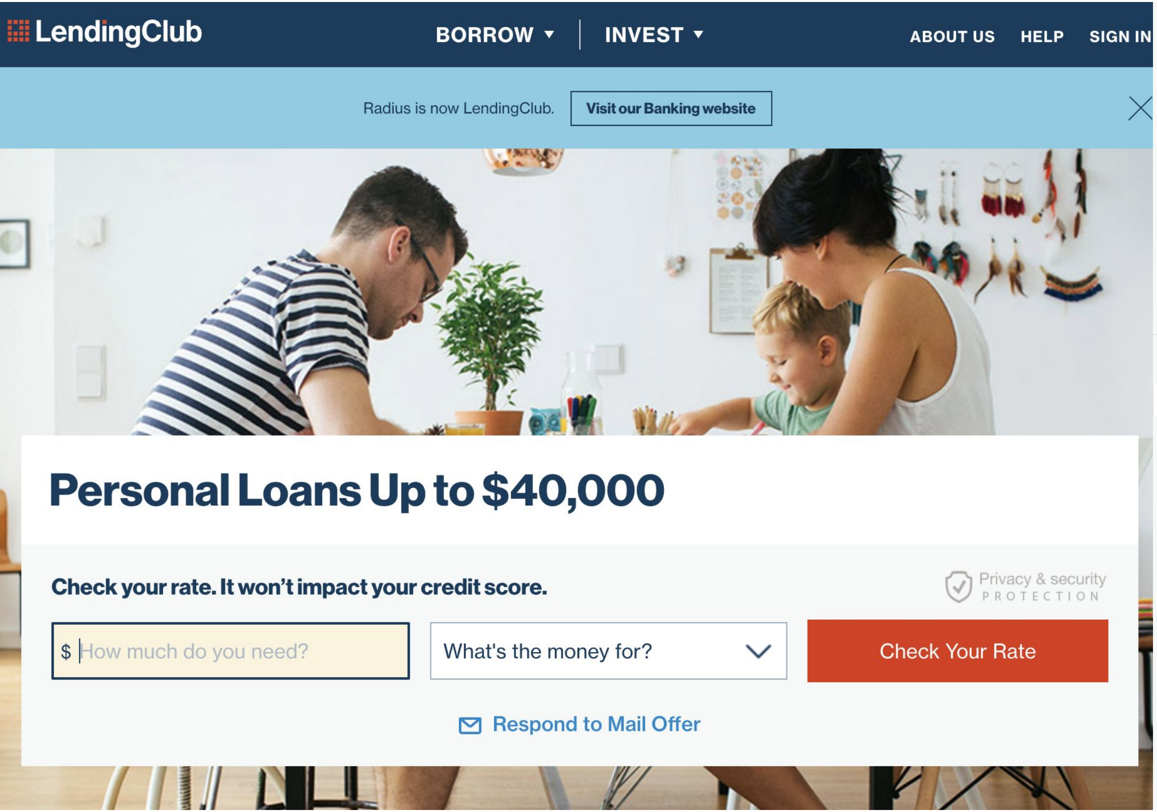
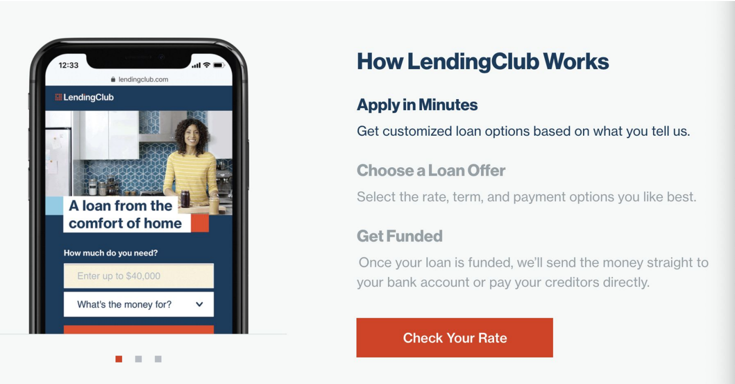
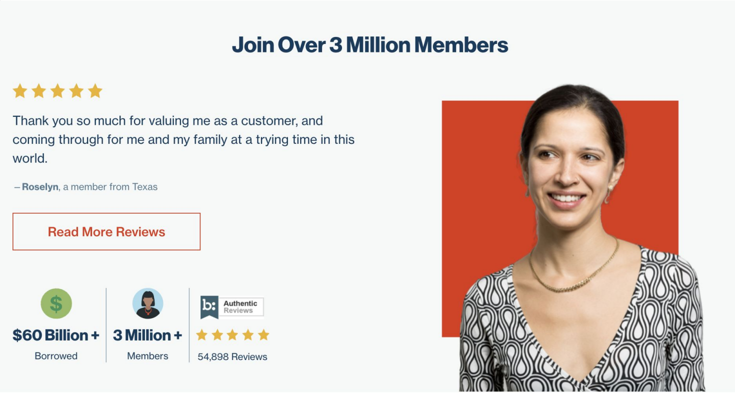
Exploratory data analysis

Exploring numerical data

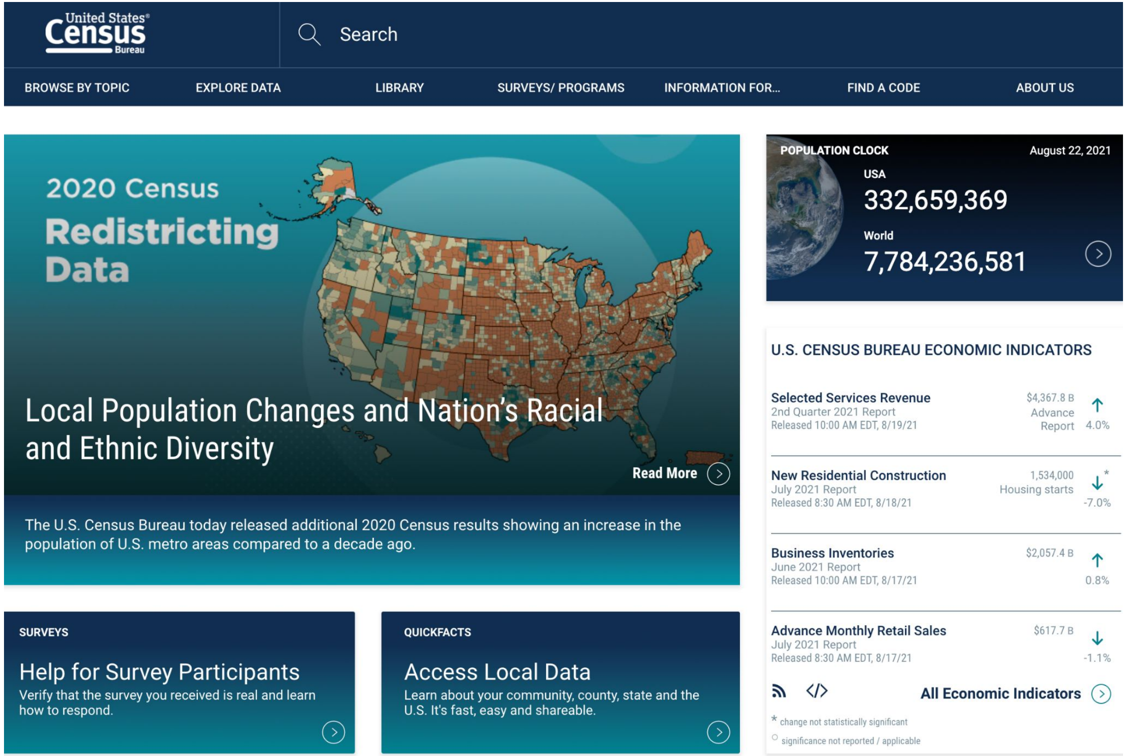
Prof. Dr. Jan Kirenz
HdM Stuttgart


LendingClub
BORROW
INVEST
ABOUT US
HELP
SIGN IN
Radius is now LendingClub.
Visit our Banking website
Personal Loans Up to $40,000
Check your rate. It won't impact your credit score.
Privacy & security
PROTECTION
$ How much do you need?
What's the money for?
Check Your Rate
Respond to Mail Offer


12:33
lendingclub.com
LendingClub
A loan from the comfort of home
How much do you need?
Enter up to $40,000
What's the money for?
How LendingClub Works
Apply in Minutes
Get customized loan options based on what you tell us.
Choose a Loan Offer
Select the rate, term, and payment options you like best.
Get Funded
Once your loan is funded, we'll send the money straight to your bank account or pay your creditors directly.
Check Your Rate


Join Over 3 Million Members
Thank you so much for valuing me as a customer, and coming through for me and my family at a trying time in this world.
– Roselyn, a member from Texas
Read More Reviews
$60 Billion +
Borrowed
3 Million +
Members
Authentic Reviews
54,898 Reviews

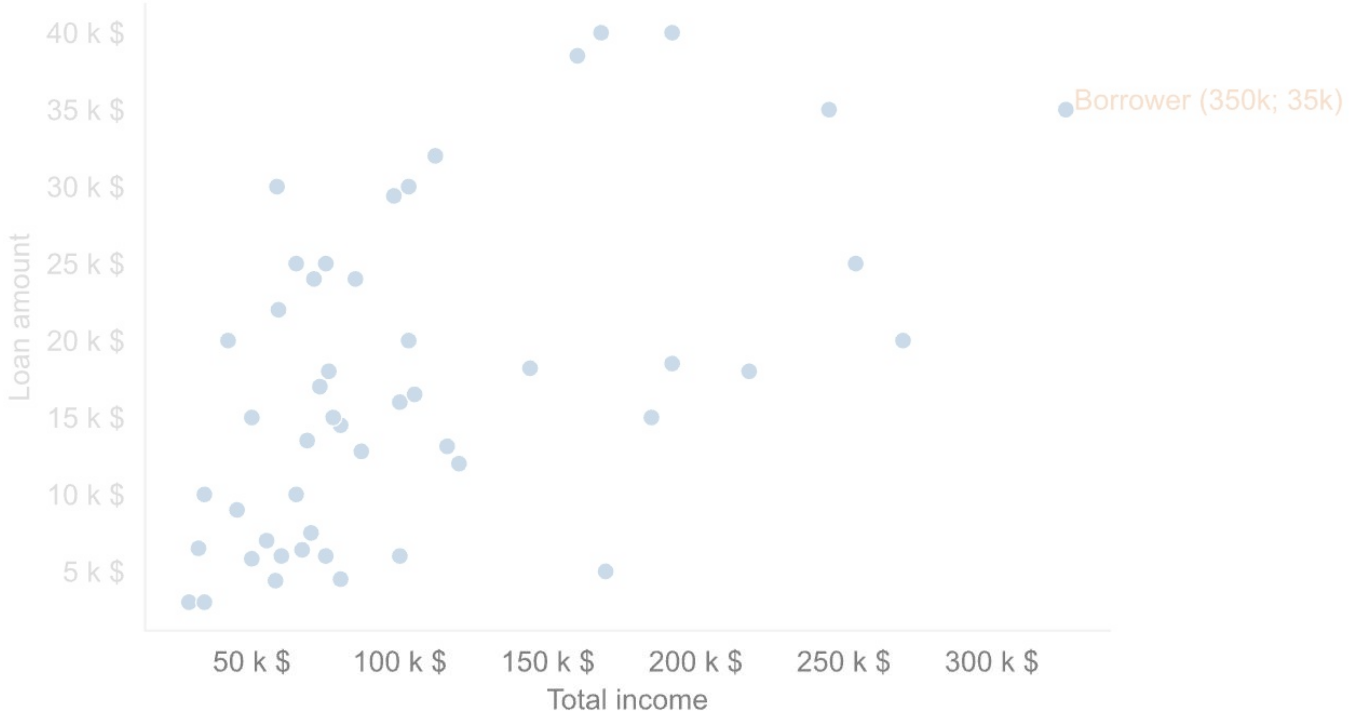
Data for 3142 counties in the United States



County dataset: https://www.openintro.org/data/index.php?data=county

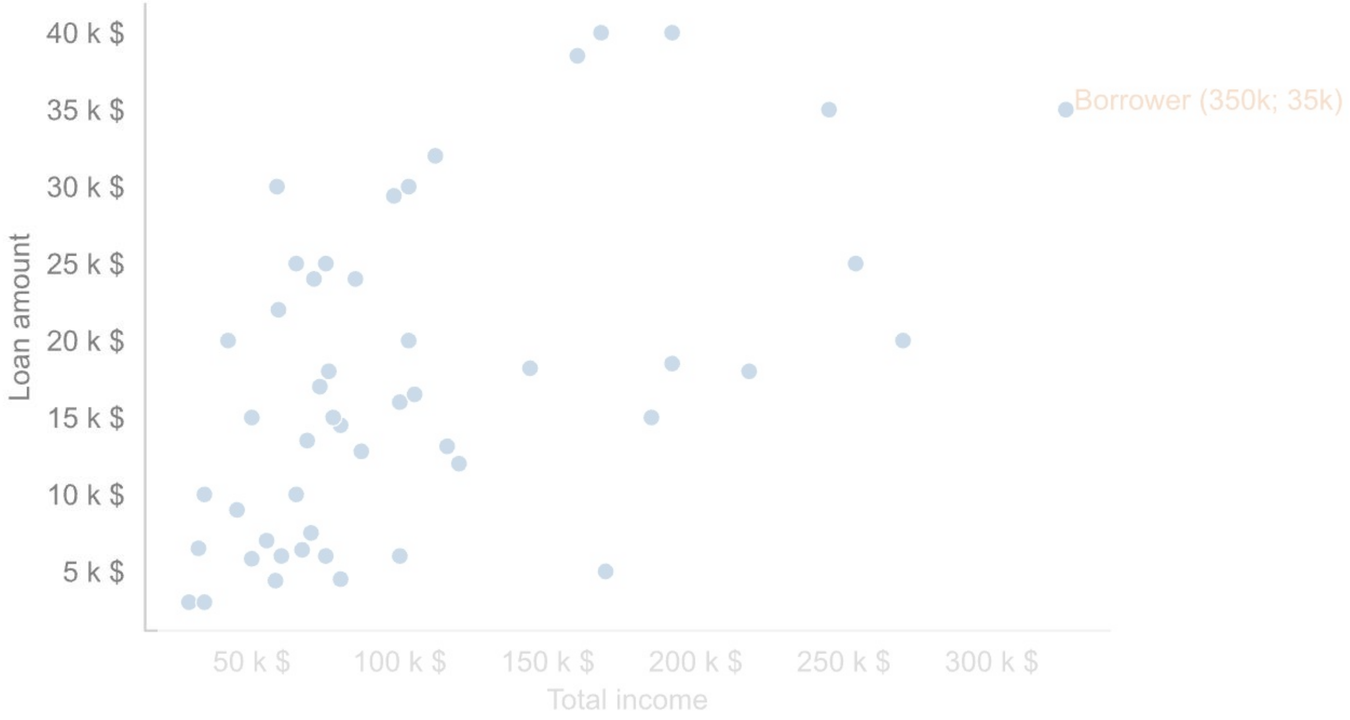
Scatterplots for paired data

Total income is on the x-axis



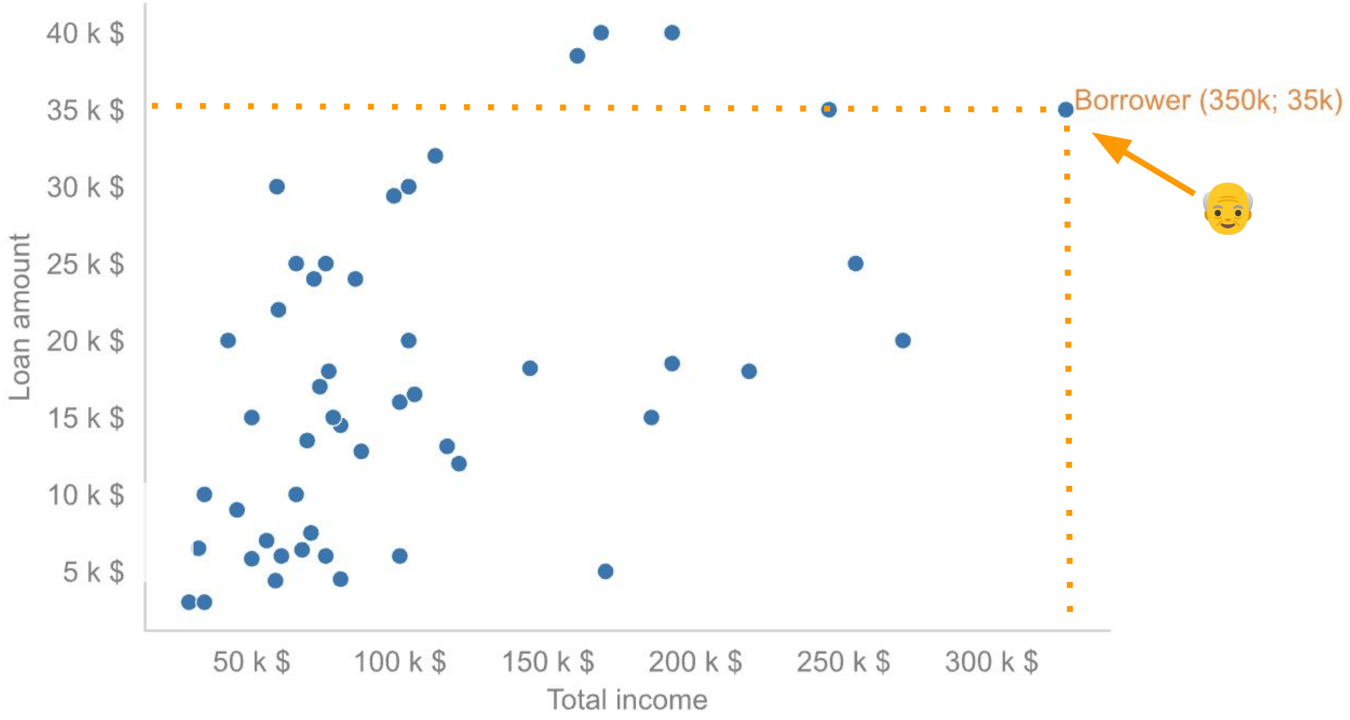
Data: loans50

Loan amount is on the y-axis



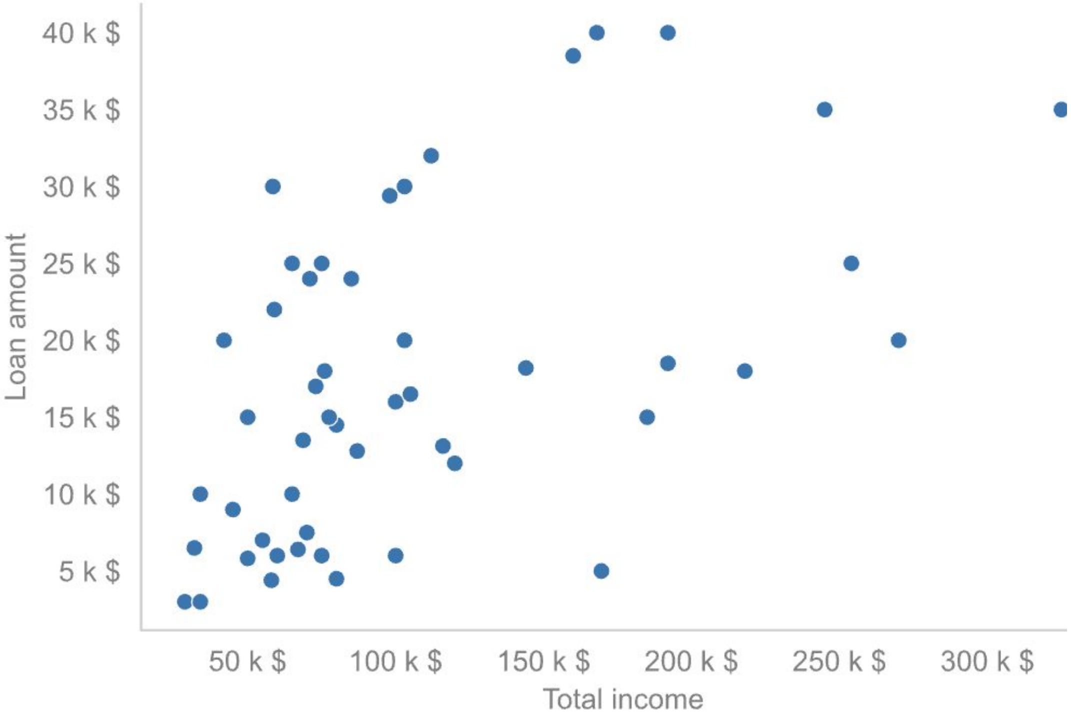
Data: loans50

This is a borrower with a **total income** of 350 k and a **loan amount** of 35 k



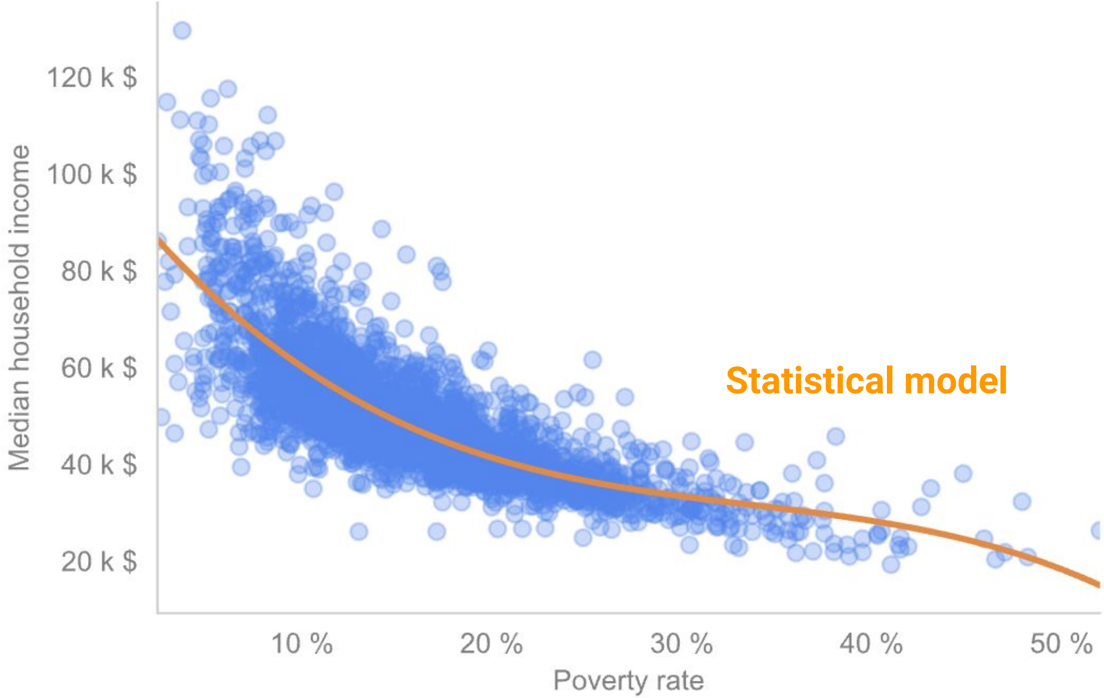
Data: loans50

Each point represents a single case (borrower)



Data: loans50

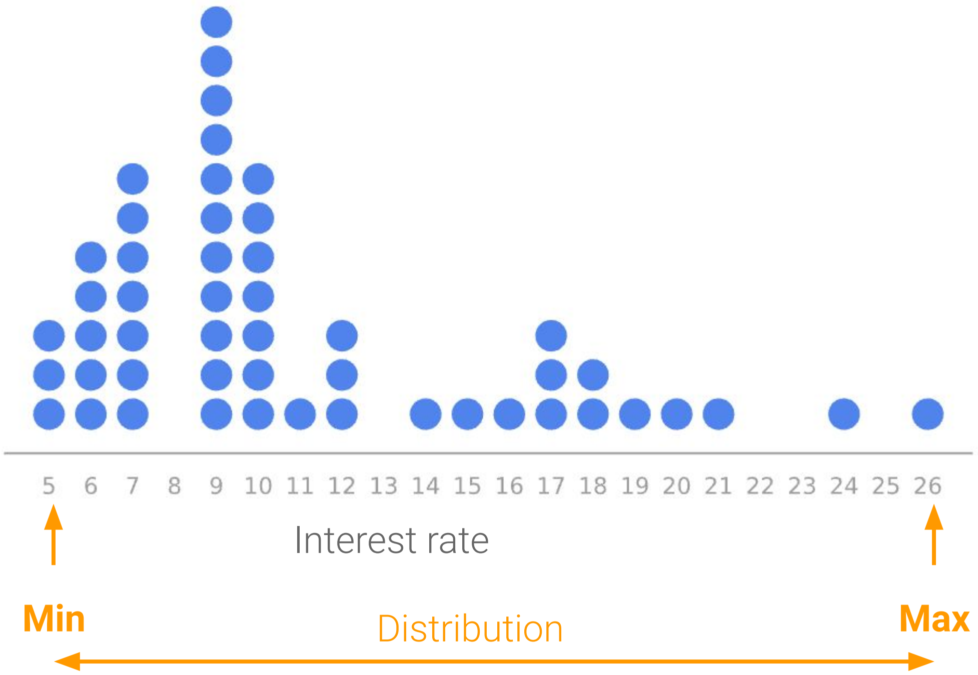
A scatterplot of the median household income against the poverty rate for the county dataset



Data: county

Dot plots

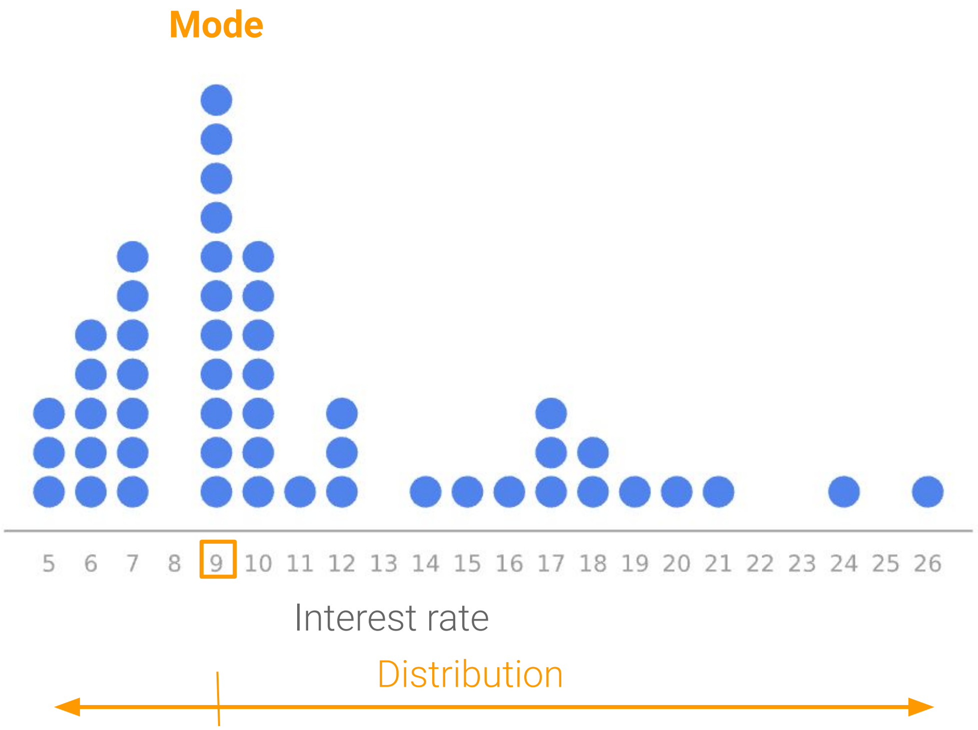
A dot plot of interest rate (numbers are rounded)



n = 50 (there are 50 dots)

Data: loan50

The **mode** is the value with the most occurences



Median

- If the data are ordered from smallest to largest, the median is the observation right in the middle.
- If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

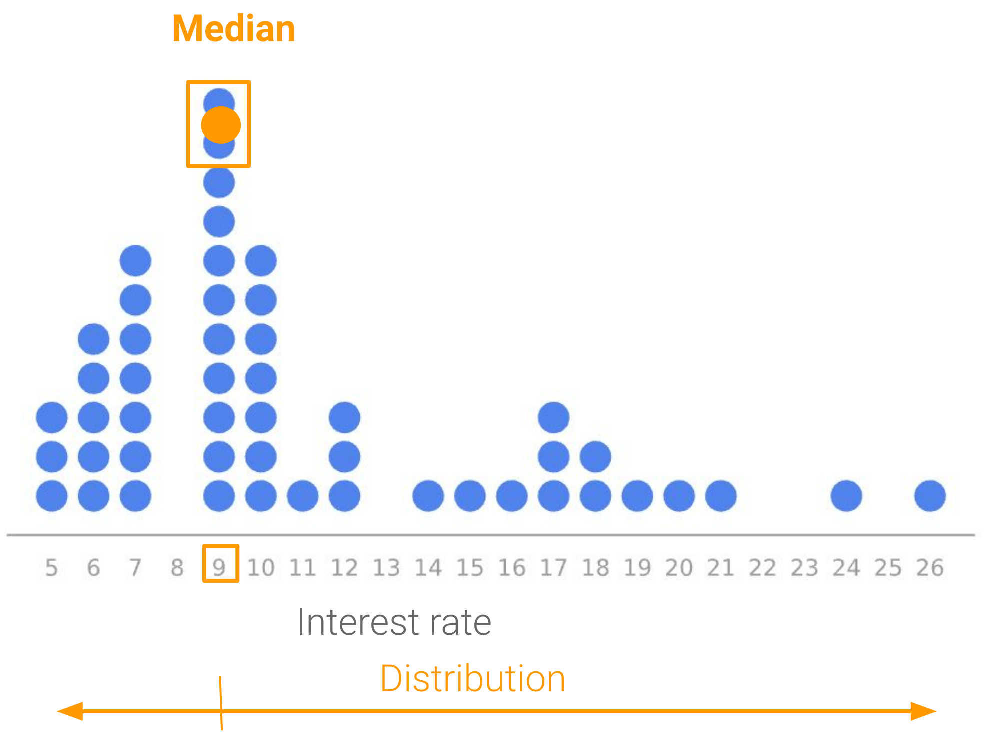
Interest rates from the loan50 dataset, arranged in ascending order (original values)

	1	2	3	4	5	6	7	8	9	10
1	5.31	5.31	5.32	6.08	6.08	6.08	6.71	6.71	7.34	7.35
10	7.35	7.96	7.96	7.96	7.97	9.43	9.43	9.44	9.44	9.44
20	9.92	9.92	9.92	9.92	9.93	9.93	10.42	10.42	10.90	10.90
30	10.91	10.91	10.91	11.98	12.62	12.62	12.62	14.08	15.04	16.02
40	17.09	17.09	17.09	18.06	18.45	19.42	20.00	21.45	24.85	26.30

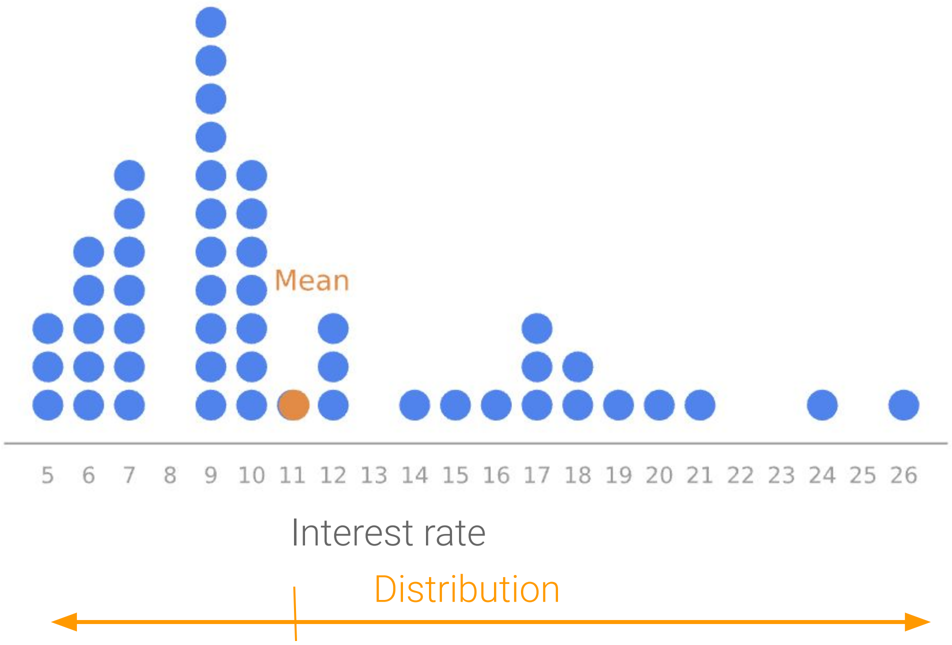
The middle

Median = (9.93 + 9.93) / 2 = 9.93

The **median** is the number in the middle



The **mean** is the center of the distribution (the average)



The **sample mean**

The sample mean, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where x_1, x_2, ..., x_n represent the n observed values.

- The sample mean is a sample statistic, and serves as a **point estimate** of the population mean.
- This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

Results of a trial of 1500 adults that suffer from asthma

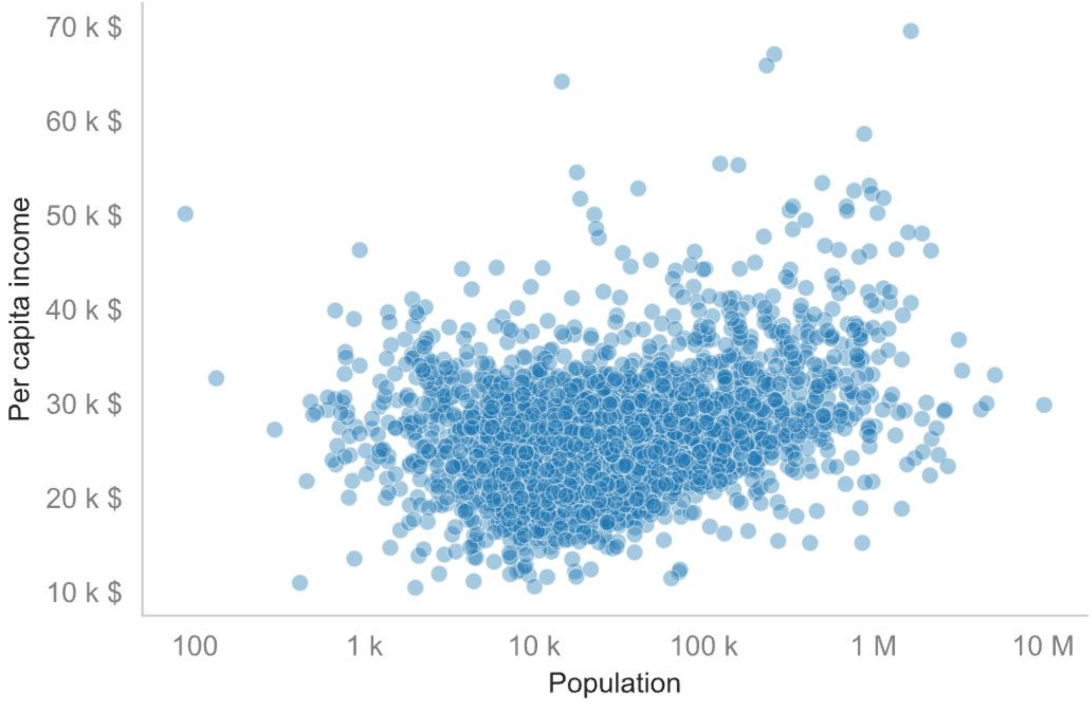
	Description	New drug	Standard drug
0	Number of patients	500	1000
1	Total asthma attacks	200	300

- New drug: 200/500 = 0.4 asthma attacks per patient
- Standard drug: 300/1000 = 0.3 asthma attacks per patient

The **population mean**

- The population mean is also computed the same way but is denoted as **μ.**
- It is often not possible to calculate μ since population data are rarely available

Per capita income against population size in 3,143 US counties



The **weighted mean**

The weighted mean can be calculated as

$$\text{weighted mean of } x_i\text{s} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$$

x_1, x_2, ..., x_n represent the n observed values.

w_1, w_2, ..., w_n represent the n weights.

- The simple mean is a weighted mean where all the weights are 1:

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1 \times x_1 + 1 \times x_2 + 1 \times x_3 + \cdots + 1 \times x_n}{1 + 1 + 1 + \cdots + 1}$$

Histograms

A histogram for interest rate with 3 **bins**, a **bin width** of 10 and a **range of scores** of 30

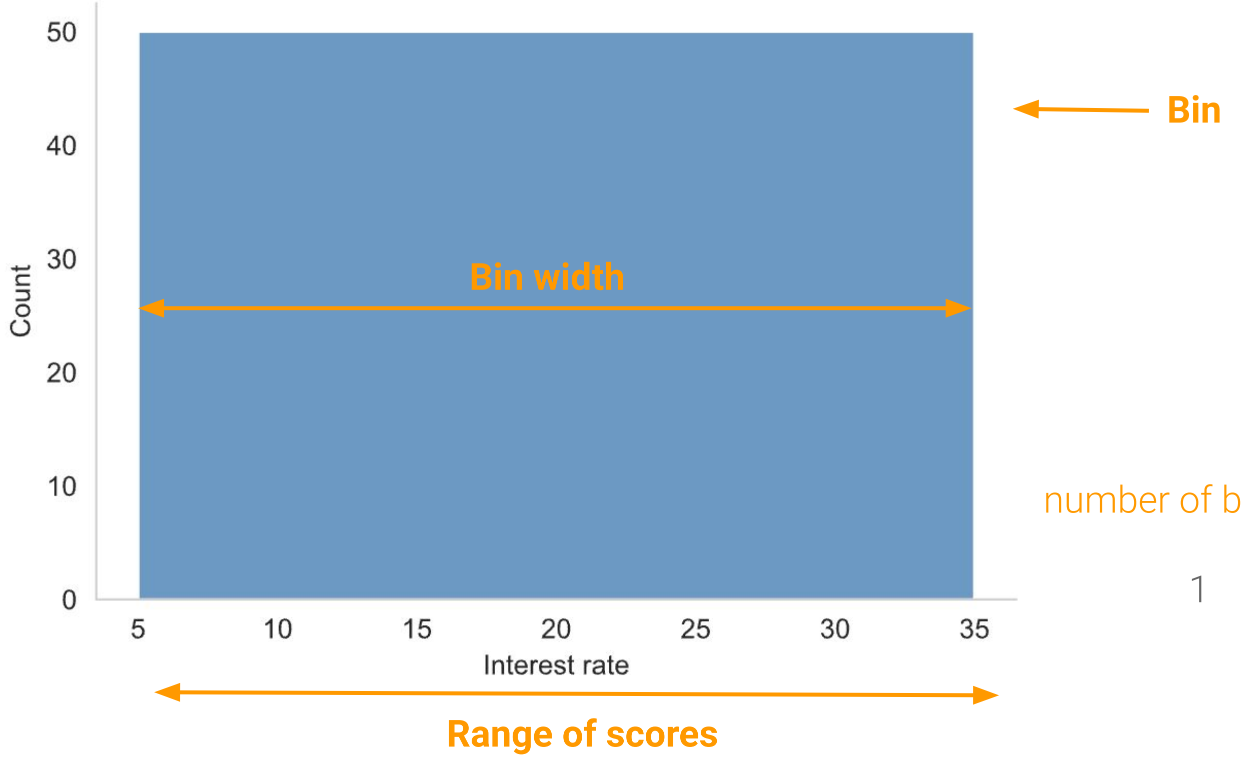
bin width = range of scores / number of bins

10 = (35 - 5) / 3

number of bins = range of scores / bin width

3 = (35 - 5) / 10

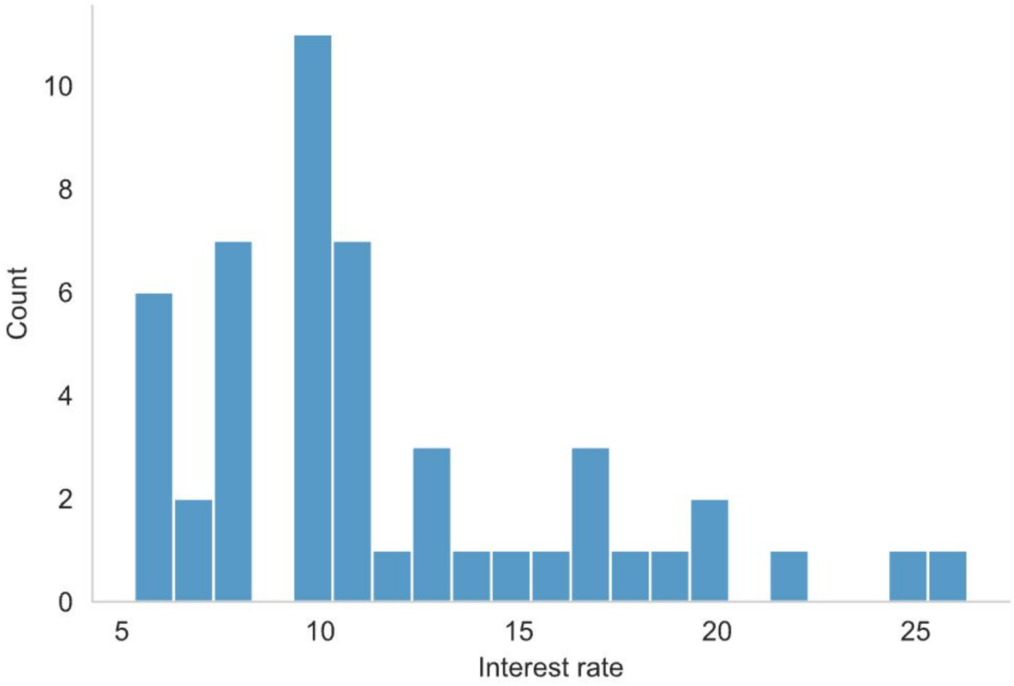
A histogram for interest rate with 1 **bin**, **bin width** of 30 and a **range of scores** of 30



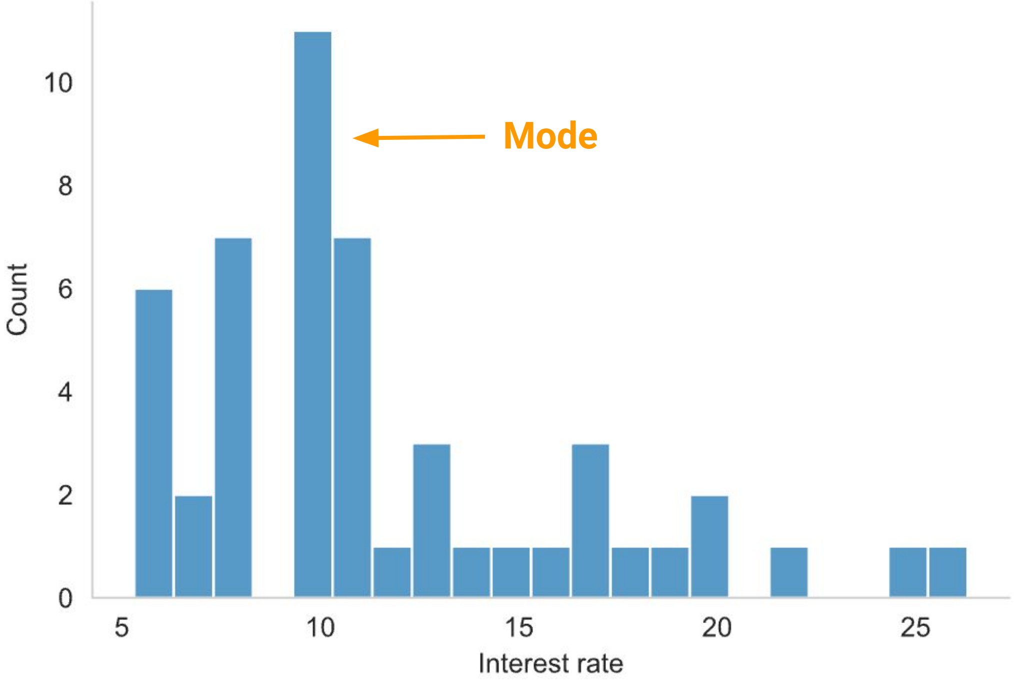
number of bins = range of scores / bin width

1 = (35 - 5) / 30

A histogram for interest rate with **bin width** 1



Histograms can be used to identify **modes**

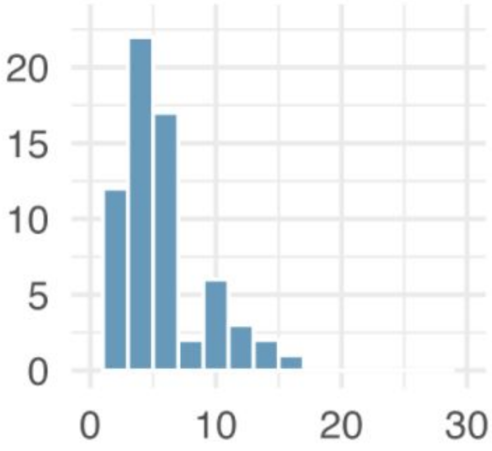


- A mode is represented by a prominent peak in the distribution.

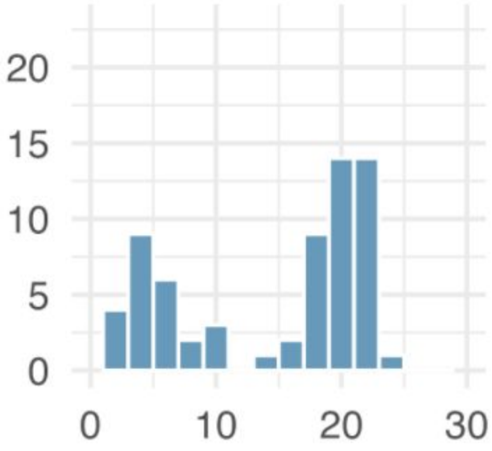
A histogram for interest rate with bin width 1.

Counting only prominent peaks.

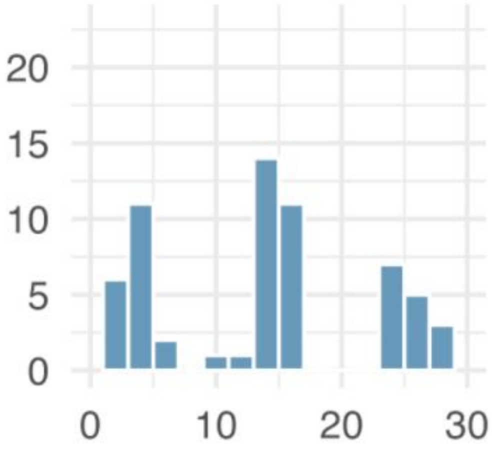
Unimodal



Bimodal

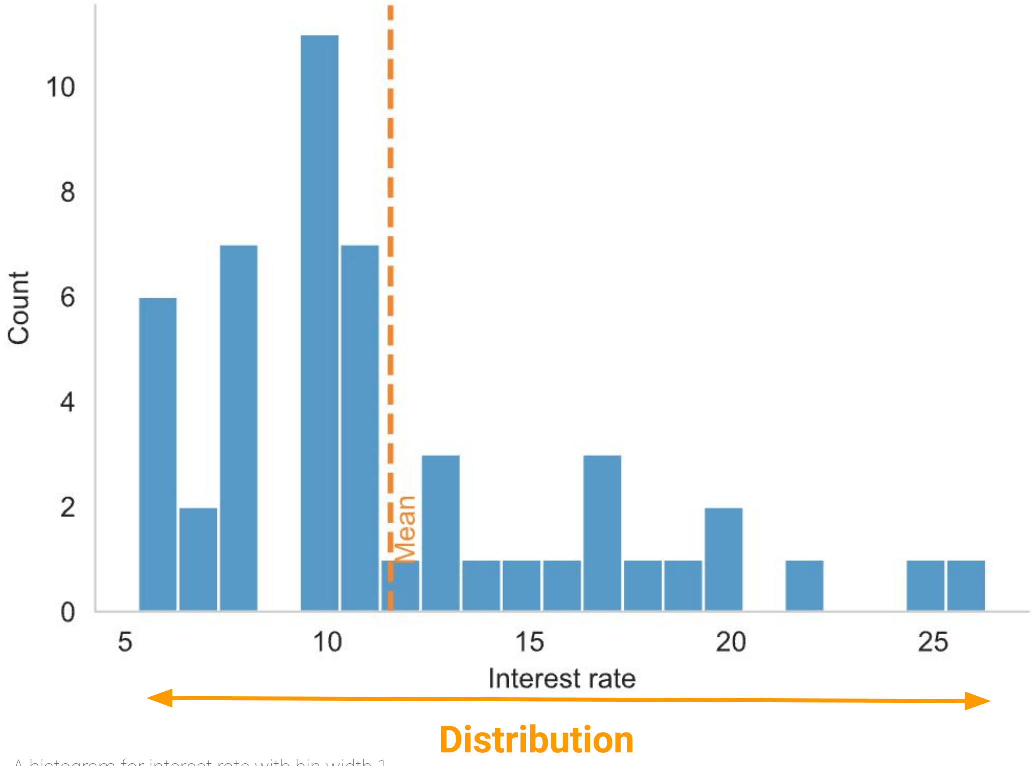


Multimodal



Variance and standard deviation

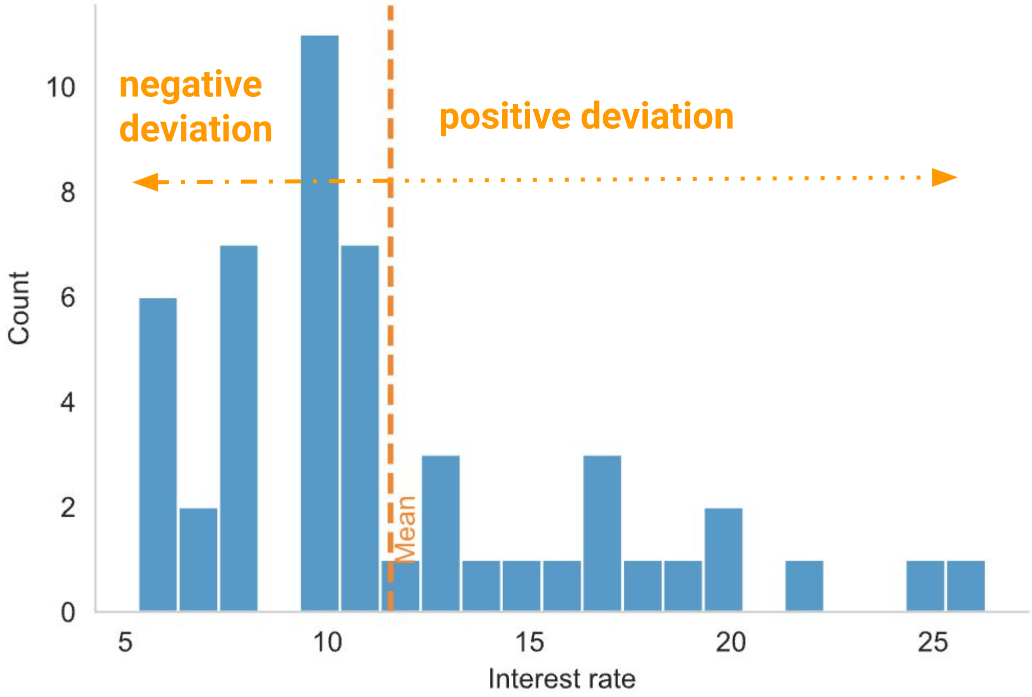
Understanding the shape of the data distribution.



A histogram for interest rate with bin width 1.

- When data trail off in one direction, the distribution has a **long tail.**
- If a distribution has a long left tail, it is **left skewed.**
- If a distribution has a long right tail, it is **right skewed.**

We call the distance of an observation from its mean its **deviation**.



$$
\begin{aligned}
x_1 - \bar{x} &= 10.9 - 11.57 = -0.67 \\
x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\
x_3 - \bar{x} &= 26.3 - 11.57 = 14.73 \\
&\vdots \\
x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49
\end{aligned}
$$

A histogram for interest rate with bin width 1.

The **sample variance**

Deviation:

$$
\begin{aligned}
x_1 - \bar{x} &= 10.9 - 11.57 = -0.67 \\
x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\
x_3 - \bar{x} &= 26.3 - 11.57 = 14.73 \\
&\vdots \\
x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49
\end{aligned}
$$

$$
\begin{aligned}
s^2 &= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \cdots + (-5.49)^2}{50 - 1} \\
&= \frac{0.45 + 2.72 + \cdots + 30.14}{49} \\
&= 25.52
\end{aligned}
$$

If we square these deviations and then take an average, the result is equal to the **sample variance**, denoted by s^2

Variance & standard deviation

- The **variance** is the average squared distance from the mean.
- The **standard deviation** is the square root of the variance.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

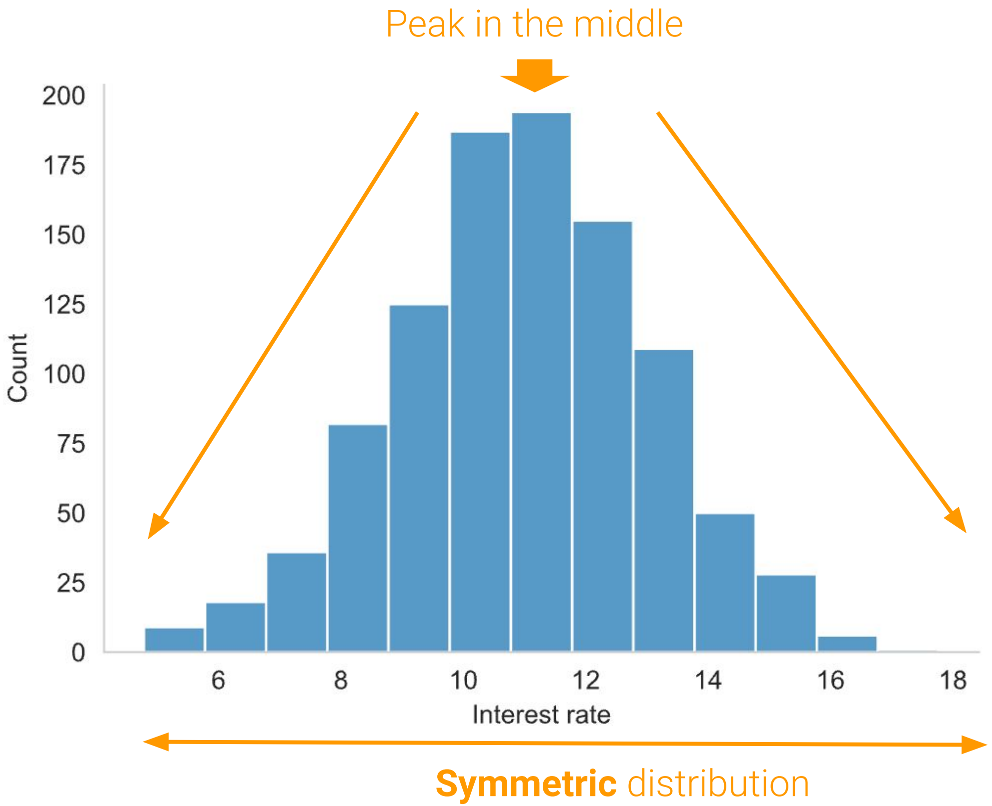
$$s = \sqrt{25.52} = 5.05$$

Standard deviation

- The standard deviation is useful when considering how far the data are **distributed** from the mean.
- The standard deviation represents the **typical deviation** of observations from the mean.

- Often about **68%** of the data will be within **one standard deviation** of the mean
- and about **95%** will be within **two standard deviations**.
- However, these percentages are not strict rules.

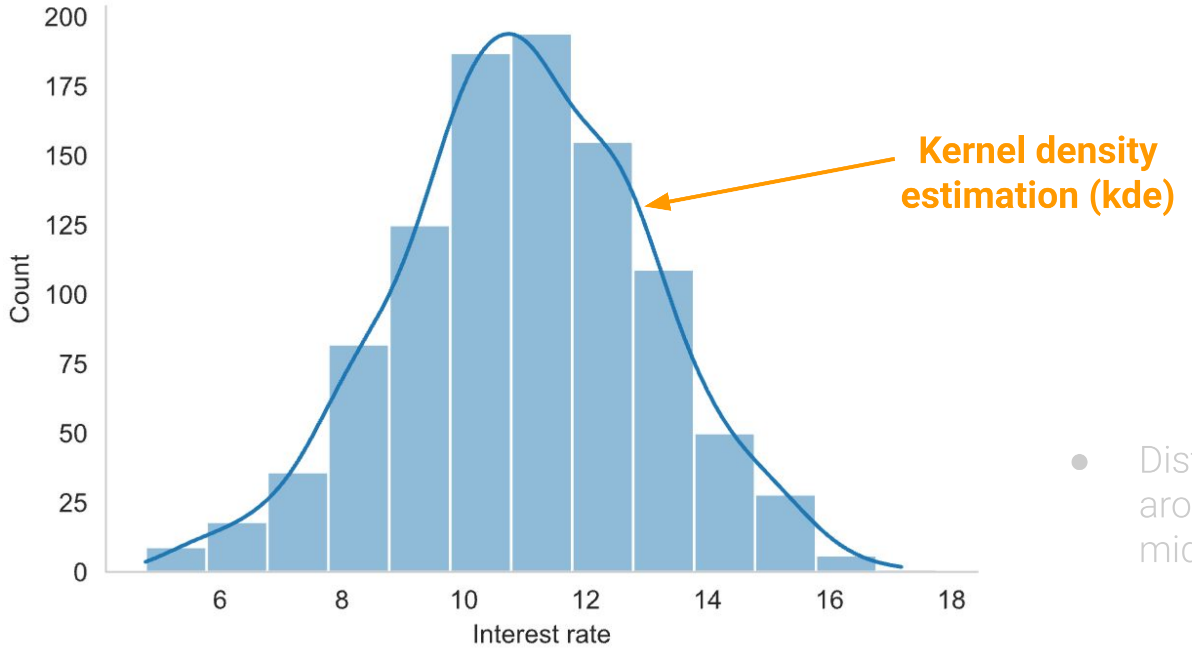
Synthetic generated data with symmetric distribution



A histogram with synthetic data (normal distribution)

- Variables that show roughly equal trailing off in both directions are called **symmetric**.
- Normal (or Gaussian) distribution

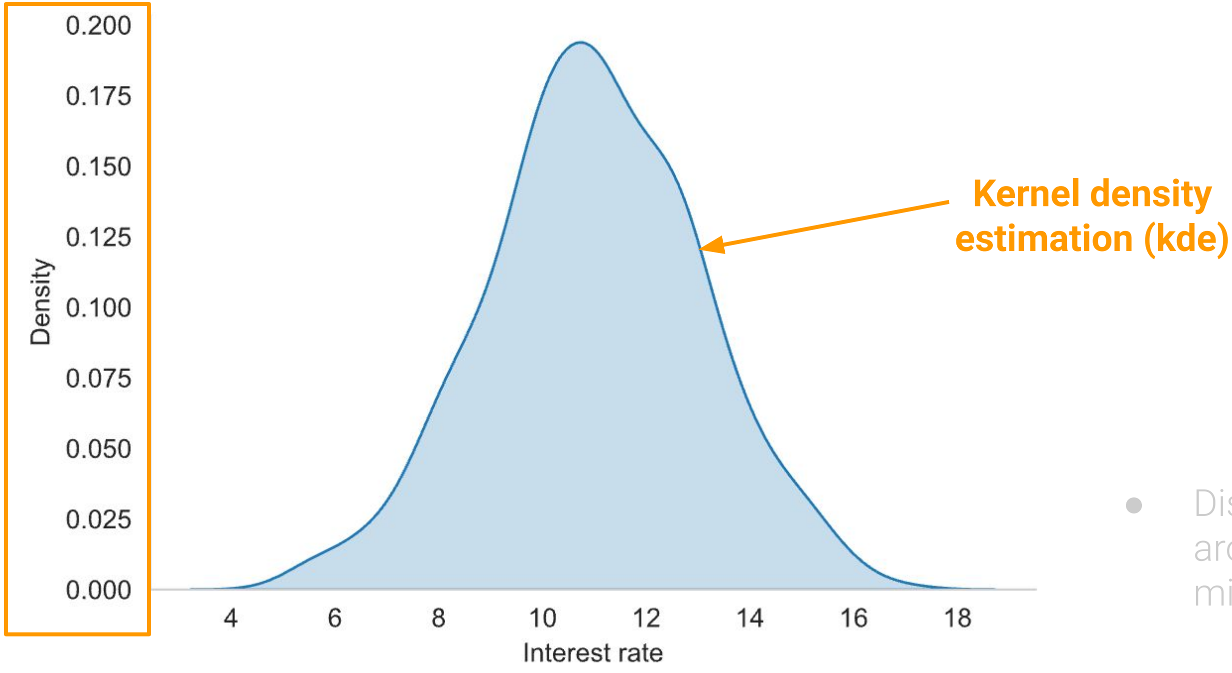
Histogram with kernel density estimation (KDE)



A histogram with synthetic data (normal distribution)

- Distribution is relatively symmetric around a rounded peak in the middle.

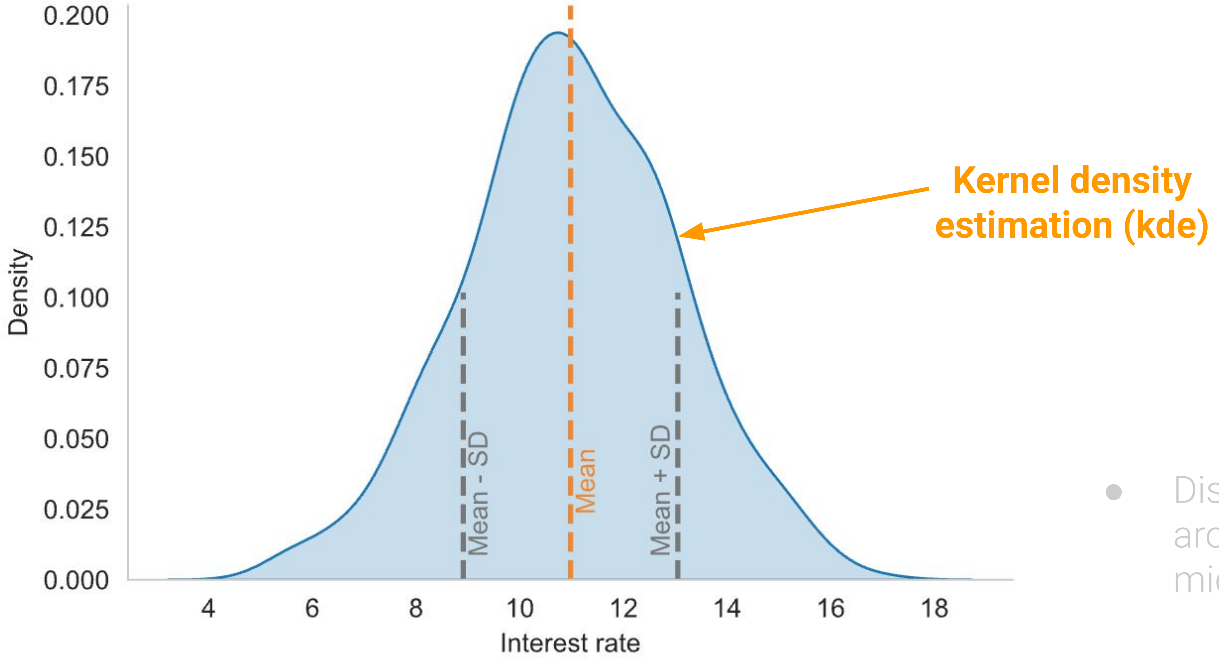
Histogram with kernel density estimation (KDE)



- Distribution is relatively symmetric around a rounded peak in the middle.

A histogram with synthetic data (normal distribution)

Histogram with kernel density estimation (KDE)



- Distribution is relatively symmetric around a rounded peak in the middle.

A histogram with synthetic data (normal distribution)

Example in Google Sheets