Describing numerical distributions

STA 032: Gateway to data science Lecture 9

Jingwei Xiong

April 21, 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.

Today

- Describing numerical distributions
  - Histograms
  - Measures of central tendency: mean, median, mode
  - Shape: skewness and modality
  - Spread: variance and standard deviation, range and interquartile range
  - Boxplots
  - Unusual observations
  - Density plot

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-
```

Selected variables

Variable	Description
`loan_amount`	Amount of the loan received, in US dollars
`interest_rate`	Interest rate on the loan, in an annual percentage
`term`	The length of the loan, which is always set as a whole number of months
`grade`	Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid
`state`	US state where the borrower resides
`annual_income`	Borrower’s annual income, including any second income, in US dollars
`homeownership`	Indicates whether the person owns, owns but has a mortgage, or rents
`debt_to_income`	Debt-to-income ratio
`issue_month`	Month the loan was issued

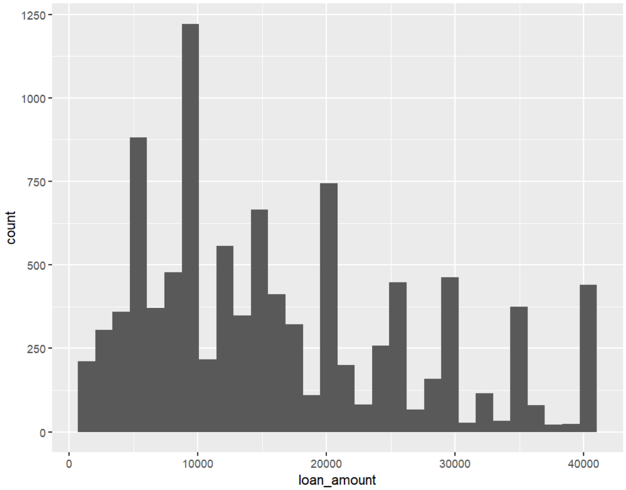
Map of Describing numerical distributions

- **Visual summaries**:
    - Histogram
    - Boxplot
    - Density plot
    - Line graph
- Measures of **central tendency**: mean, median, mode
- **Shape**:
    - Skewness: right-skewed, left-skewed, symmetric
    - Modality: unimodal, bimodal, multimodal, uniform
- **Spread**: variance and standard deviation, range and interquartile range
- **Unusual observations**
- A **summary statistic** is a single number summarizing a large amount of data

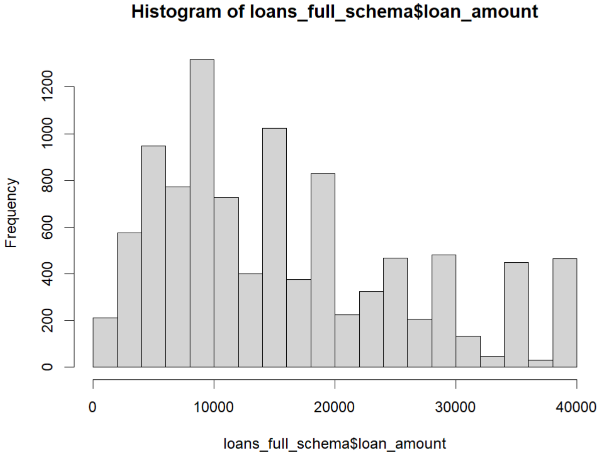
Histogram

- Shows **shape, center, and spread** of the data
- Contiguous (adjoining) boxes
  - Horizontal axis: what the data represents
  - Vertical axis: frequency or relative frequency

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



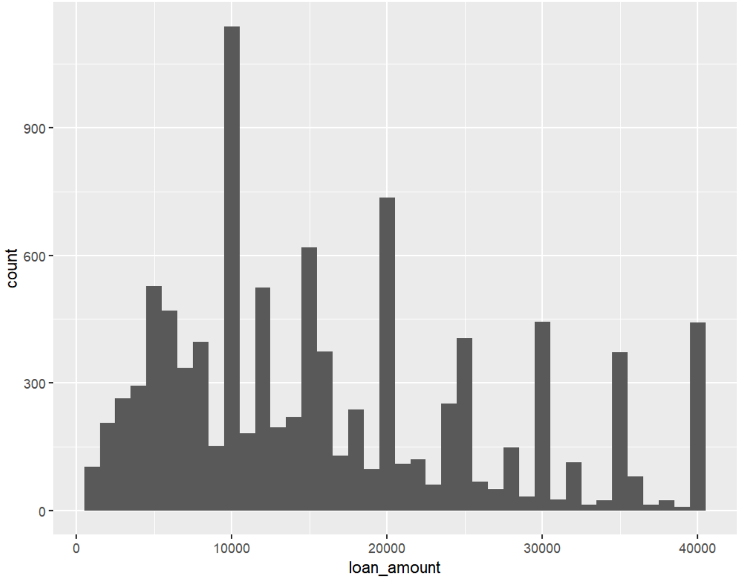
```
hist(loans_full_schema$loan_amount)
```



Histograms and binwidth

binwidth = 1000 binwidth = 5000 binwidth = 20000 Words

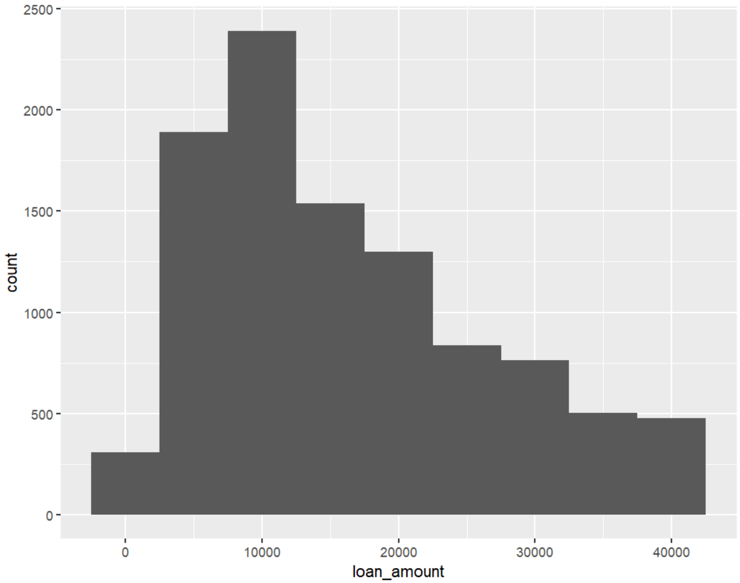
```
ggplot(loans, aes(x = loan_amount)) +
  geom_histogram(binwidth = 1000)
```



Histograms and binwidth

binwidth = 1000 | **binwidth = 5000** | binwidth = 20000 | Words

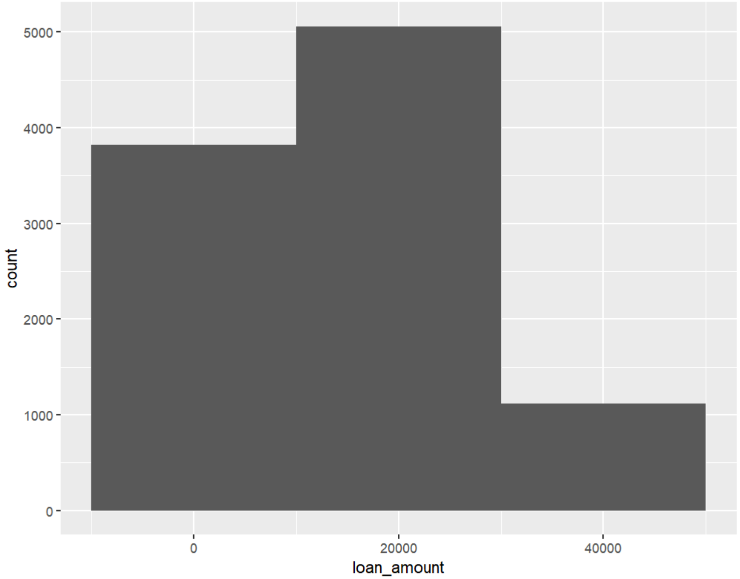
```
ggplot(loans, aes(x = loan_amount)) +
  geom_histogram(binwidth = 5000)
```



Histograms and binwidth

binwidth = 1000 binwidth = 5000 **binwidth = 20000** Words

```
ggplot(loans, aes(x = loan_amount)) +
  geom_histogram(binwidth = 20000)
```



Histograms and binwidth

binwidth = 1000 binwidth = 5000 binwidth = 20000 **Words**

Here we shows how the binwidth parameter affects the appearance of a histogram.

Binwidth is the width of each bin or bar in the histogram. You can easily change it inside `geom_histogram` function.

A small binwidth creates more bars that reflect more variation in the data but can also cause overplotting, making it difficult to see patterns.

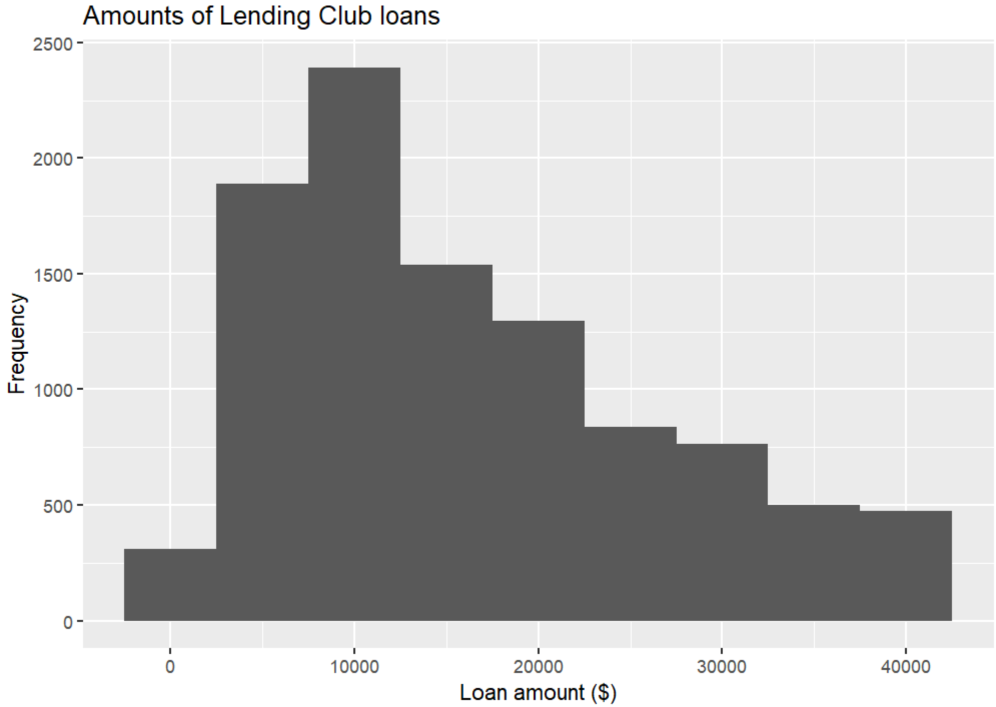
A large binwidth creates fewer bars and provides a smoother visualization but may obscure some patterns.

Choosing the appropriate binwidth depends on the data and the research question.

Adding labels

The labs() function can be an alternated way used to customize the labels and title of the plot.

Plot Code



Adding labels

The labs() function can be an alternated way used to customize the labels and title of the plot.

Plot Code

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

Population vs. sample (2nd time, more later)

- A **sample** is a portion or **subset** of the larger **population**
- E.g., population may be UC Davis students; randomly sample 300 students on the Quad this morning
- Population **parameter**, e.g., population mean
  - This is a fixed quantity
- Sample **statistic**, e.g., sample mean
  - Depends on the sample

> Use sample statistics to make **inferences** about **unknown** population parameters

Measures of central tendency (R funcion)

- **Mean**: "Average", sum the numbers and divide by the count (`mean()`)

  $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$, where x is the variable of interest, the subscripts index the n observations, and $\bar{x}$ denotes the sample mean.

  The population mean is often denoted by μ.

- **Median**: "Middle value", arrange in ascending order (`median()`)

- **Mode**: Most frequent value (`mode()` does not do what you might think)

  > We won't test you how to calculate them without using R.

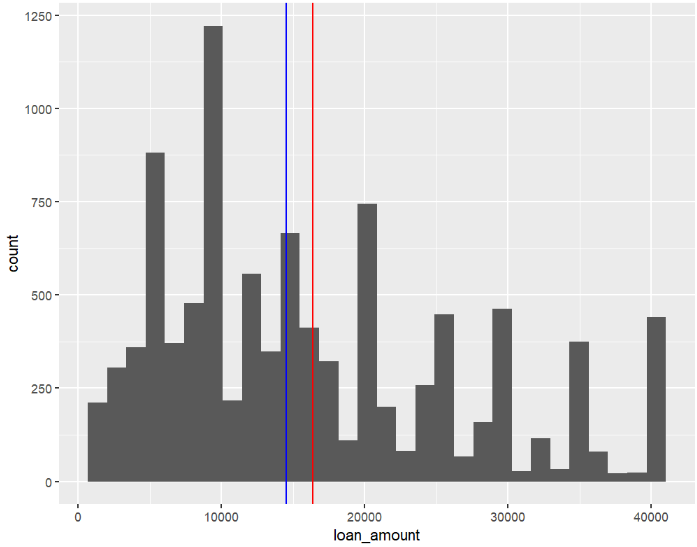
- Note: you will sometimes need the `na.rm = TRUE` option where the data include `NA` (missing) values.

Measures of central tendency: Example

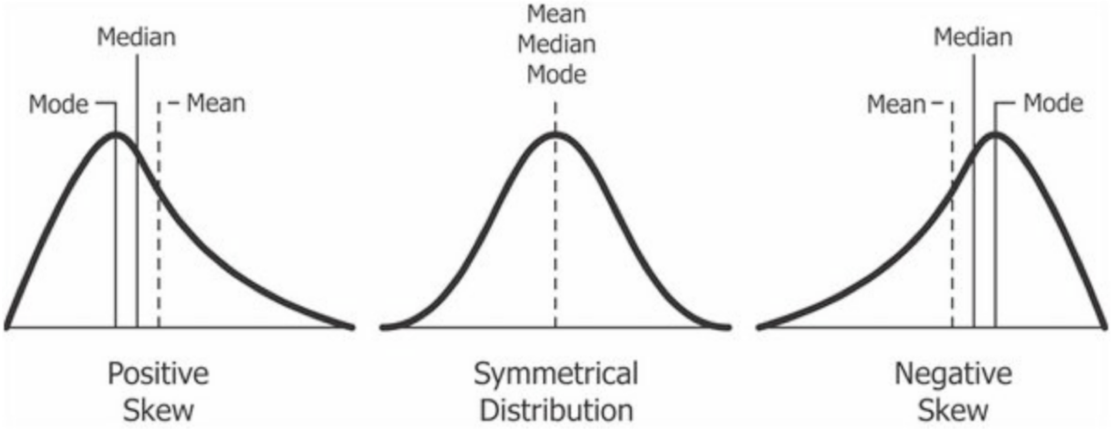
- `mean(loans$loan_amount)` = 16361.92; `median(loans$loan_amount)` = 14500
- Mode is 10000 (Exercise: how to code this in R?)

> Solution

```
ggplot(loans, aes(x = loan_amount)) +
  geom_histogram() +
  geom_vline(xintercept = median(loans$loan_amou
             col = "blue") +
  geom_vline(xintercept = mean(loans$loan_amount
             col = "red")
```



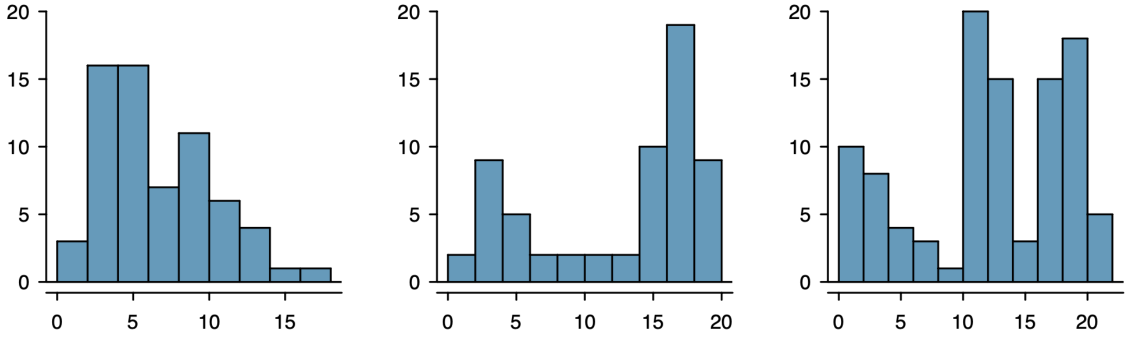
Measures of shape: Skewness



- **Skewness** is the degree to which the distribution is off-centered
  - **Positive skew/right skew**: mean > median
  - **Negative skew/left skew**: mean < median

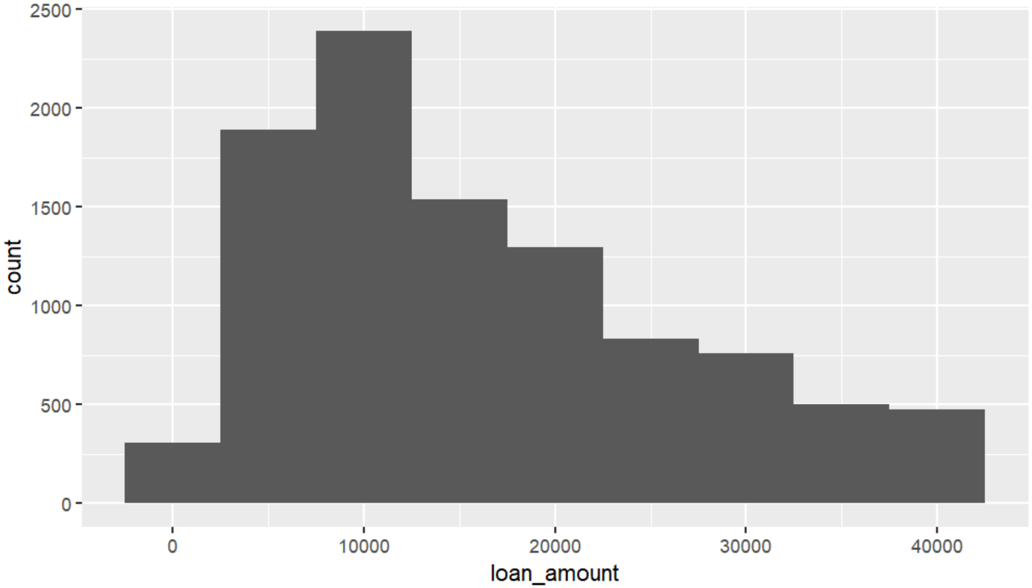
Measures of shape: Modality

- **Mode** is the most frequent value, but in real-world data sets, there might not be any observations with the same value.
- A mode is represented by a **prominent peak in the distribution**
- **Unimodal** = one prominent peak, **bimodal** = two prominent peaks, **multimodal** = more than two prominent peaks, uniform means no peaks



Loans data

```
ggplot(loans, aes(x = loan_amount)) +
  geom_histogram(binwidth = 5000)
```



Question: What is the skewness and modality?

Measures of spread: Variance and standard deviation

- **Standard deviation** measures how far data values are from their mean

- **Deviation** is the distance of an observation from its mean, $x_i - \bar{x}$

- **Sample variance**: Take the square of deviations and find the mean, denote as s

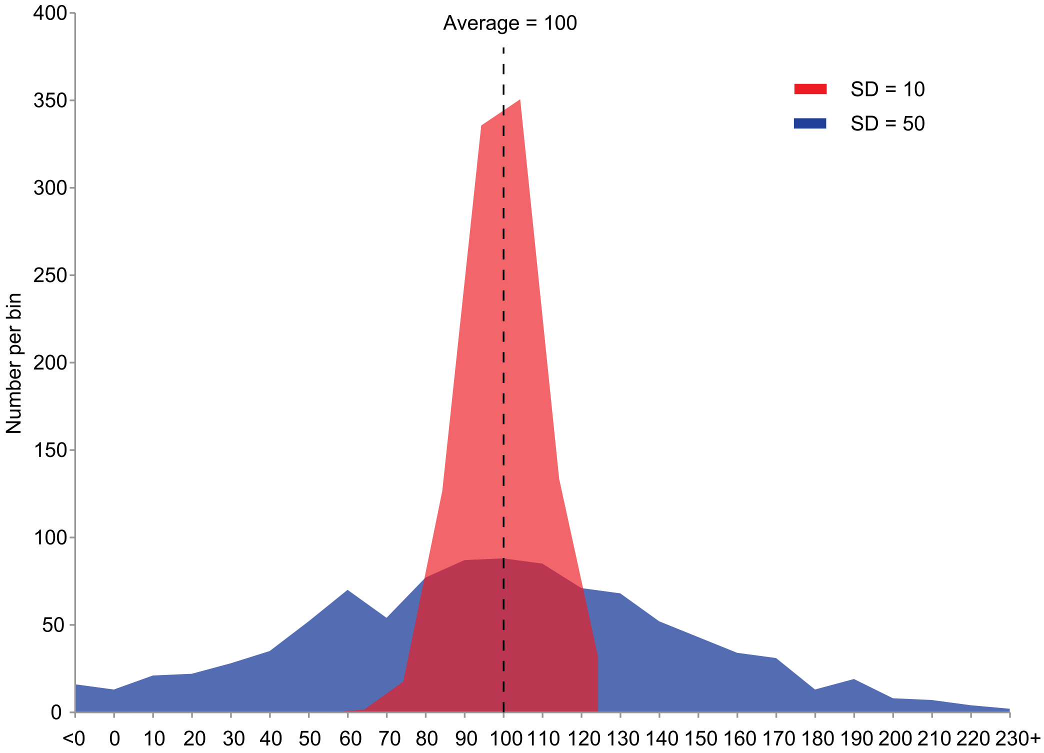
  - $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$
  - For the denominator, use $n-1$ instead of n to make it an *unbiased estimator of the population mean*

- **Sample standard deviation**, $s = \sqrt{s^2}$

- In R, `sd()` for sample standard deviation, `var()` for sample variance

- Population variance and standard deviation are often denoted by σ^2 and σ

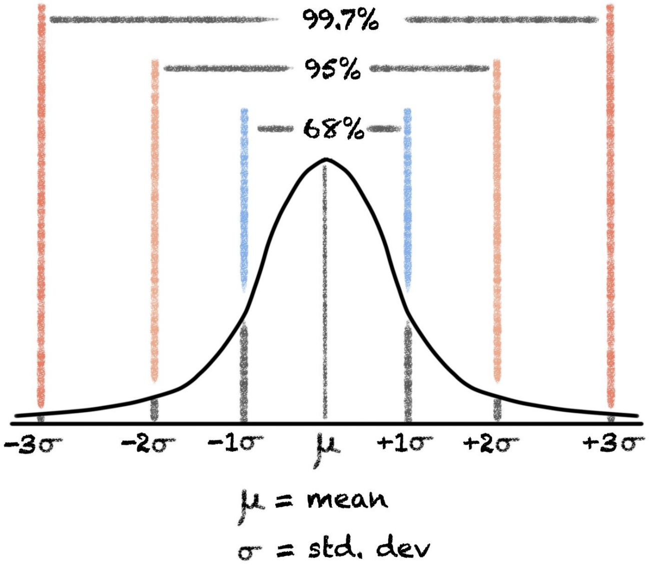
Measures of spread: Example



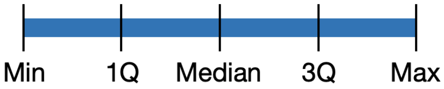
- **Red distribution**: concentrated closely near the mean
- **Blue distribution**: more widely spread out from the mean
- They have the same mean, skewness, modality

Measures of spread: Variance and standard deviation

- Standard deviation can roughly be interpreted as the mean distance from mean
- **Rules of thumb** for symmetric, bell-shaped distributions: 68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively



Measures of spread: Range and interquartile range



- **Percentile**: a number that divides ordered data into hundredths
  - Median = 50th percentile
- **Quartile**: a number that divides ordered data into quarters
  - First quartile = 25th percentile
  - Second quartile = Median = 50th percentile
  - Third quartile = 75th percentile
- **Range** = Max - min; sensitive to outliers
- **Interquartile range** (IQR) = 3rd - 1st quartile
  - `IQR()` in R
- **Five-number summary**: Min, 1Q, Median, 3Q, Max
  - `summary()` in R (also gives mean)

Loans data

```
sd(loans$loan_amount)
```

```
[1] 10301.96
```

```
var(loans$loan_amount)
```

```
[1] 106130313
```

```
sqrt(var(loans$loan_amount))
```

```
[1] 10301.96
```

```
summary(loans$loan_amount)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1000    8000   14500   16362   24000   40000 
```

```
IQR(loans$loan_amount)
```

```
[1] 16000
```

???

Here we use R functions to get these summary statistics of the loan_amount variable.

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

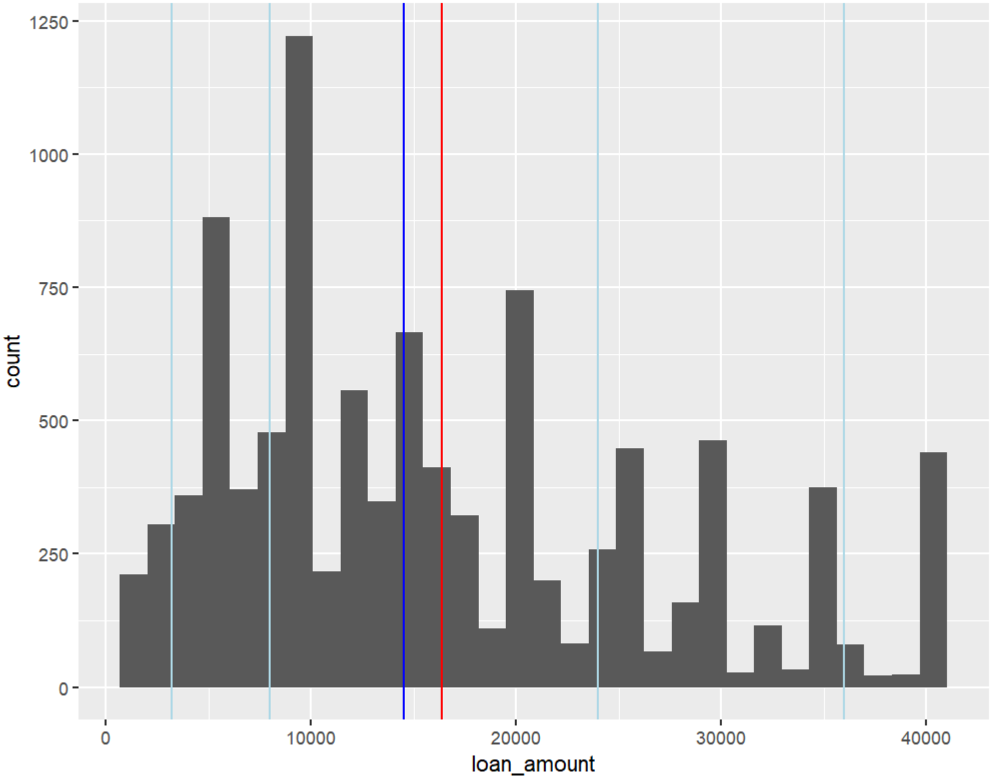
```
loans %>%
  group_by(homeownership) %>%
  summarise(var = var(loan_amount),
            mean = mean(loan_amount))
```

```
# A tibble: 3 x 3
  homeownership        var   mean
  <fct>              <dbl>  <dbl>
1 MORTGAGE      113425575. 18129.
2 OWN           100698933. 15684.
3 RENT           91169018. 14406.
```

Percentiles

Vertical lines for 5th, 25th percentile, median, mean, 75th and 95th percentiles

Plot Code



Percentiles

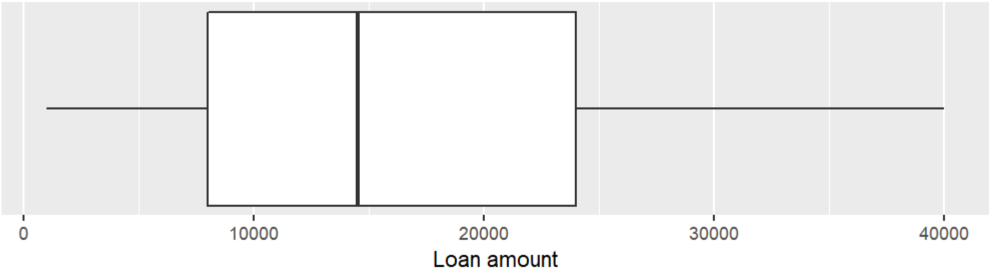
Vertical lines for 5th, 25th percentile, median, mean, 75th and 95th percentiles

Plot Code

```
ggplot(loans, aes(x = loan_amount)) +
  geom_histogram() +
  geom_vline(xintercept = median(loans$loan_amount),
             col = "blue") +
  geom_vline(xintercept = quantile(loans$loan_amount, .05),
             col = "lightblue") +
  geom_vline(xintercept = quantile(loans$loan_amount, .25),
           col = "lightblue") +
  geom_vline(xintercept = quantile(loans$loan_amount, .75),
           col = "lightblue") +
  geom_vline(xintercept = quantile(loans$loan_amount, .95),
           col = "lightblue") +
  geom_vline(xintercept = mean(loans$loan_amount),
             col = "red")
```

Boxplots

```
ggplot(loans, aes(x = loan_amount)) +
  geom_boxplot() +
  labs(x = "Loan amount") +
  scale_y_continuous(breaks = NULL)
```

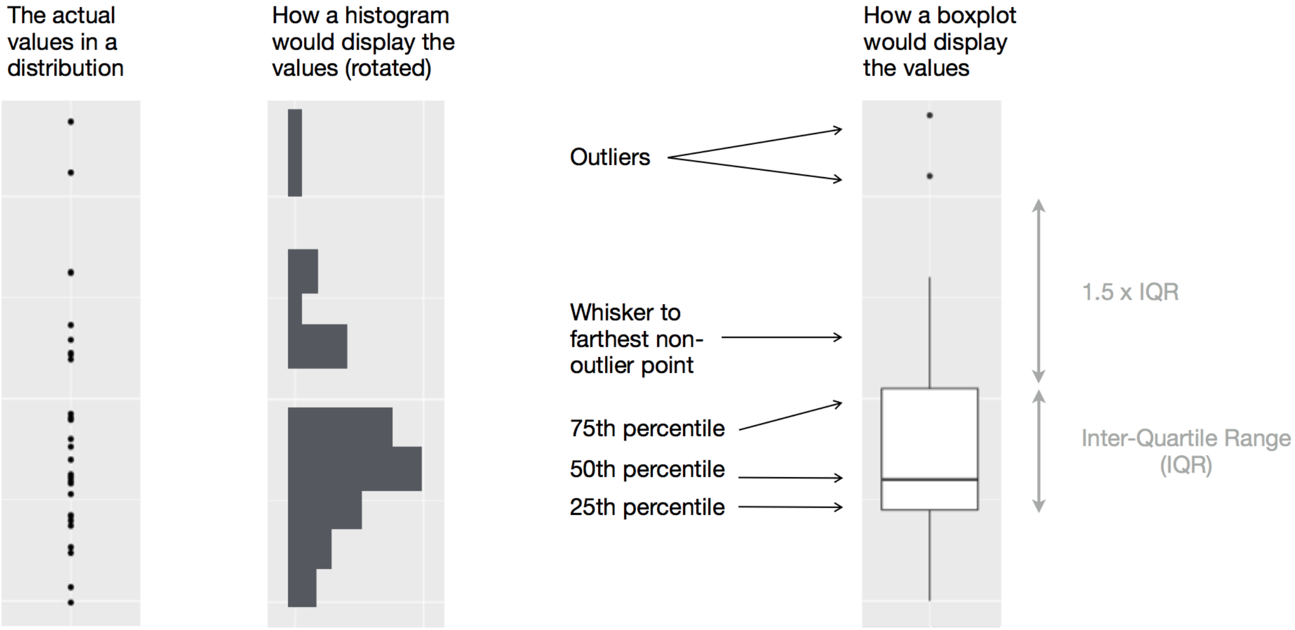


- Lower whisker, box (1Q, median, 3Q), upper whisker
- Total length of the box is IQR
- The length of each whisker is up to 1.5*IQR
- Any points beyond that are **outliers**, observations that are unusually far from the rest of the data
- Outliers appear as points

> Boxplots are particularly useful for identifying outliers and understanding the spread of the data.

Boxplots

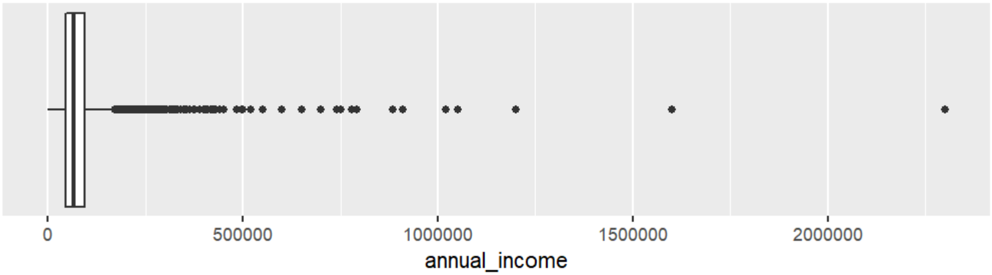
- Boxplot provide a concise summary of the data


The actual values in a distribution
How a histogram would display the values (rotated)
How a boxplot would display the values
Outliers
Whisker to farthest non-outlier point
75th percentile
50th percentile
25th percentile
1.5 x IQR
Inter-Quartile Range (IQR)

Box plot and outliers

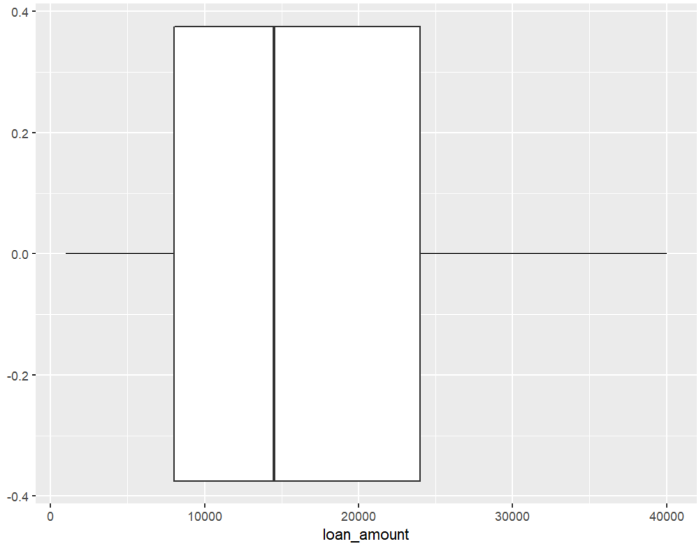
Income data are often skewed (right or left?)

```
ggplot(loans, aes(x = annual_income)) +
  geom_boxplot()+
  scale_y_continuous(breaks = NULL)
```

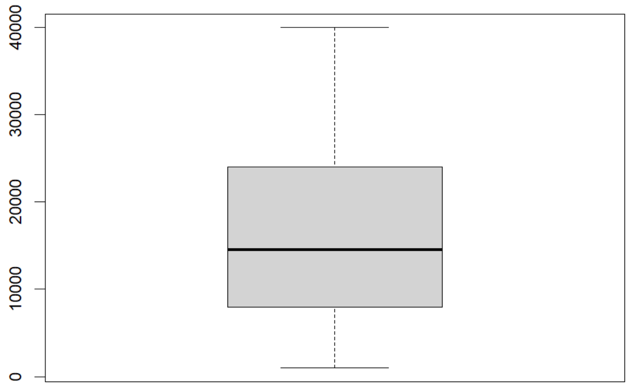


Boxplots in base R

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



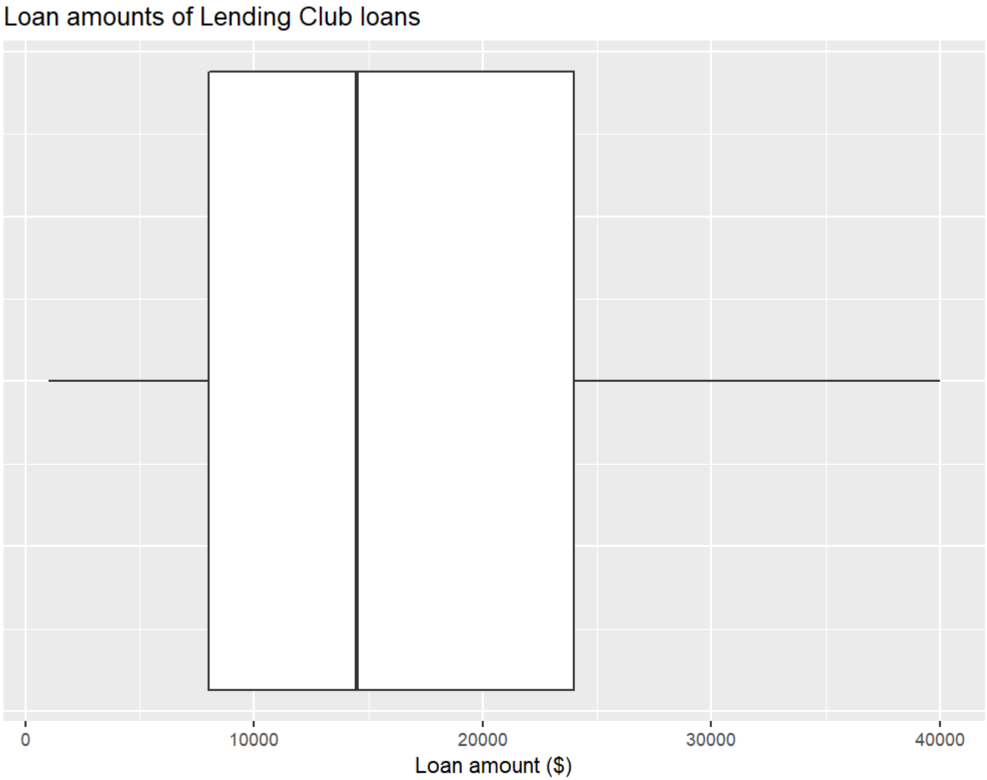
```r
boxplot(loans$loan_amount)
```



Another way to remove y-axis labels

We saw `scale_y_continuous(breaks = NULL)` earlier

Plot Code



Another way to remove y-axis labels

We saw `scale_y_continuous(breaks = NULL)` earlier

Plot Code

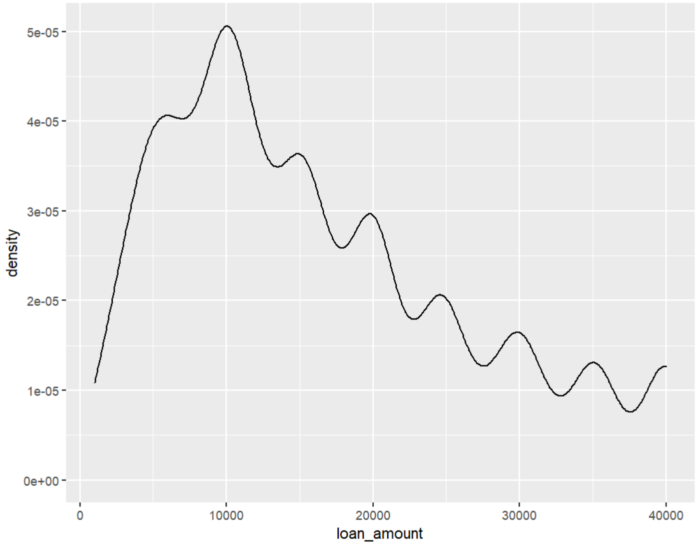
```
ggplot(loans, aes(x = loan_amount)) +
  geom_boxplot() +
  labs(
    x = "Loan amount ($)",
    y = NULL,
    title = "Loan amounts of Lending Club loans"
  ) +
  theme(
    axis.ticks.y = element_blank(),
    axis.text.y = element_blank()
  )
```

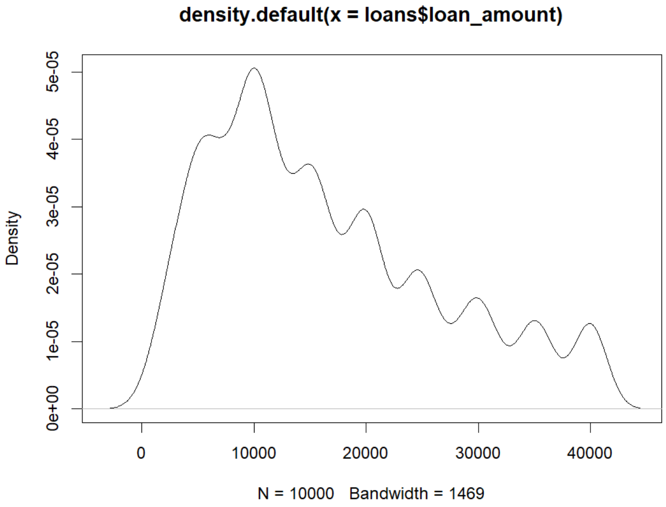
Density plot

Density plots are an alternative to histograms

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

```
plot(density(loans$loan_amount))
```



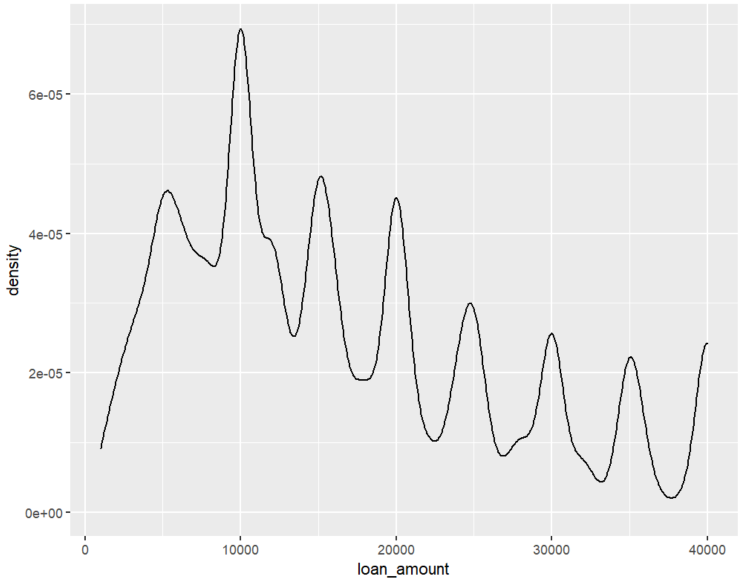


Density plots and adjusting bandwidth

- Larger adjust = smoother

adjust = 0.5 adjust = 1 adjust = 2

```
ggplot(loans, aes(x = loan_amount)) +
  geom_density(adjust = 0.5)
```

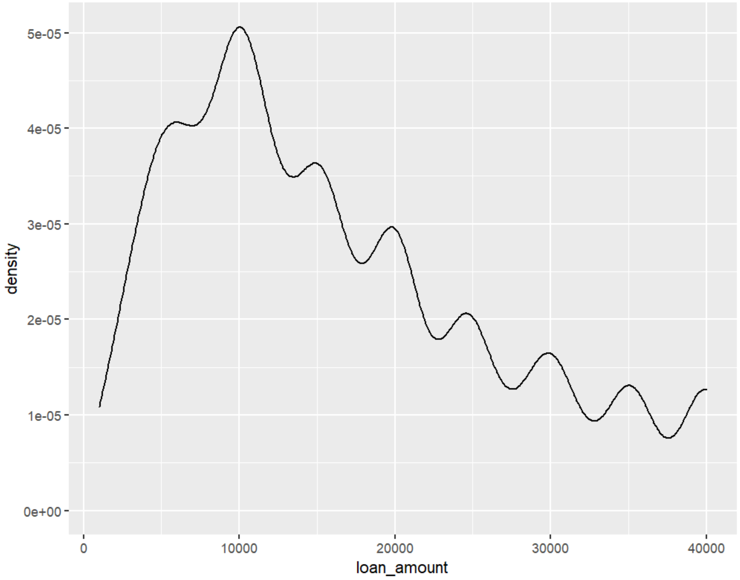


Density plots and adjusting bandwidth

- Larger adjust = smoother

adjust = 0.5 adjust = 1 adjust = 2

```
ggplot(loans, aes(x = loan_amount)) +
  geom_density(adjust = 1) # default bandwidth
```

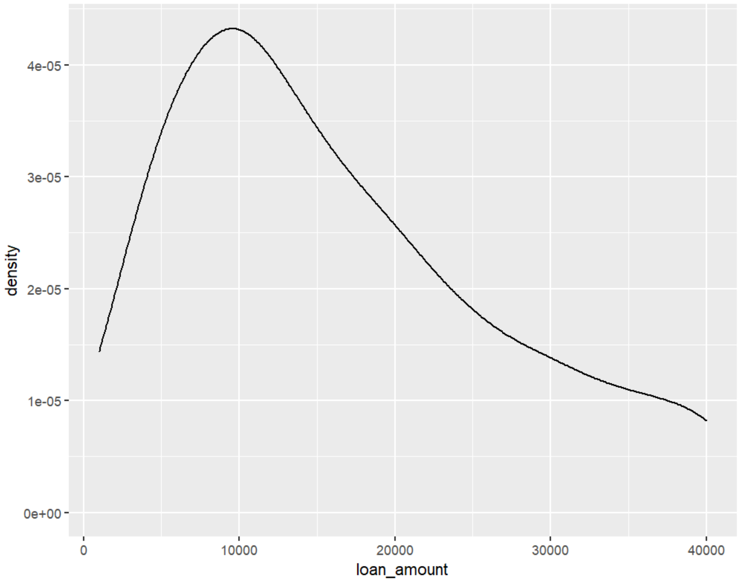


Density plots and adjusting bandwidth

- Larger adjust = smoother

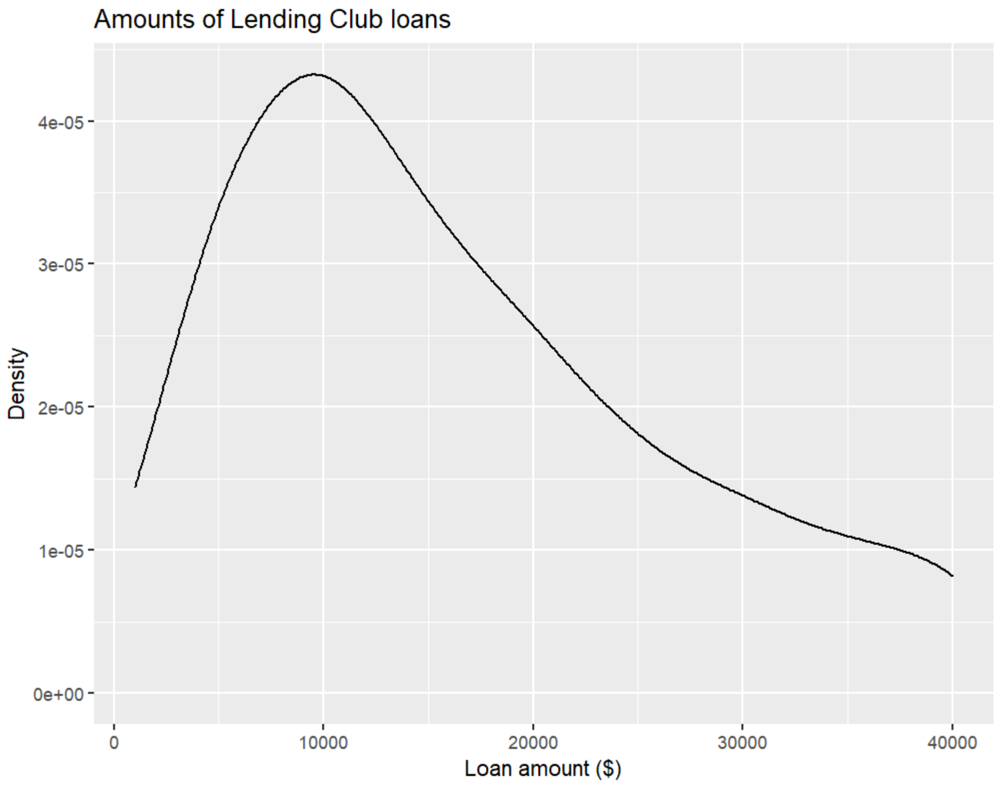
adjust = 0.5 adjust = 1 adjust = 2

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



Title and labels

Plot Code



Title and labels

Plot Code

```
ggplot(loans, aes(x = loan_amount)) +
  geom_density(adjust = 2) +
  labs(
    x = "Loan amount ($)",
    y = "Density",
    title = "Amounts of Lending Club loans"
  )
```

Next lecture: Relationships between numerical variables

- Paired or bivariate data
  - Scatterplot
  - Hexplot
  - Correlation
  - Line graph

Summary

- Describing numerical distributions
    - Histograms
    - Measures of central tendency: mean, median, mode
    - Shape: skewness and modality
    - Spread: variance and standard deviation, range and interquartile range
    - Boxplots
    - Unusual observations
    - Density plot

Reading

- Open Intro Statistics Chapter 1
- Chapter 9 Visualizing data distributions