Challenge-7

Loy Yee Keen

2023-10-04

Data: Palmer Penguins

Measurements for penguin species, island in Palmer Archipelago, size (flipper length, body mass, bill dimensions), and gender.

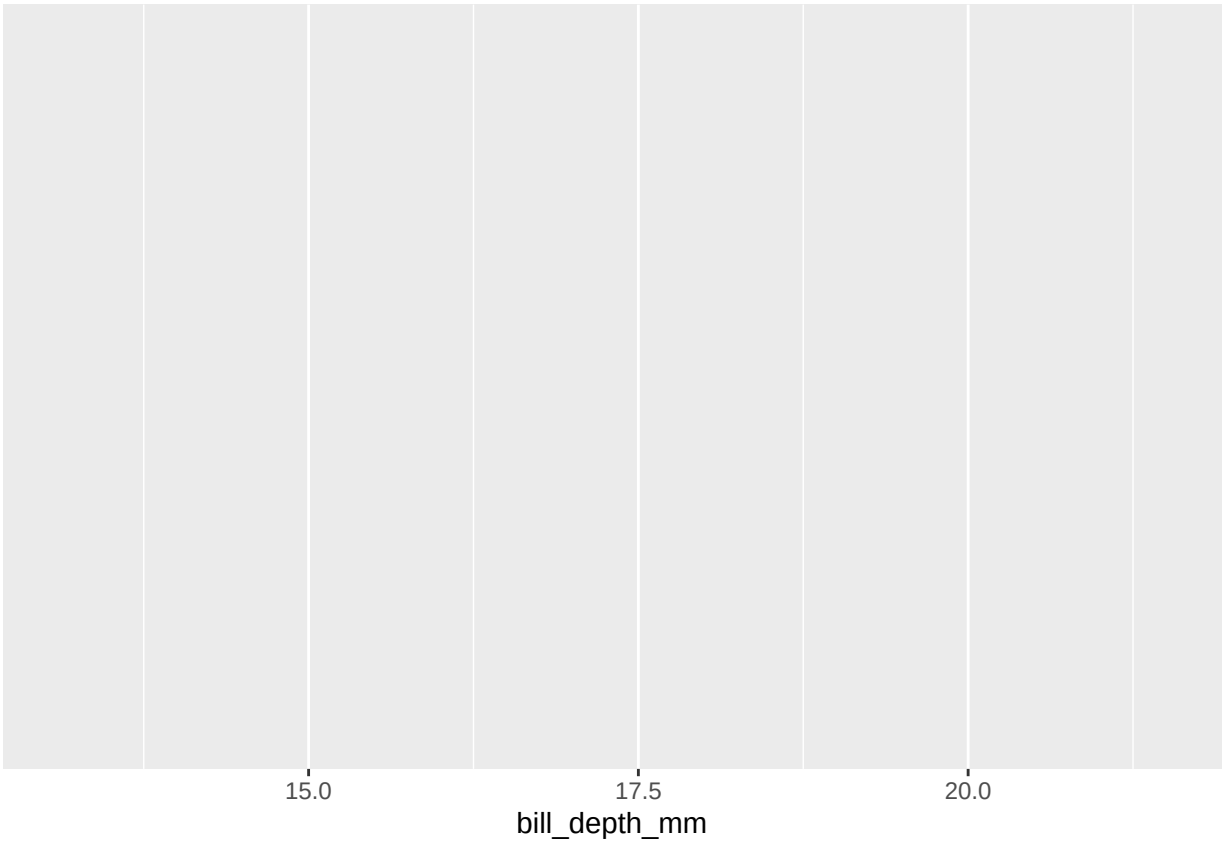
```
## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

## Warning: package 'palmerpenguins' was built under R version 4.2.3
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
## $ sex               <fct> male, female, female, NA, female, male, female, male~
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

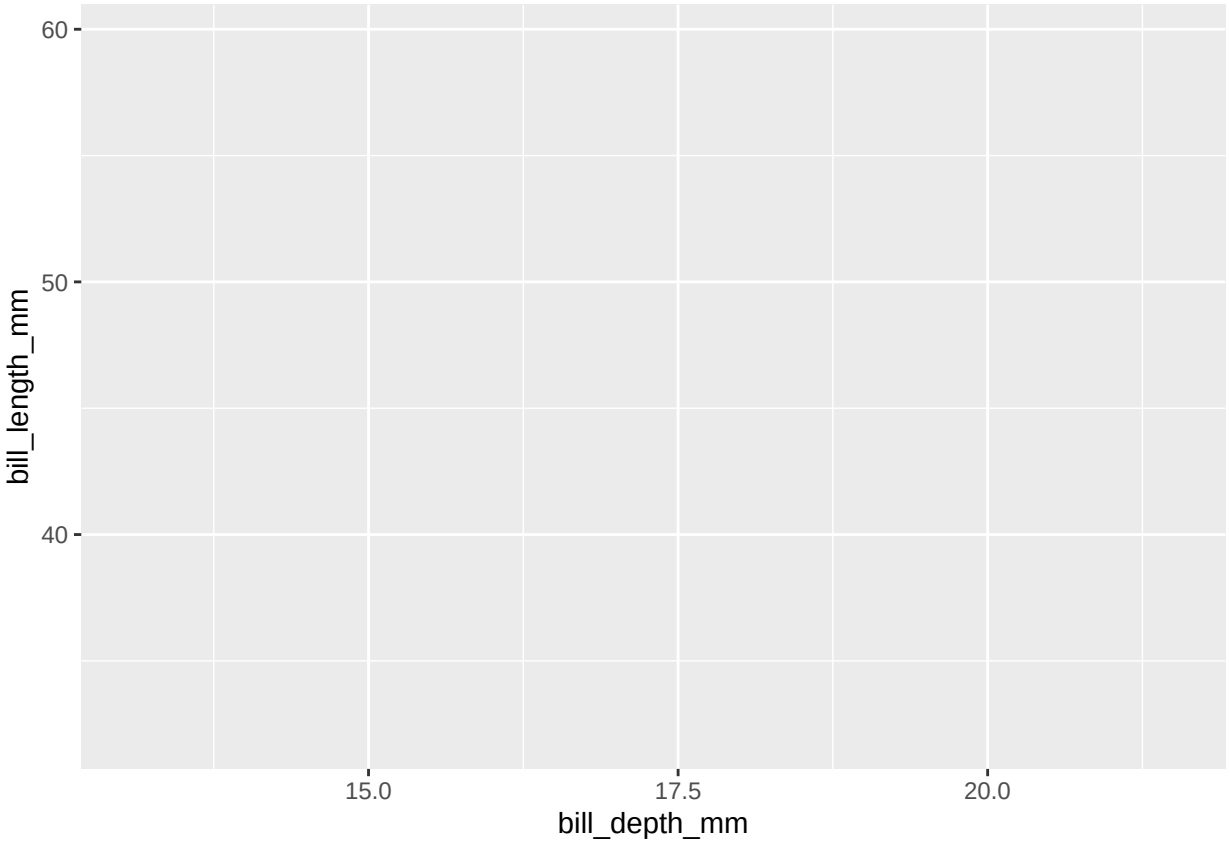
Palmer Penguins: Plot recreation

a. Start with the penguins data frame

b. Map bill depth to the x-axis

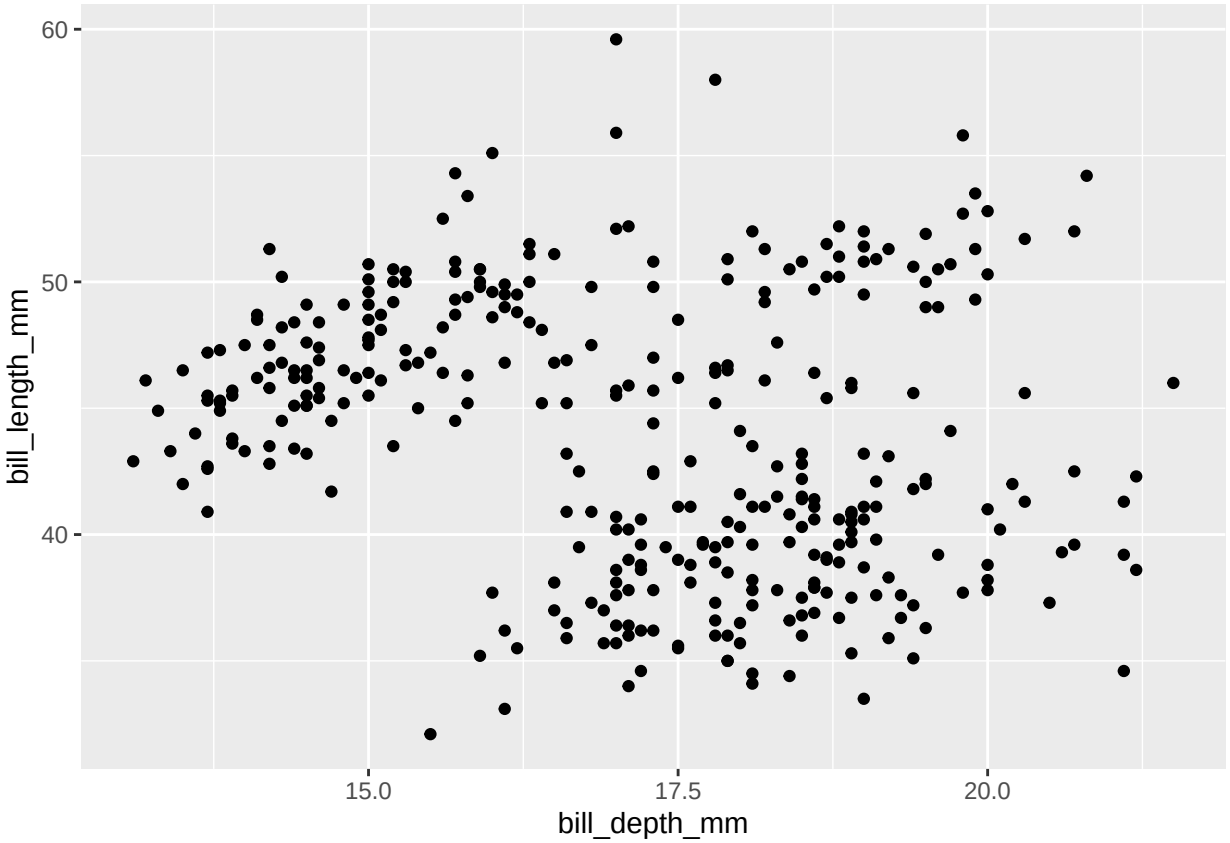


c. Map bill depth to the y-axis



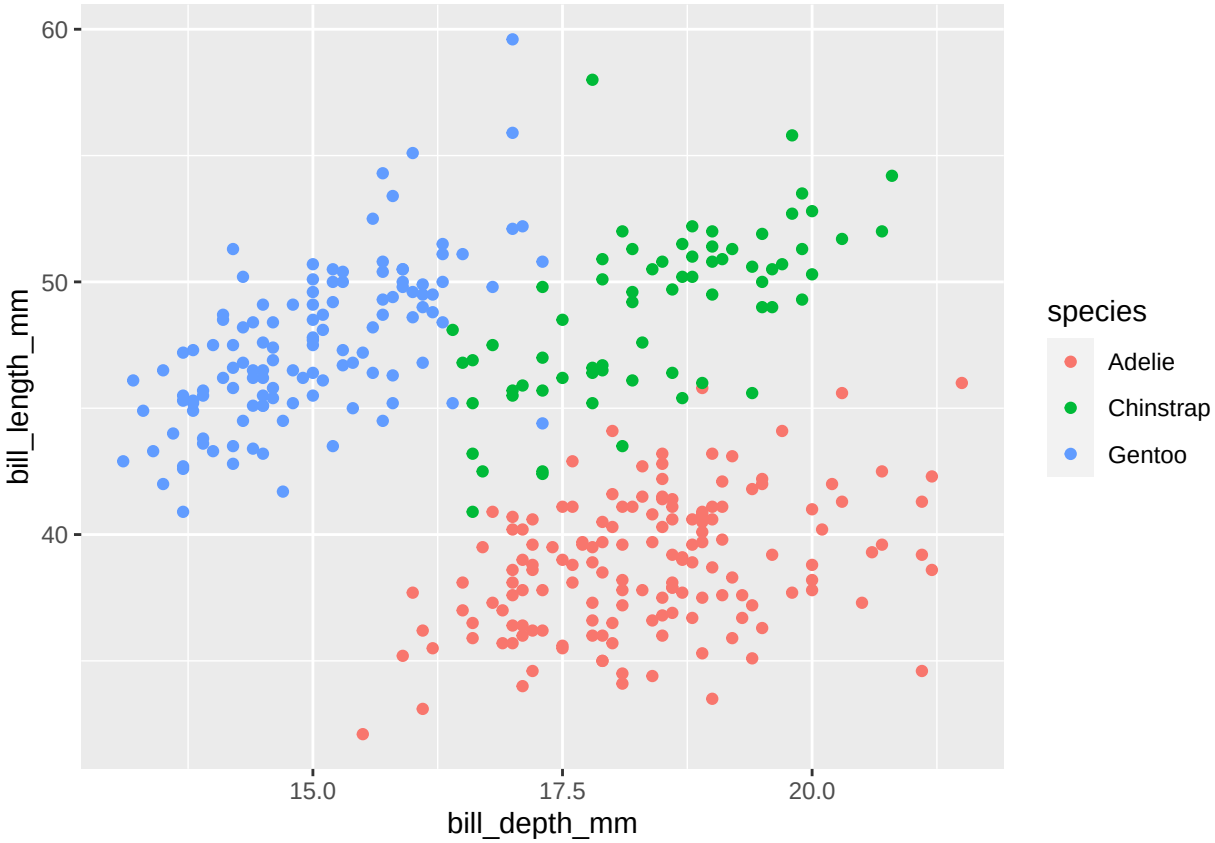
d. Represent each observation with a point

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



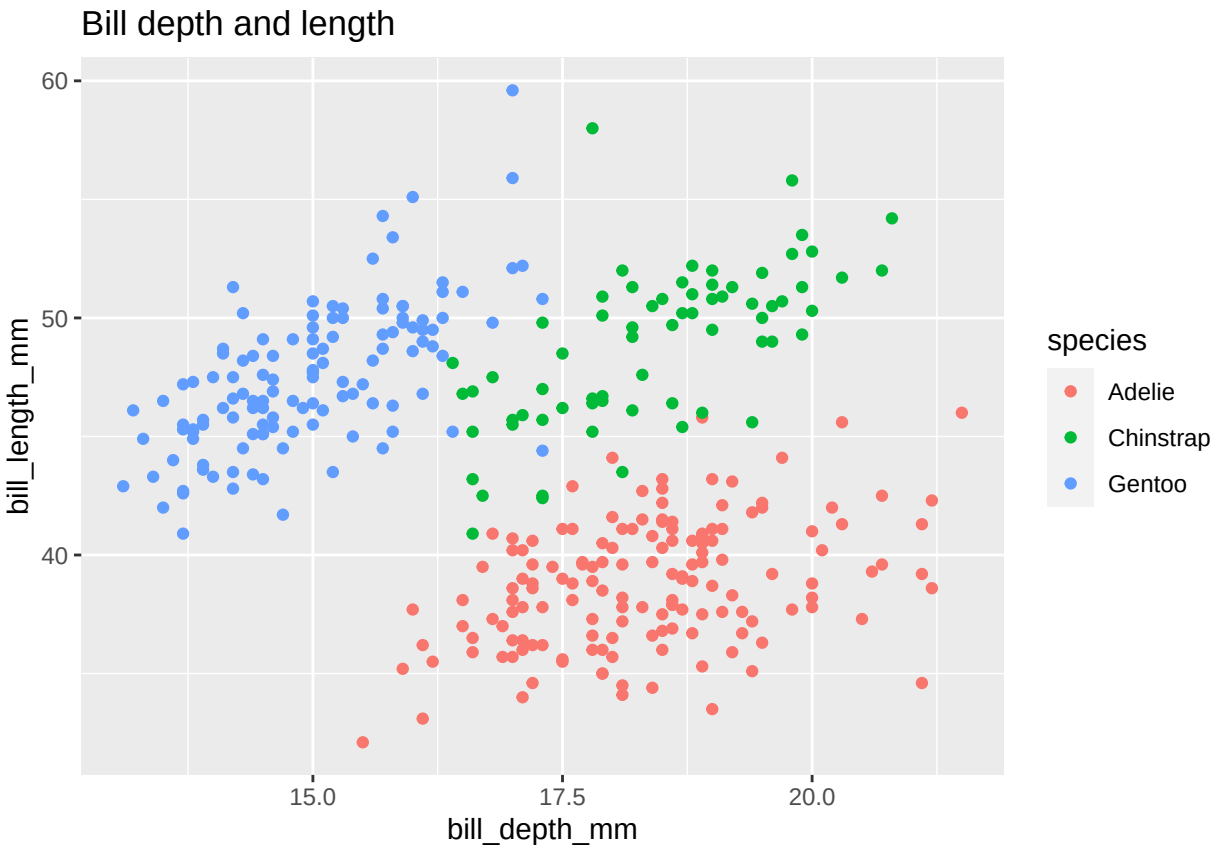
e. Map species to the colour of each point

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



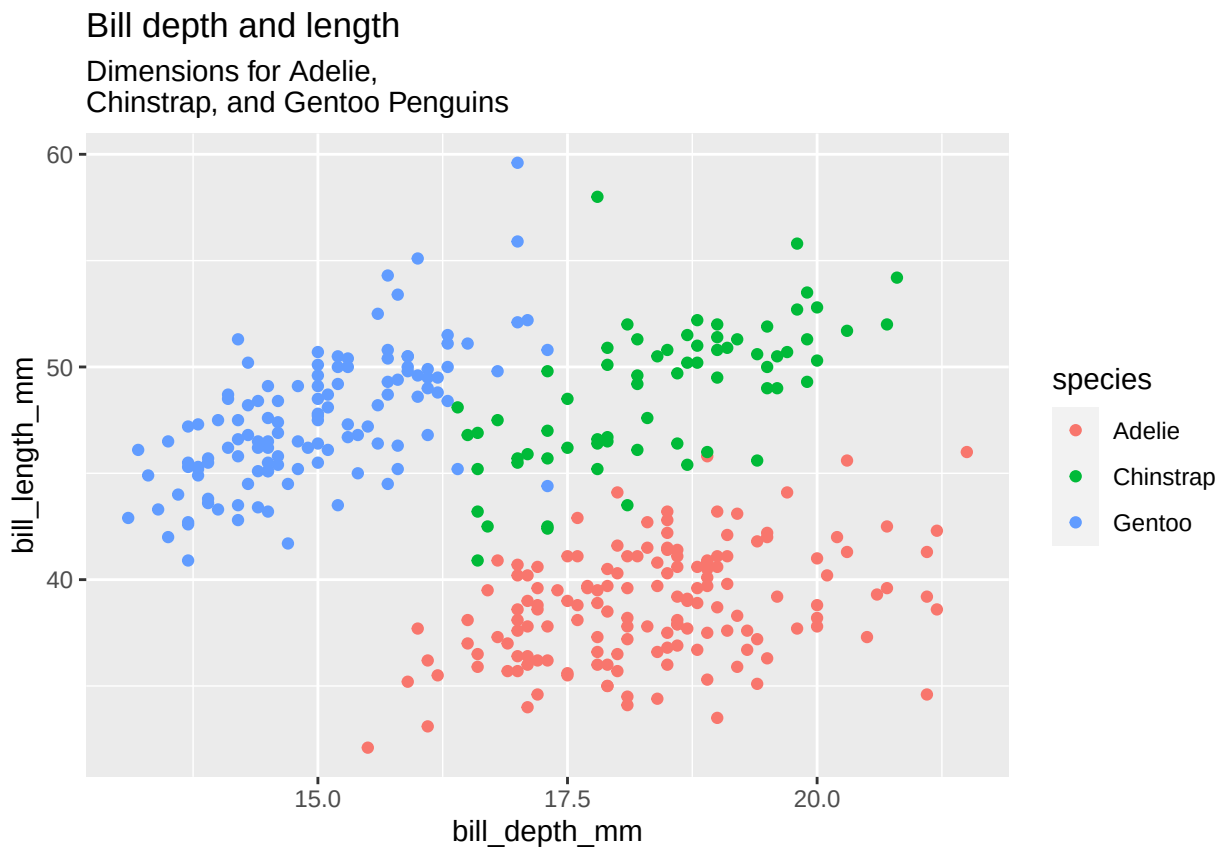
f. Title the plot “Bill depth and length”

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



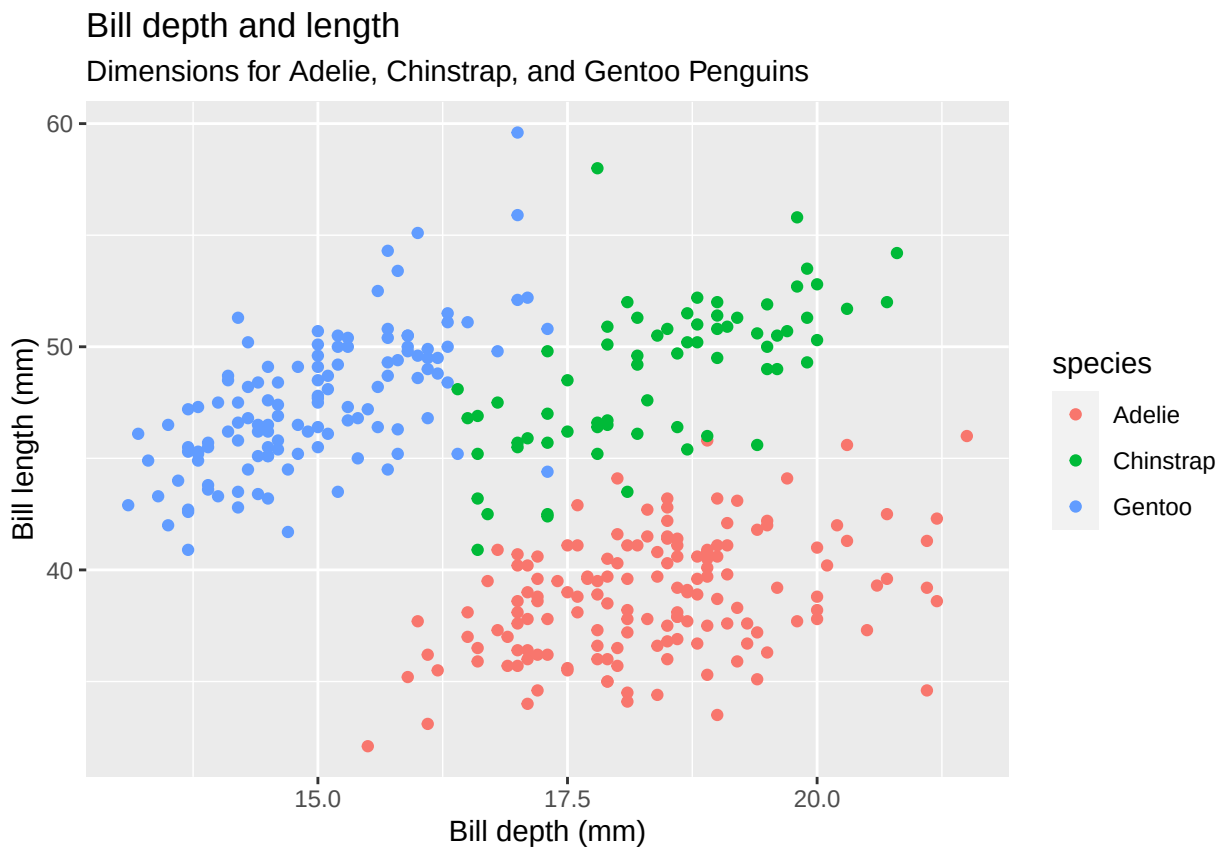
g. Add the subtitle "Dimensions for Adelie, Chinstrap, and Gentoo Penguins"

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



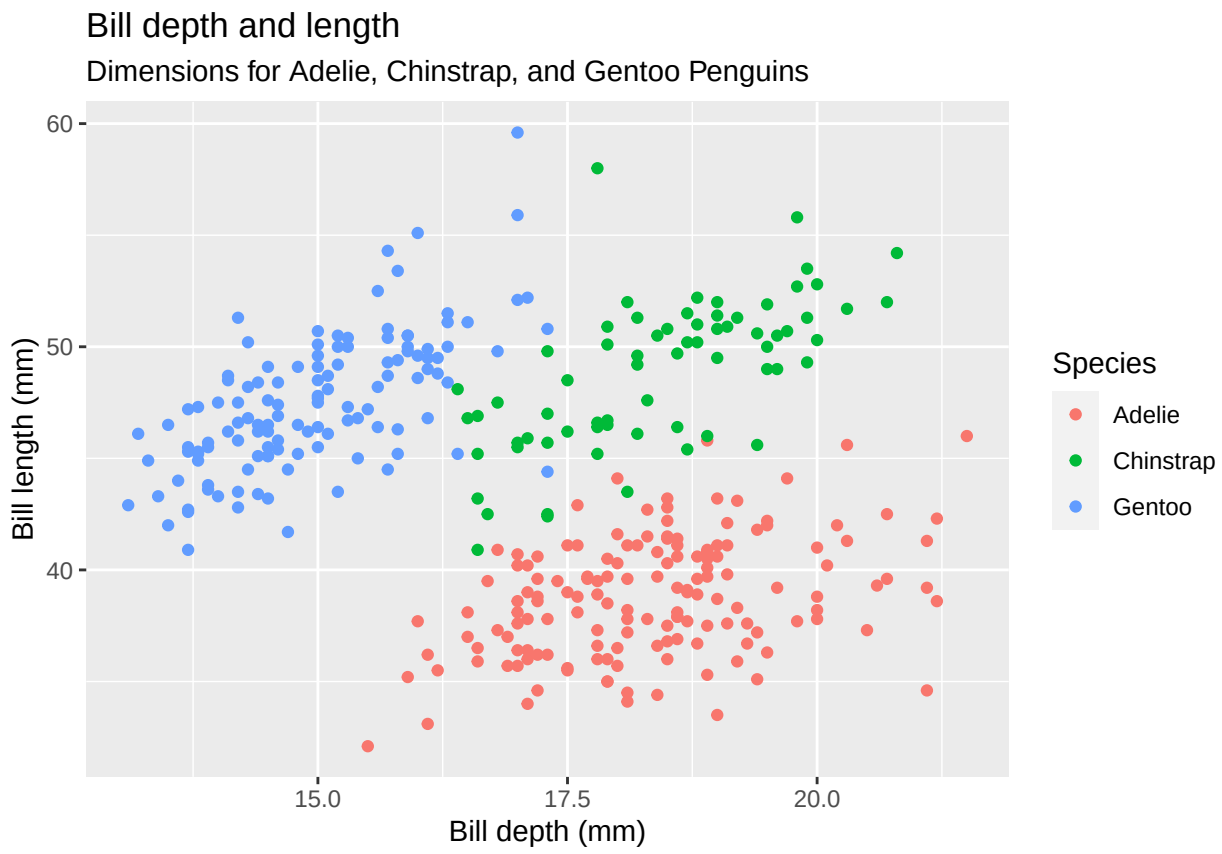
h. Label the x and y axes as “Bill depth (mm)” and “Bill length (mm)”, respectively

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



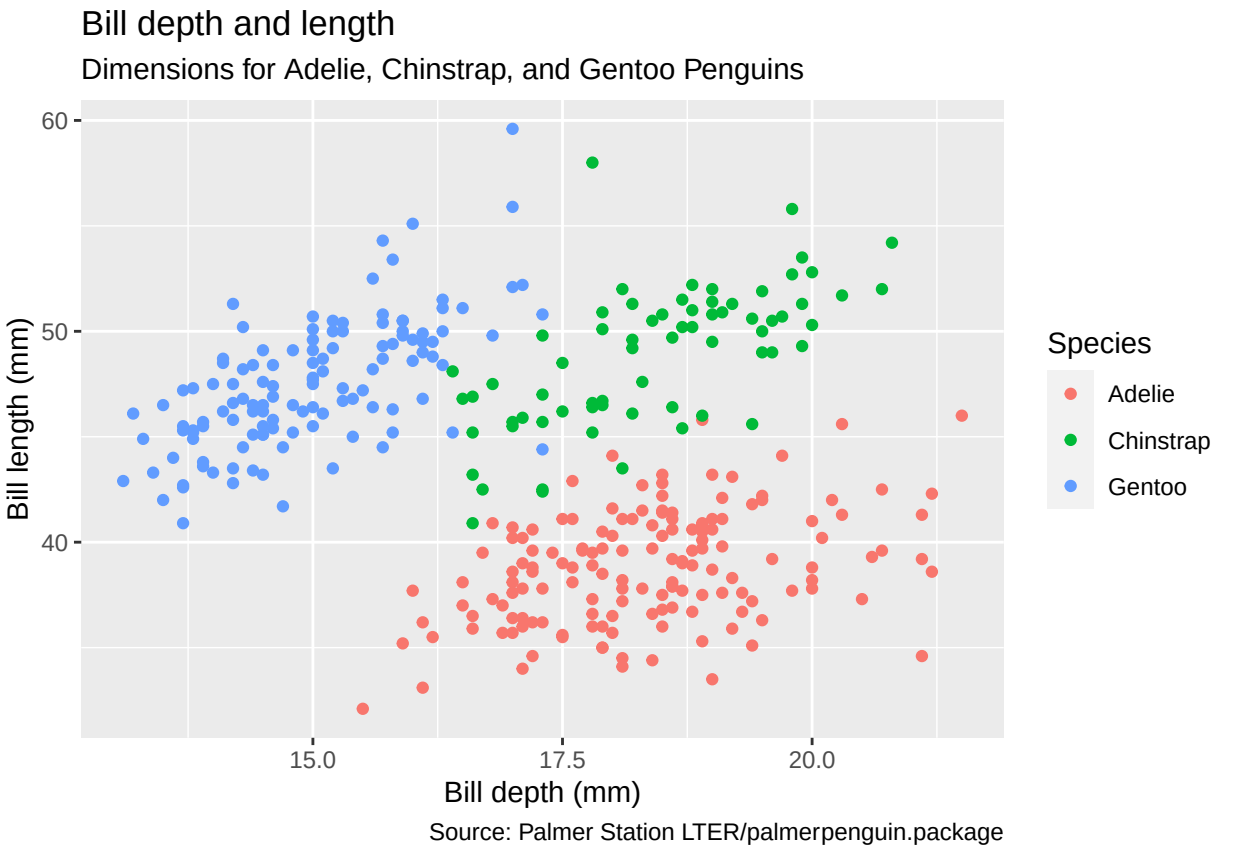
i. Label the legend “Species”

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



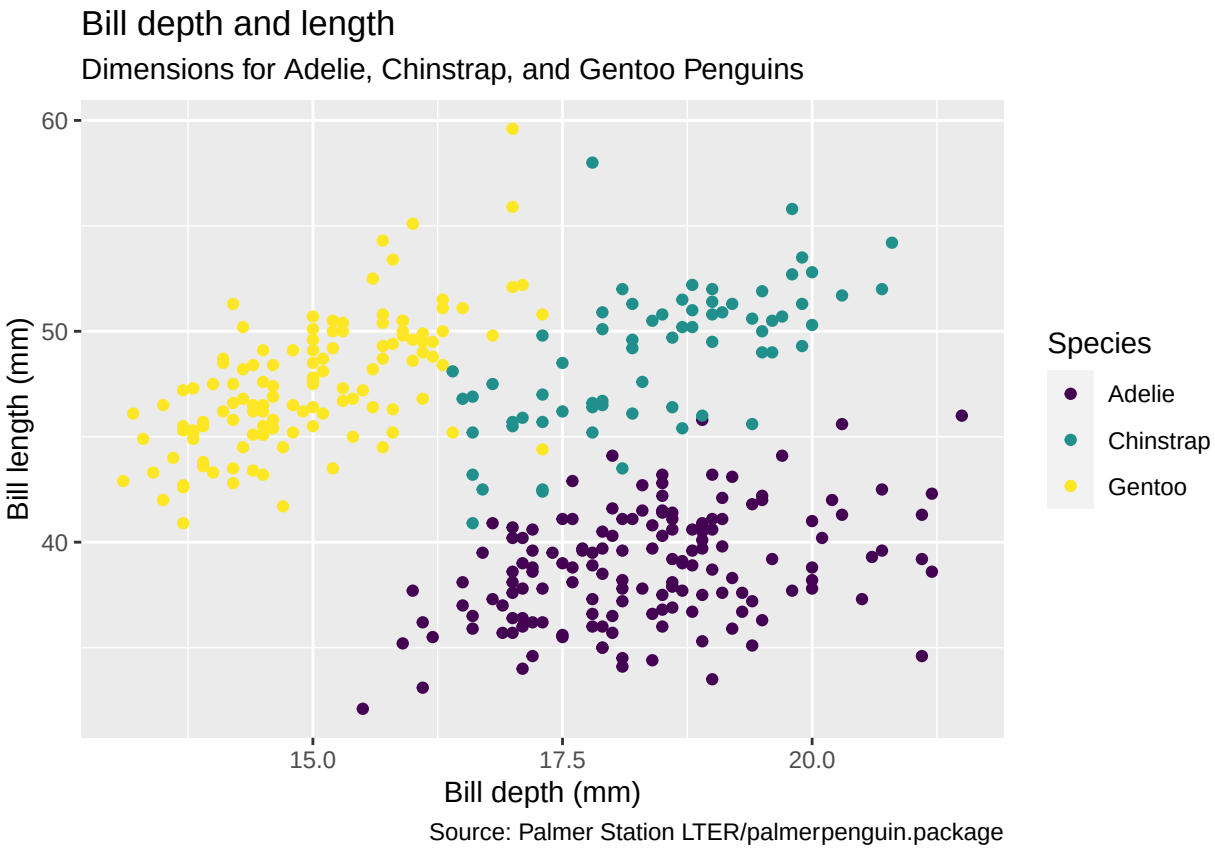
j. Add a caption for the data source

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



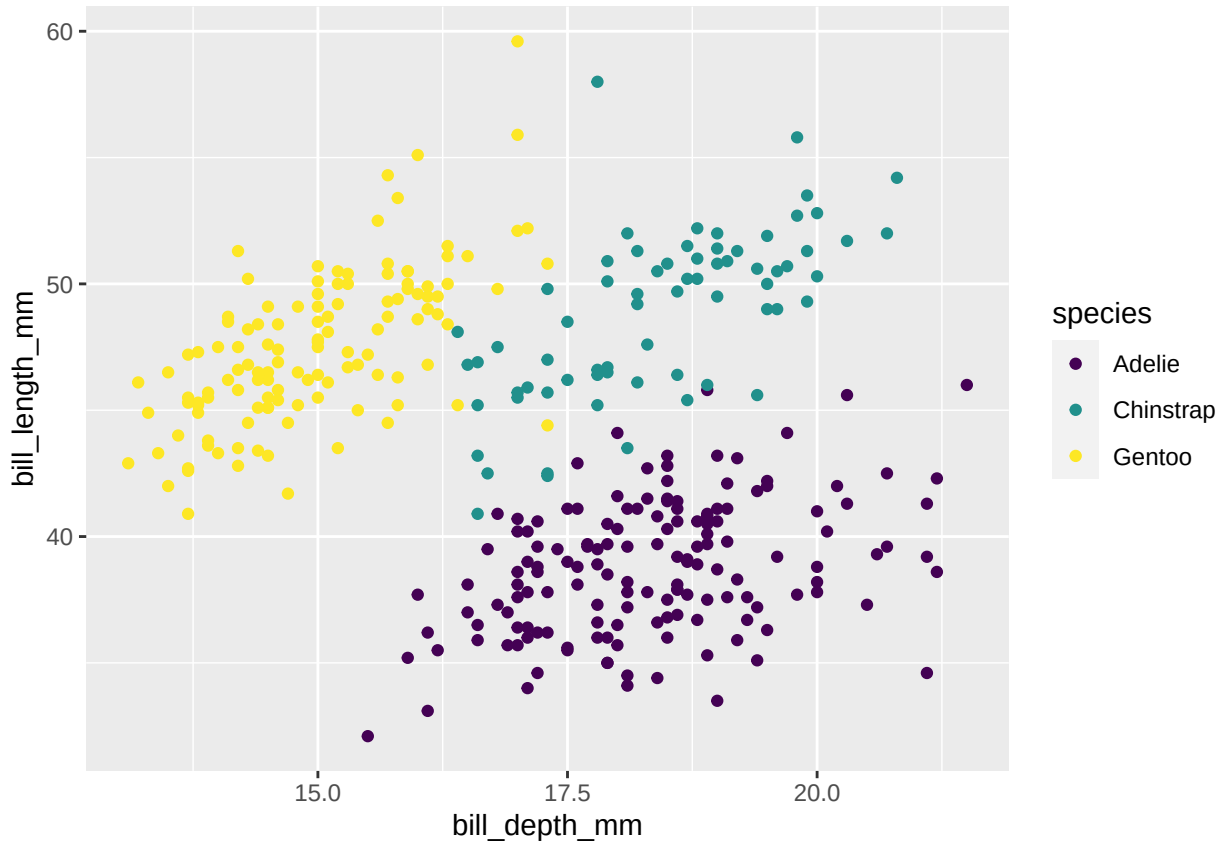
k. Finally, use a discrete colour scale that is designed to be perceived by viewers with common forms of colour blindness.

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



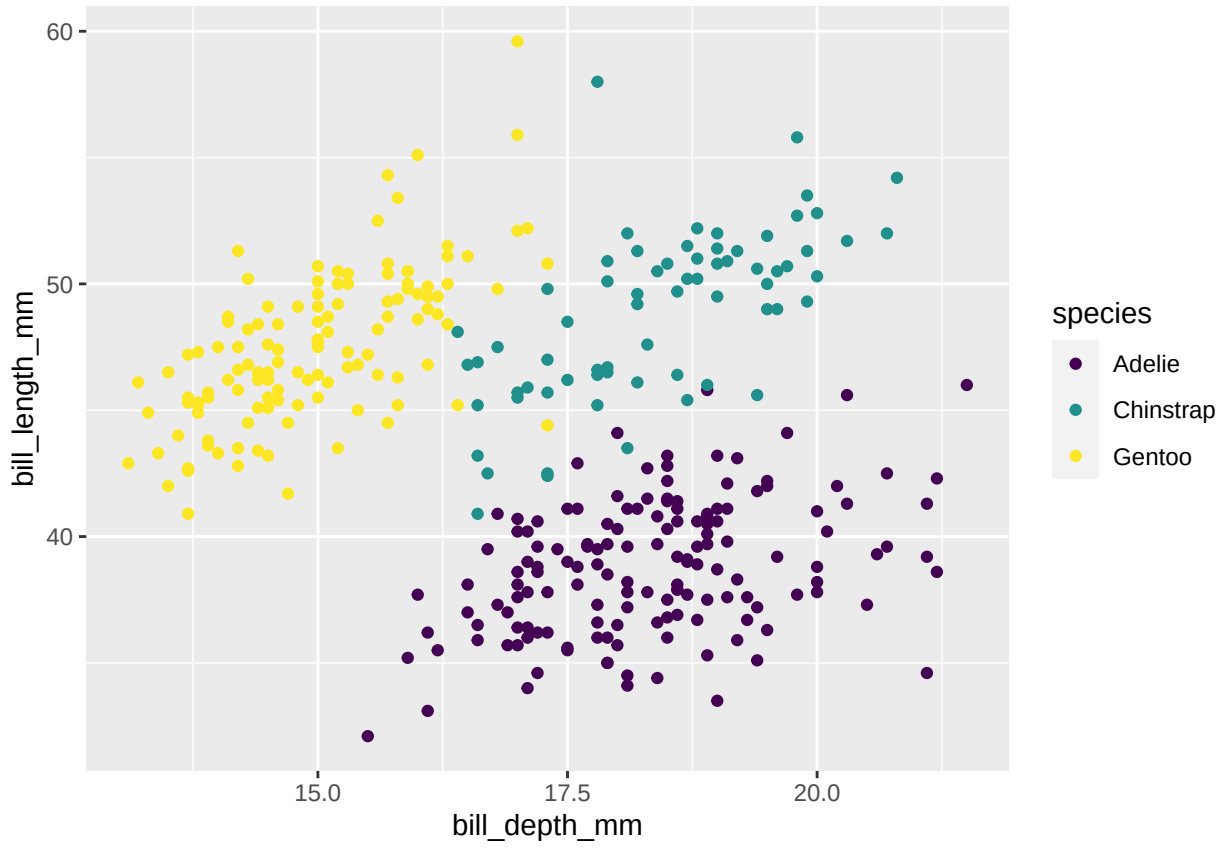
Can omit the names of first two arguments when building plots with ggplot()

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



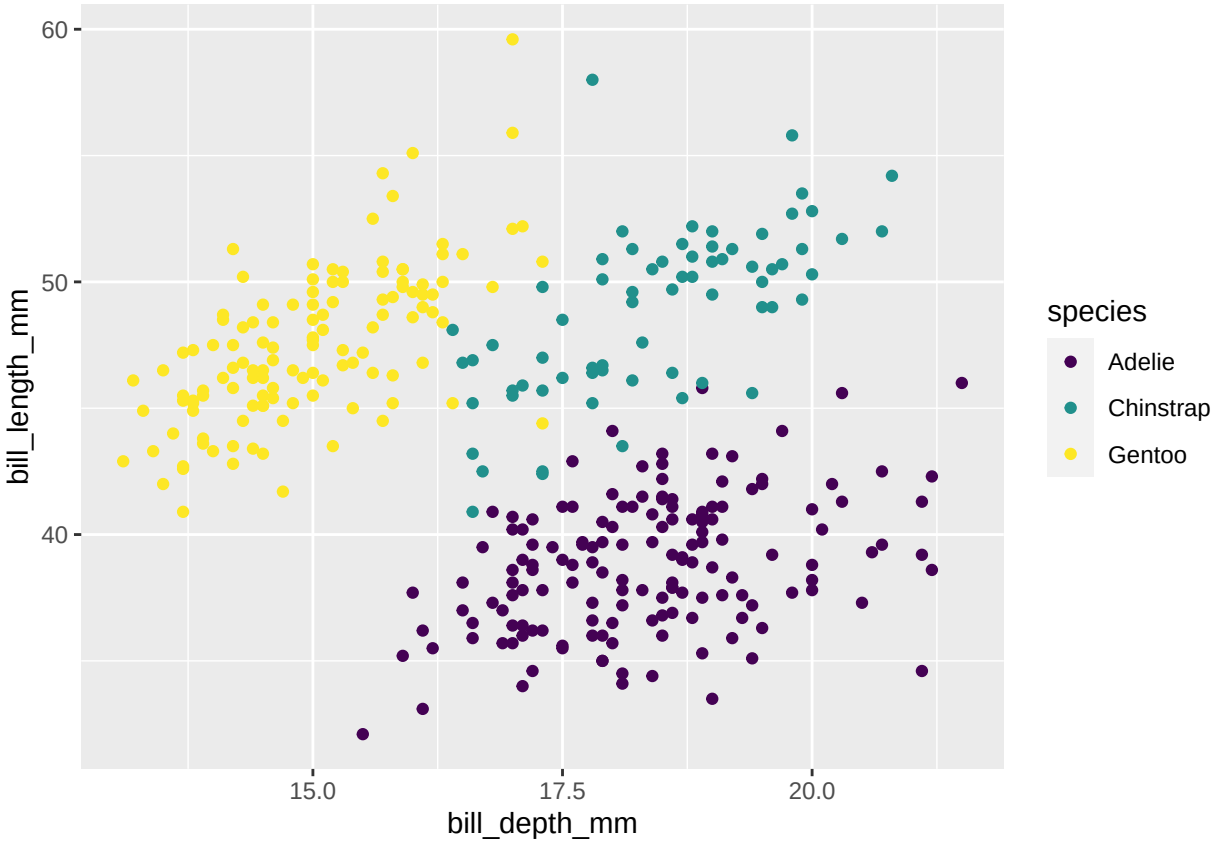
vs

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



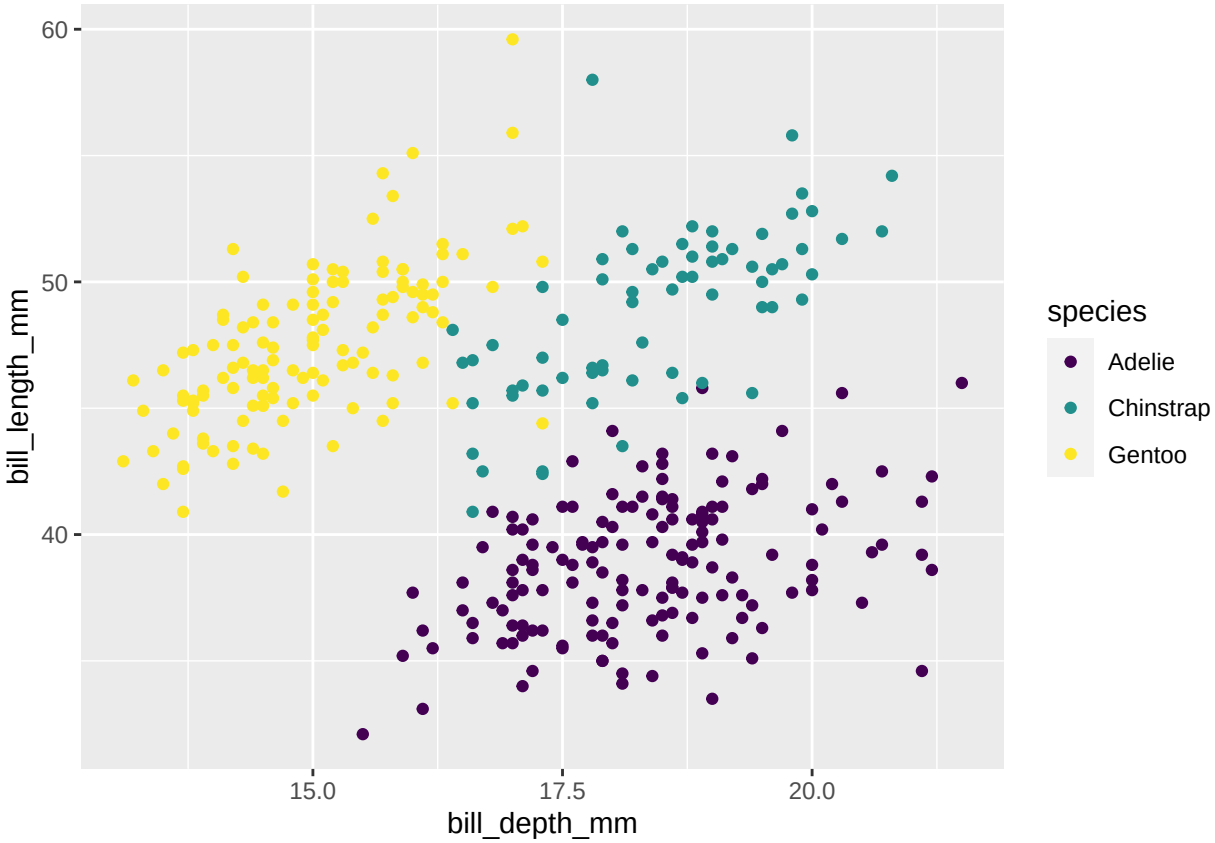
vs

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



Palmer Penguins: Colour

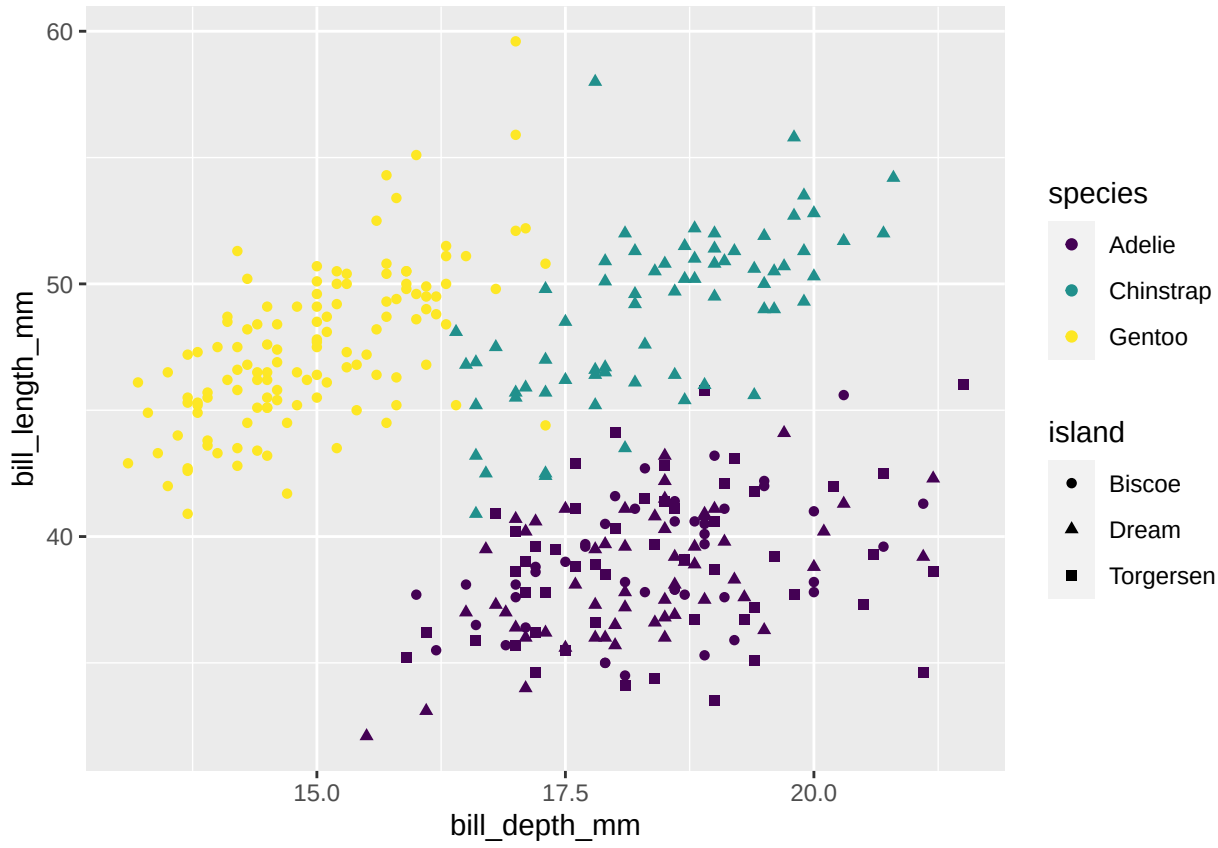
```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



Palmer Penguins: Shape

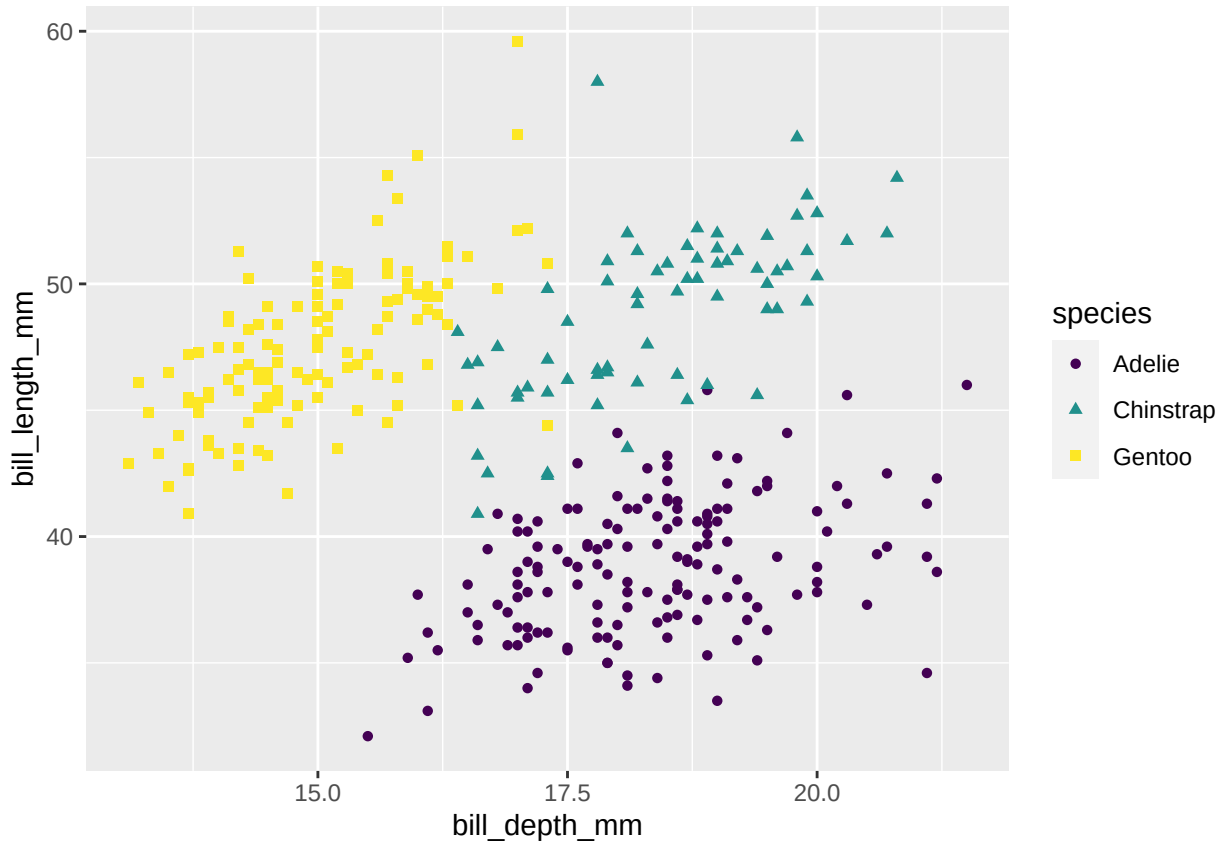
Island

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



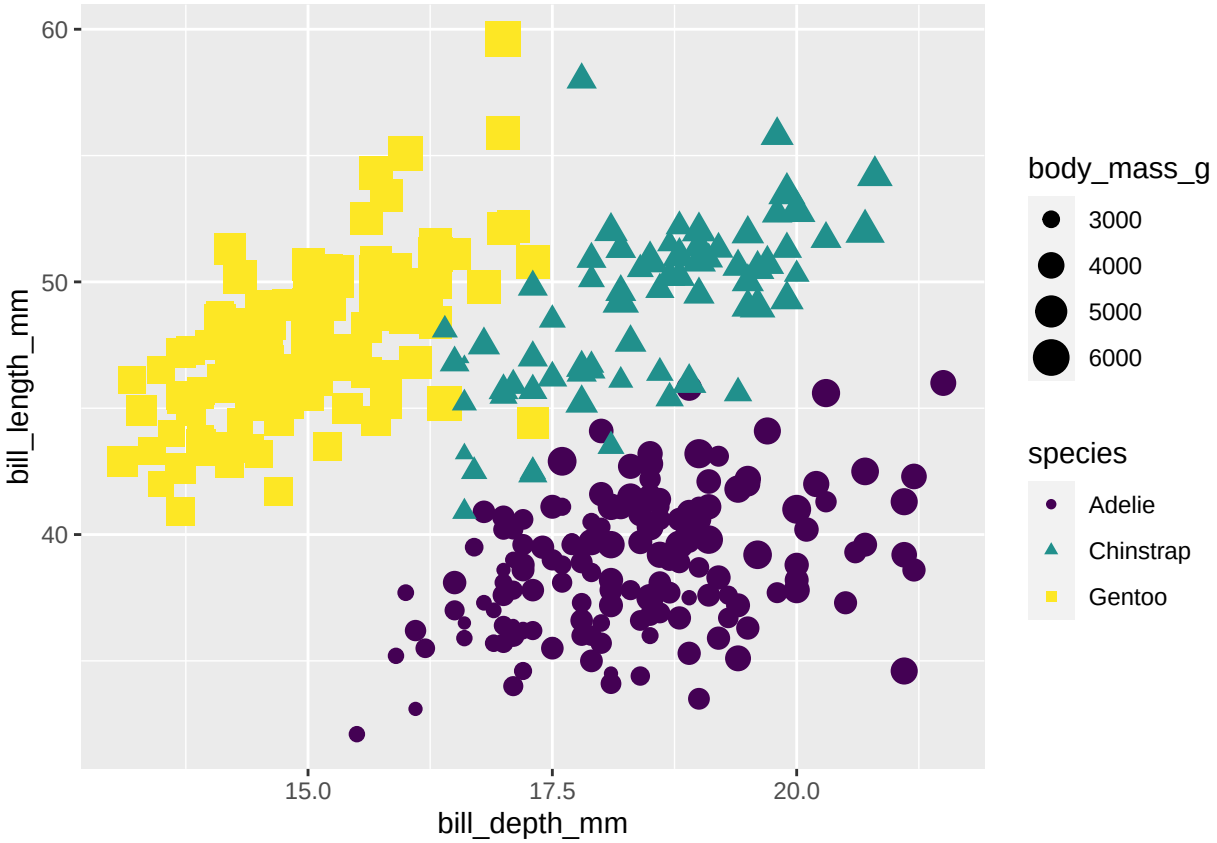
Species

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



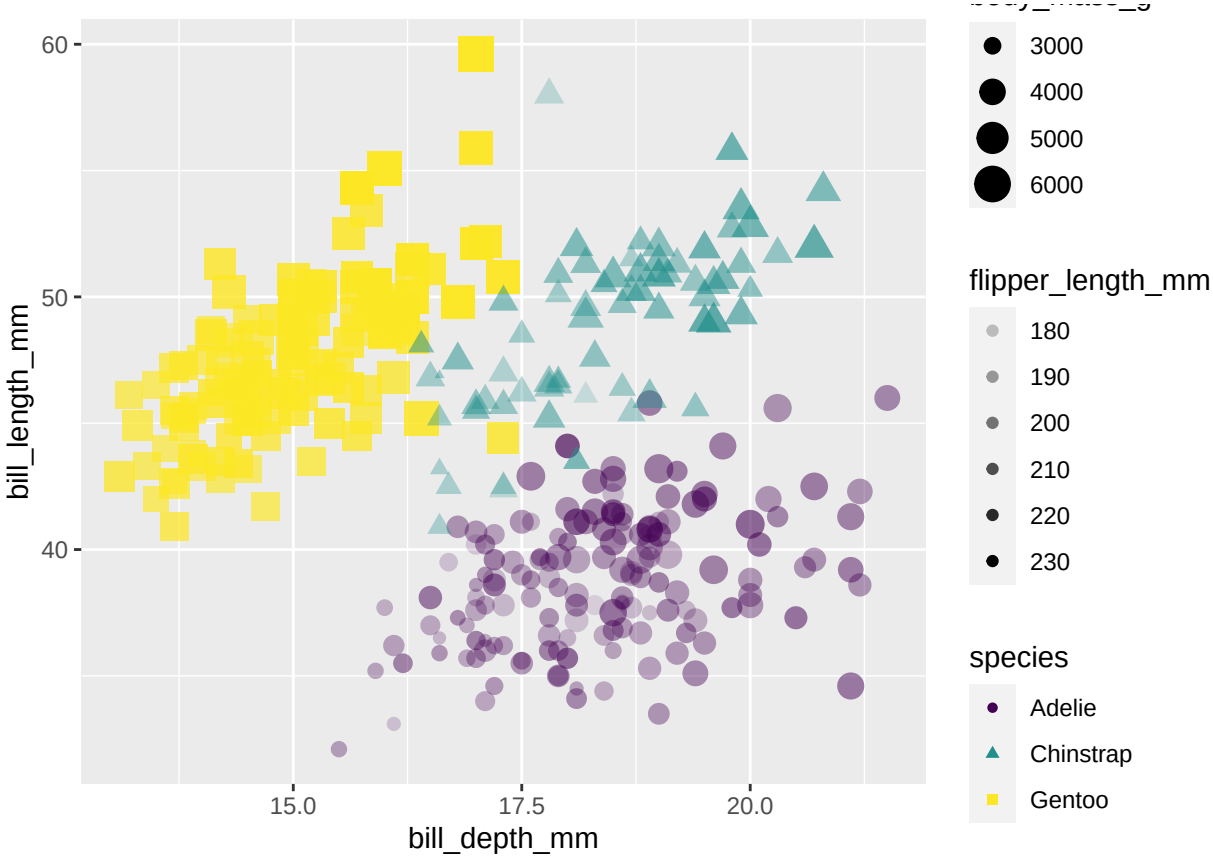
Size

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



Alpha

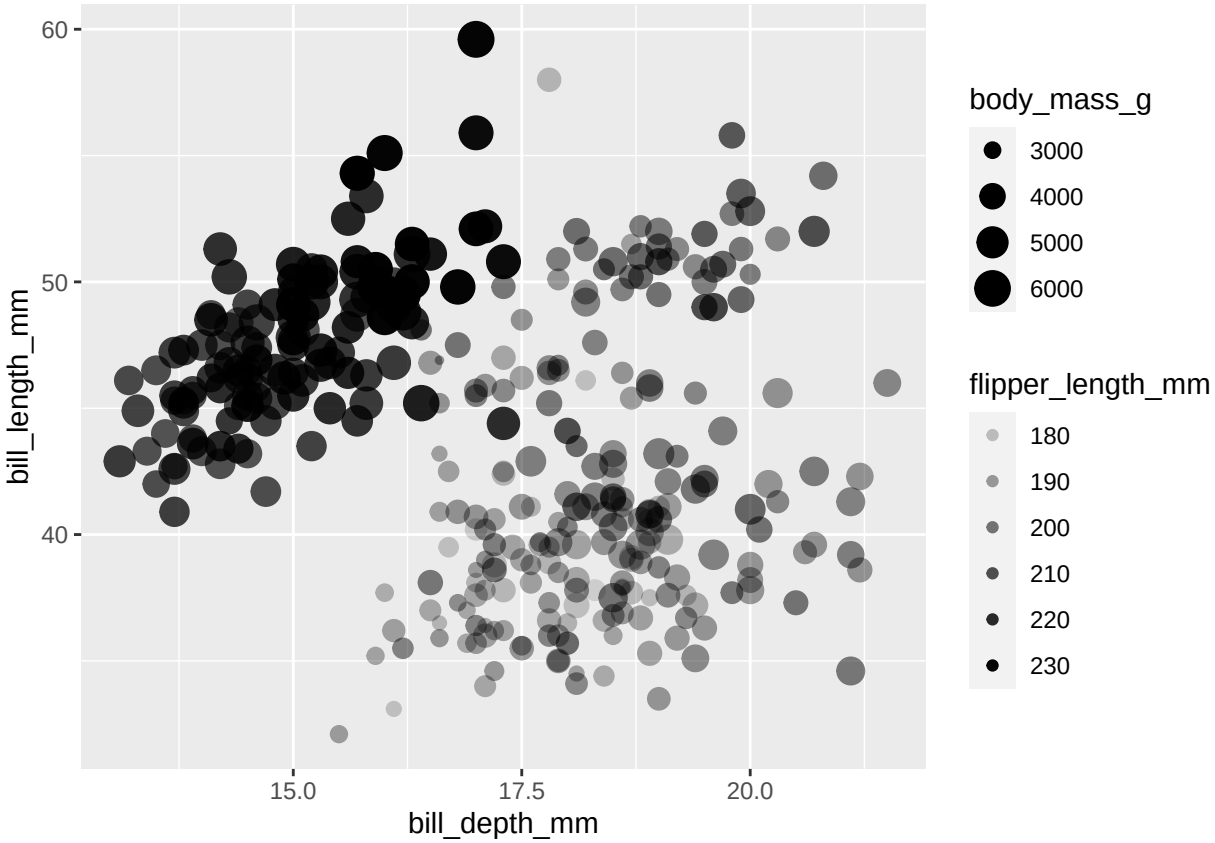
```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



Mapping vs Setting

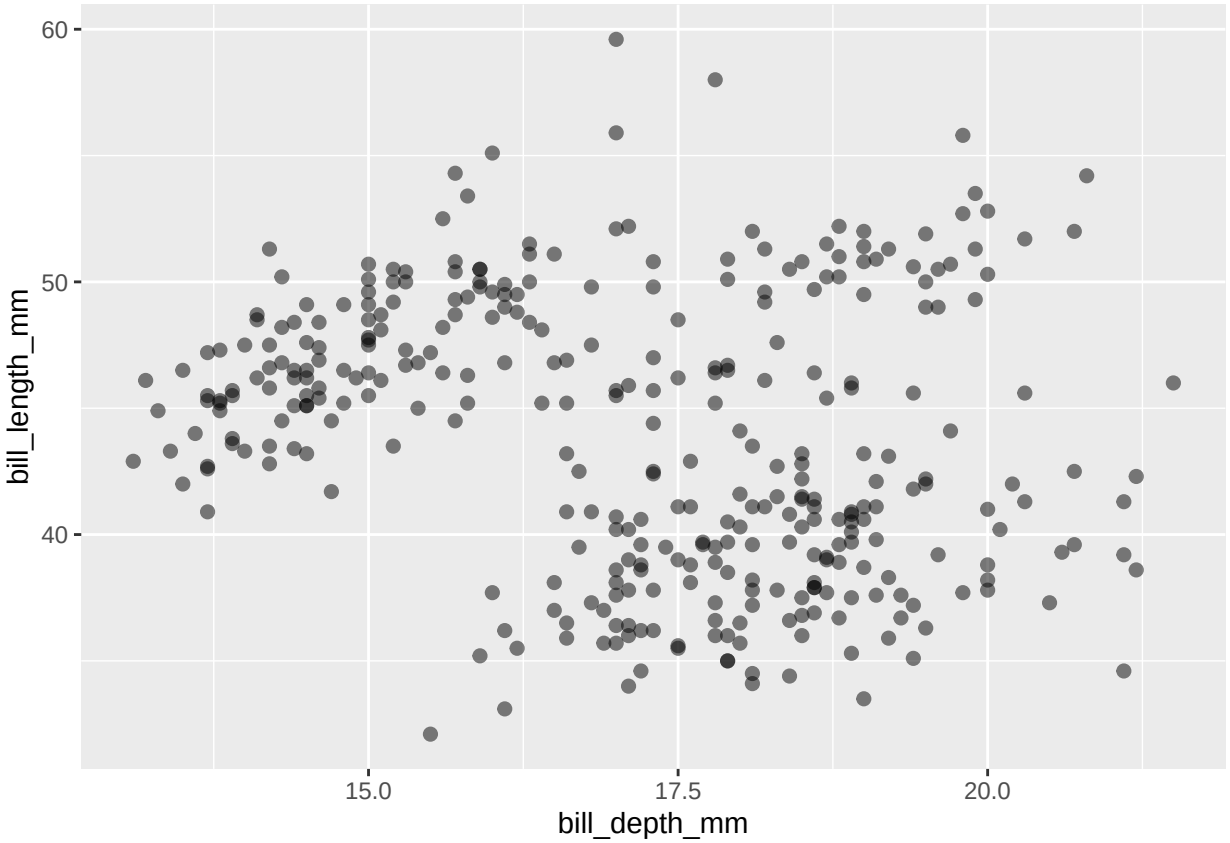
Mapping

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



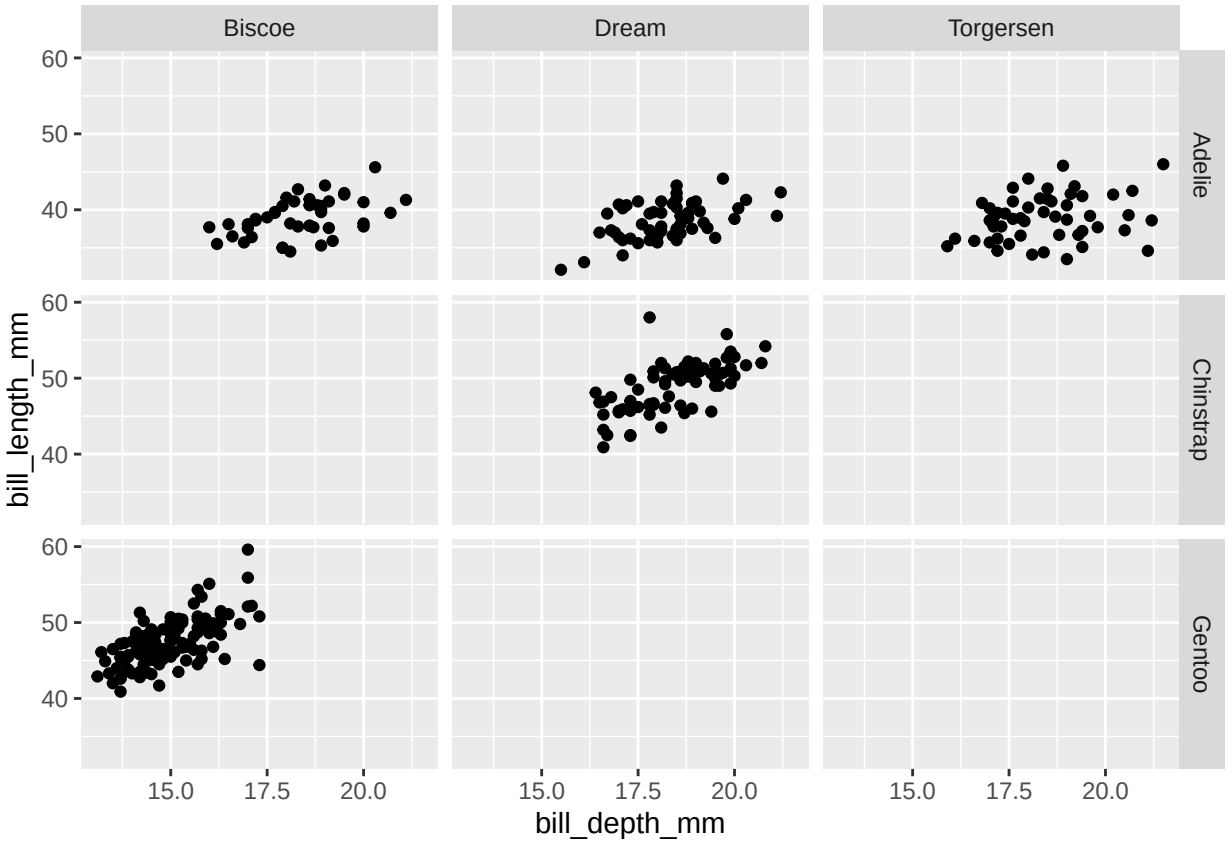
Setting

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```

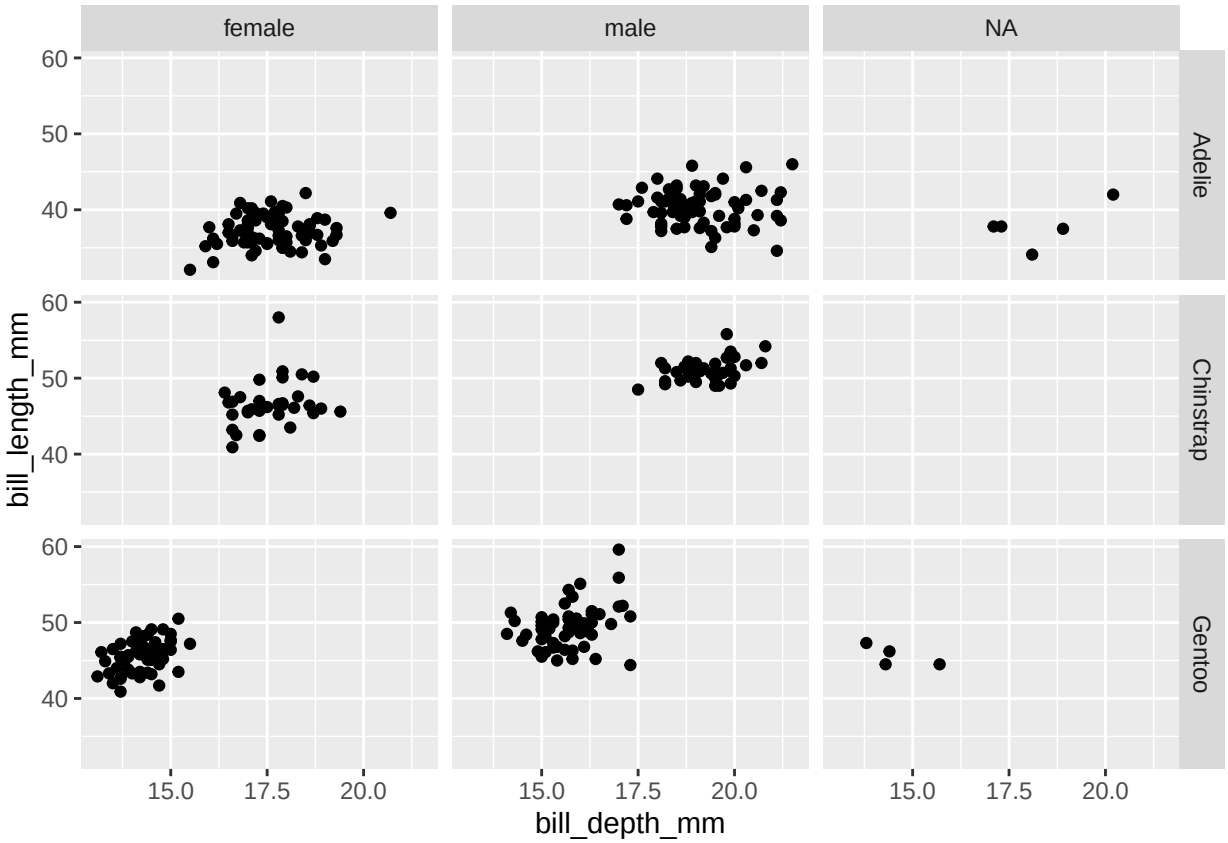


Faceting

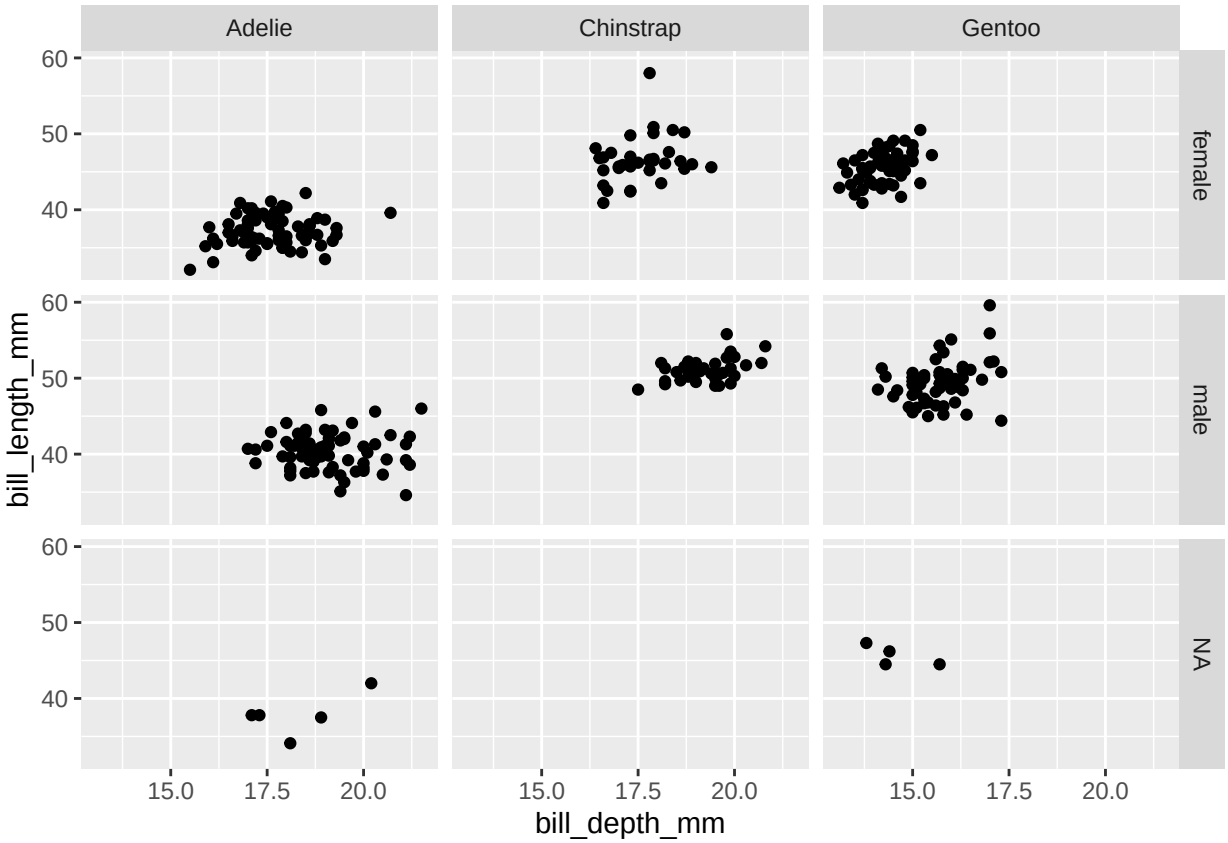
```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



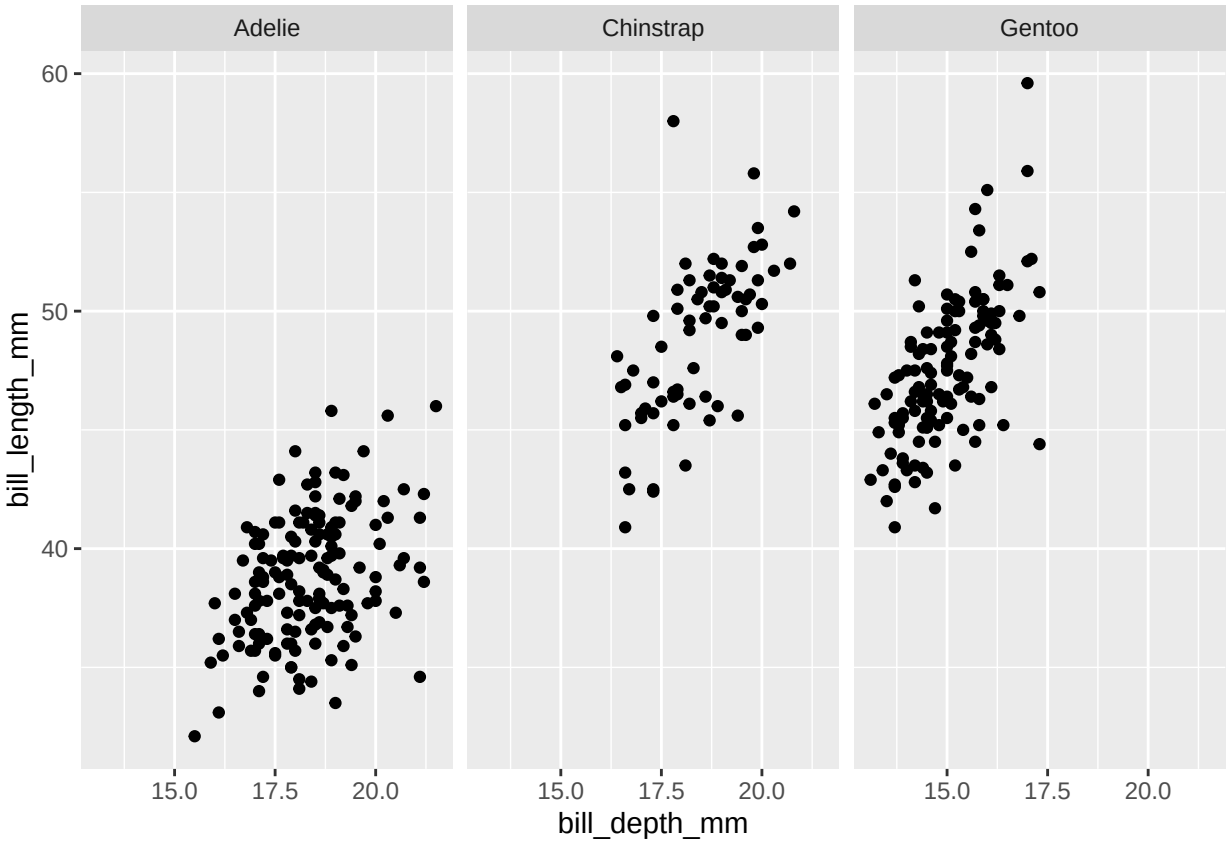

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



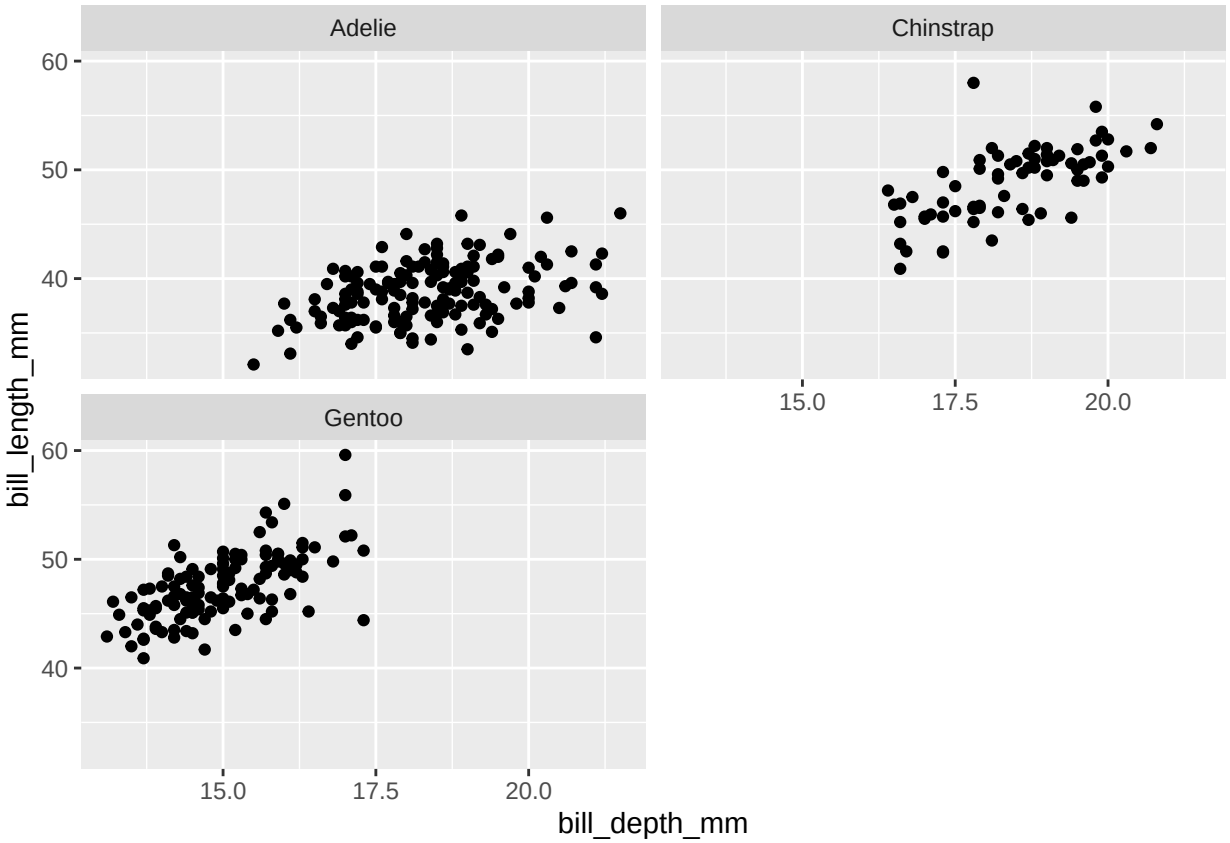

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



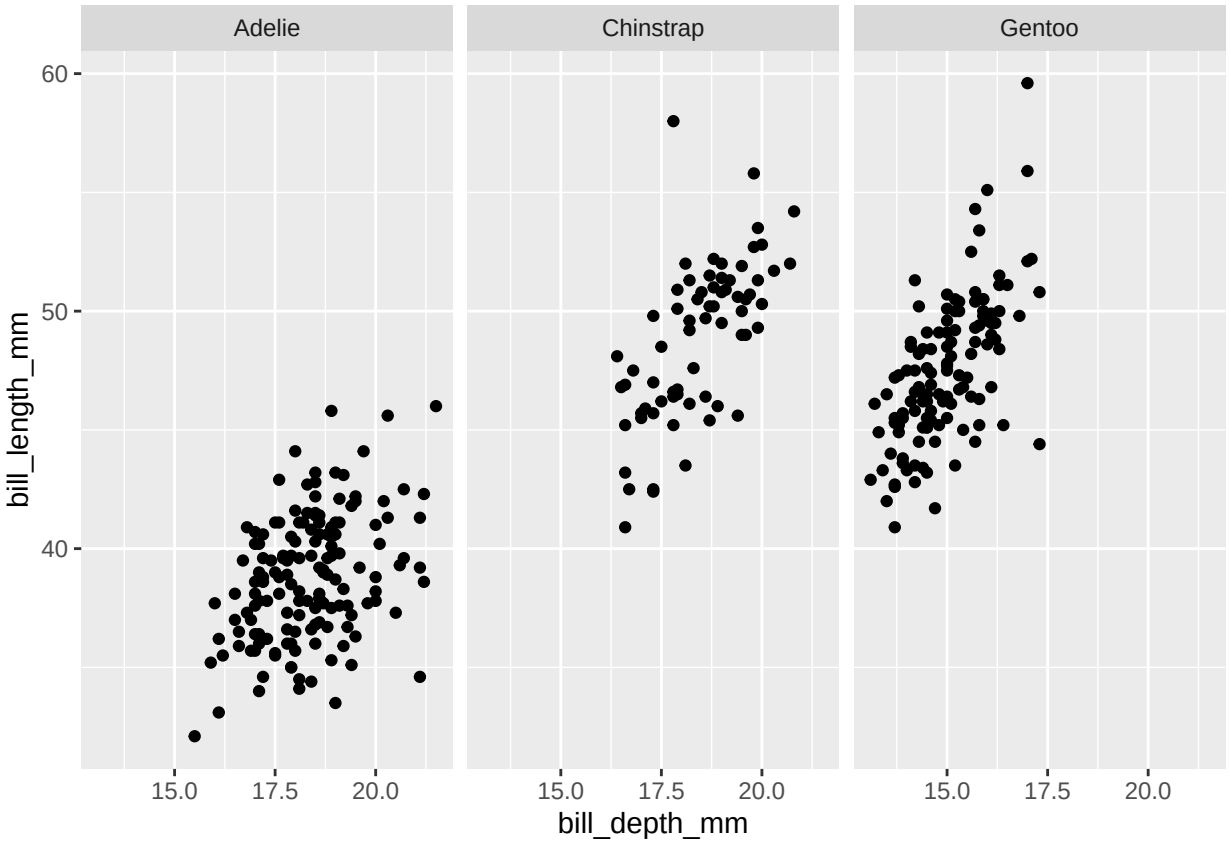

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



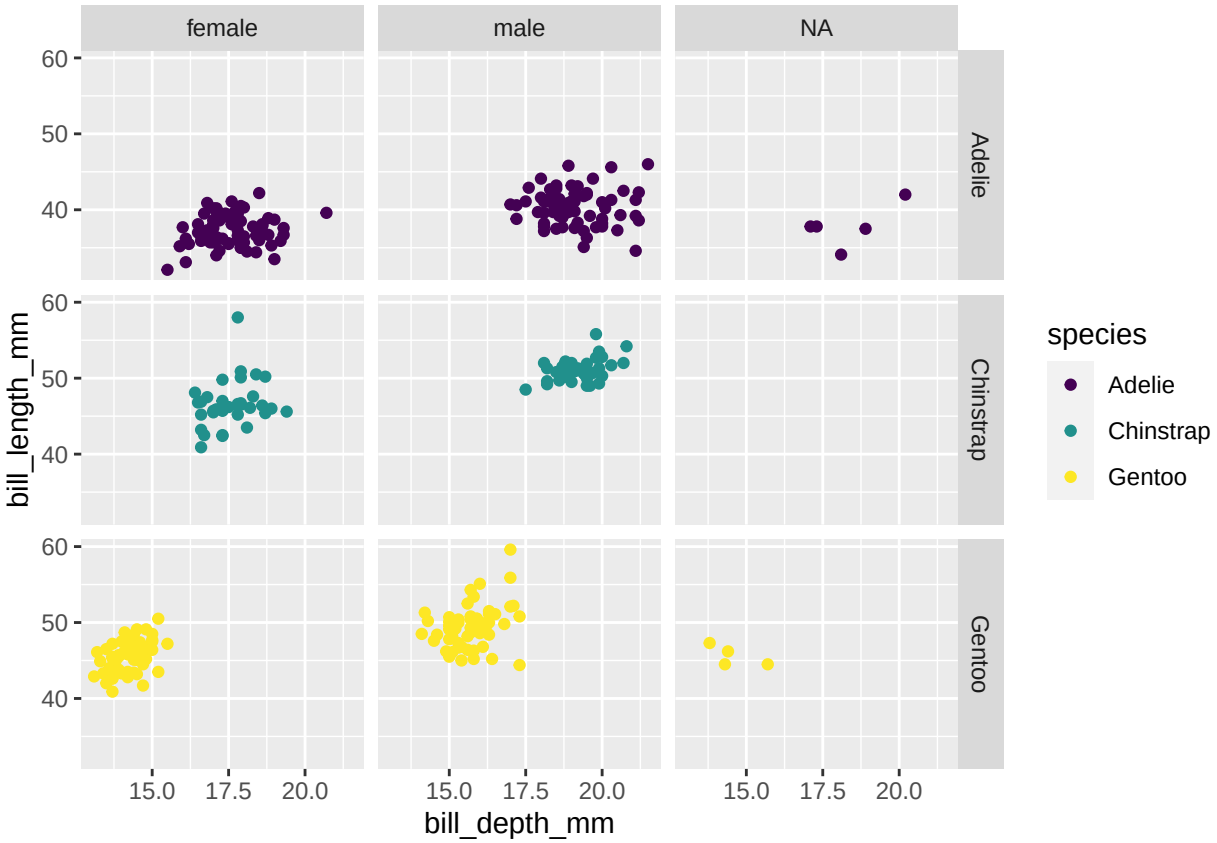

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



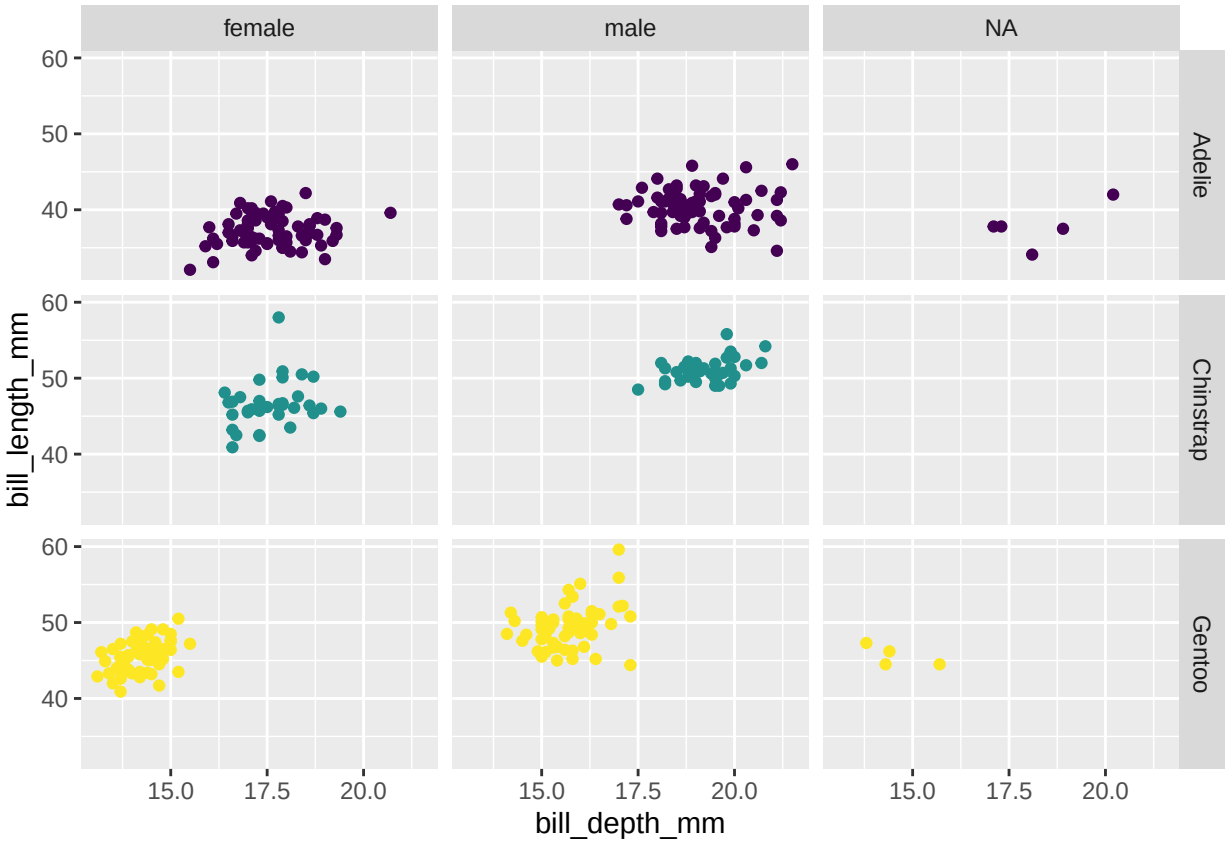

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```




```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```




```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



Data 2: Lending Club

Take a peek at data

```
## Warning: package 'openintro' was built under R version 4.2.3

## Loading required package: airports

## Warning: package 'airports' was built under R version 4.2.3

## Loading required package: cherryblossom

## Warning: package 'cherryblossom' was built under R version 4.2.3

## Loading required package: usdata

## Warning: package 'usdata' was built under R version 4.2.3

## Rows: 10,000
## Columns: 55
## $ emp_title                         <chr> "global config engineer ", "warehouse~
## $ emp_length                        <dbl> 3, 10, 3, 1, 10, NA, 10, 10, 10, 3, 1~
```

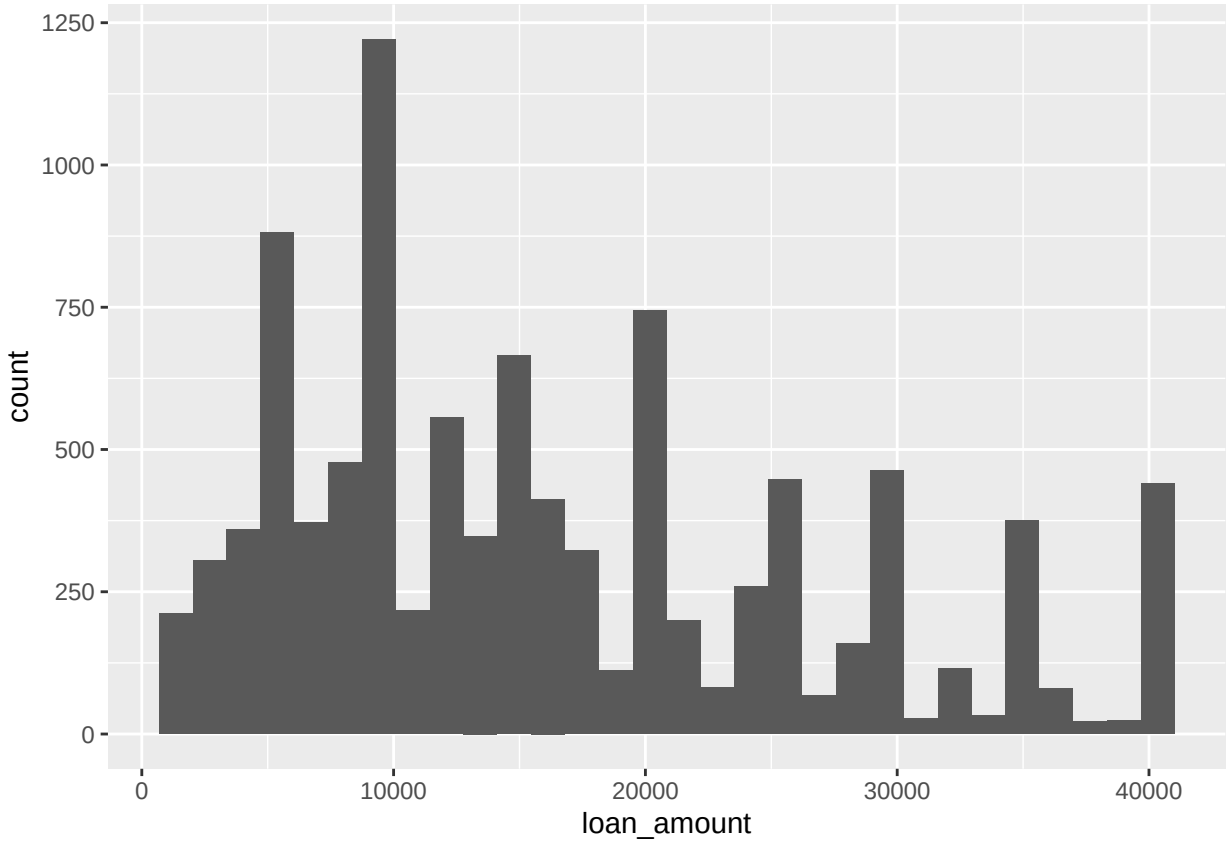
```
## $ state                              <fct> NJ, HI, WI, PA, CA, KY, MI, AZ, NV, I~
## $ homeownership                      <fct> MORTGAGE, RENT, RENT, RENT, RENT, OWN~
## $ annual_income                      <dbl> 90000, 40000, 40000, 30000, 35000, 34~
## $ verified_income                    <fct> Verified, Not Verified, Source Verifi~
## $ debt_to_income                     <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6.4~
## $ annual_income_joint                <dbl> NA, NA, NA, NA, 57000, NA, 155000, NA~
## $ verification_income_joint          <fct> , , , , Verified, , Not Verified, , ,~
## $ debt_to_income_joint               <dbl> NA, NA, NA, NA, 37.66, NA, 13.12, NA,~
## $ delinq_2y                          <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0~
## $ months_since_last_delinq           <int> 38, NA, 28, NA, NA, 3, NA, 19, 18, NA~
## $ earliest_credit_line               <dbl> 2001, 1996, 2006, 2007, 2008, 1990, 2~
## $ inquiries_last_12m                 <int> 6, 1, 4, 0, 7, 6, 1, 1, 3, 0, 4, 4, 8~
## $ total_credit_lines                 <int> 28, 30, 31, 4, 22, 32, 12, 30, 35, 9,~
## $ open_credit_lines                  <int> 10, 14, 10, 4, 16, 12, 10, 15, 21, 6,~
## $ total_credit_limit                 <int> 70795, 28800, 24193, 25400, 69839, 42~
## $ total_credit_utilized              <int> 38767, 4321, 16000, 4997, 52722, 3898~
## $ num_collections_last_12m           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ num_historical_failed_to_pay       <int> 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ months_since_90d_late              <int> 38, NA, 28, NA, NA, 60, NA, 71, 18, N~
## $ current_accounts_delinq            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ total_collection_amount_ever       <int> 1250, 0, 432, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ current_installment_accounts       <int> 2, 0, 1, 1, 1, 0, 2, 2, 6, 1, 2, 1, 2~
## $ accounts_opened_24m                <int> 5, 11, 13, 1, 6, 2, 1, 4, 10, 5, 6, 7~
## $ months_since_last_credit_inquiry   <int> 5, 8, 7, 15, 4, 5, 9, 7, 4, 17, 3, 4,~
## $ num_satisfactory_accounts          <int> 10, 14, 10, 4, 16, 12, 10, 15, 21, 6,~
## $ num_accounts_120d_past_due         <int> 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, ~
## $ num_accounts_30d_past_due          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ num_active_debit_accounts          <int> 2, 3, 3, 2, 10, 1, 3, 5, 11, 3, 2, 2,~
## $ total_debit_limit                  <int> 11100, 16500, 4300, 19400, 32700, 272~
## $ num_total_cc_accounts              <int> 14, 24, 14, 3, 20, 27, 8, 16, 19, 7, ~
## $ num_open_cc_accounts               <int> 8, 14, 8, 3, 15, 12, 7, 12, 14, 5, 8,~
## $ num_cc_carrying_balance            <int> 6, 4, 6, 2, 13, 5, 6, 10, 14, 3, 5, 3~
## $ num_mort_accounts                  <int> 1, 0, 0, 0, 0, 3, 2, 7, 2, 0, 2, 3, 3~
## $ account_never_delinq_percent       <dbl> 92.9, 100.0, 93.5, 100.0, 100.0, 78.1~
## $ tax_liens                          <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ public_record_bankrupt             <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ loan_purpose                       <fct> moving, debt_consolidation, other, de~
## $ application_type                   <fct> individual, individual, individual, i~
## $ loan_amount                        <int> 28000, 5000, 2000, 21600, 23000, 5000~
## $ term                               <dbl> 60, 36, 36, 36, 36, 36, 60, 60, 36, 3~
## $ interest_rate                      <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6.7~
## $ installment                        <dbl> 652.53, 167.54, 71.40, 664.19, 786.87~
## $ grade                              <fct> C, C, D, A, C, A, C, B, C, A, C, B, C~
## $ sub_grade                          <fct> C3, C1, D1, A3, C3, A3, C2, B5, C2, A~
## $ issue_month                        <fct> Mar-2018, Feb-2018, Feb-2018, Jan-201~
## $ loan_status                        <fct> Current, Current, Current, Current, C~
## $ initial_listing_status             <fct> whole, whole, fractional, whole, whol~
## $ disbursement_method                <fct> Cash, Cash, Cash, Cash, Cash, Cash, C~
## $ balance                            <dbl> 27015.86, 4651.37, 1824.63, 18853.26,~
## $ paid_total                         <dbl> 1999.330, 499.120, 281.800, 3312.890,~
## $ paid_principal                     <dbl> 984.14, 348.63, 175.37, 2746.74, 1569~
## $ paid_interest                      <dbl> 1015.19, 150.49, 106.43, 566.15, 754.~
## $ paid_late_fees                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Selected variables

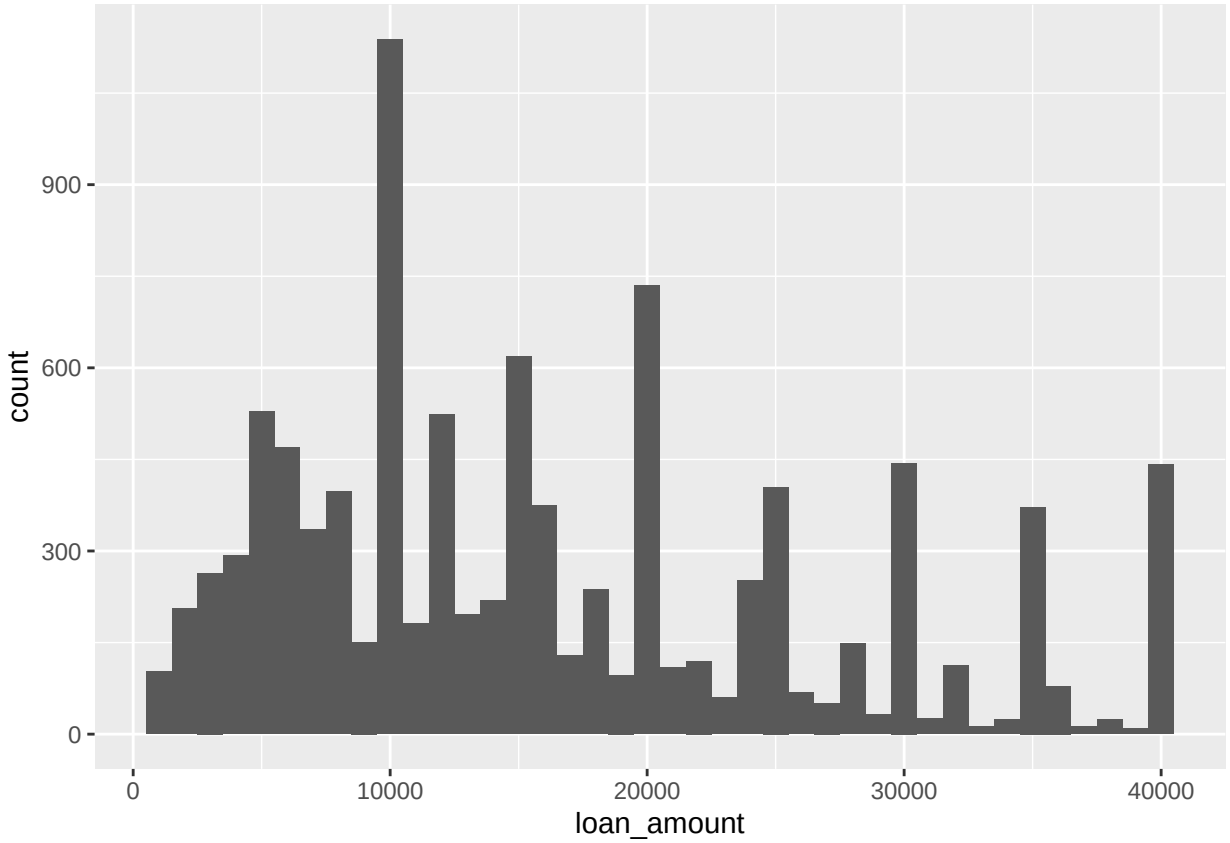
```
## Rows: 10,000
## Columns: 8
## $ loan_amount    <int> 28000, 5000, 2000, 21600, 23000, 5000, 24000, 20000, 20~
## $ interest_rate  <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6.72, 13.59, 11.99, 1~
## $ term           <dbl> 60, 36, 36, 36, 36, 36, 60, 60, 36, 36, 60, 60, 36, 60,~
## $ grade          <fct> C, C, D, A, C, A, C, B, C, A, C, B, C, B, D, D, D, F, E~
## $ 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, 110000~
## $ homeownership  <fct> MORTGAGE, RENT, RENT, RENT, RENT, OWN, MORTGAGE, MORTGA~
## $ debt_to_income <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6.46, 23.66, 16.19, 3~
```

Histogram

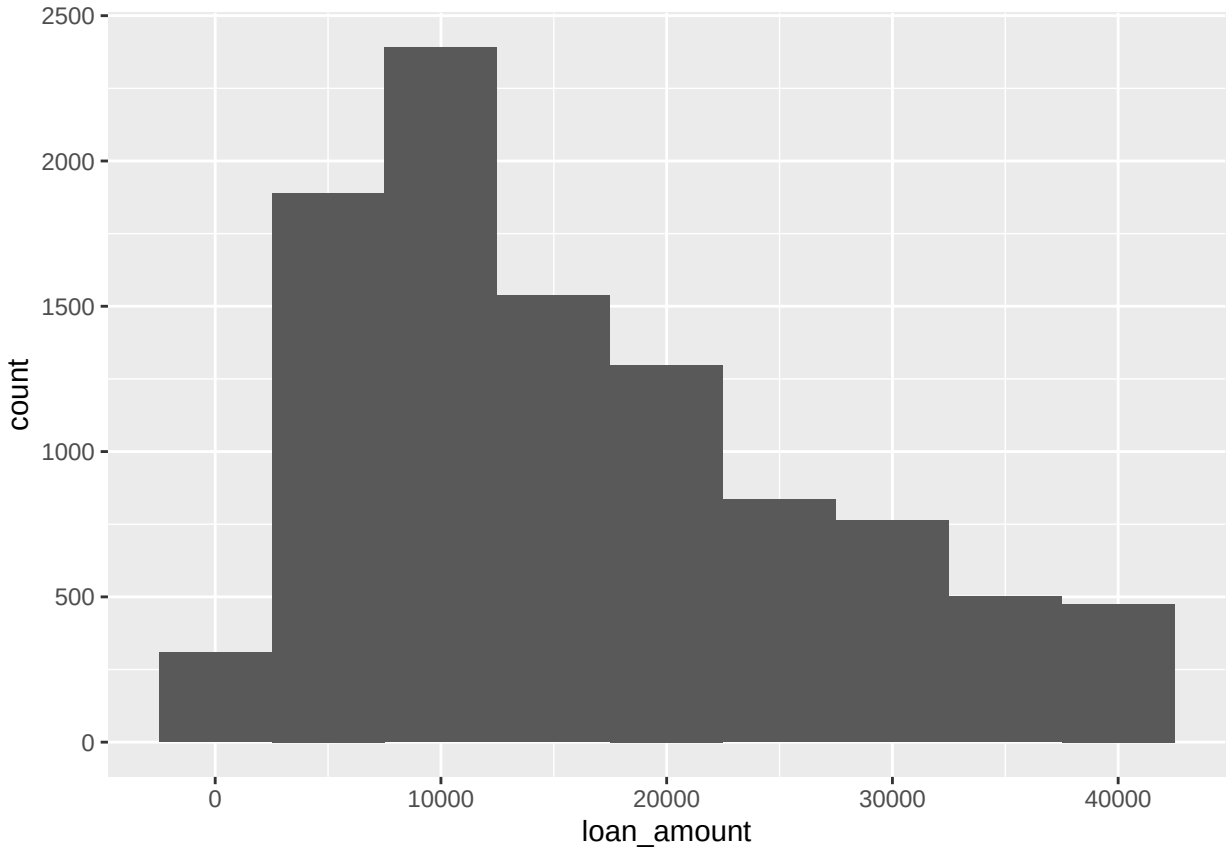
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



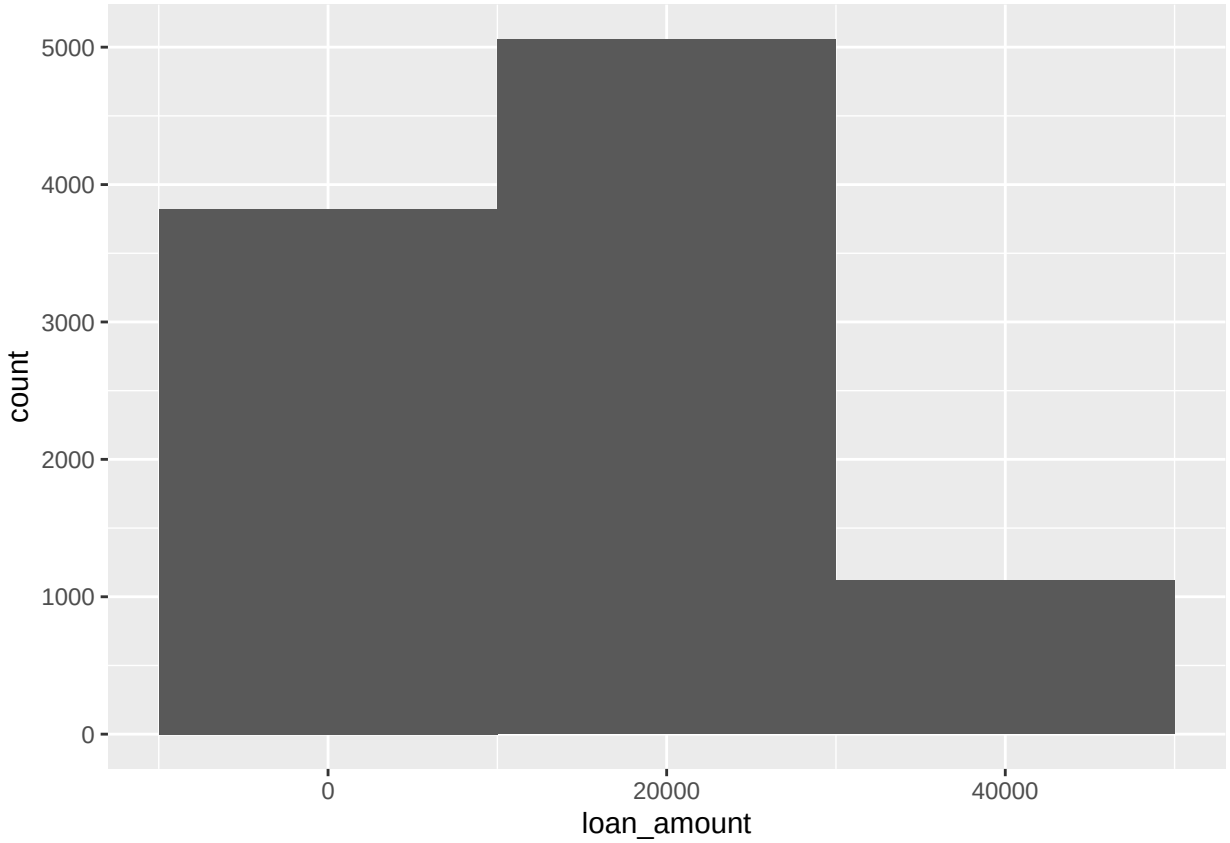
Set Binwidth to 1000



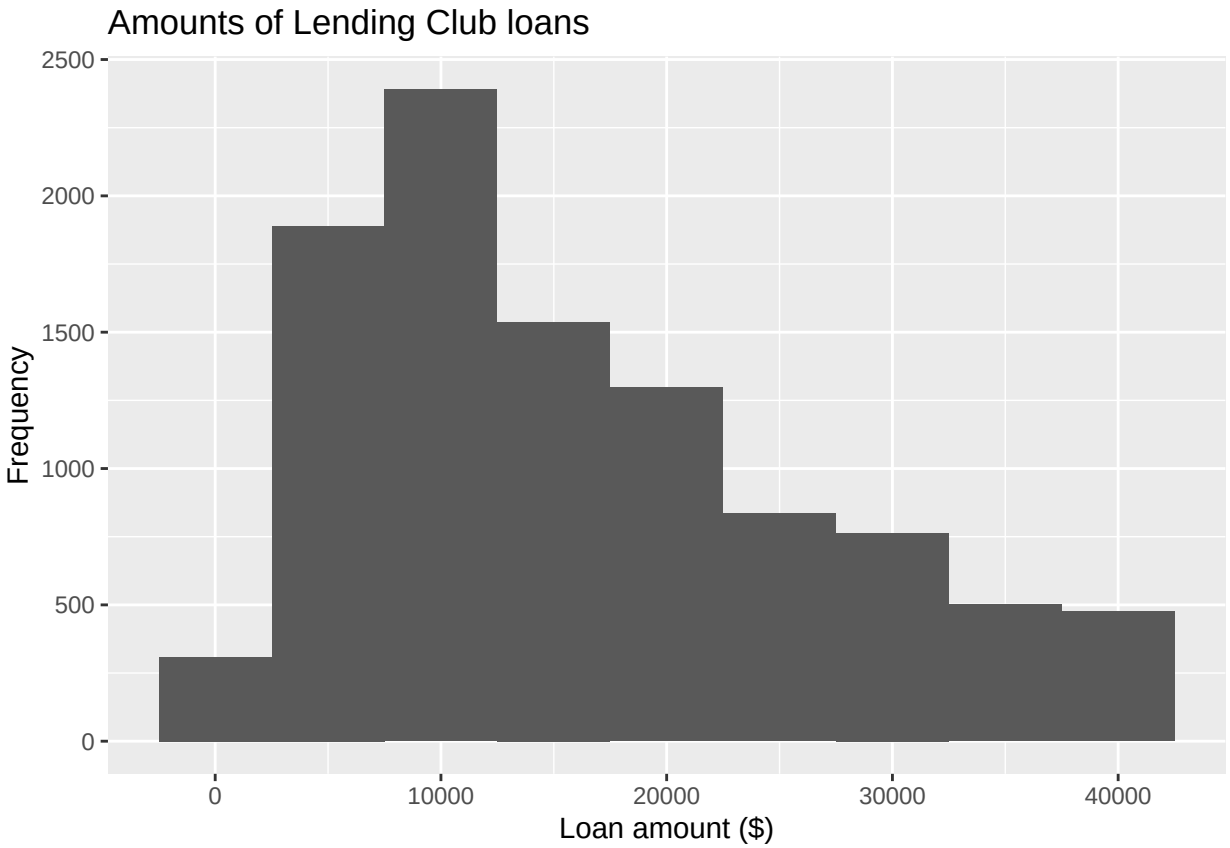
Set Binwidth to 5000



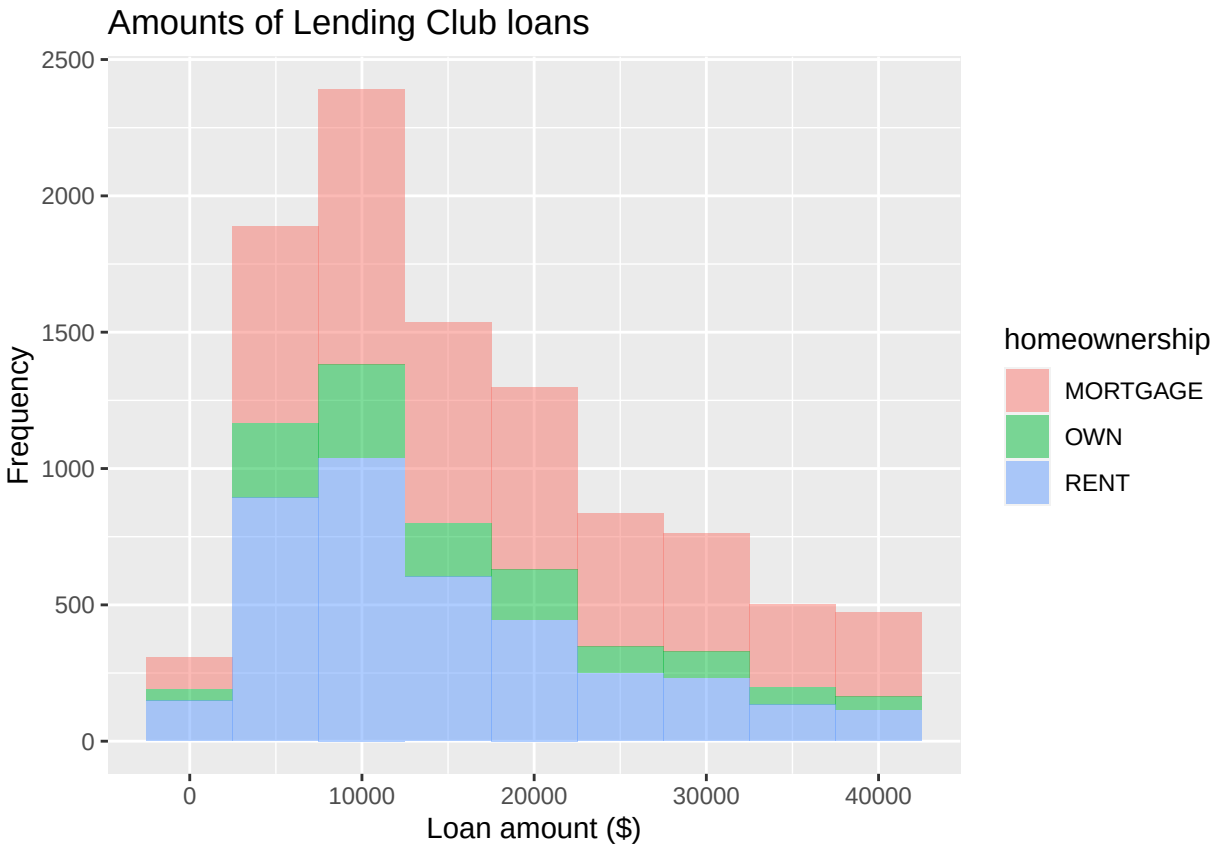
Set Binwidth to 20000


5000
4000
3000
2000
1000
0
count
0
20000
40000
loan_amount

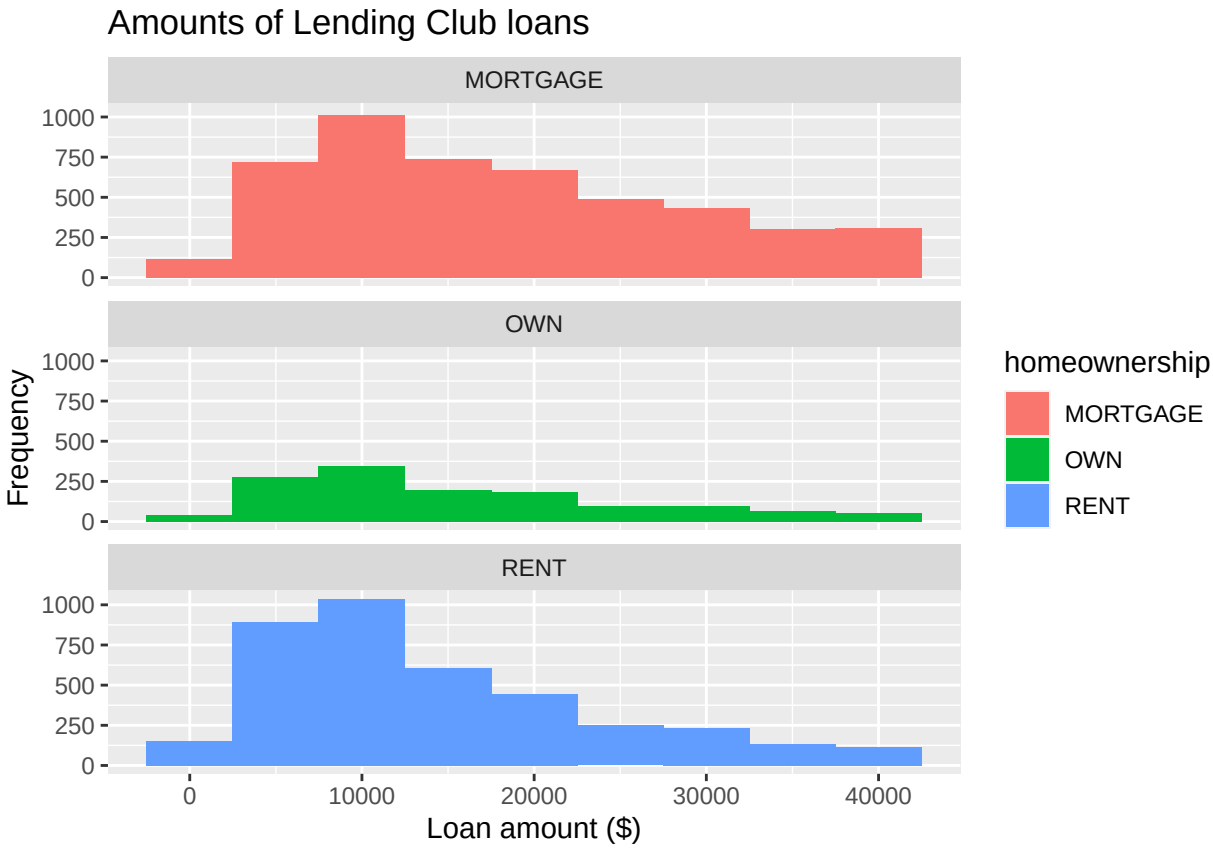
Customising Histograms



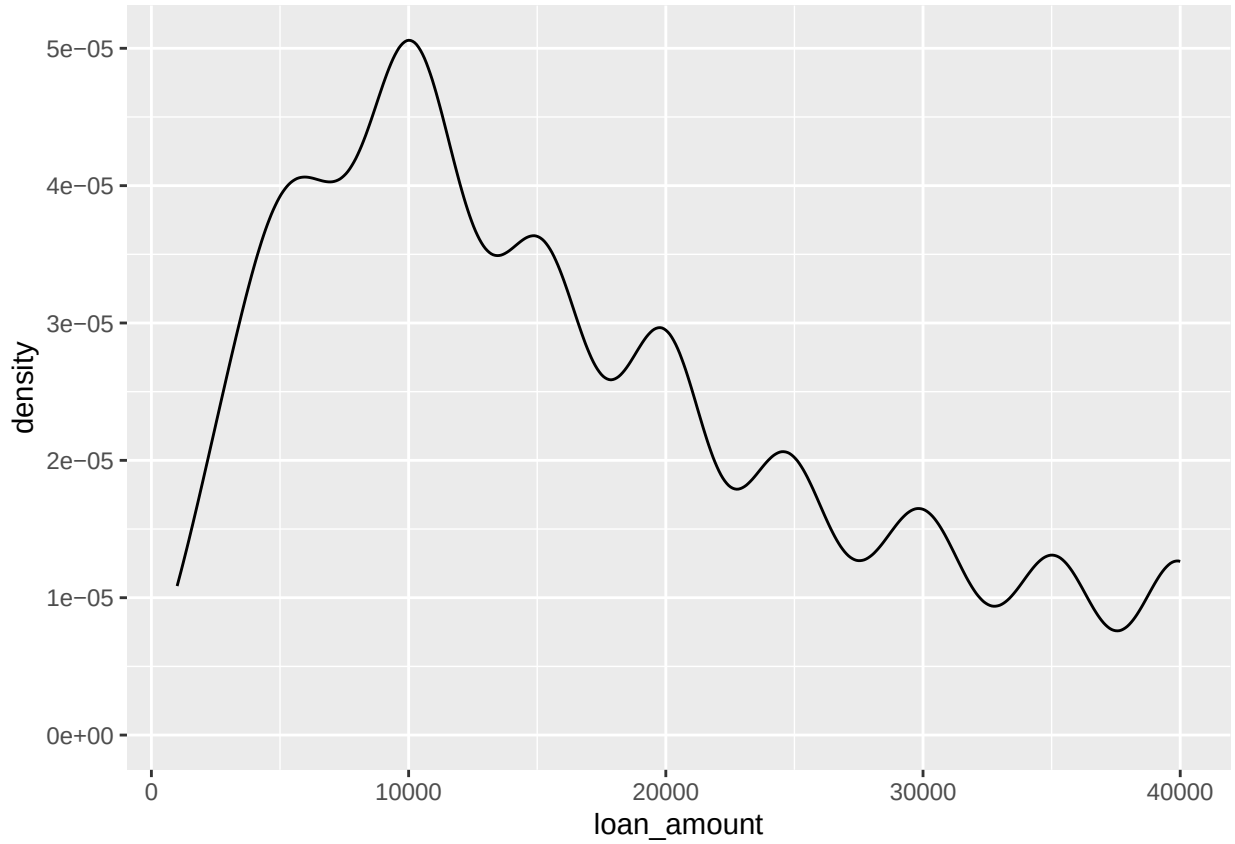
Fill with a categorical variable



Facet with a categorical variable

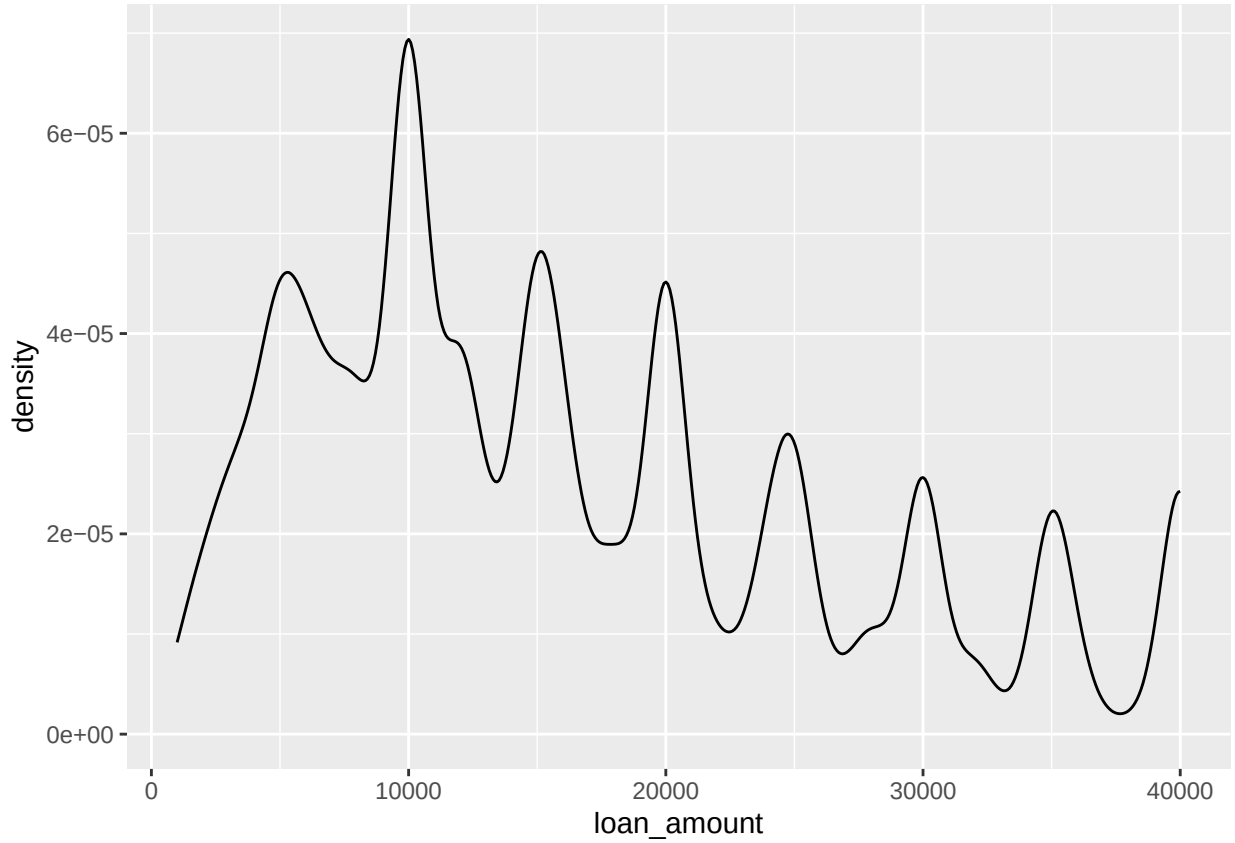


Density plot

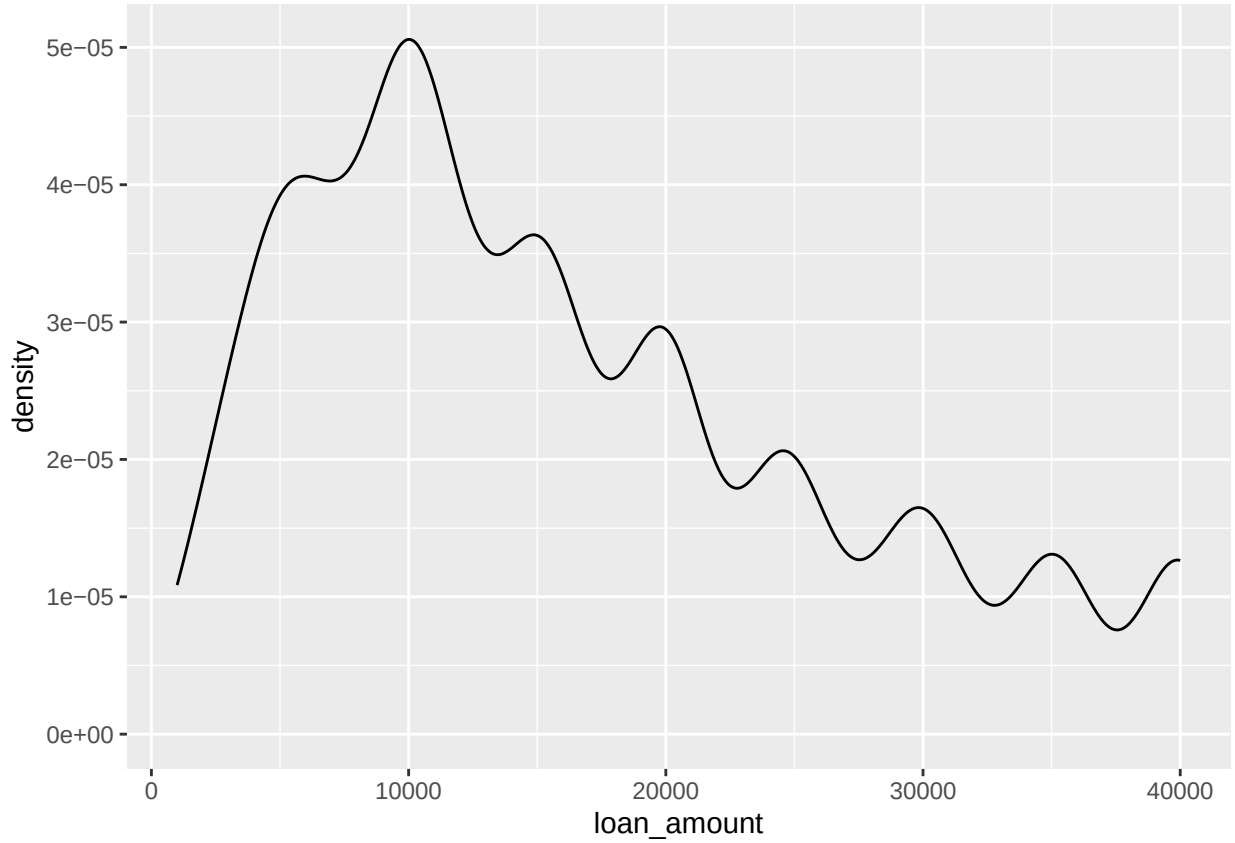


Density plots and adjusting bandwidth

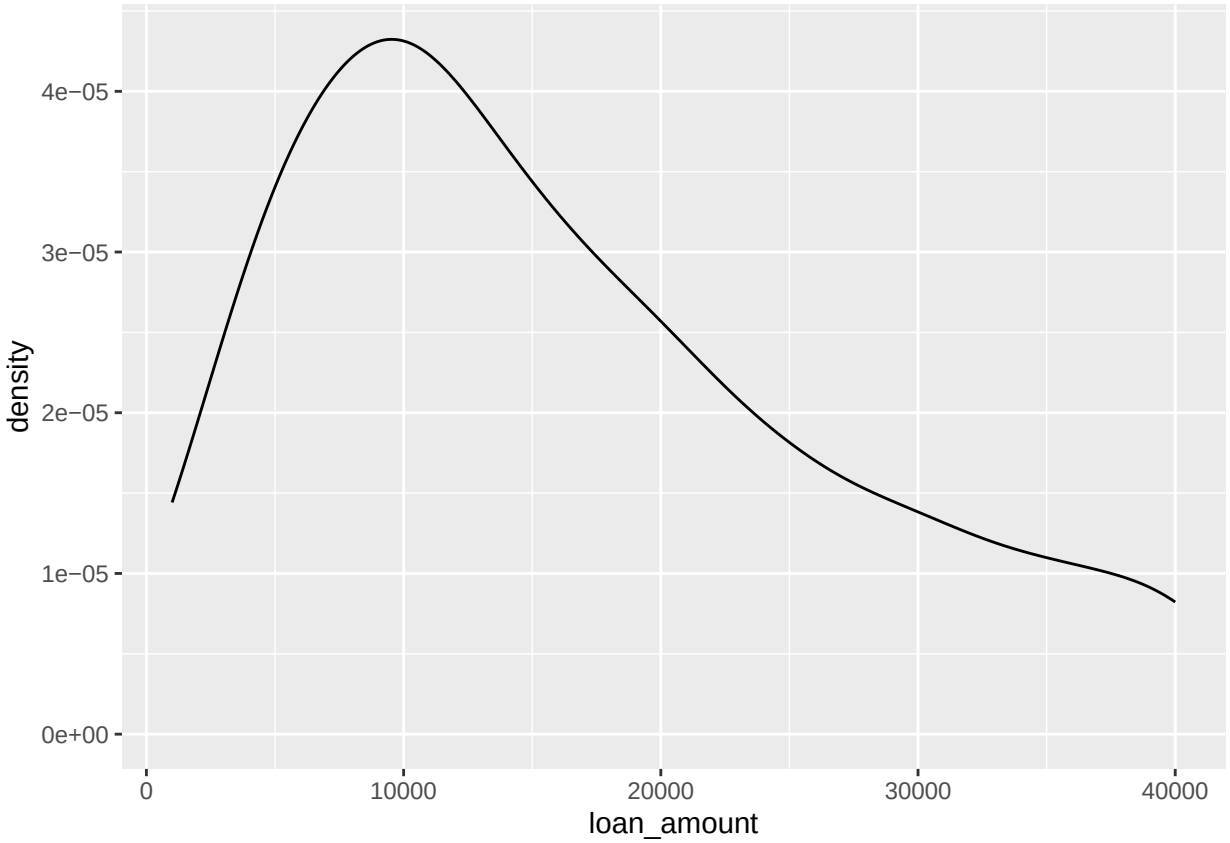
Adjust to 0.5


6e−05
4e−05
2e−05
0e+00
density
0
10000
20000
30000
40000
loan_amount

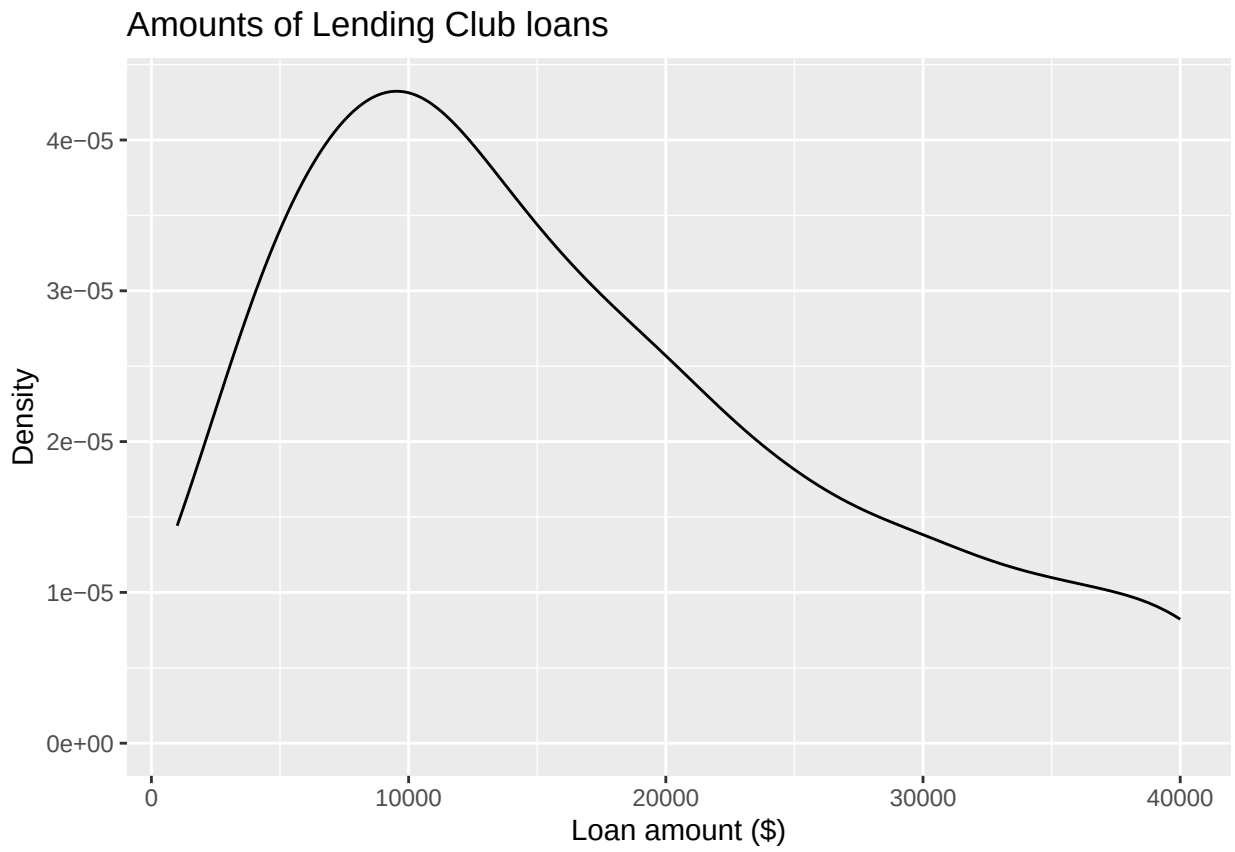
Adjust to 1


5e−05
4e−05
3e−05
2e−05
1e−05
0e+00
density
0
10000
20000
30000
40000
loan_amount

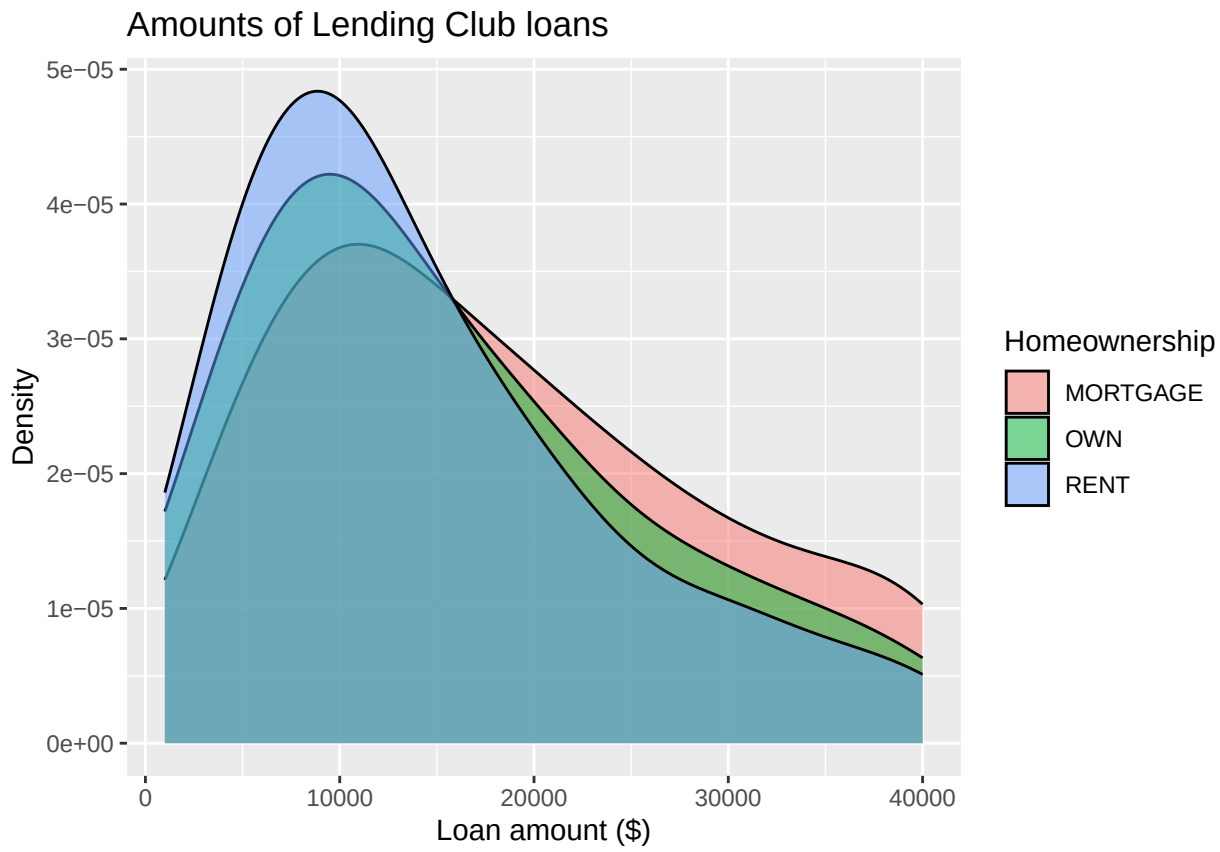
Adjust to 2


4e−05
3e−05
2e−05
1e−05
0e+00
density
0
10000
20000
30000
40000
loan_amount

Customising density plots

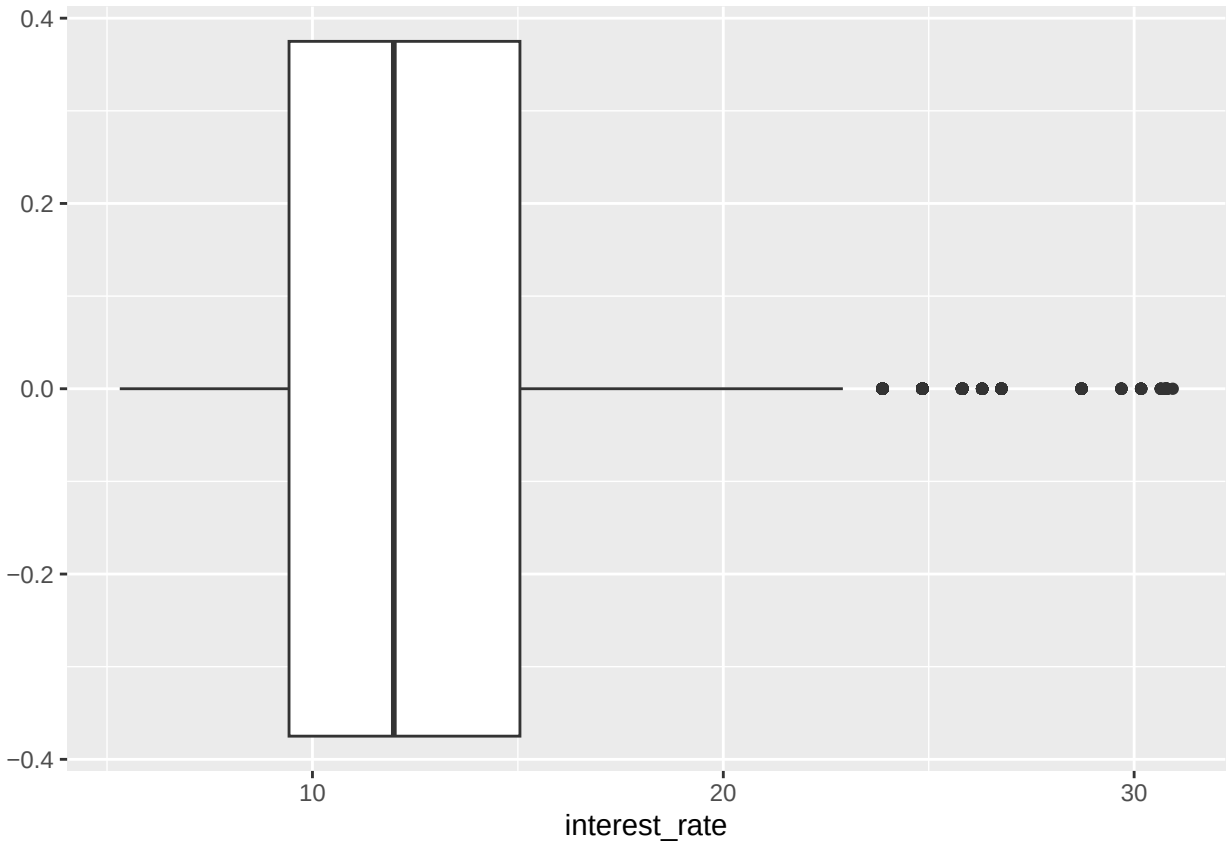


Adding a categorical variable

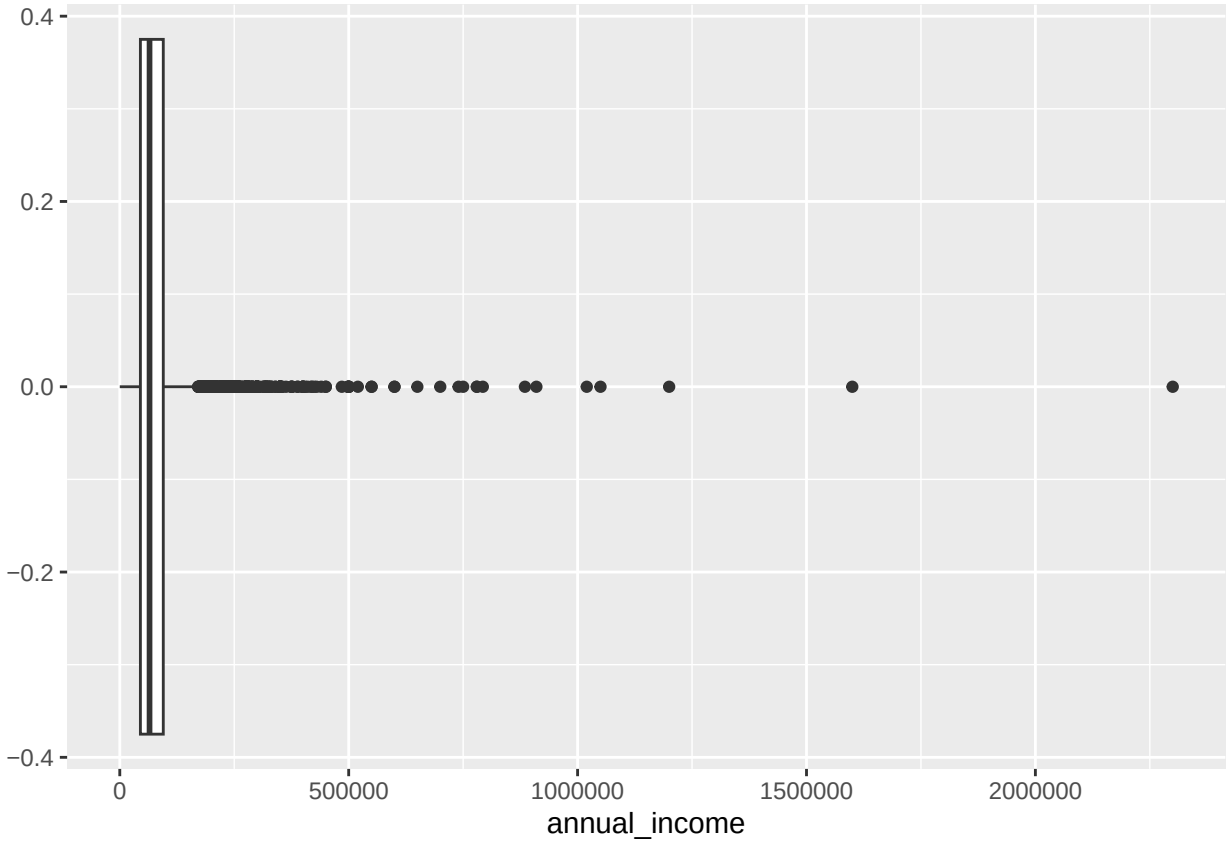


#Box Plot

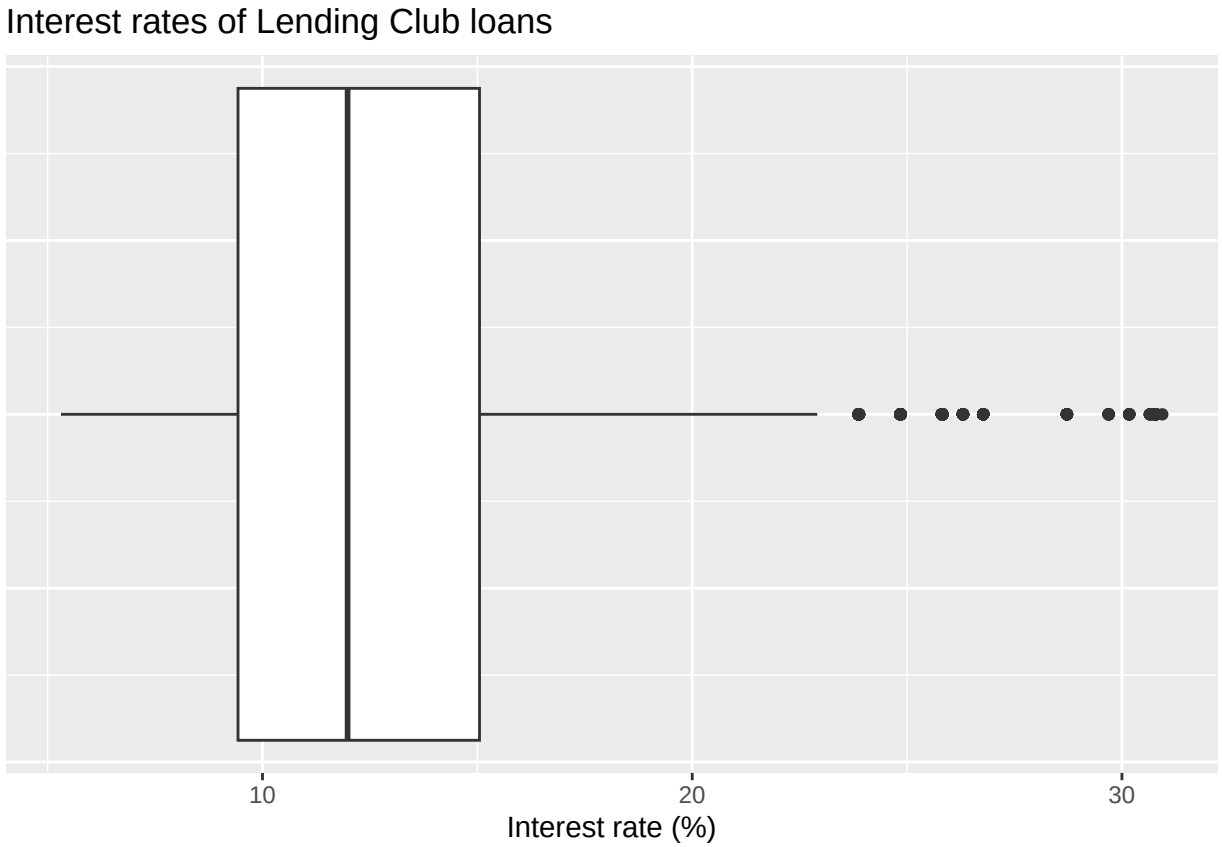
Box Plot of Interest Rate



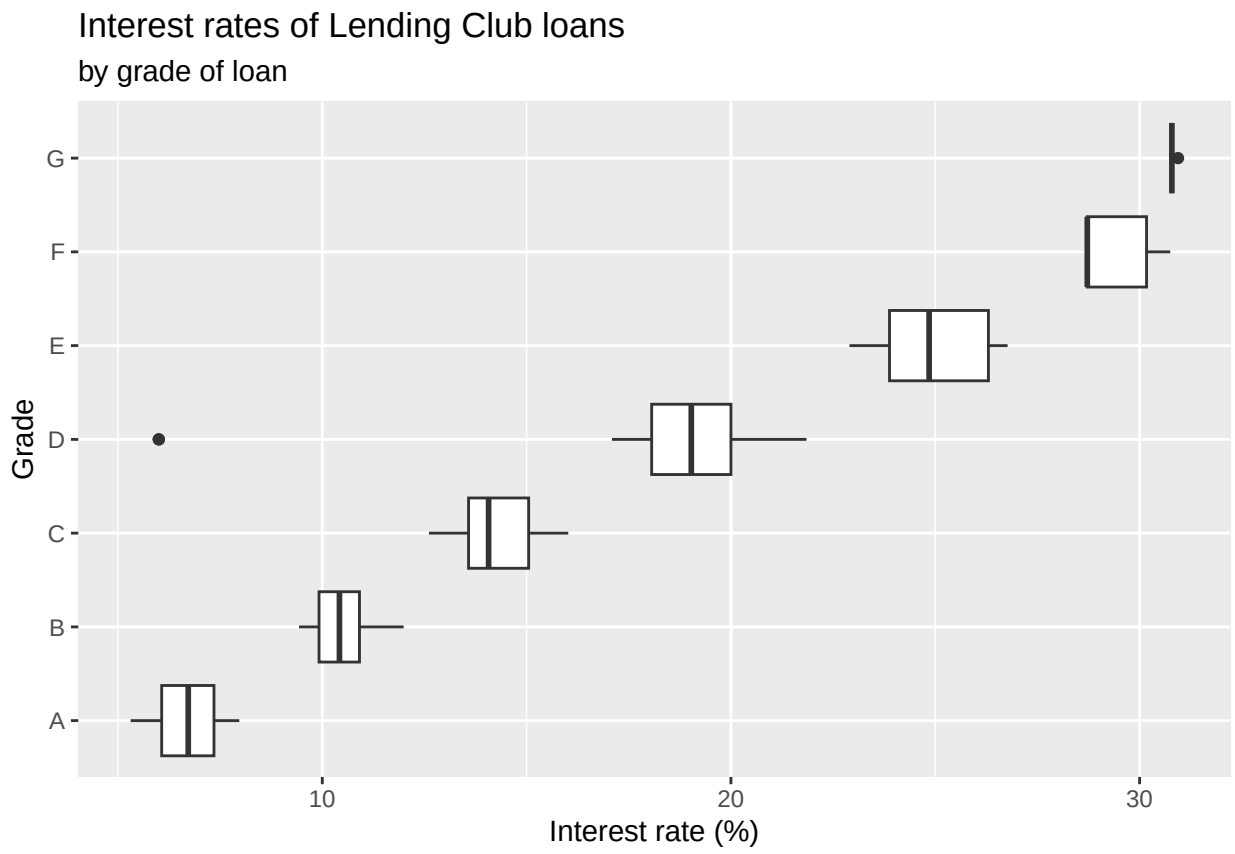
Box Plot of Annual Income



#Customising Box Plots

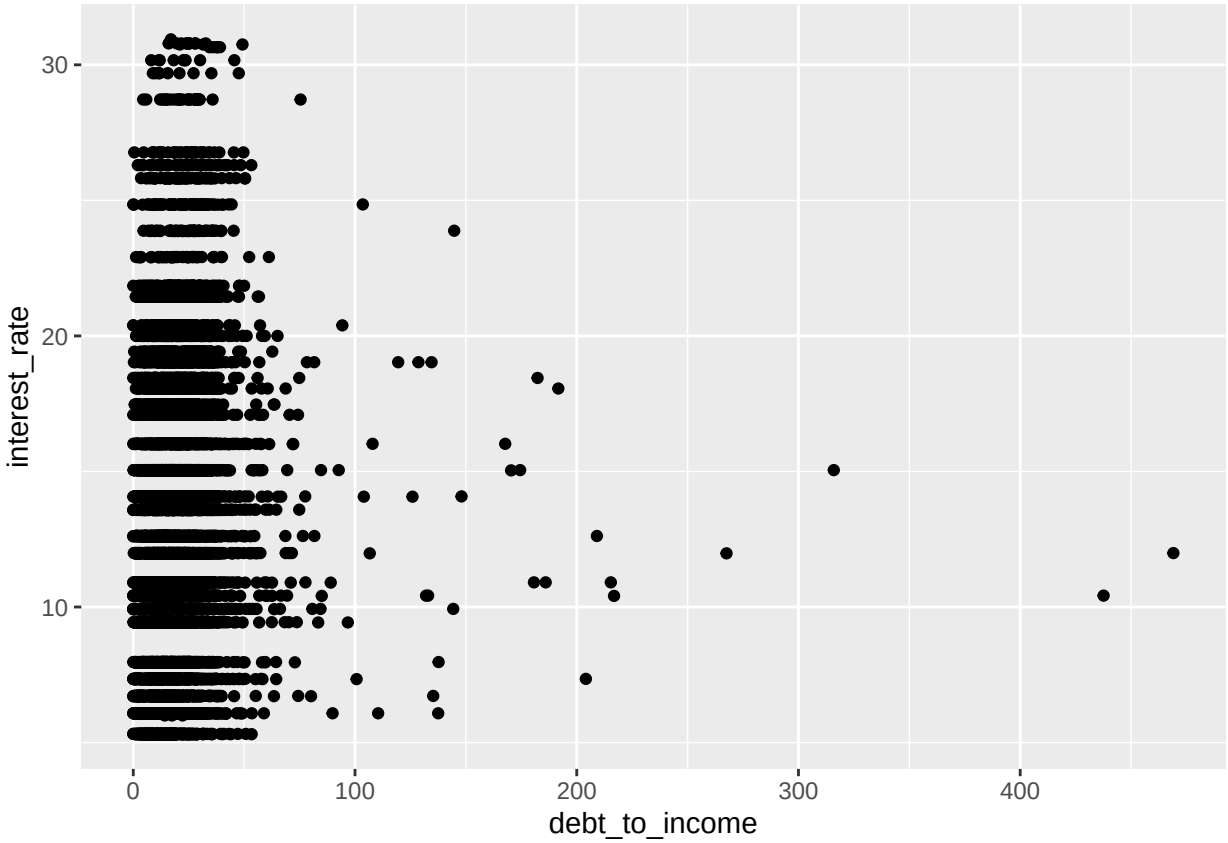


#Adding a categoric variable



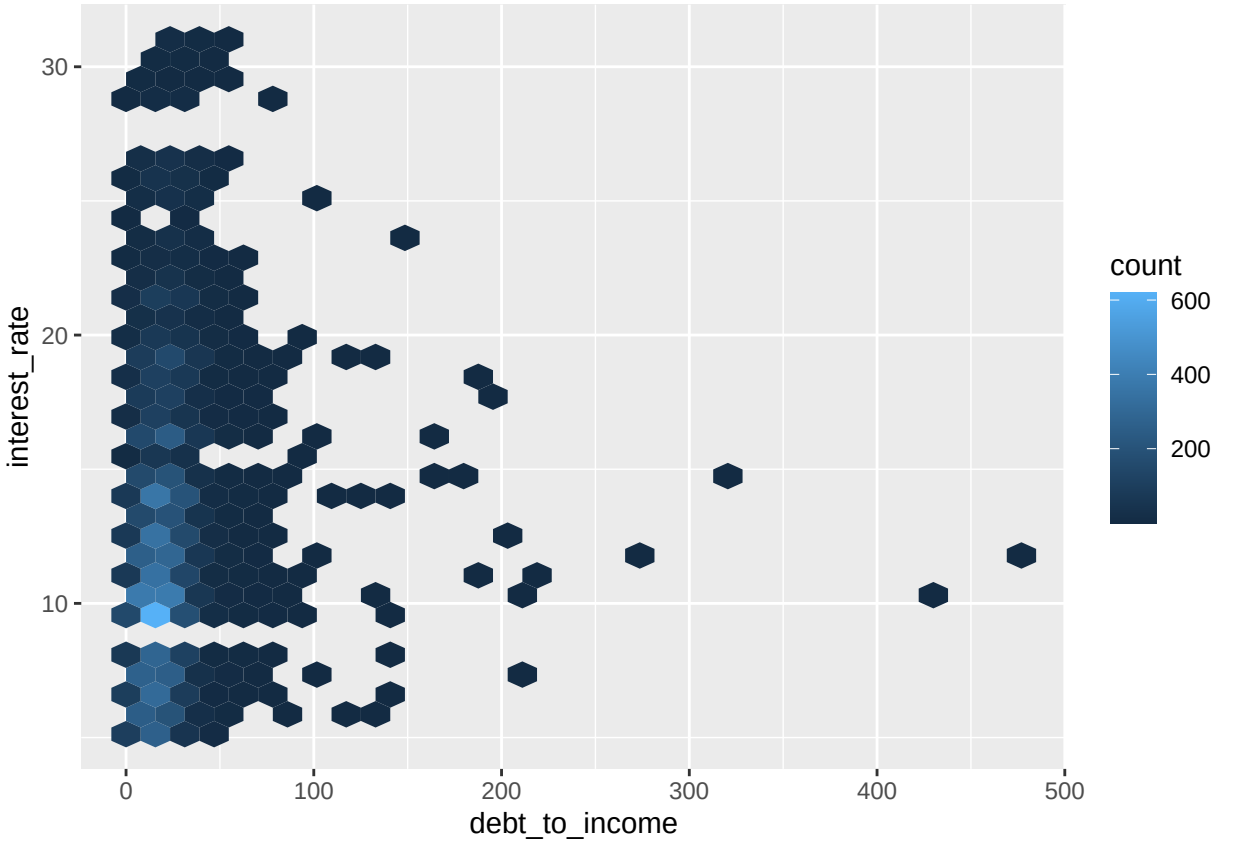
Scatterplot

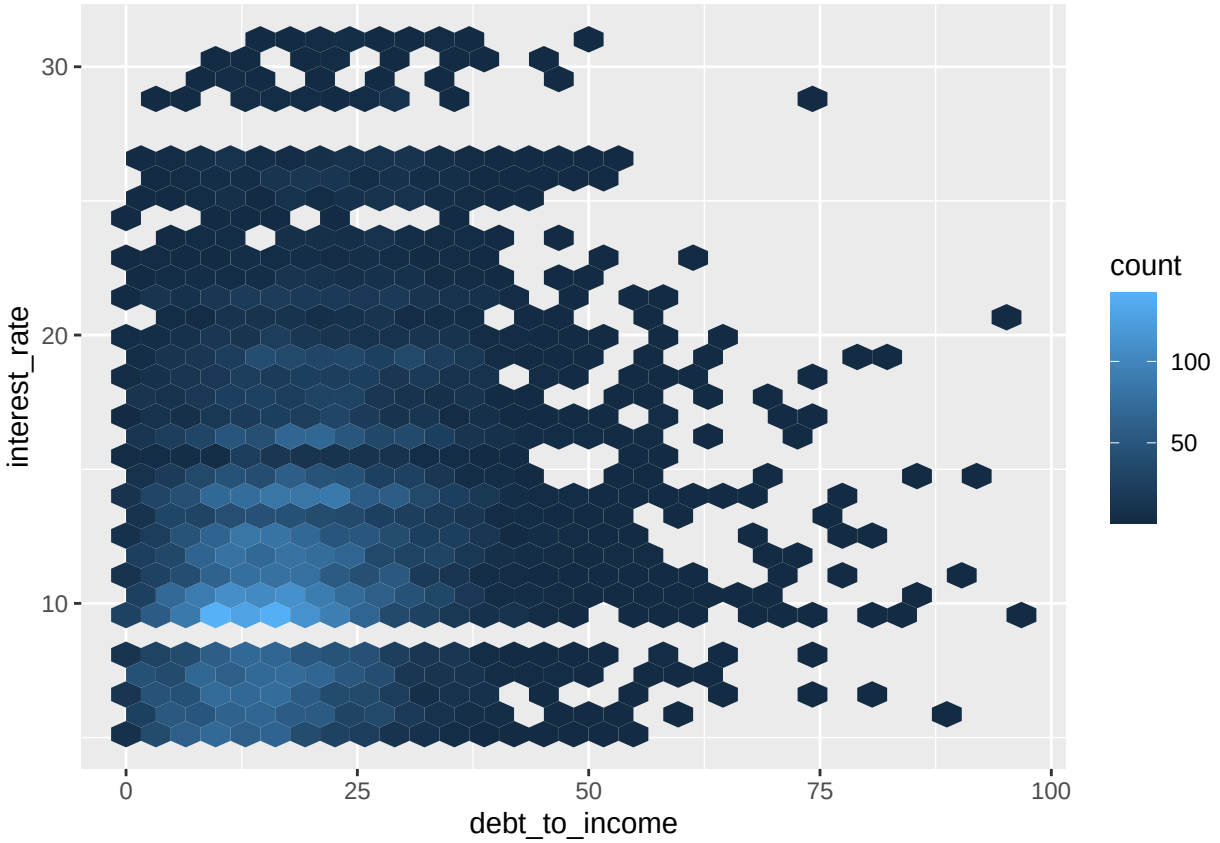
```
## Warning: Removed 24 rows containing missing values (‘geom_point()‘).
```



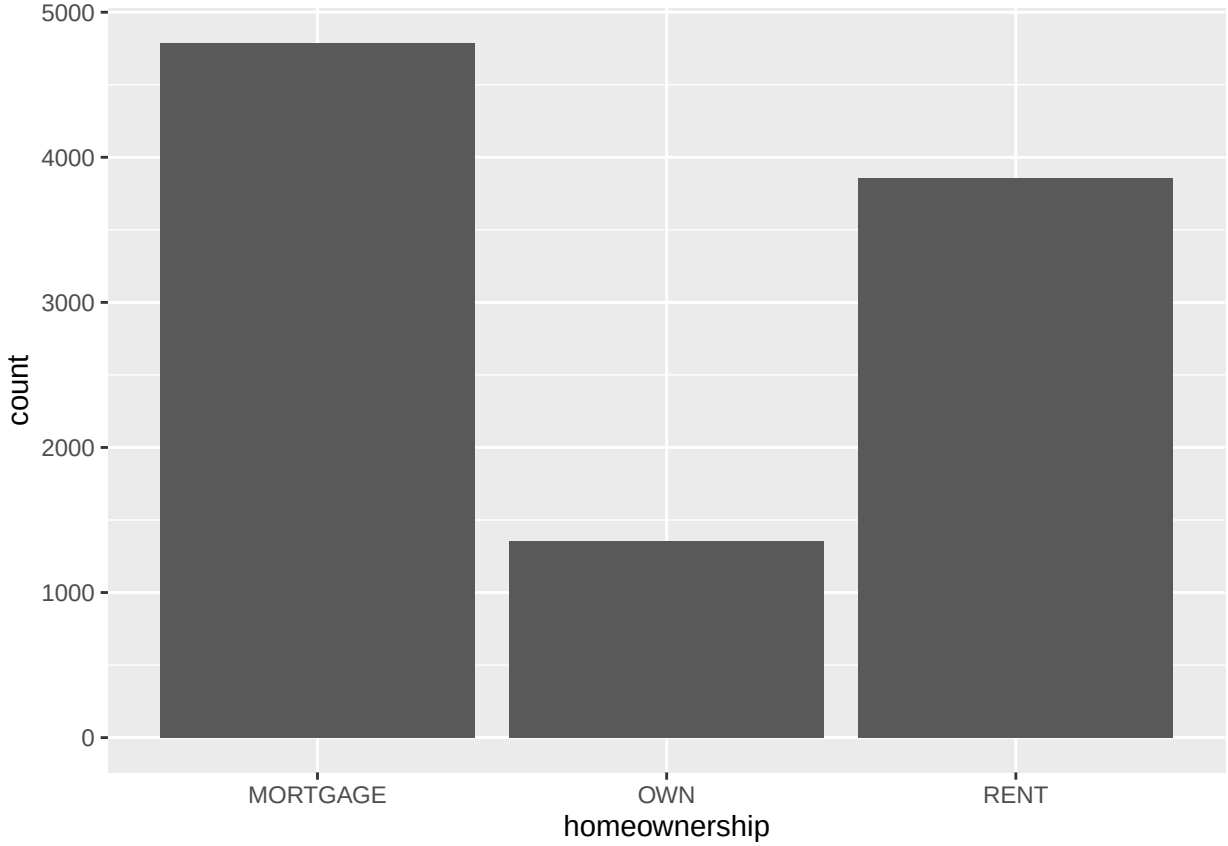
#Hex Plot

```
## Warning: Removed 24 rows containing non-finite values (‘stat_binhex()‘).
```

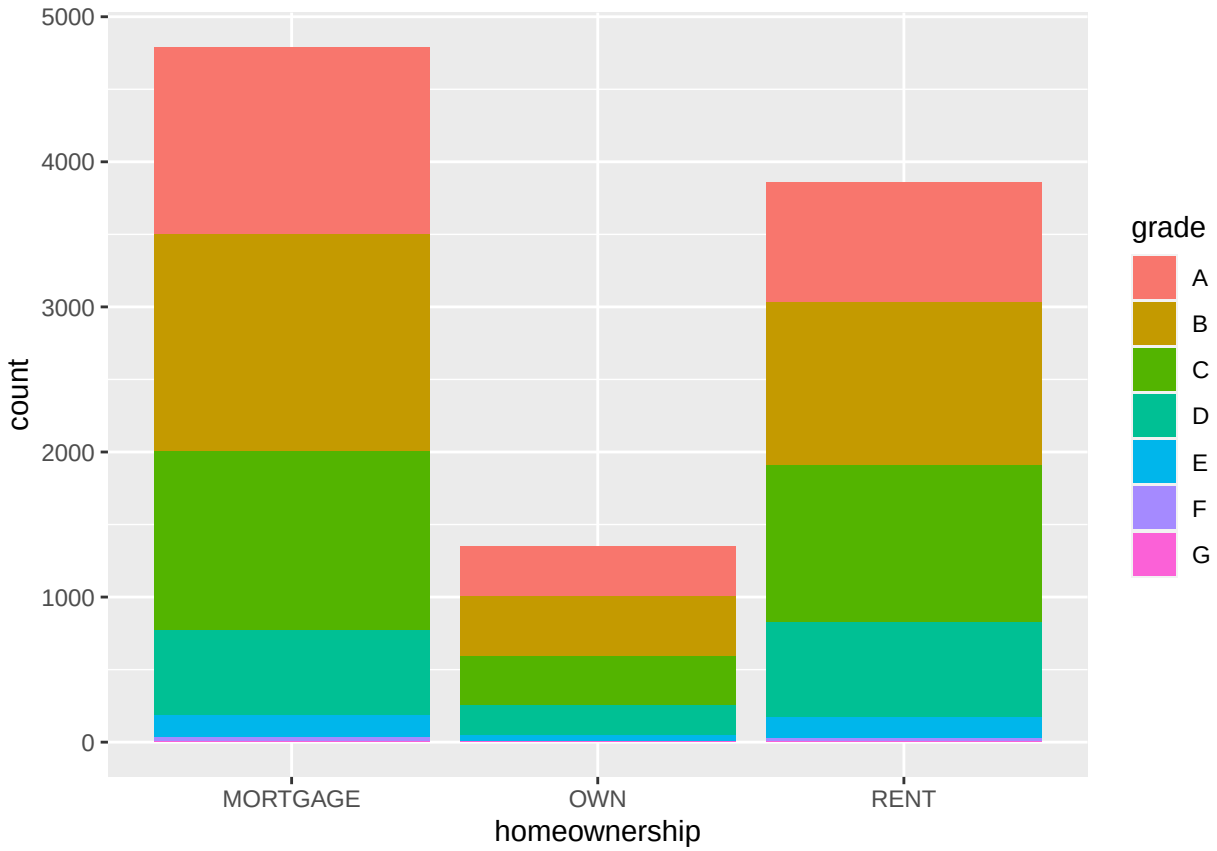

30
20
10
interest_rate
0
100
200
300
400
500
debt_to_income
count
600
400
200

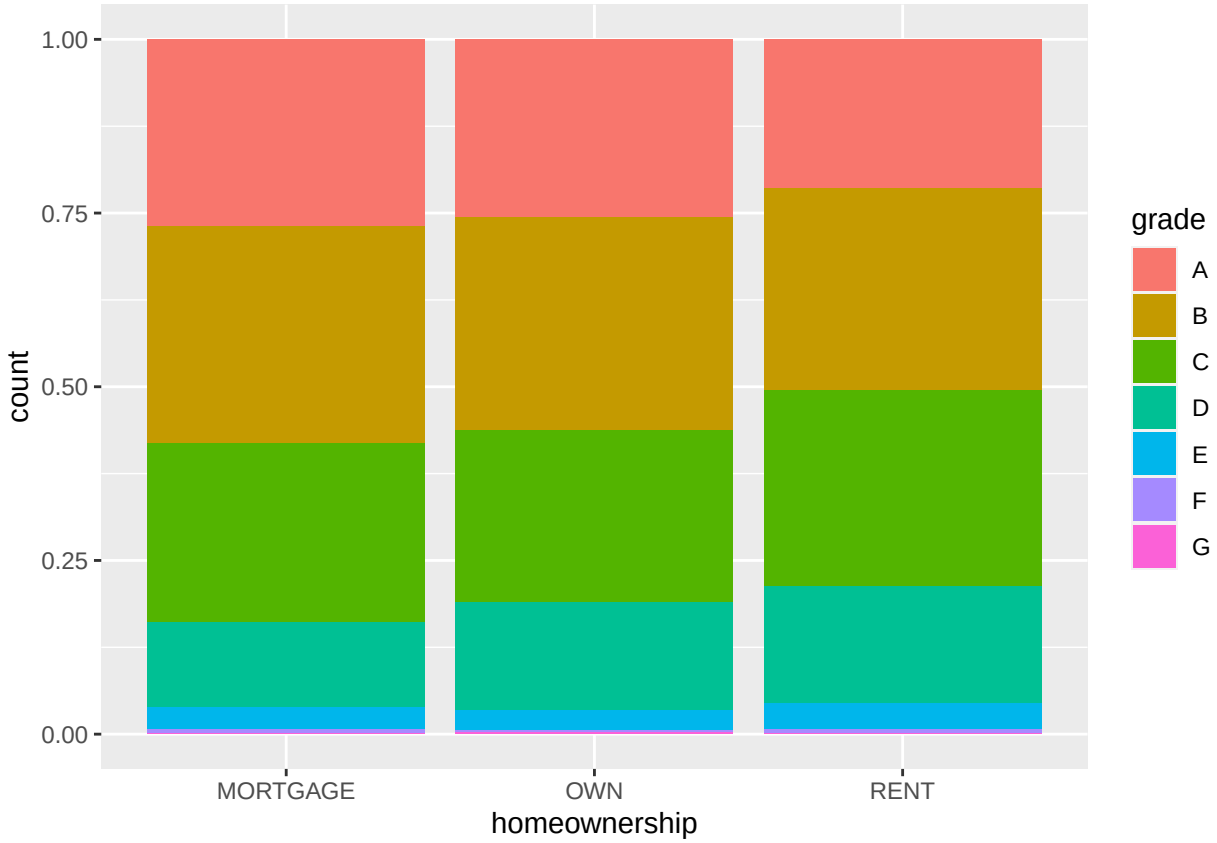

30
20
10
interest_rate
0
25
50
75
100
debt_to_income
count
100
50

Bar Plot

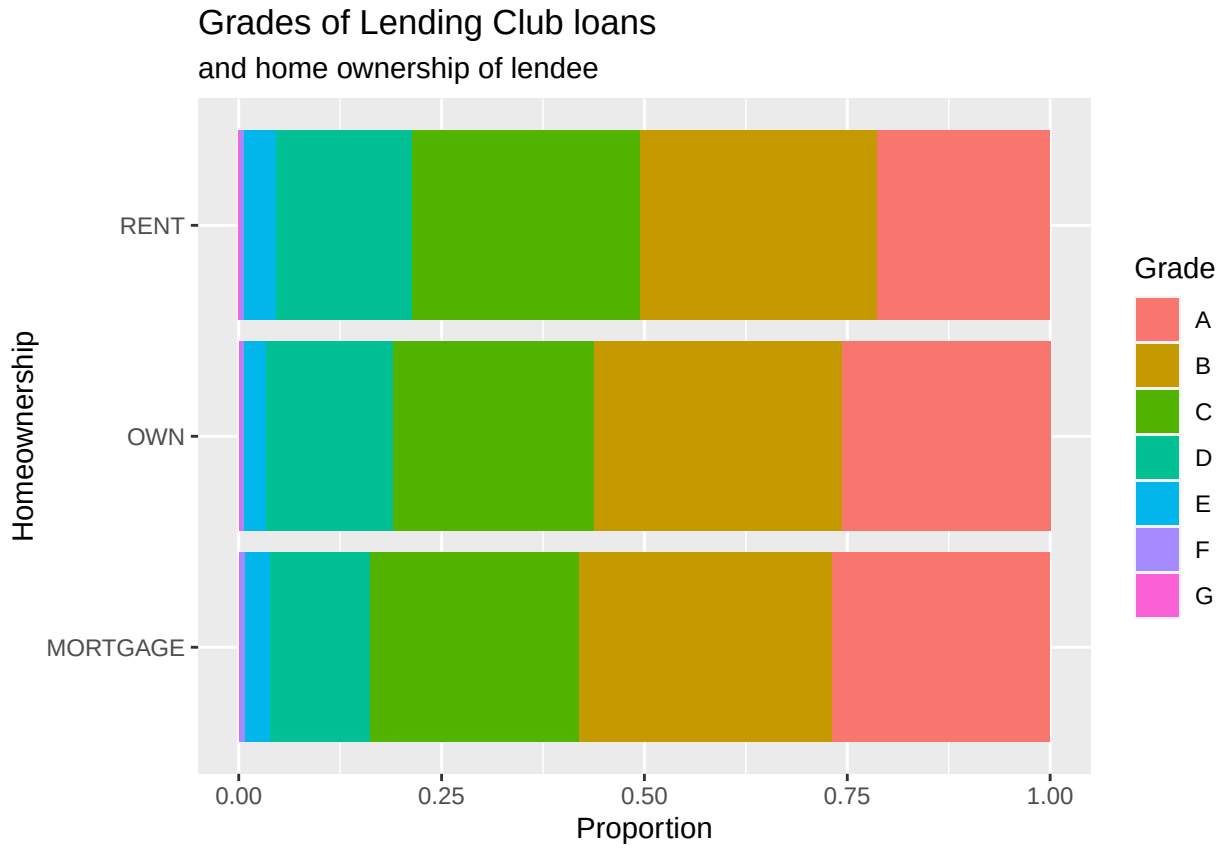


Segmented Bar Plot

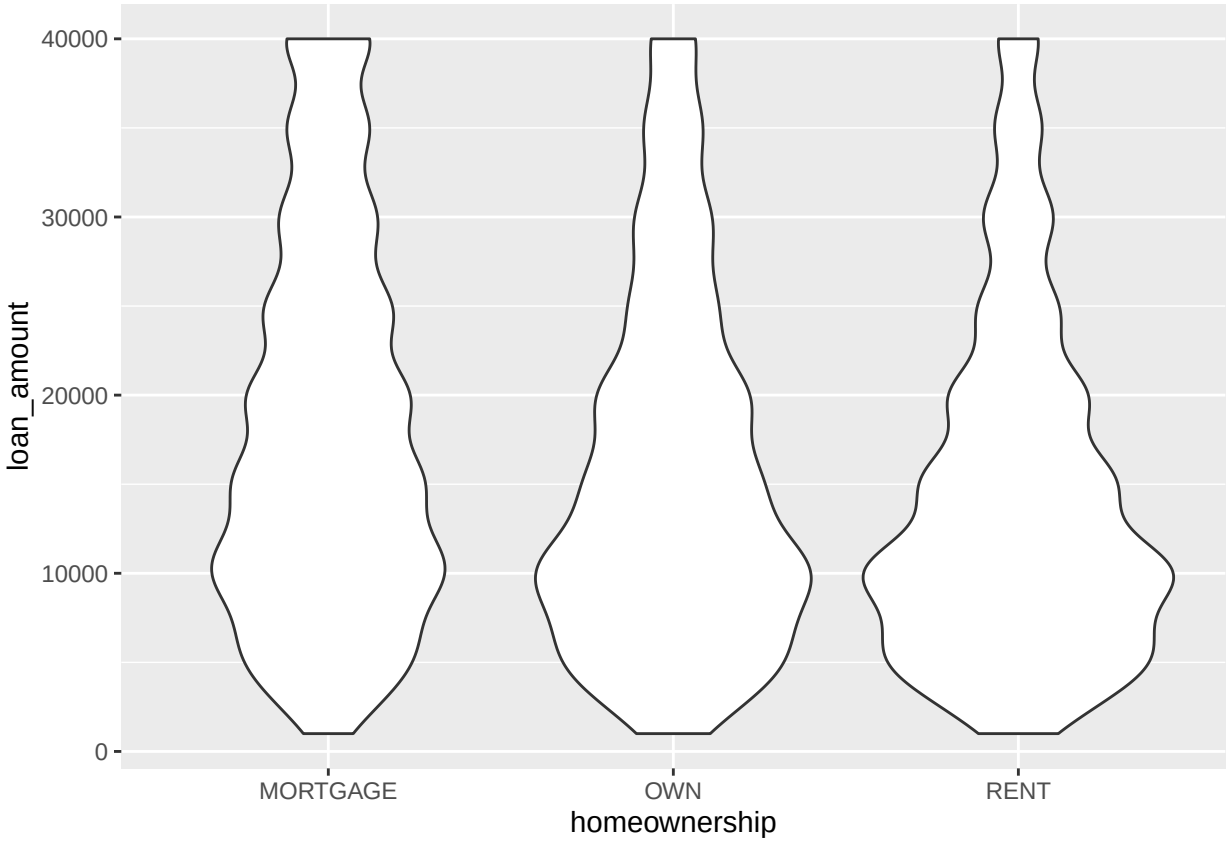

5000
4000
3000
2000
1000
0
count
MORTGAGE
OWN
RENT
homeownership
grade
A
B
C
D
E
F
G


1.00
0.75
0.50
0.25
0.00
count
MORTGAGE
OWN
RENT
homeownership
grade
A
B
C
D
E
F
G

Customising Bar Plots



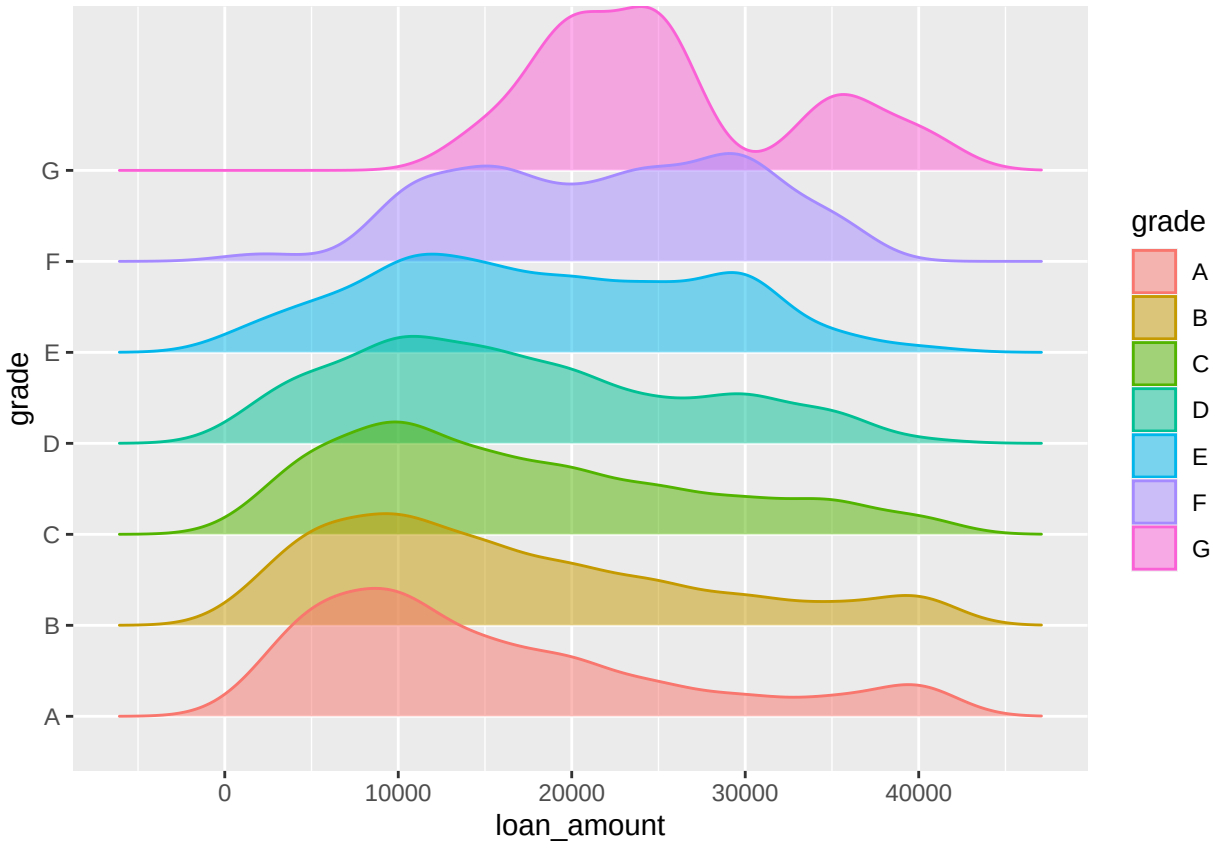
Violin Plots



Ridge Plots

```
## Warning: package 'ggridges' was built under R version 4.2.3
```

```
## Picking joint bandwidth of 2360
```


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