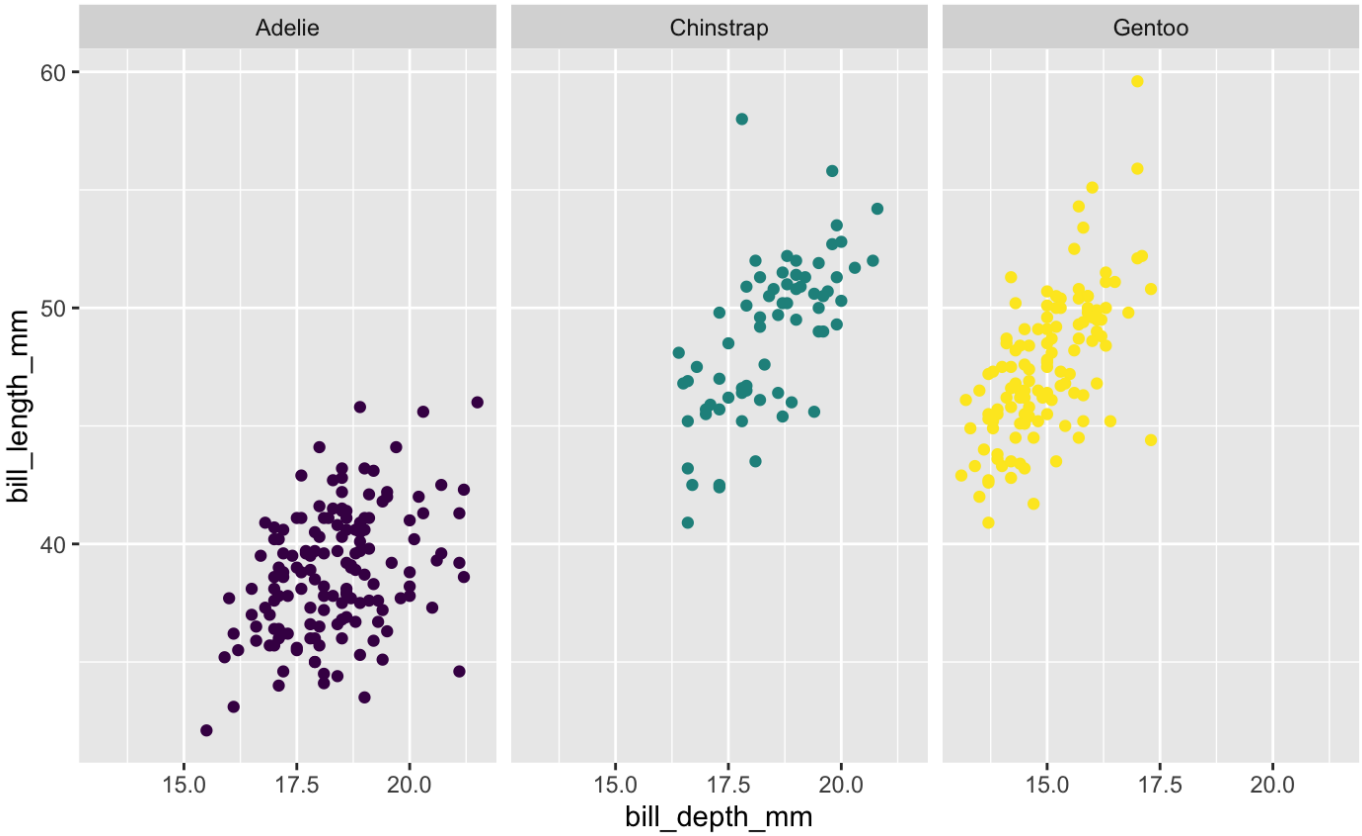
Hide

```
#Faceting: Smaller plots that display different subsets of the data. Useful for explo
ring conditional relationships and large data.

ggplot (data = penguins) +
  aes( x = bill_depth_mm,
       y = bill_length_mxm) +
  geom_point() +
  facet_grid(species ~ island) +

ggplot (data = penguins) +
  aes( x = bill_depth_mm,
       y = bill_length_mm) +
  geom_point() +
  facet_grid(species ~ sex)
```

```
Error in `ggplot_add()`:
! Can't add `ggplot(data = penguins)` to a <ggplot>
  object.
Backtrace:
 1. ggplot2:::`+.gg`(...)
 2. ggplot2:::add_ggplot(e1, e2, e2name)
 4. ggplot2:::ggplot_add.default(object, p, objectname)
```

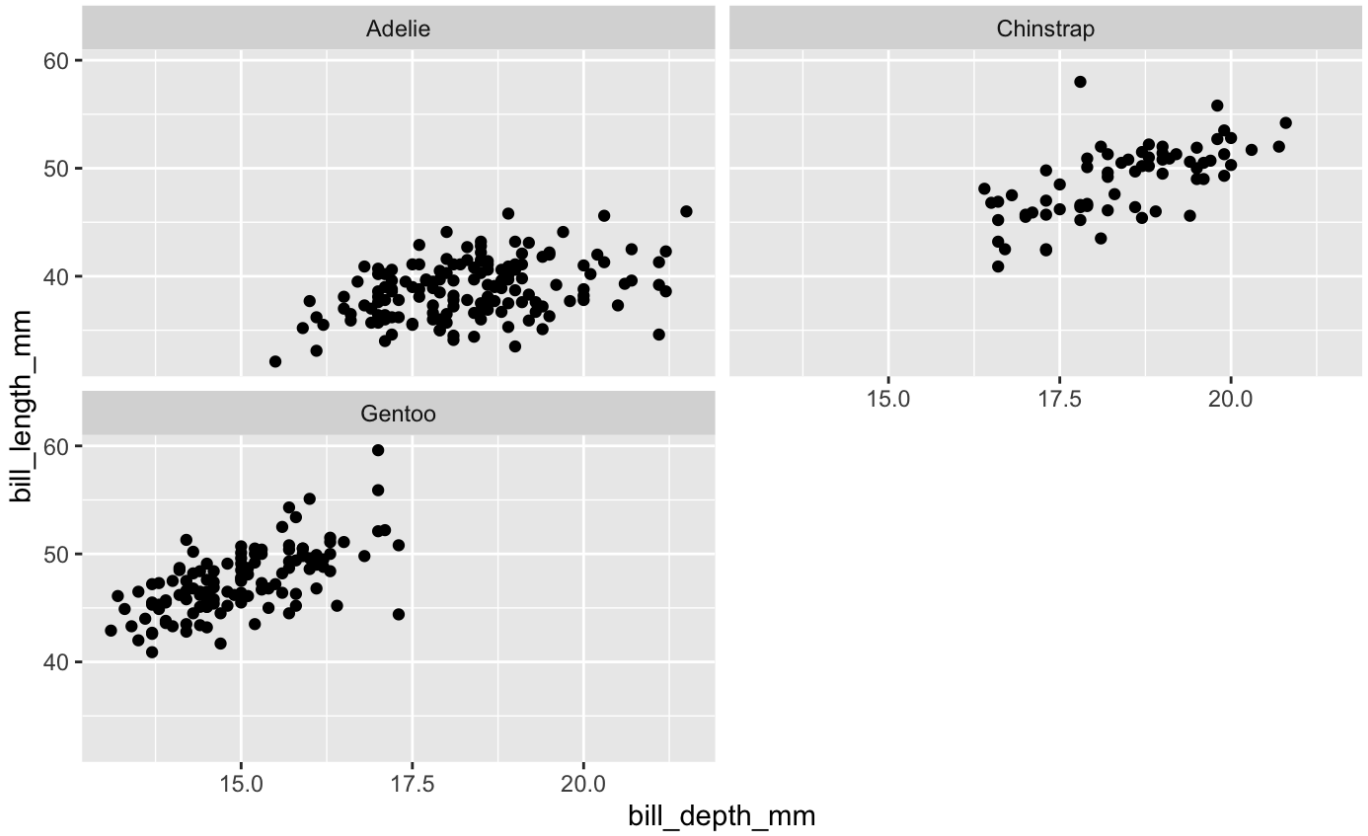
Hide

```
#Faceting: Smaller plots that display different subsets of the data. Useful for explo
ring conditional relationships and large data.
#Faceting for single variable to sort data by:
ggplot (data = penguins) +
  aes( x = bill_depth_mm,
       y = bill_length_mm,
       color = species) +
  geom_point() +
  facet_wrap( ~ species) +
  scale_colour_viridis_d() +
  guides(color = "none")
```



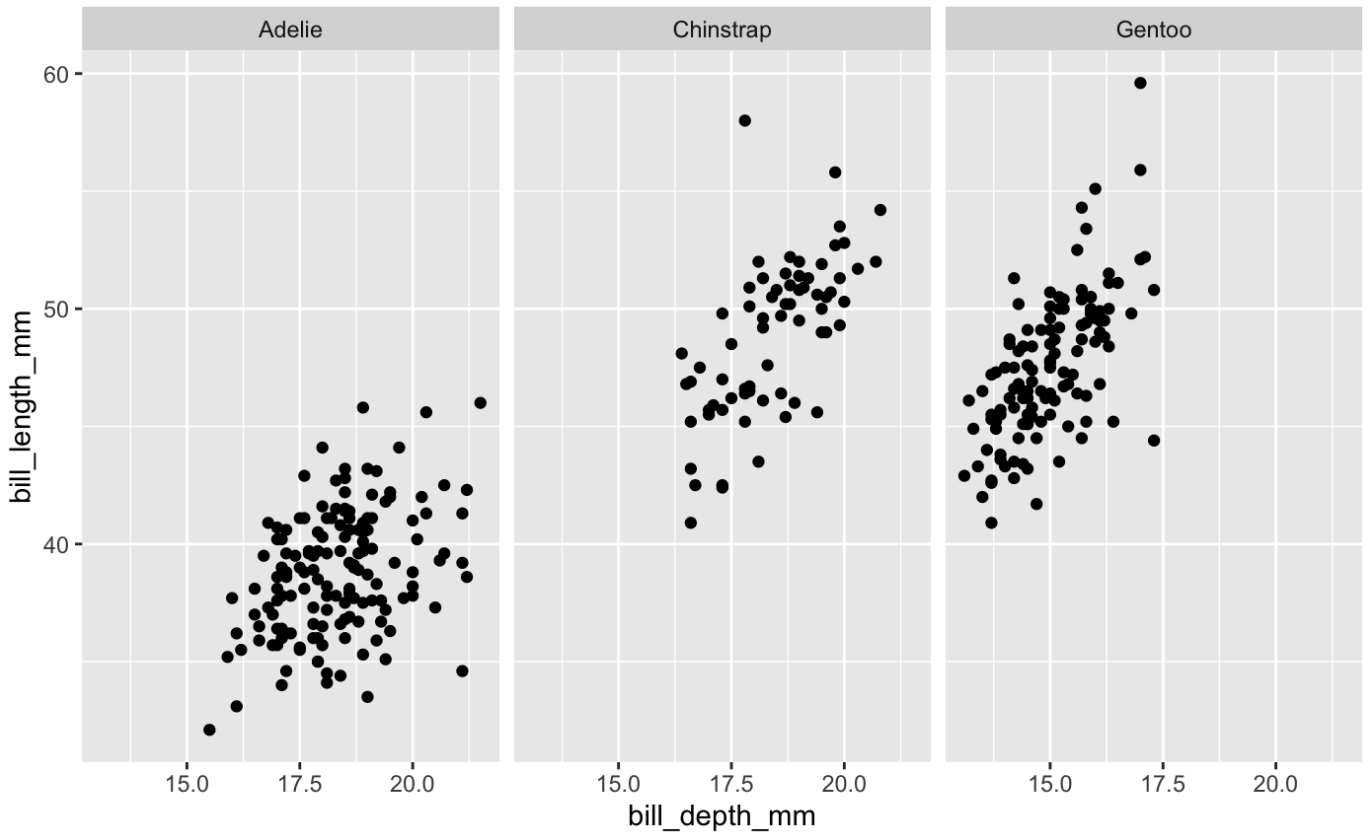
Hide

```
ggplot (data = penguins) +
  aes( x = bill_depth_mm,
       y = bill_length_mm) +
  geom_point() +
  facet_wrap( ~ species,
              ncol = 2)
```



Hide

```
ggplot (data = penguins) +
  aes( x = bill_depth_mm,
       y = bill_length_mm) +
  geom_point() +
  facet_grid(. ~ species) #Identical to a basic facet_wrap
```



Hide

```
library(openintro)
```

```
Loading required package: airports
Loading required package: cherryblossom
Loading required package: usdata
```

Hide

```
glimpse(loans_full_schema)
```

```
Rows: 10,000
Columns: 55
$ emp_title                          <chr> "global config en…
$ emp_length                         <dbl> 3, 10, 3, 1, 10, …
$ state                              <fct> NJ, HI, WI, PA, C…
$ homeownership                      <fct> MORTGAGE, RENT, R…
$ annual_income                      <dbl> 90000, 40000, 400…
$ verified_income                    <fct> Verified, Not Ver…
$ debt_to_income                     <dbl> 18.01, 5.04, 21.1…
$ annual_income_joint                <dbl> NA, NA, NA, NA, 5…
$ verification_income_joint          <fct> , , , , Verified,…
$ debt_to_income_joint               <dbl> NA, NA, NA, NA, 3…
$ delinq_2y                          <int> 0, 0, 0, 0, 0, 1,…
$ months_since_last_delinq           <int> 38, NA, 28, NA, N…
$ earliest_credit_line               <dbl> 2001, 1996, 2006,…
$ inquiries_last_12m                 <int> 6, 1, 4, 0, 7, 6,…
$ total_credit_lines                 <int> 28, 30, 31, 4, 22…
$ open_credit_lines                  <int> 10, 14, 10, 4, 16…
$ total_credit_limit                 <int> 70795, 28800, 241…
$ total_credit_utilized              <int> 38767, 4321, 1600…
$ num_collections_last_12m           <int> 0, 0, 0, 0, 0, 0,…
$ num_historical_failed_to_pay       <int> 0, 1, 0, 1, 0, 0,…
$ months_since_90d_late              <int> 38, NA, 28, NA, N…
$ current_accounts_delinq            <int> 0, 0, 0, 0, 0, 0,…
$ total_collection_amount_ever       <int> 1250, 0, 432, 0, …
$ current_installment_accounts       <int> 2, 0, 1, 1, 1, 0,…
$ accounts_opened_24m                <int> 5, 11, 13, 1, 6, …
$ months_since_last_credit_inquiry <int> 5, 8, 7, 15, 4, 5…
$ num_satisfactory_accounts          <int> 10, 14, 10, 4, 16…
$ num_accounts_120d_past_due         <int> 0, 0, 0, 0, 0, 0,…
$ num_accounts_30d_past_due          <int> 0, 0, 0, 0, 0, 0,…
$ num_active_debit_accounts          <int> 2, 3, 3, 2, 10, 1…
$ total_debit_limit                  <int> 11100, 16500, 430…
$ num_total_cc_accounts              <int> 14, 24, 14, 3, 20…
$ num_open_cc_accounts               <int> 8, 14, 8, 3, 15, …
$ num_cc_carrying_balance            <int> 6, 4, 6, 2, 13, 5…
$ num_mort_accounts                  <int> 1, 0, 0, 0, 0, 3,…
$ account_never_delinq_percent       <dbl> 92.9, 100.0, 93.5…
$ tax_liens                          <int> 0, 0, 0, 1, 0, 0,…
$ public_record_bankrupt             <int> 0, 1, 0, 0, 0, 0,…
$ loan_purpose                       <fct> moving, debt_cons…
$ application_type                   <fct> individual, indiv…
$ loan_amount                        <int> 28000, 5000, 2000…
$ term                               <dbl> 60, 36, 36, 36, 3…
$ interest_rate                      <dbl> 14.07, 12.61, 17.…
$ installment                        <dbl> 652.53, 167.54, 7…
$ grade                              <fct> C, C, D, A, C, A,…
$ sub_grade                          <fct> C3, C1, D1, A3, C…
$ issue_month                        <fct> Mar-2018, Feb-201…
$ loan_status                        <fct> Current, Current,…
$ initial_listing_status             <fct> whole, whole, fra…
$ disbursement_method                <fct> Cash, Cash, Cash,…
$ balance                            <dbl> 27015.86, 4651.37…
$ paid_total                         <dbl> 1999.330, 499.120…
$ paid_principal                     <dbl> 984.14, 348.63, 1…
```

```
$ paid_interest                    <dbl> 1015.19, 150.49, …
$ paid_late_fees                   <dbl> 0, 0, 0, 0, 0, 0,…
```

Hide

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

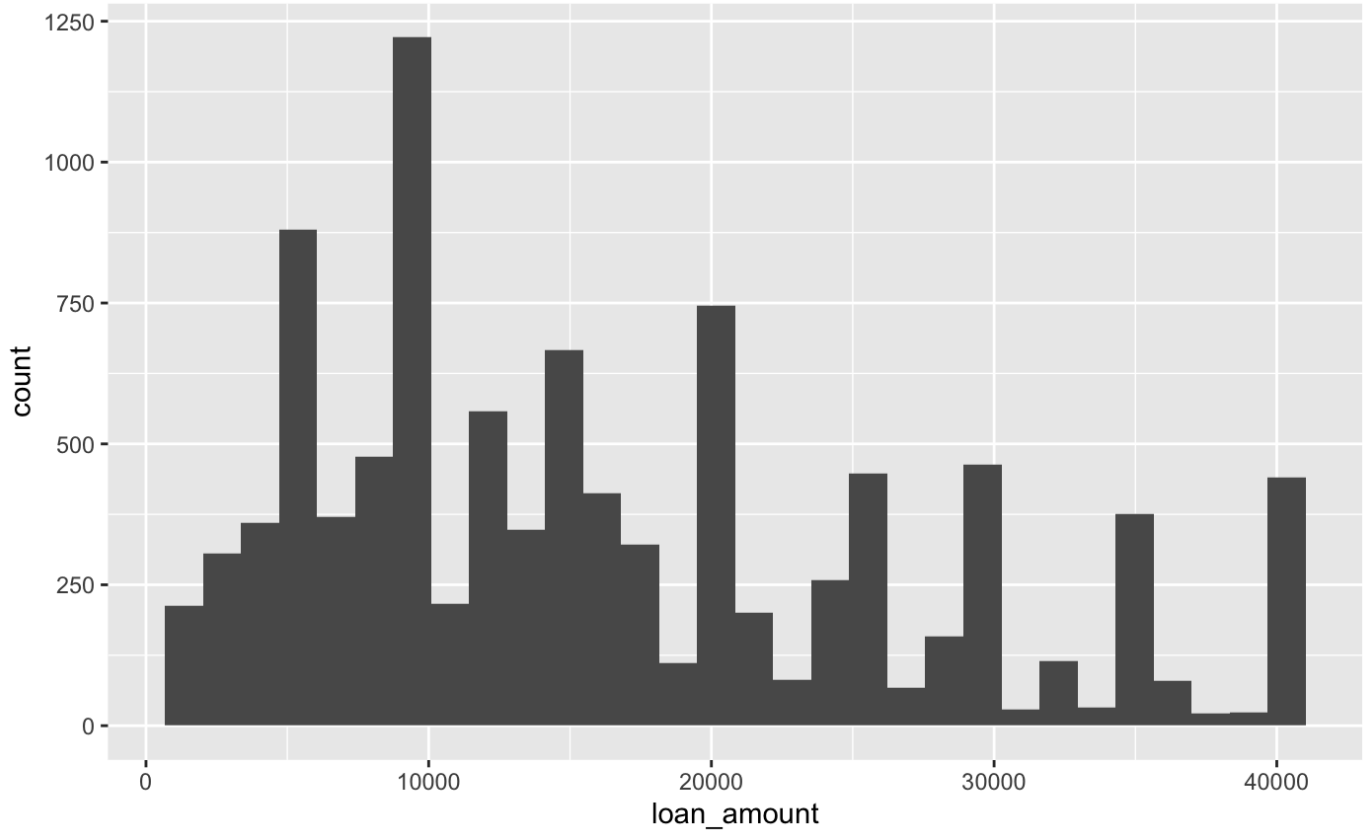
```
Rows: 10,000
Columns: 8
$ loan_amount    <int> 28000, 5000, 2000, 21600, 23000, 50…
$ interest_rate  <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6…
$ term           <dbl> 60, 36, 36, 36, 36, 36, 60, 60, 36,…
$ grade          <fct> C, C, D, A, C, A, C, B, C, A, C, B,…
$ state          <fct> NJ, HI, WI, PA, CA, KY, MI, AZ, NV,…
$ annual_income  <dbl> 90000, 40000, 40000, 30000, 35000, …
$ homeownership  <fct> MORTGAGE, RENT, RENT, RENT, RENT, O…
$ debt_to_income <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6…
```

Hide

```
#Shape:Skewness (right, left, symmetric) and Modality (unimodal, bimodal, multimodal,
uniform)

#Center: centered at mean, median, or mode

#Spread: range (range), standard deviation (sd), inter-quartile range (IQR)

#Anomaly: Unusual observations
```

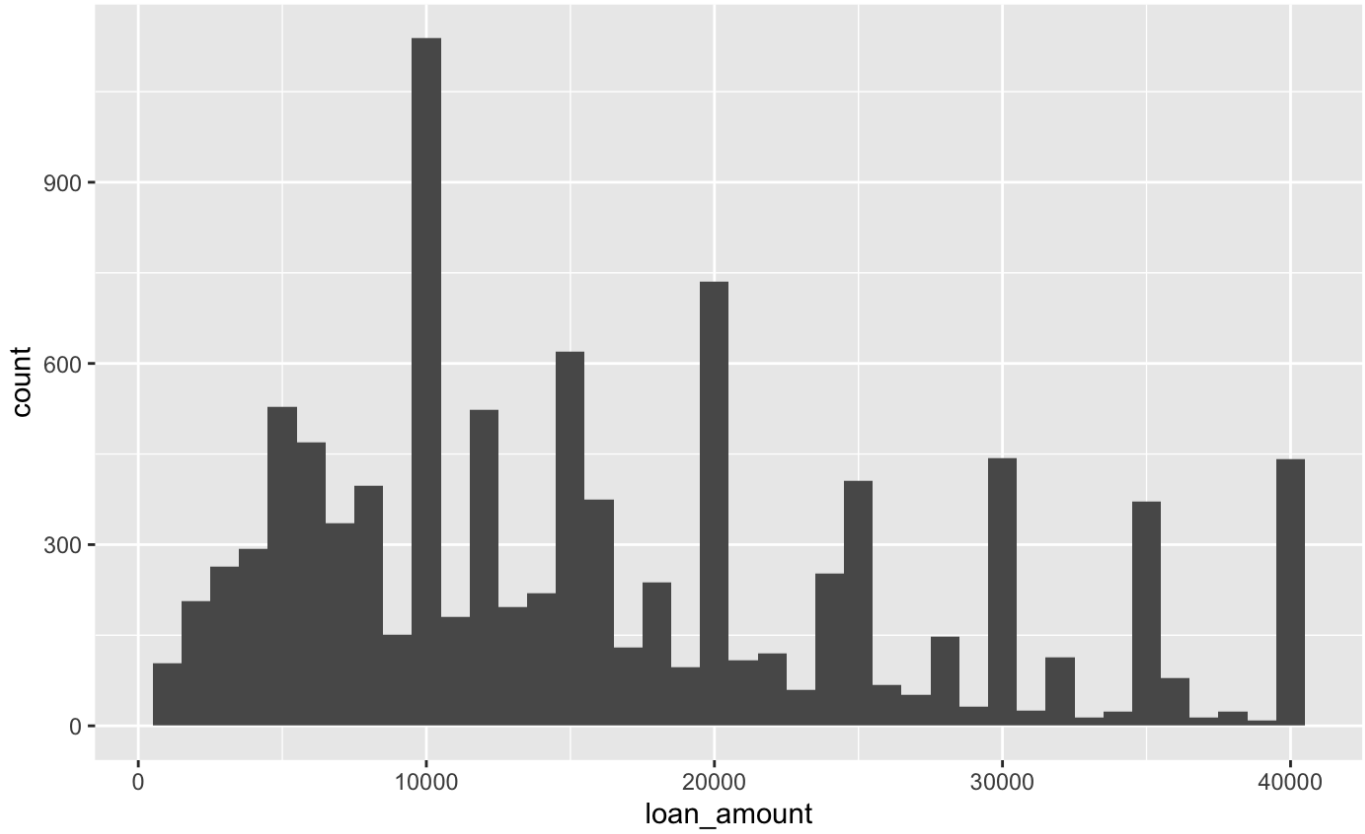
Hide

```
#Frequency of value: histogram, where x is the variable of interest

ggplot(loans) +
  aes(x = loan_amount) +
  geom_histogram()
```



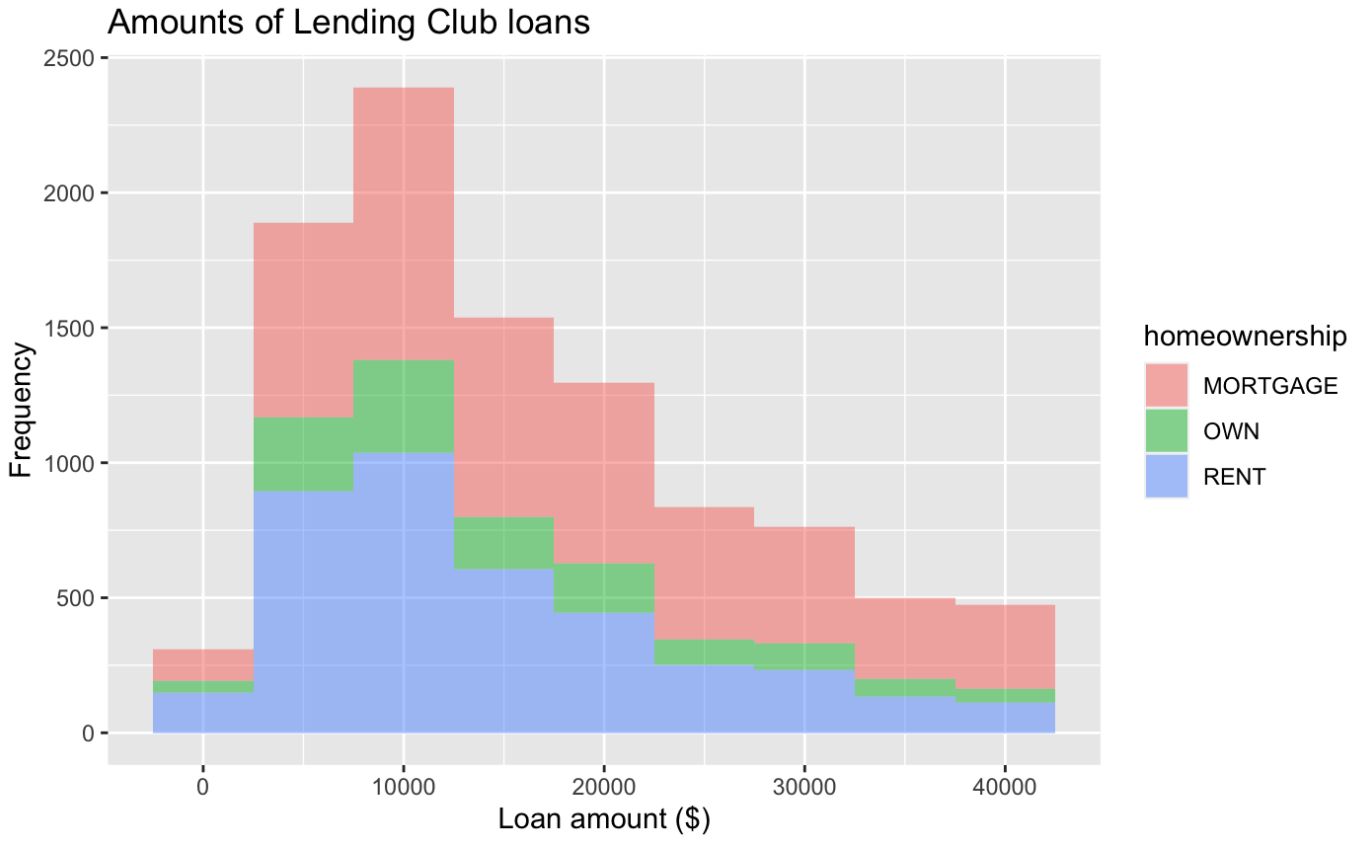
Hide

```
#Changing bin width; each bin contains 1 frequency of its value.
ggplot(loans) +
  aes(x = loan_amount) +
  geom_histogram( binwidth = 1000)
```



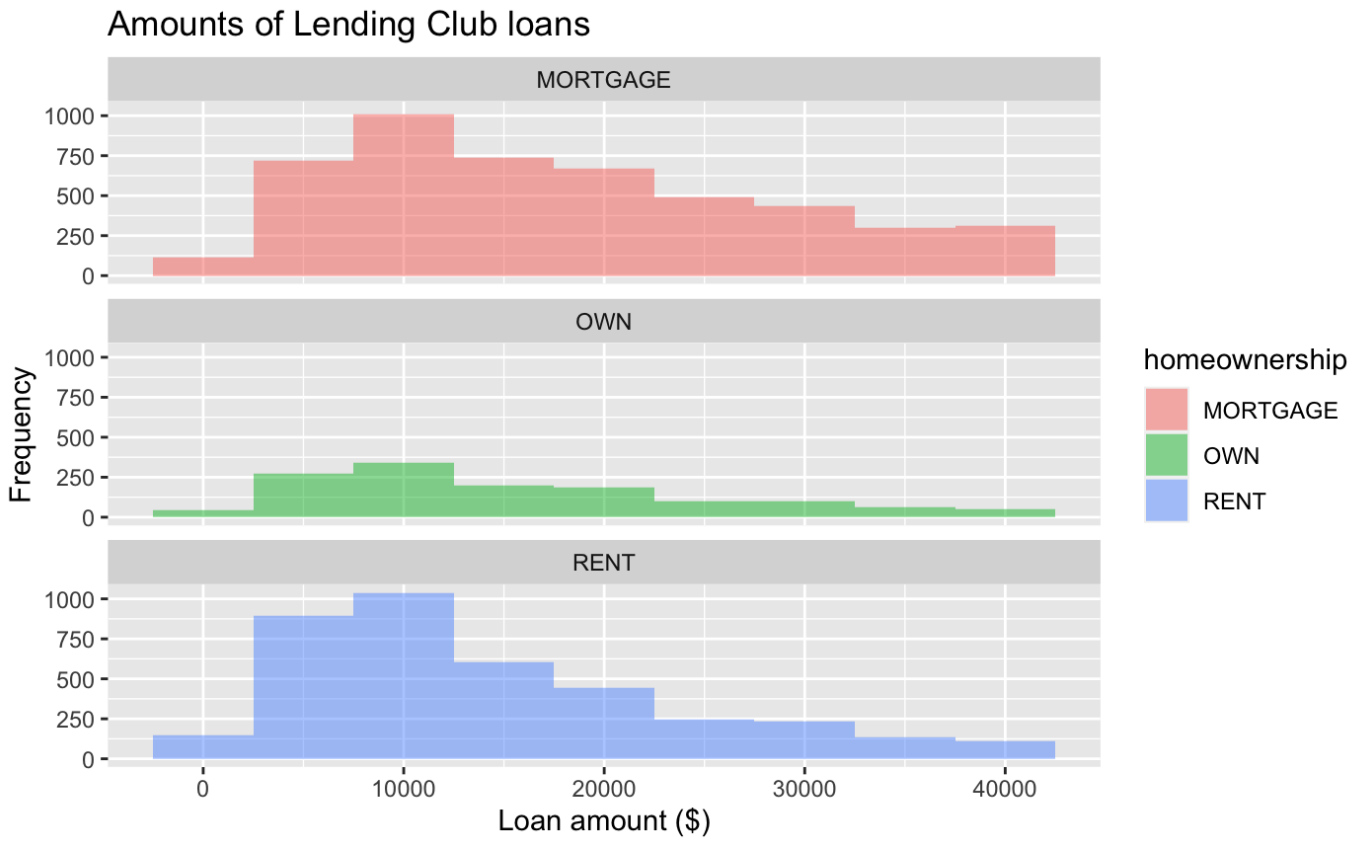
Hide

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



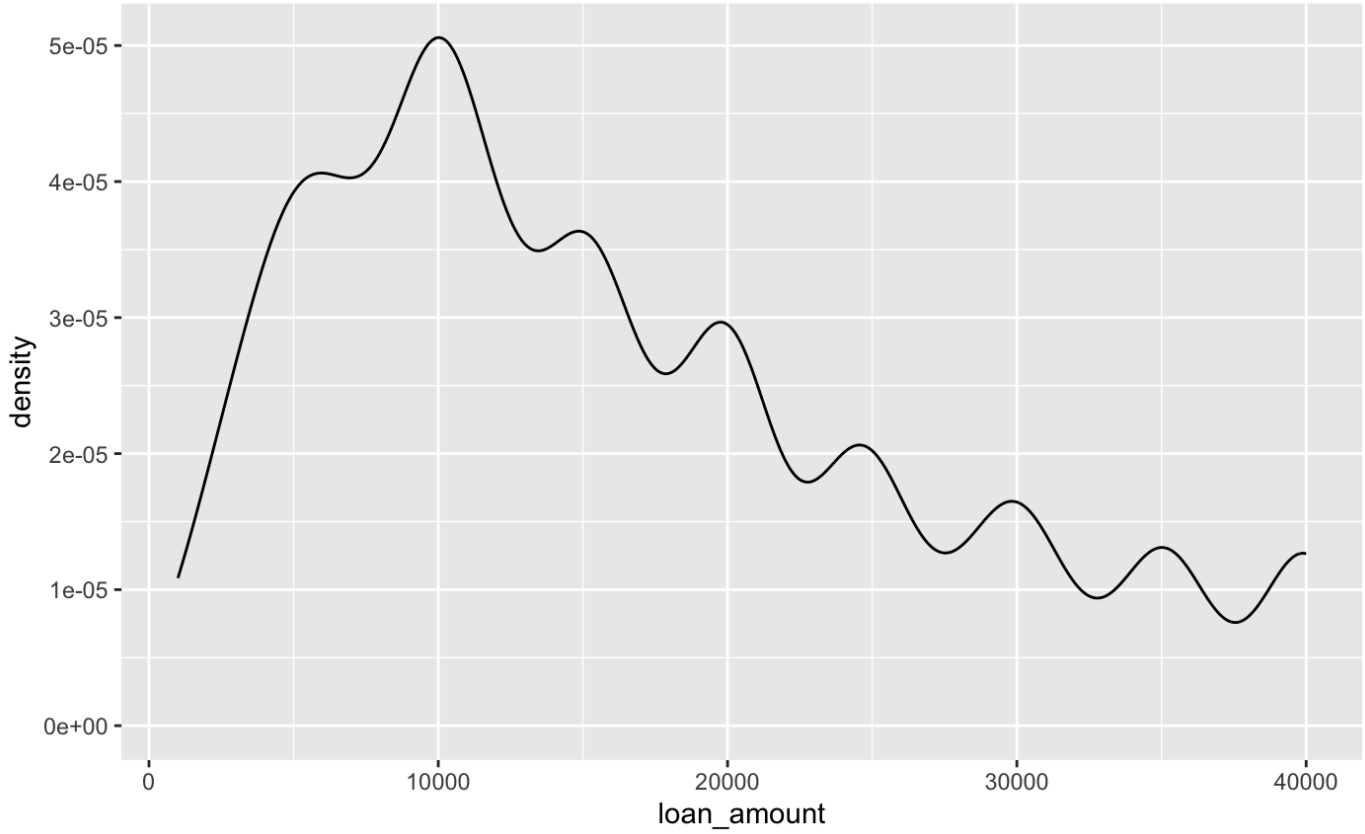
Hide

```
ggplot(loans) +
  aes(x = loan_amount, fill = homeownership) + #Mapping fill/color
  geom_histogram( binwidth = 5000, alpha = 0.5) +
  labs( x = "Loan amount ($)",
        y = "Frequency",
        title = "Amounts of Lending Club loans") +
  facet_wrap( ~ homeownership, ncol = 1)
```



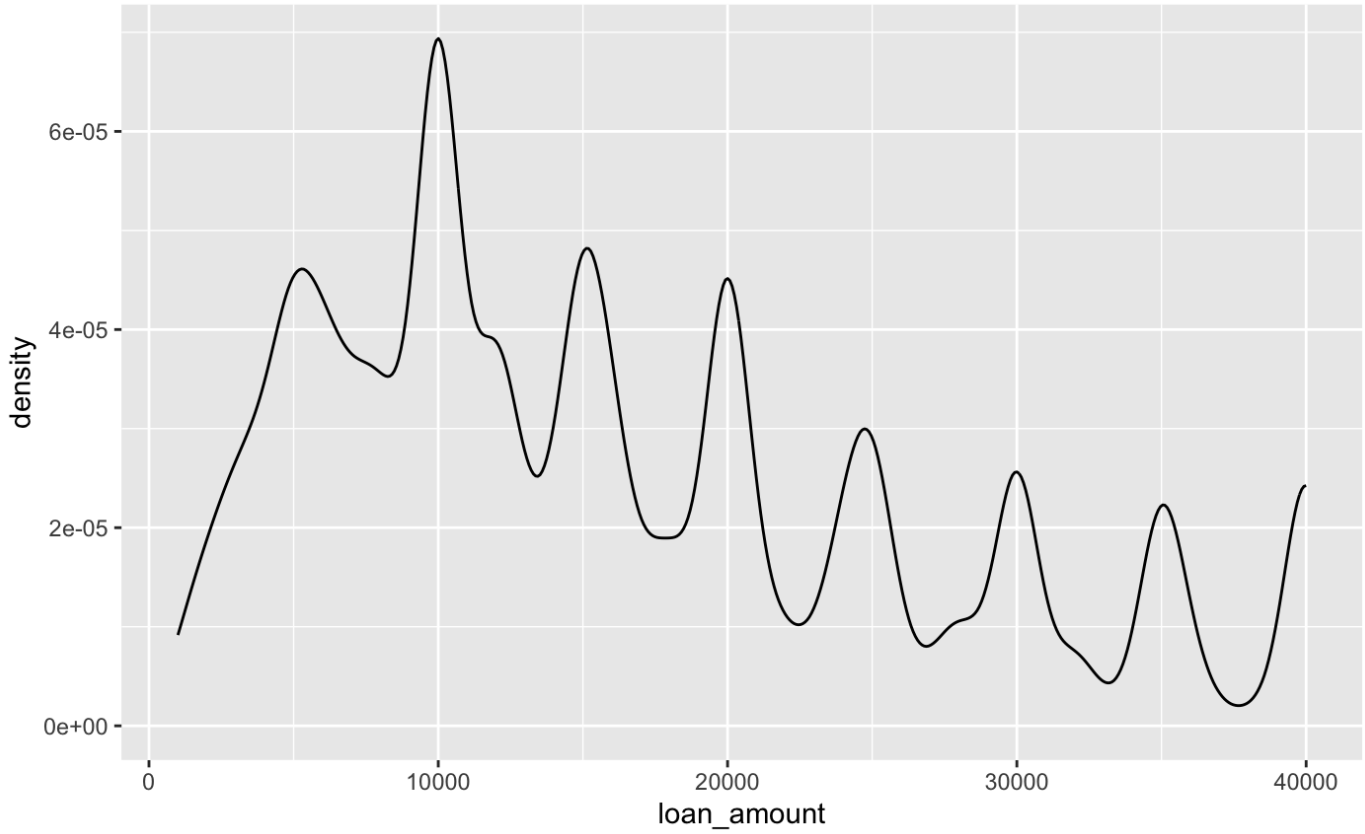
Hide

```
#Probability density (within a range) [ie. number of times values in a certain range occur over the total number of values]: density plot producing smooth curve, where x is the variable of interest

ggplot(loans) +
  aes (x = loan_amount) +
  geom_density()
```



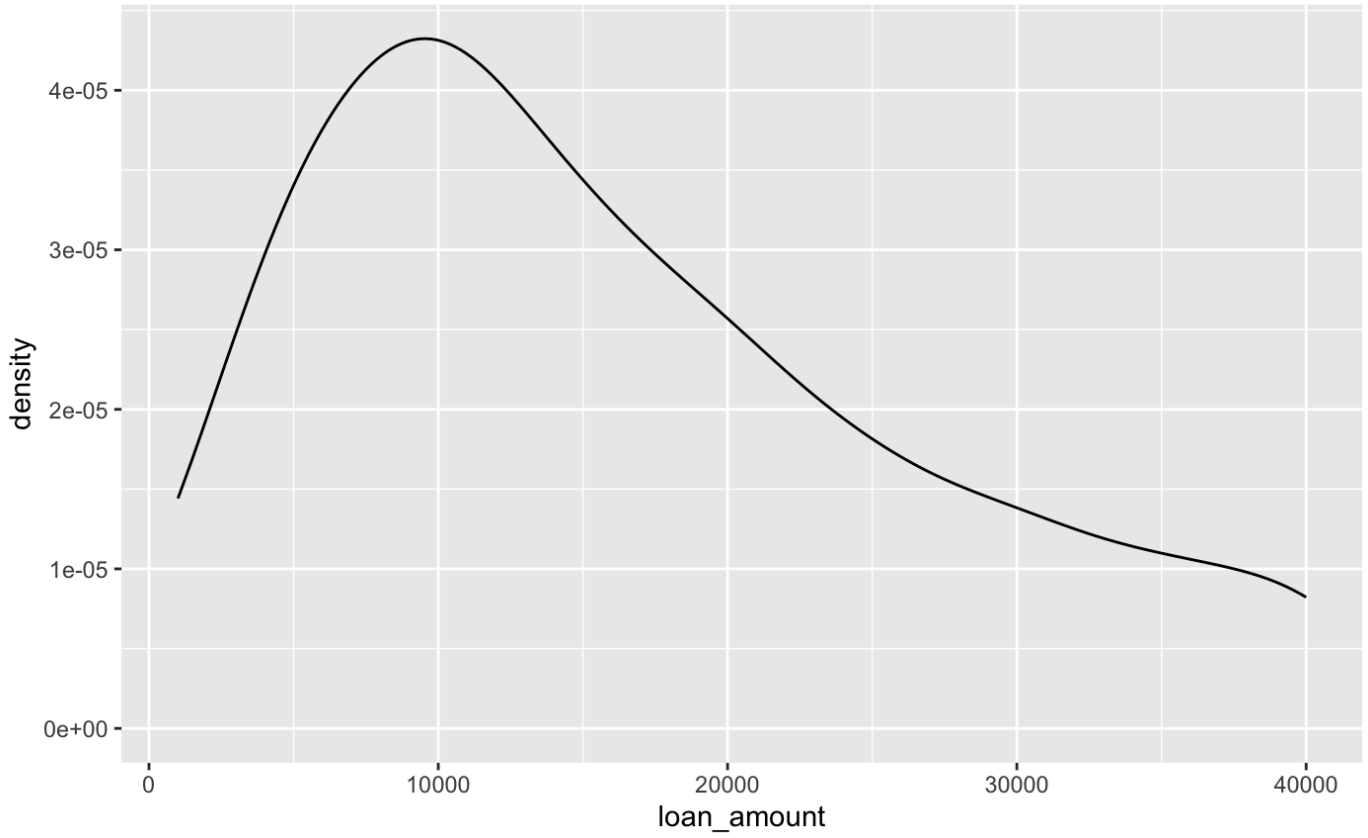
Hide

```
ggplot(loans) +
  aes (x = loan_amount) +
  geom_density( adjust = 0.5) #Adjust bandwidth, higher is smoother because it has a
lowered res.
```



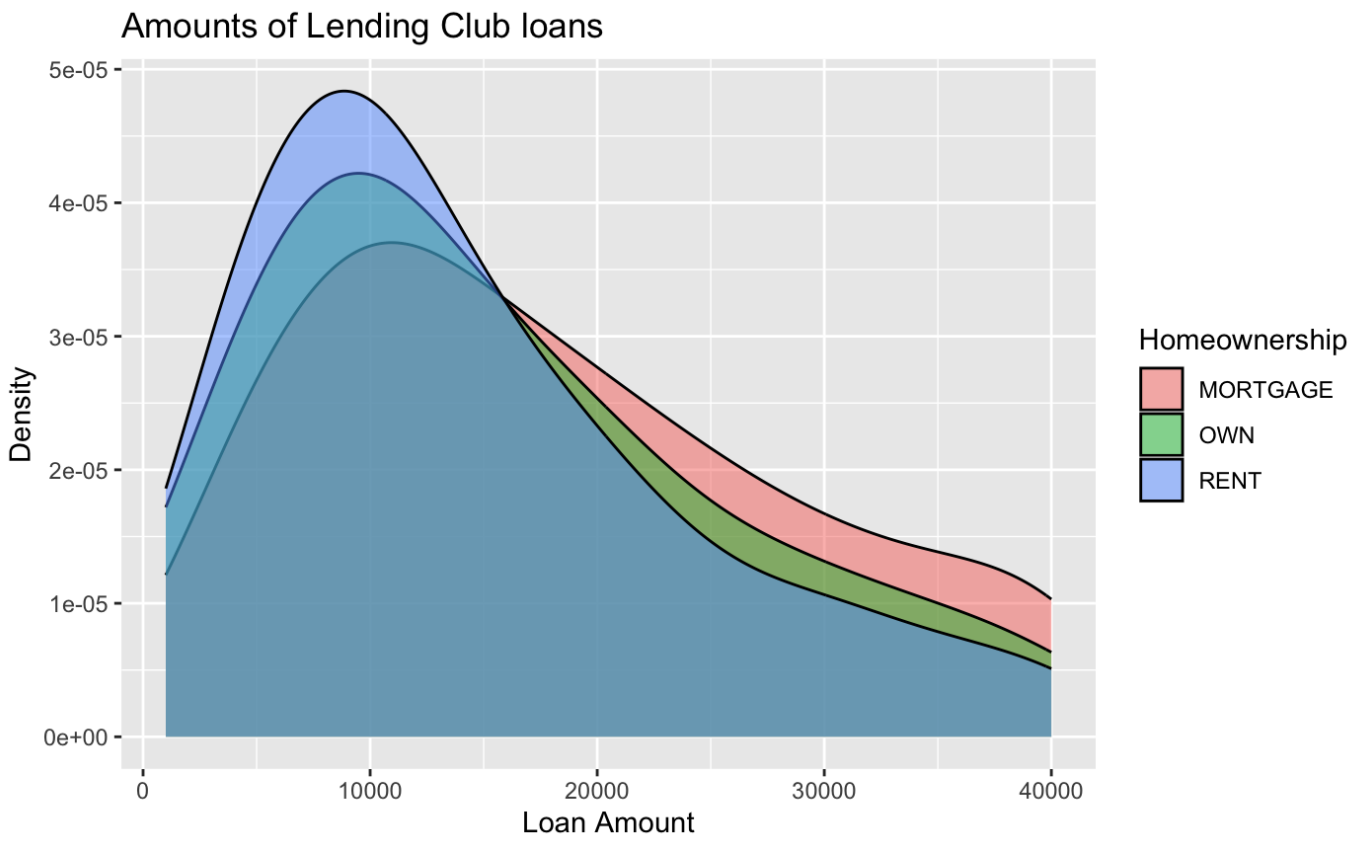
Hide

```
ggplot(loans) +
  aes (x = loan_amount) +
  geom_density( adjust = 2)
```



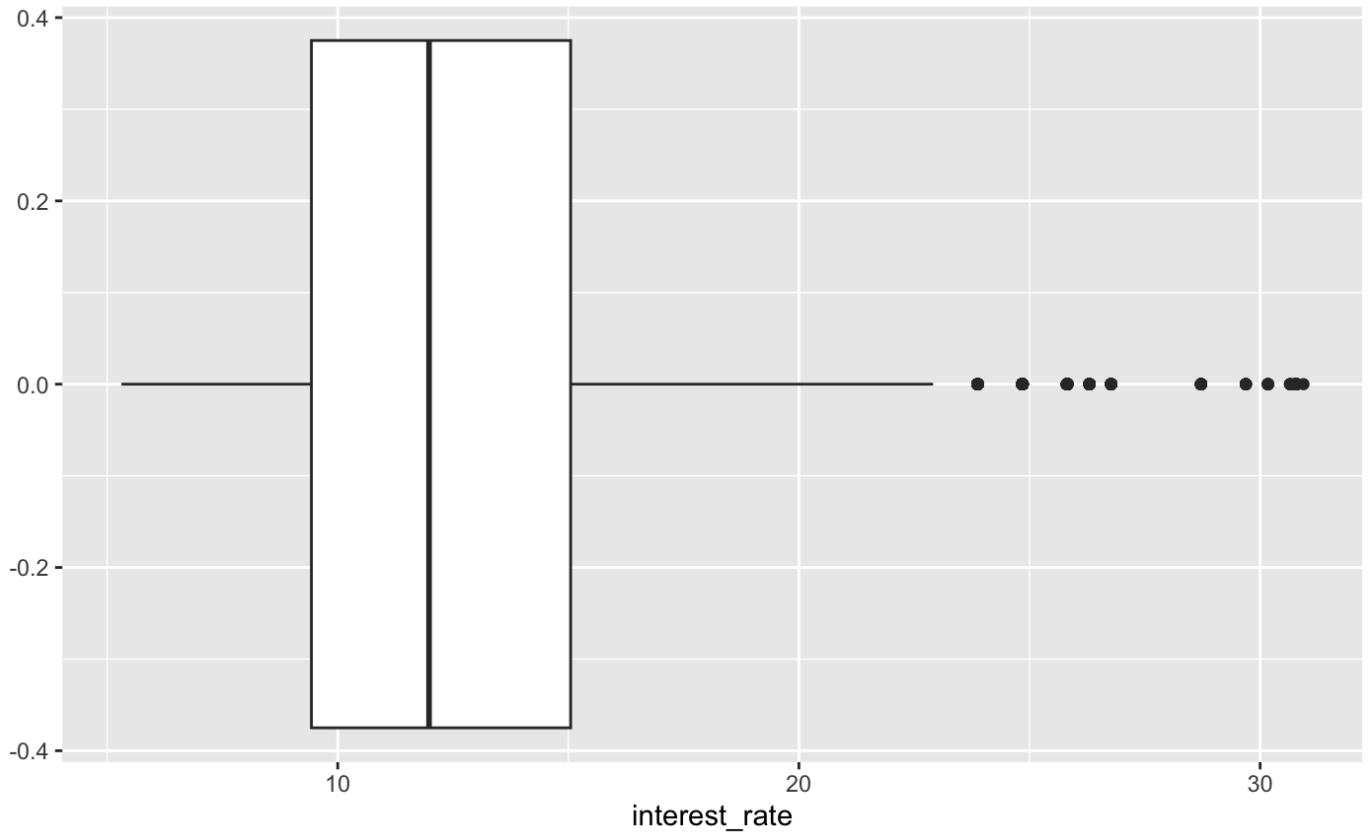
Hide

```
#Adding a categorical value:
ggplot(loans) +
  aes (x = loan_amount,
       fill = homeownership) +
  geom_density( adjust = 2, alpha = 0.5) +
  labs (x = "Loan Amount",
        y = "Density",
        title = "Amounts of Lending Club loans",
        fill = "Homeownership")
```



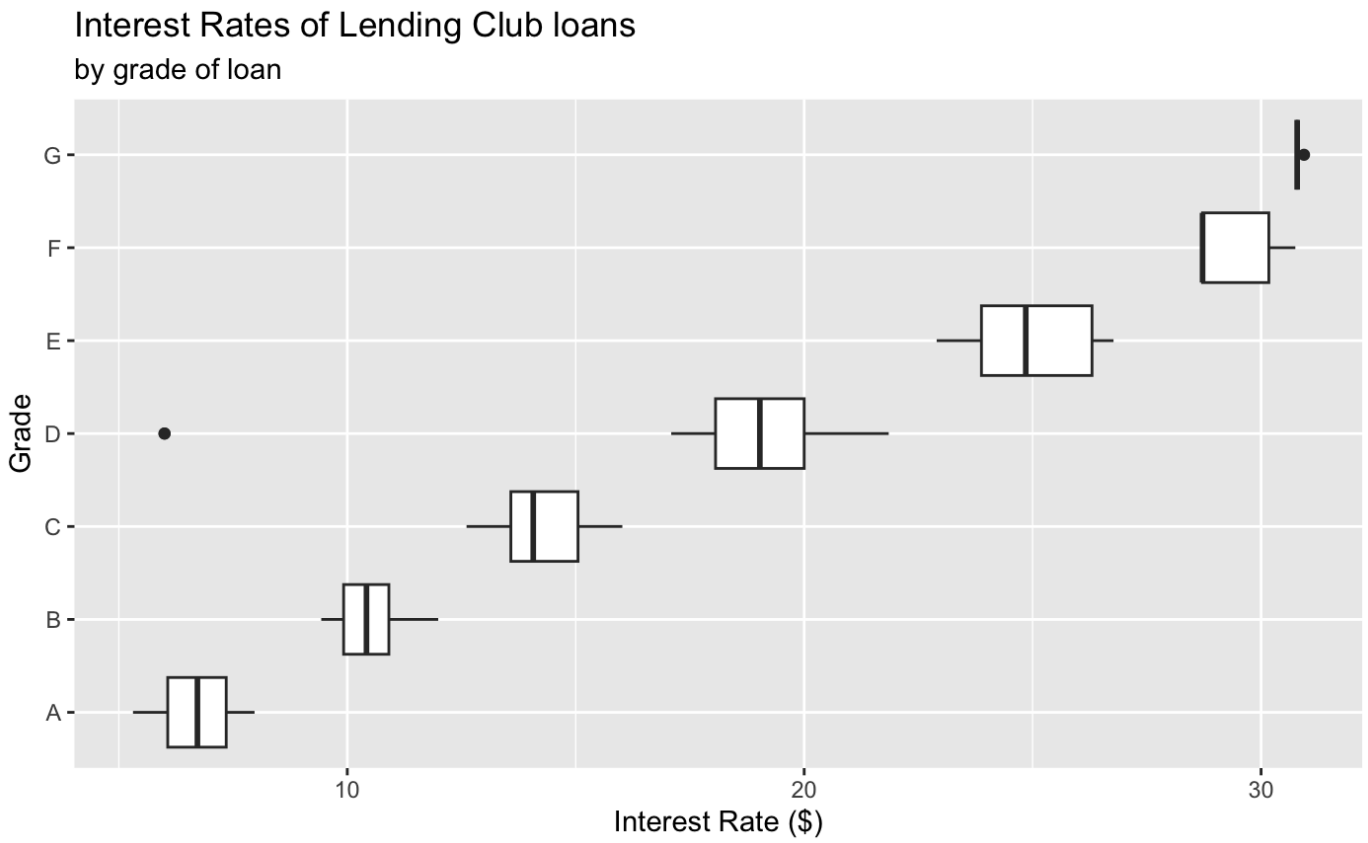
Hide

```
#Boxplots: Indicate important values wrt distribution, eg. median (thick line), IQR
(boundaries of box plot), outliers (dots outside the line)

ggplot(loans) +
  aes (x = interest_rate) +
  geom_boxplot()
```



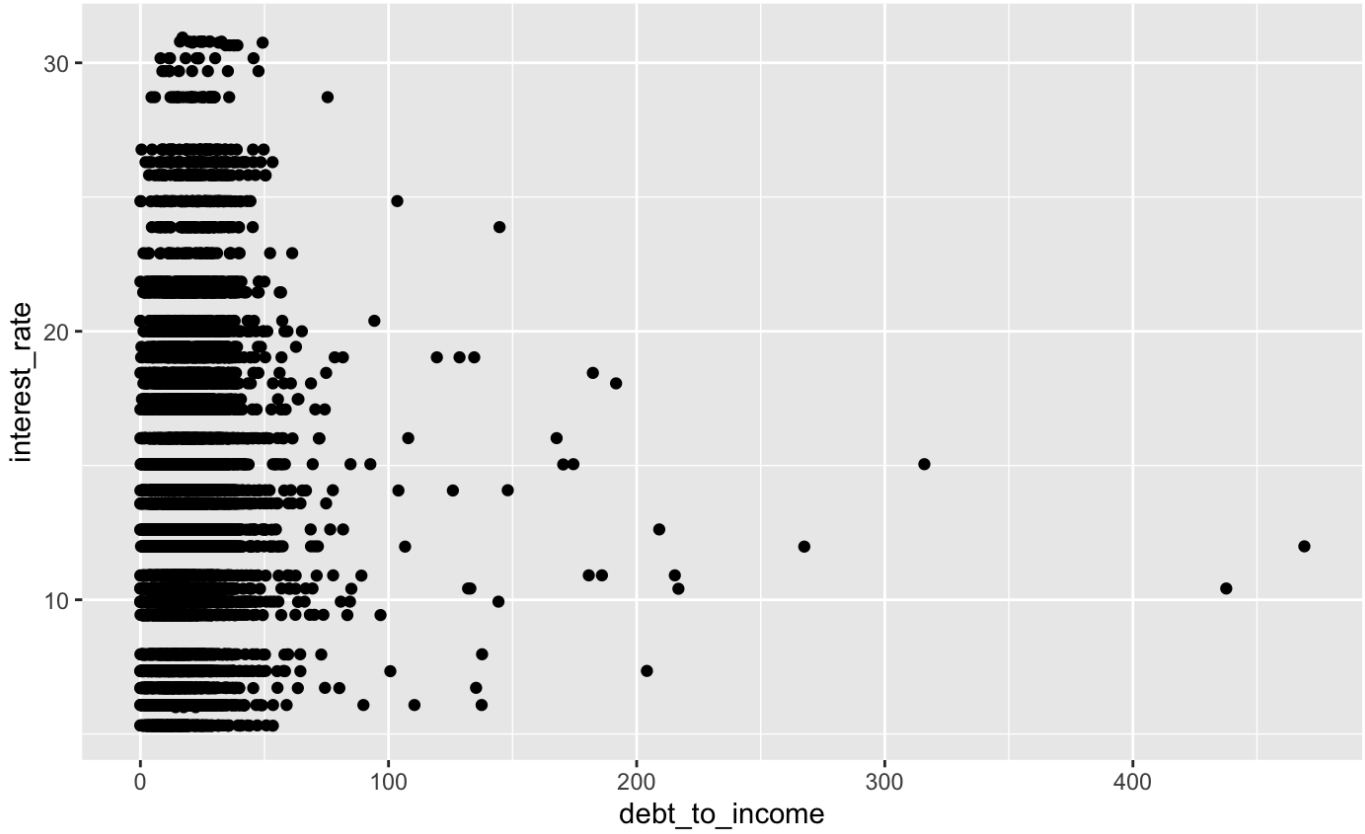
Hide

```
ggplot(loans) +
  aes (x = interest_rate,
       y = grade) +
  geom_boxplot() +
  labs (x = "Interest Rate ($)",
        y = "Grade",
        title = "Interest Rates of Lending Club loans",
        subtitle = "by grade of loan")
```



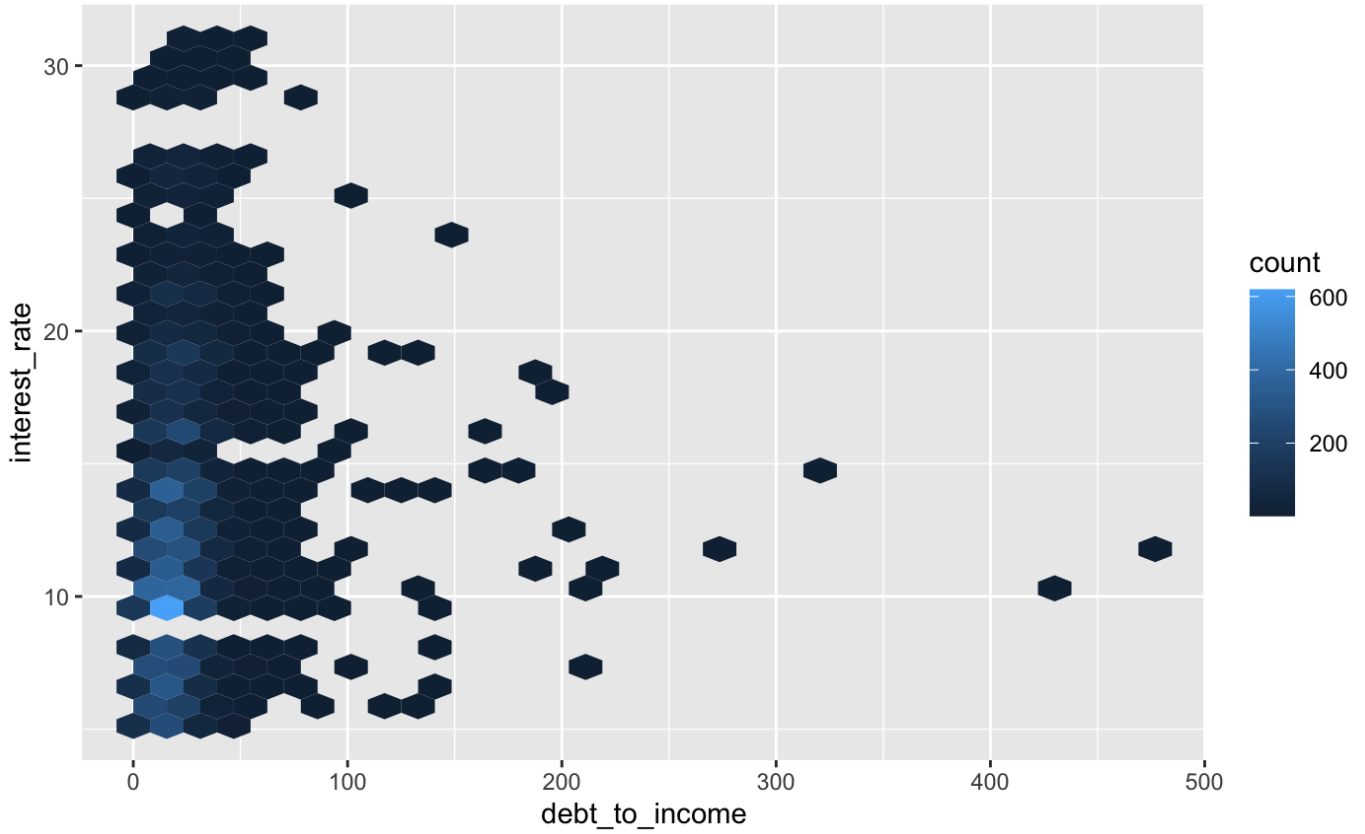
Hide

```
#Scatterplots: Allow you to infer details about the r/s between chosen variables. Alt
ernative to this is hex plot which uses color to also accommodate representation of c
oncentrations of datapoints.

ggplot(loans) +
  aes (x = debt_to_income,
       y = interest_rate) +
  geom_point()
```



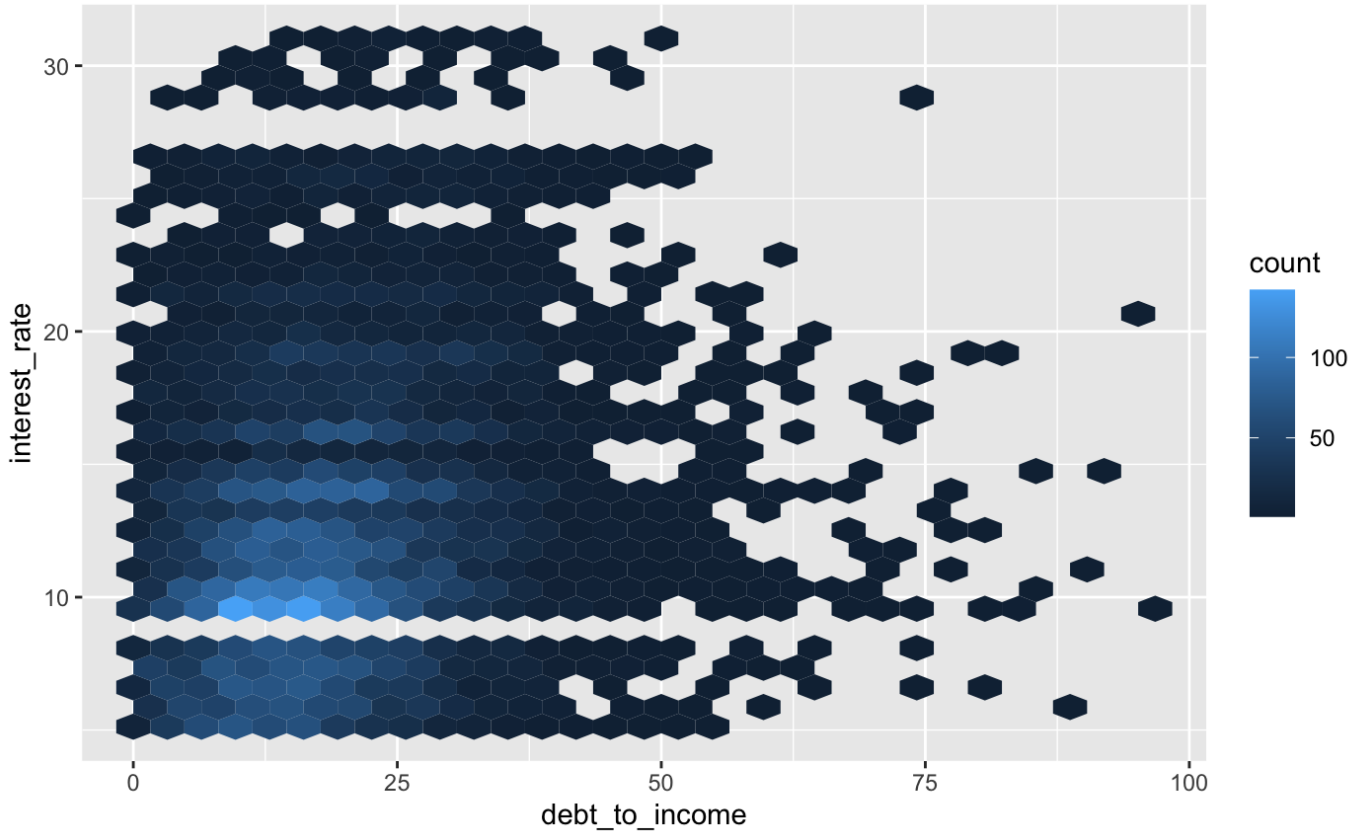
Hide

```
ggplot(loans) +
  aes (x = debt_to_income,
       y = interest_rate) +
  geom_hex()
```



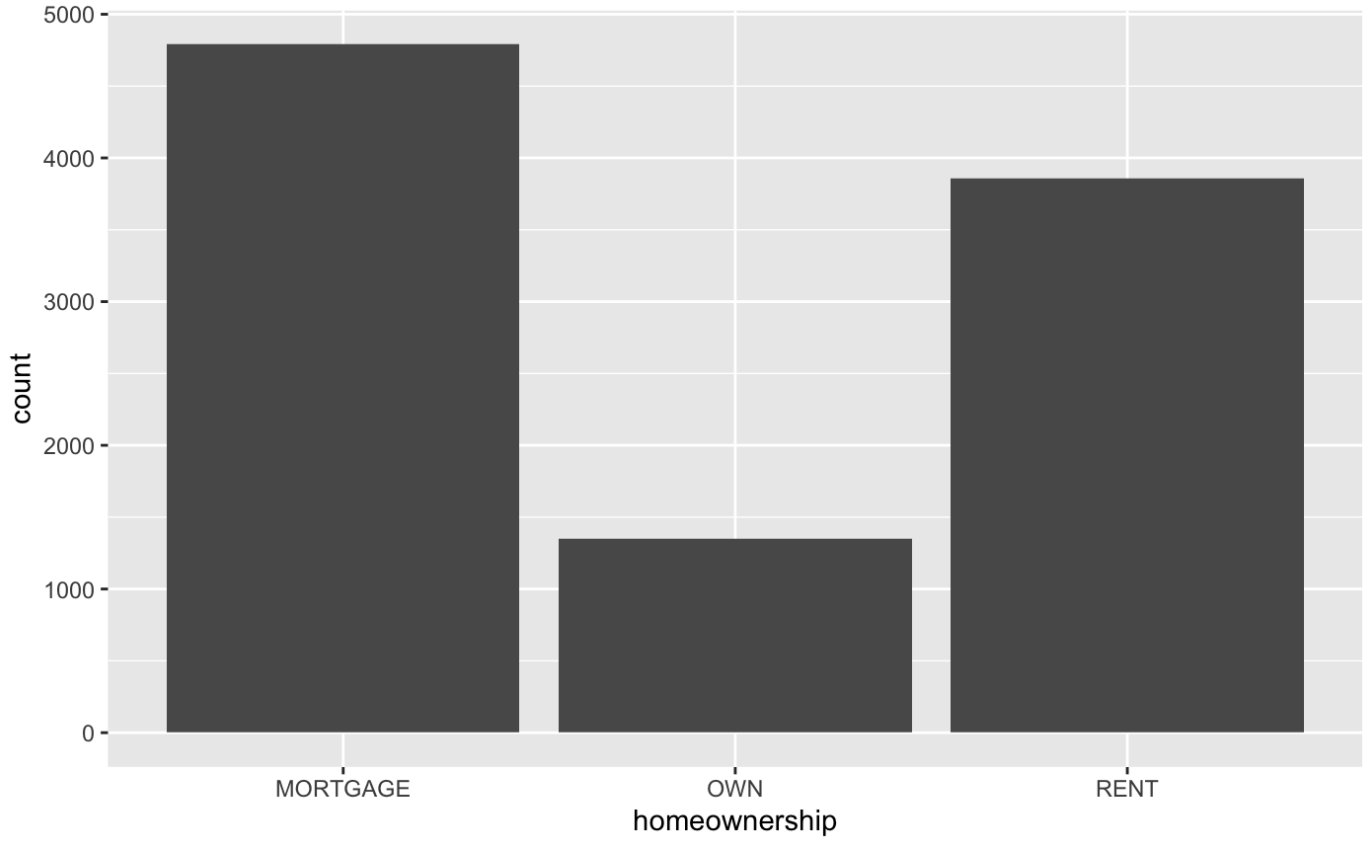
Hide

```
ggplot(loans %>% filter(debt_to_income < 100)) +
  aes (x = debt_to_income,
       y = interest_rate) +
  geom_hex()
```



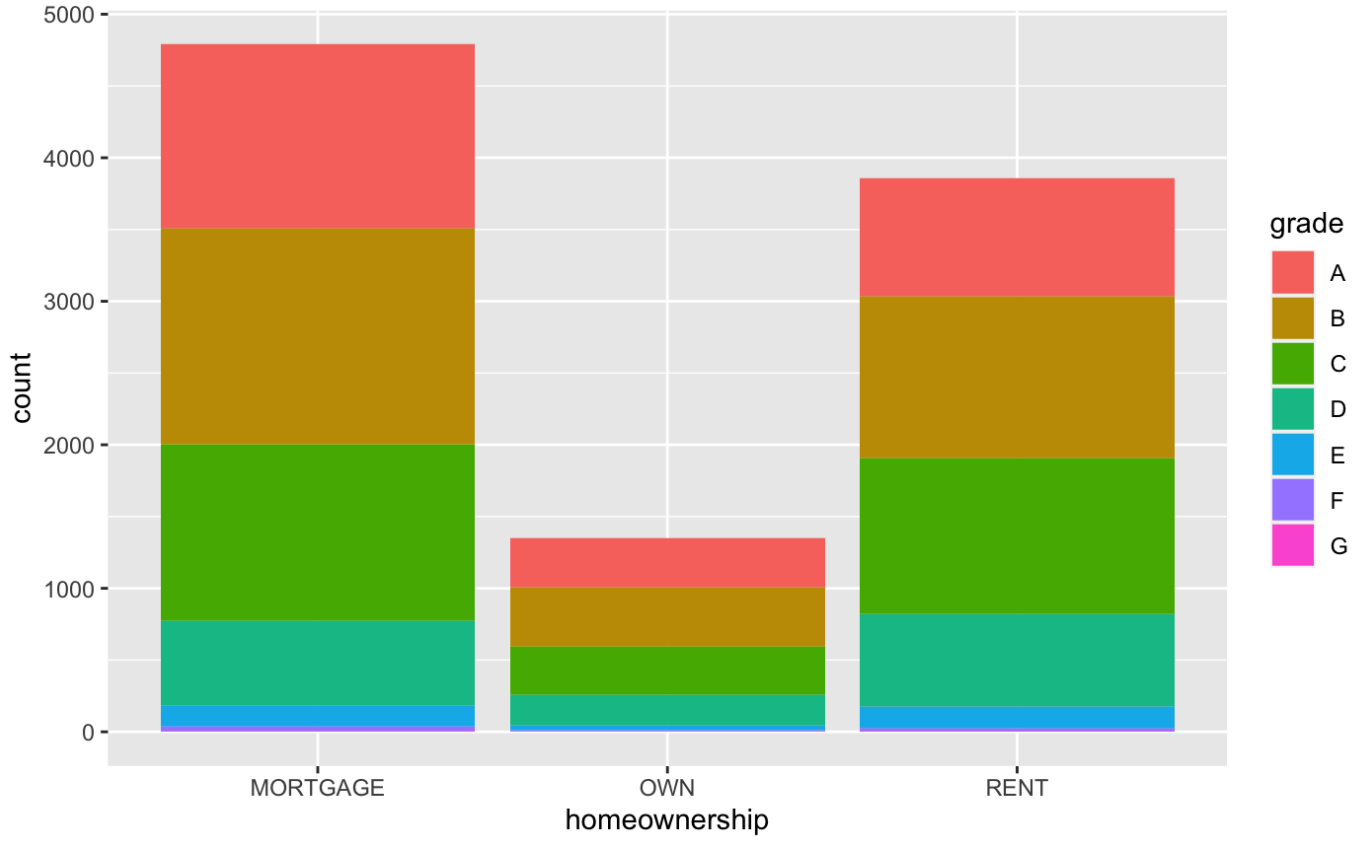
Hide

```
#Barplots: R extracts the unique values // count
ggplot(loans) +
  aes (x = homeownership) +
  geom_bar()
```



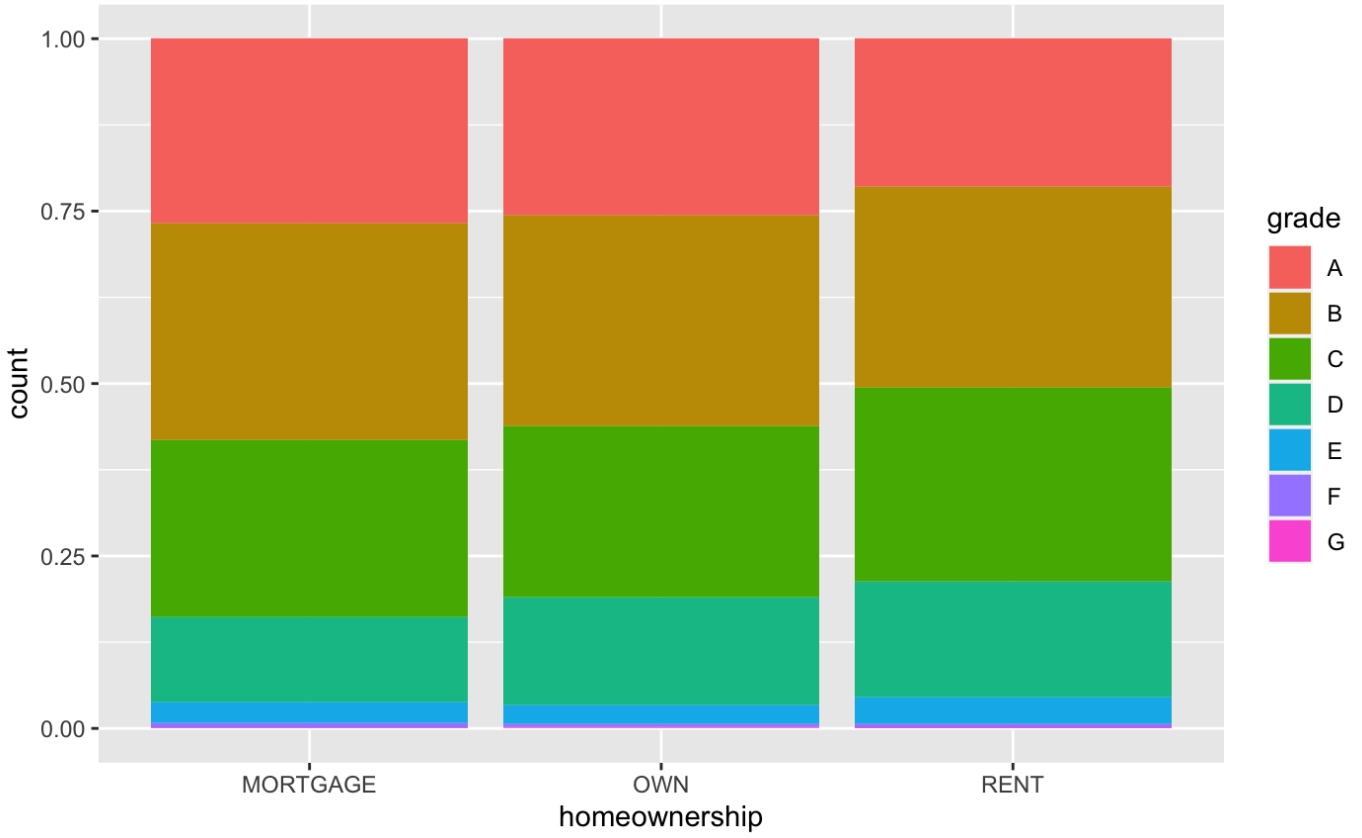
Hide

```
#Segmented bar plot
ggplot(loans) +
  aes (x = homeownership,
       fill = grade) +
  geom_bar()
```



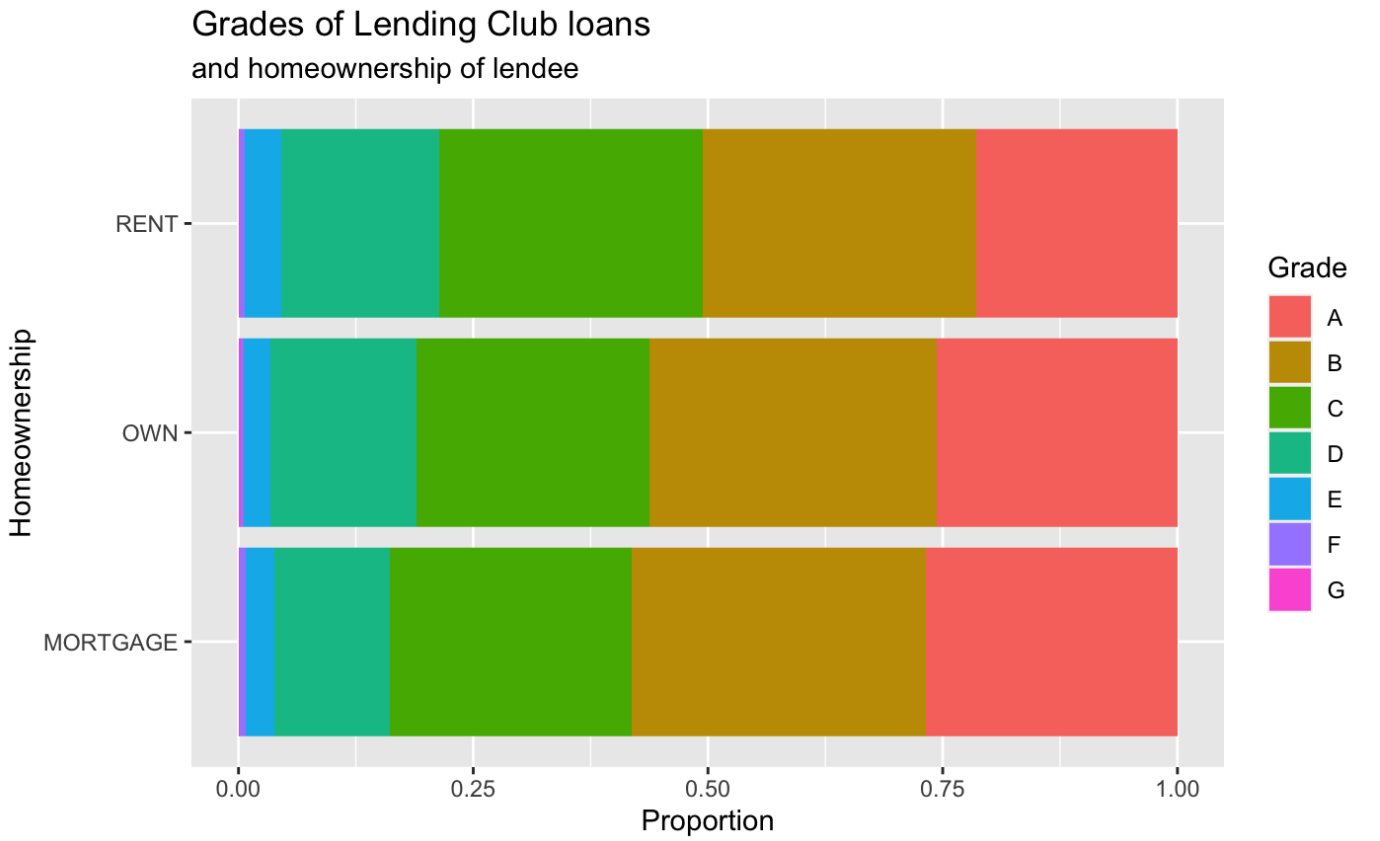
Hide

```
#Same height = Easier comparison
ggplot(loans) +
  aes (x = homeownership,
       fill = grade) +
  geom_bar( position = "fill")
```



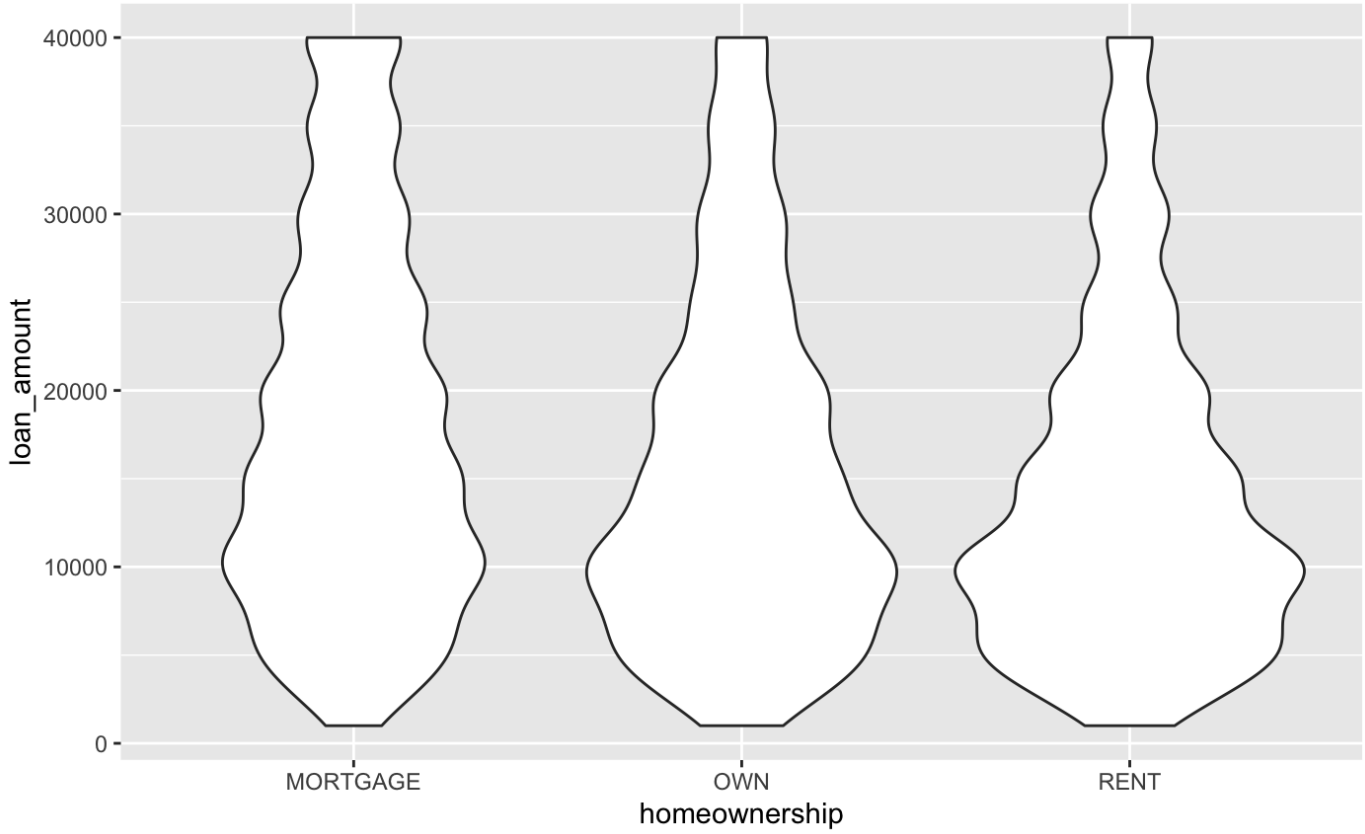
Hide

```
#Swap x to y to change the orientation
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")
```



Hide

```
#Scenarios involving a numeric and a categoric variable or in general variables of mo
re than one type

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



Hide

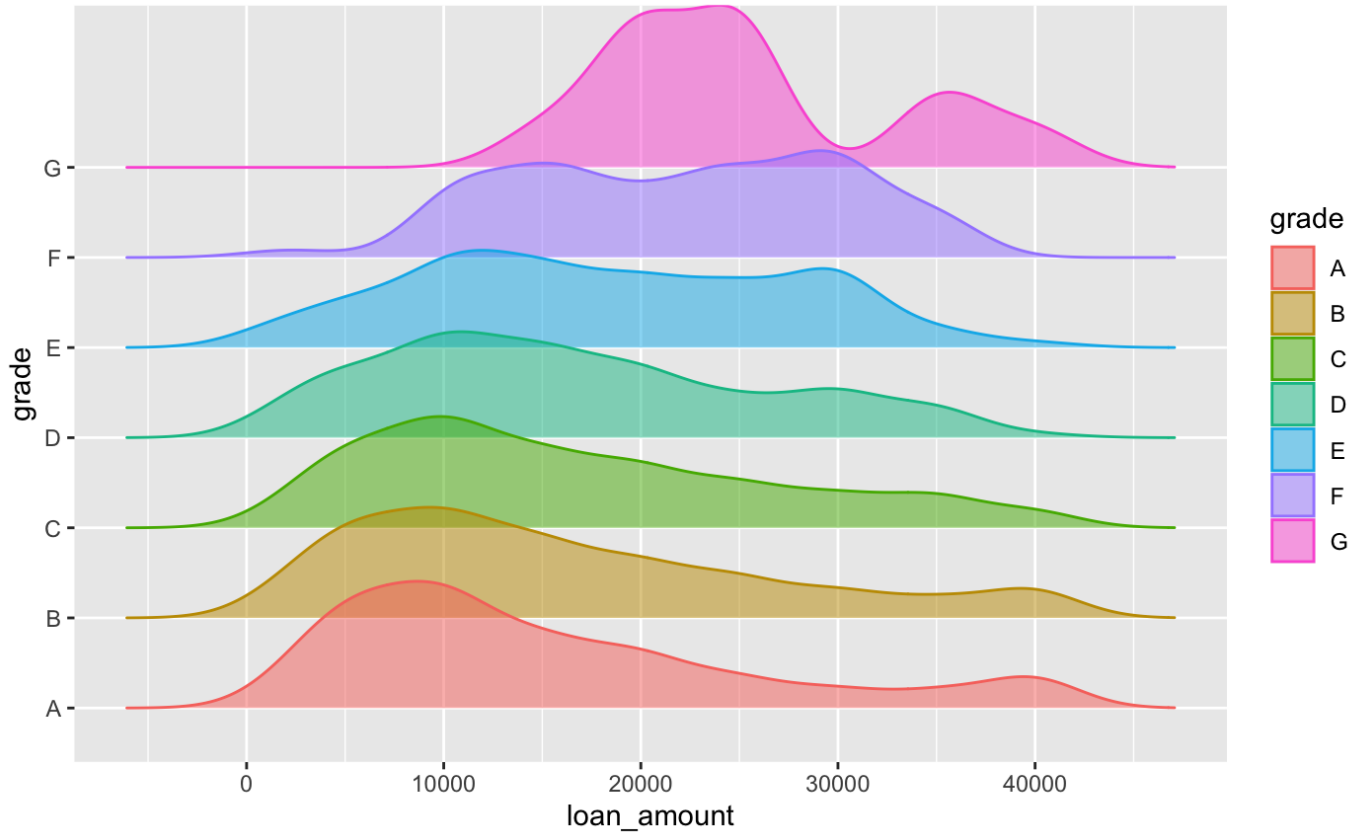
```
#Ridge plot
install.packages("ggridges")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/ggridges_0.
5.4.tgz'
Content type 'application/x-gzip' length 2240459 bytes (2.1 MB)
==================================================
downloaded 2.1 MB
```

```
The downloaded binary packages are in
    /var/folders/8w/_k66pvq90sncpcr1ps95370h0000gn/T//RtmpOhXpy6/downloaded_packages
```

Hide

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


grade
G
F
E
D
C
B
A
0
10000
20000
30000
40000
loan_amount
grade
A
B
C
D
E
F
G

Hide

```
NA
NA
```