

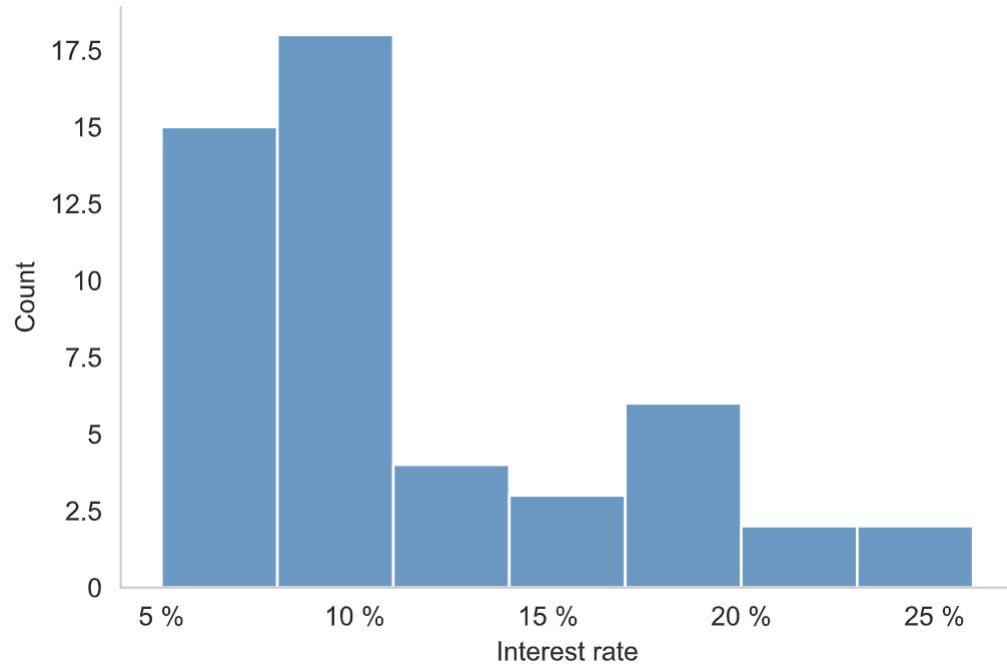
Exploratory data analysis

Exploring numerical data

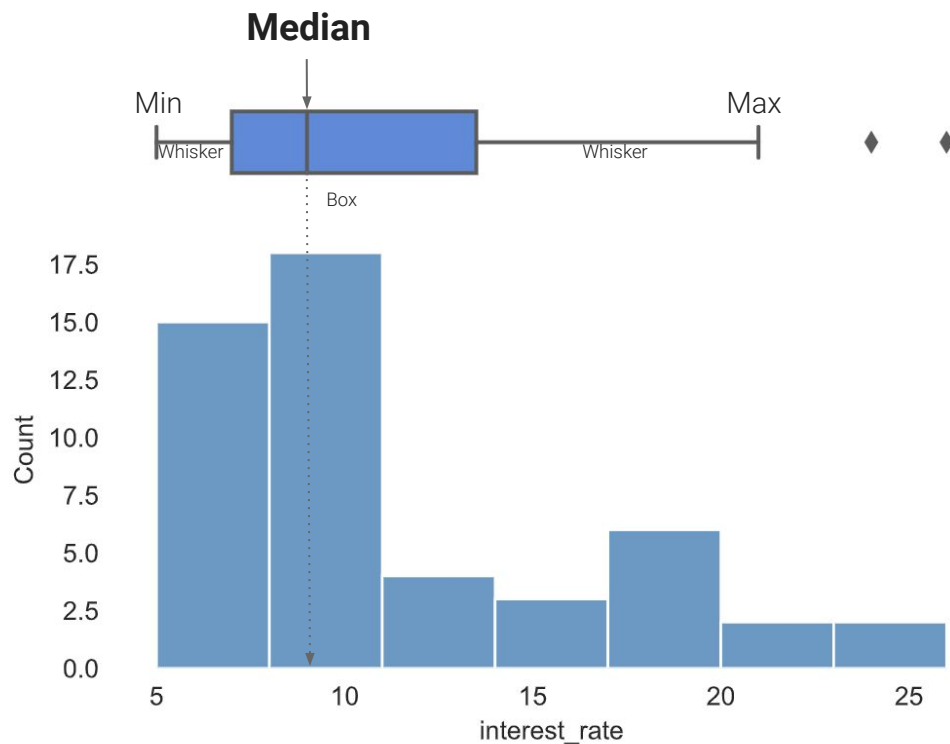
Prof. Dr. Jan Kirenz
HdM Stuttgart

Box plots, quartiles, and the median

Histogram for variable interest rate



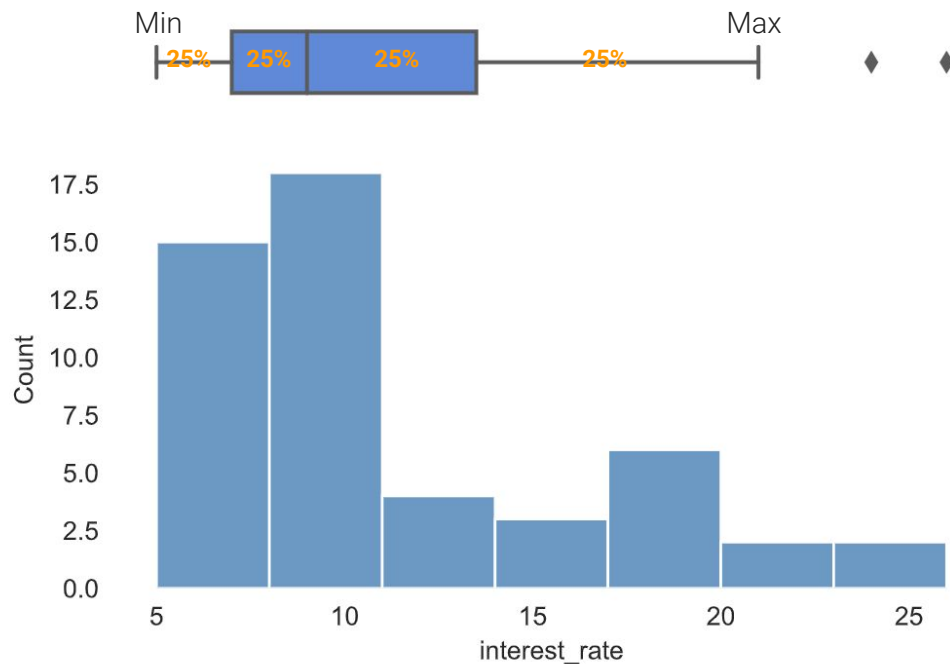
Boxplot and histogram for interest rate



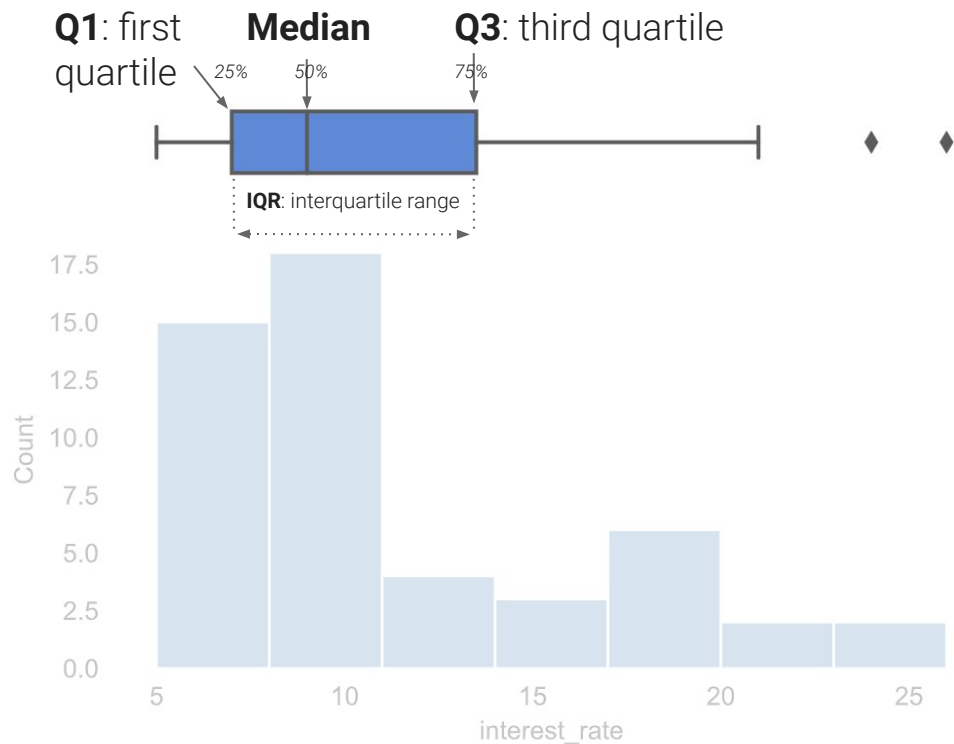
A box plot summarizes a dataset using five statistics while also identifying unusual observations

It includes the median as a bar inside a box with the shape of a rectangle.

Boxplot and quartiles



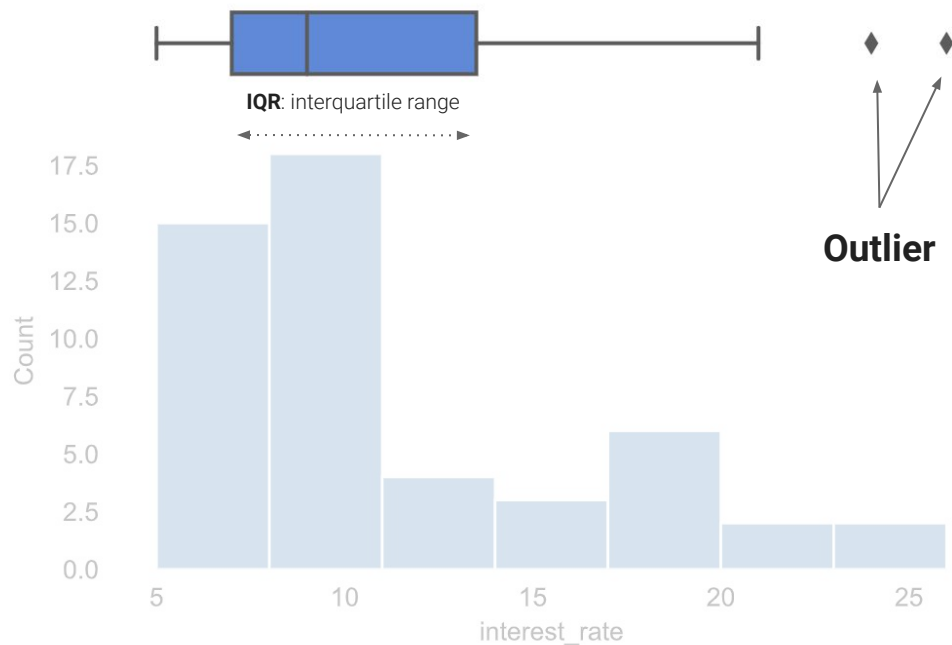
Boxplot, first quartile, third quartile and IQR



The length of the box is called the interquartile range, or IQR.

The two boundaries of the box are called the first quartile (the 25th percentile, i.e., 25% of the data fall below this value) and the third quartile (the 75th percentile, i.e., 75% of the data fall below this value), and these are often labeled Q1 and Q3, respectively.

Boxplot and outlier



An outlier is an observation that appears extreme relative to the rest of the data.

A commonly used formula is that any observation beyond **$1.5 \times \text{IQR}$** away from the first or the third quartile is considered an outlier.

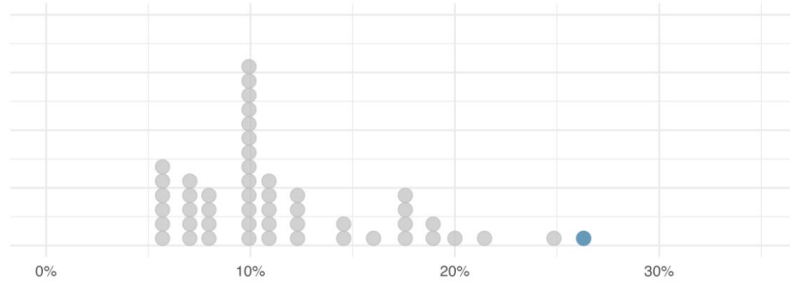
Outliers are extreme

- Examining data for outliers serves many useful purposes, including
 - identifying strong **skew** in the distribution,
 - identifying possible data collection or data entry **errors**, and
 - providing **insight** into interesting properties of the data.

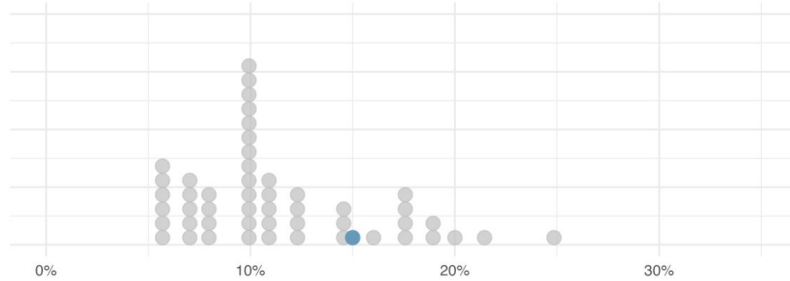
Keep in mind, however, that some datasets have a naturally long skew and outlying points do not represent any sort of problem in the dataset.

Robust statistics

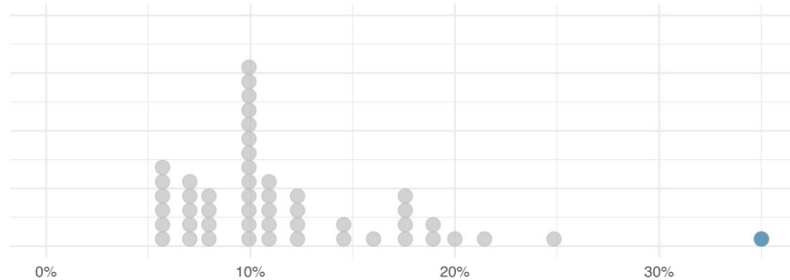
Original data



Move 26.3% to 15%



Move 26.3% to 35%




Scenario	Robust		Not robust	
	Median	IQR	Mean	SD
Original data	9.93	5.75	11.6	5.05
Move 26.3% to 15%	9.93	5.75	11.3	4.61
Move 26.3% to 35%	9.93	5.75	11.7	5.68

A comparison of how the median, IQR, mean, and standard deviation change as the value of an extreme observation from the original interest data changes.

Transforming data

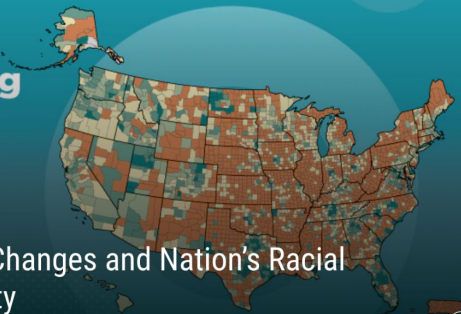
Data for 3142 counties in the United States



Search

[BROWSE BY TOPIC](#)[EXPLORE DATA](#)[LIBRARY](#)[SURVEYS/ PROGRAMS](#)[INFORMATION FOR...](#)[FIND A CODE](#)[ABOUT US](#)

2020 Census Redistricting Data



Local Population Changes and Nation's Racial and Ethnic Diversity

Read More

The U.S. Census Bureau today released additional 2020 Census results showing an increase in the population of U.S. metro areas compared to a decade ago.

SURVEYS

Help for Survey Participants

Verify that the survey you received is real and learn how to respond.

QUICKFACTS

Access Local Data

Learn about your community, county, state and the U.S. It's fast, easy and shareable.

POPULATION CLOCK

August 22, 2021

USA
332,659,369

World
7,784,236,581

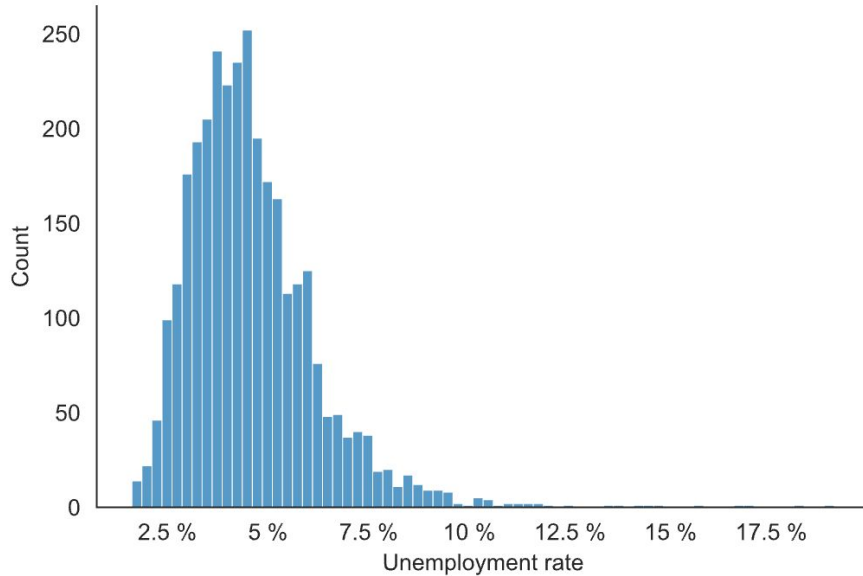
U.S. CENSUS BUREAU ECONOMIC INDICATORS

Selected Services Revenue 2nd Quarter 2021 Report Released 10:00 AM EDT, 8/19/21	\$4,367.8 B Advance Report 4.0%
New Residential Construction July 2021 Report Released 8:30 AM EDT, 8/18/21	1,534,000 Housing starts -7.0%
Business Inventories June 2021 Report Released 10:00 AM EDT, 8/17/21	\$2,057.4 B 0.8%
Advance Monthly Retail Sales July 2021 Report Released 8:30 AM EDT, 8/17/21	\$617.7 B -1.1%

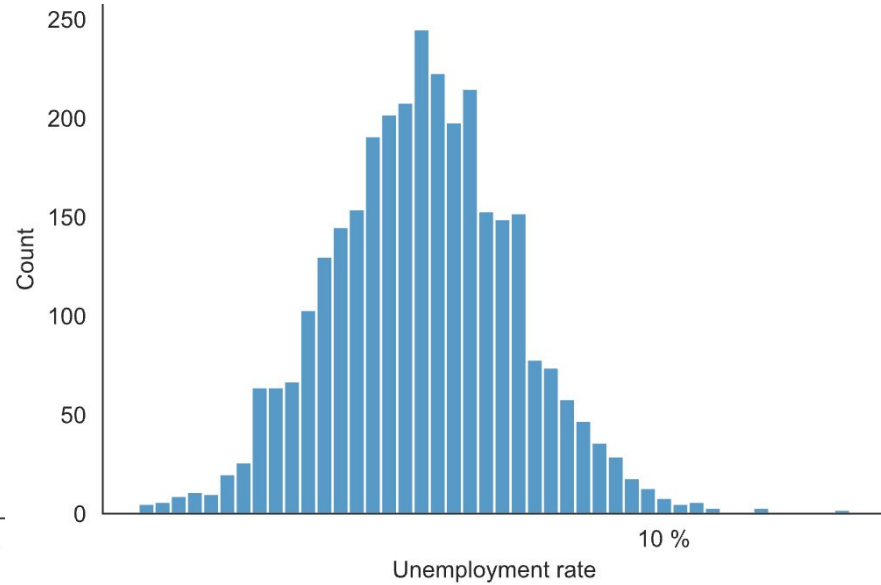
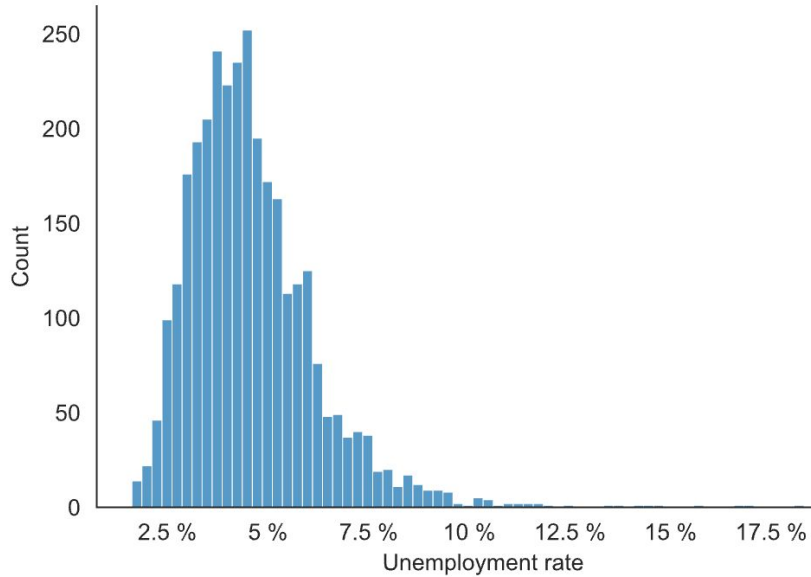
[All Economic Indicators](#)

* change not statistically significant
○ significance not reported / applicable

Skewed data: histogram of the percentage of unemployed in all US counties.



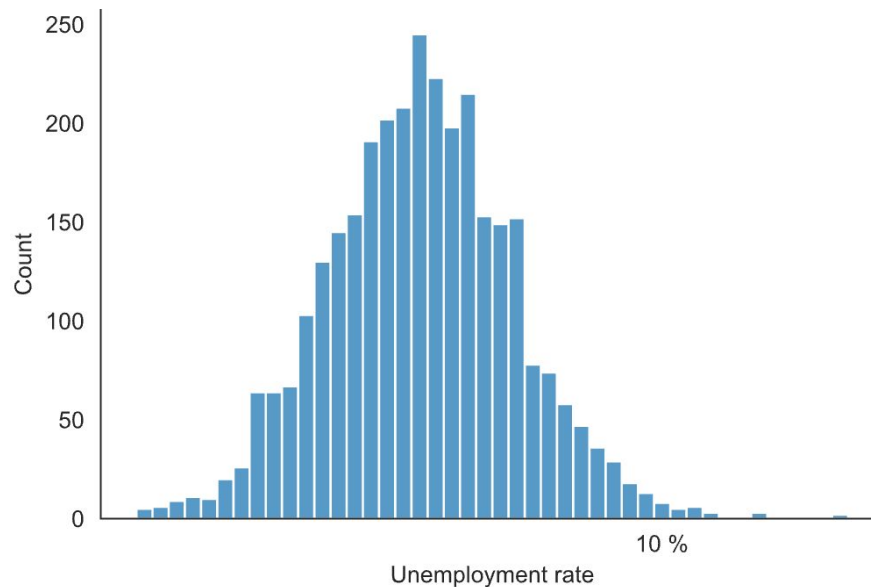
Skewed data and **log transformed** data



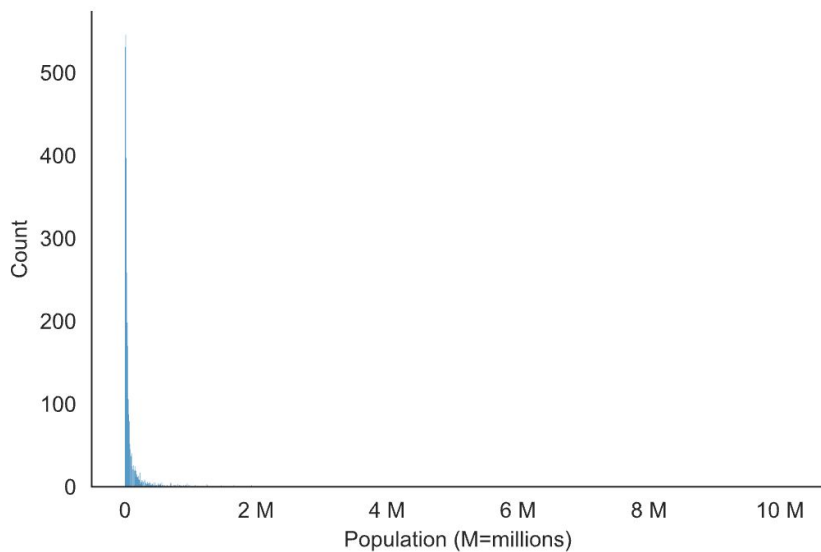
Skewed data and **log transformed** data

The x-value corresponds to the power of 10, e.g.,

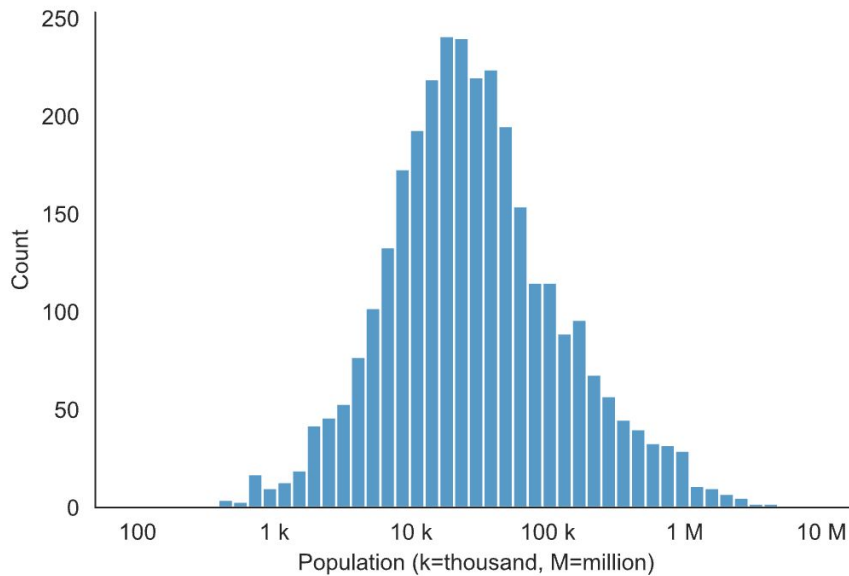
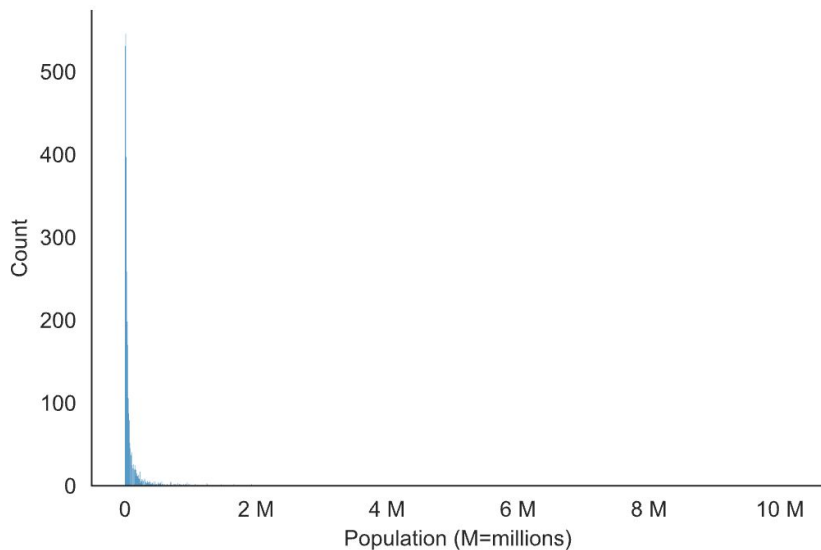
- 1 on the x-axis corresponds to $10^1 = 10$
- 5 on the x-axis corresponds to $10^5 = 100,000$.



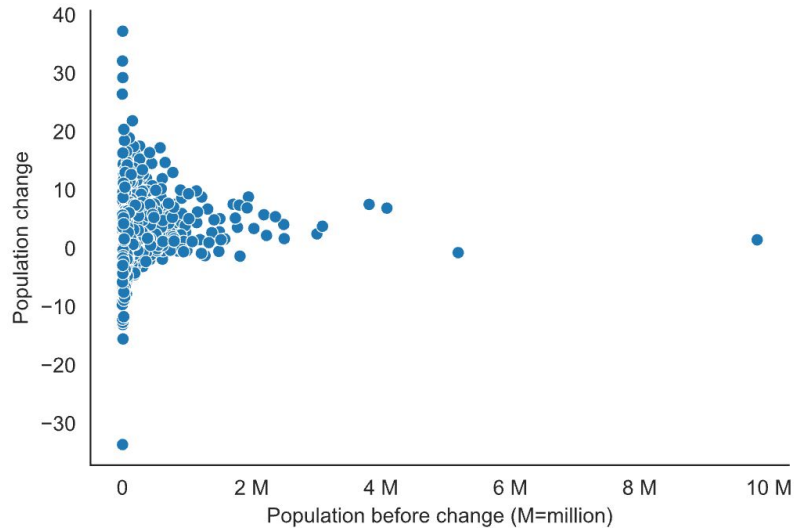
A histogram of population in all US counties



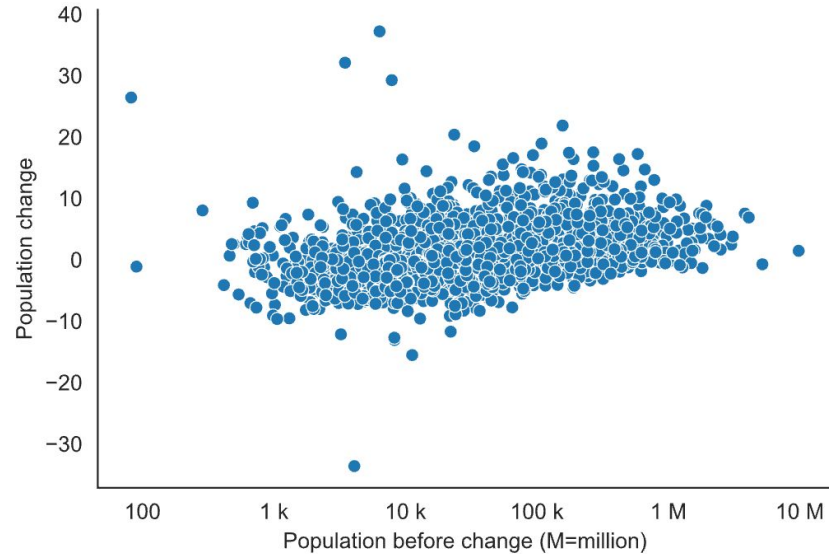
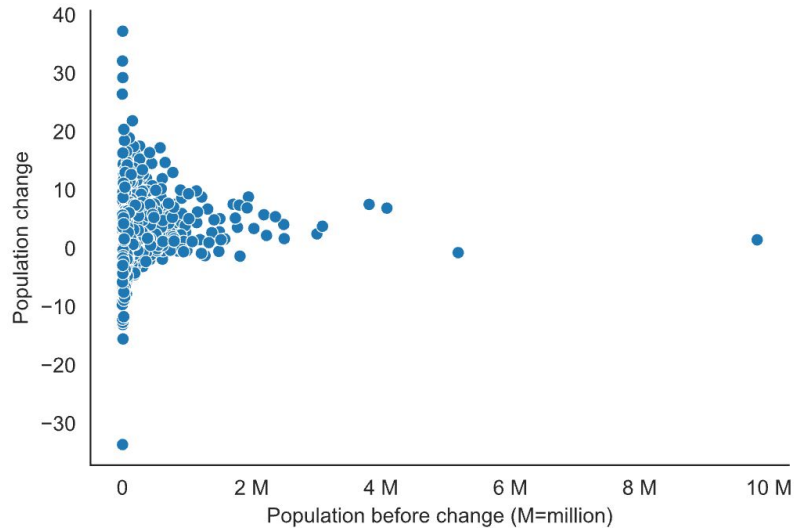
Skewed and **log transformed** data



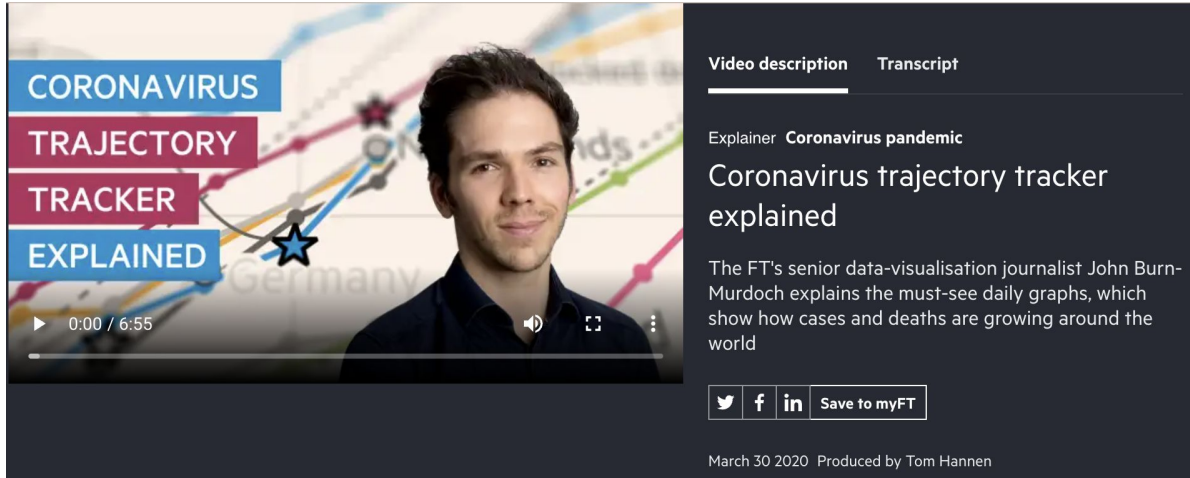
Scatterplot of population change against the population before the change



Result of **log transformation**



Example for usage of log-scales



The video player interface is split into two main sections. The left section is a video player with a dark blue background. It features a portrait of John Burn-Murdoch, a man with dark hair and a beard, wearing a dark blue shirt. Behind him is a blurred background of colorful lines and dots, representing data trajectories. On the left side of the video player, there are four stacked rectangular buttons: 'CORONAVIRUS' (blue), 'TRAJECTORY' (red), 'TRACKER' (red), and 'EXPLAINED' (blue). Below these buttons is a play button icon and a progress bar showing '0:00 / 6:55'. The right section is a dark grey area containing the video description. At the top, there are two tabs: 'Video description' (active) and 'Transcript'. Below the tabs, the text 'Explainer Coronavirus pandemic' is displayed. The main title 'Coronavirus trajectory tracker explained' is in a large, white font. Below the title, a paragraph of text explains that the FT's senior data-visualisation journalist John Burn-Murdoch explains the must-see daily graphs, which show how cases and deaths are growing around the world. At the bottom of the right section, there are social media sharing icons for Twitter, Facebook, and LinkedIn, followed by a 'Save to myFT' button. Below these icons, the date 'March 30 2020' and the producer 'Produced by Tom Hannen' are listed.

CORONAVIRUS
TRAJECTORY
TRACKER
EXPLAINED

0:00 / 6:55

Video description Transcript

Explainer **Coronavirus pandemic**

Coronavirus trajectory tracker explained

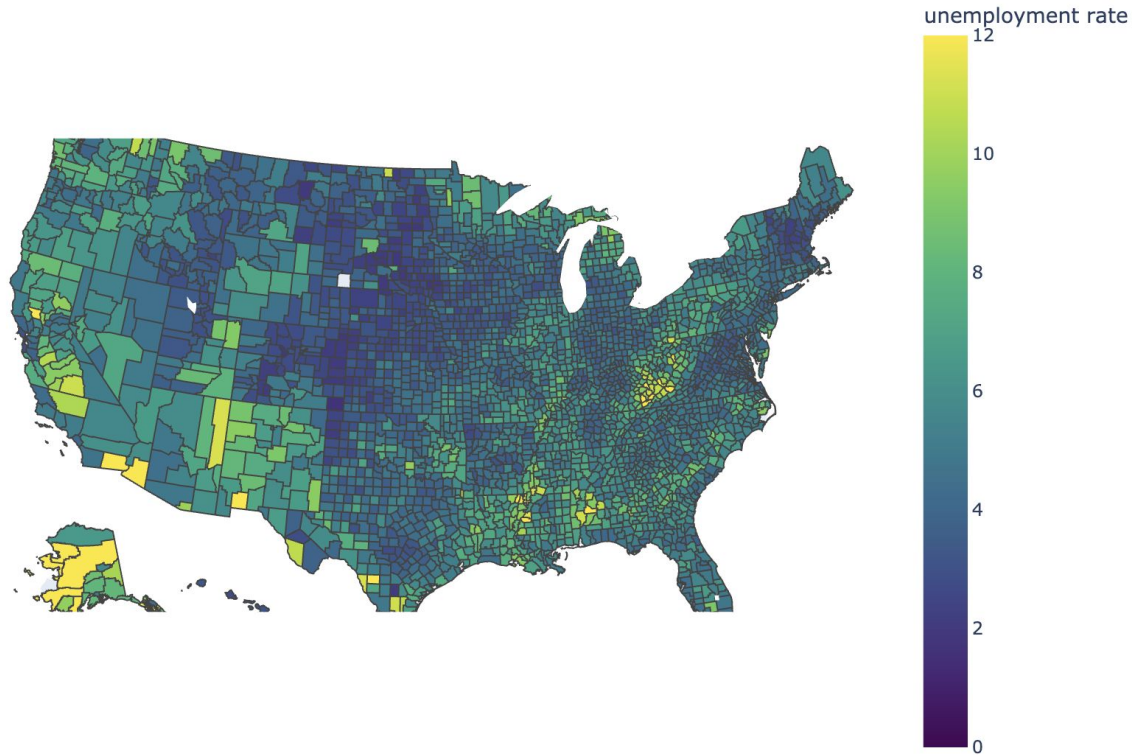
The FT's senior data-visualisation journalist John Burn-Murdoch explains the must-see daily graphs, which show how cases and deaths are growing around the world

[Twitter](#) [Facebook](#) [LinkedIn](#) [Save to myFT](#)

March 30 2020 Produced by Tom Hannen

Mapping

Intensity map of the unemployment rate (percent)



Terms you should know

average	IQR	scatterplot
bimodal	left skewed	standard deviation
box plot	mean	symmetric
data density	median	tail
deviation	multimodal	third quartile
distribution	nonlinear	transformation
dot plot	outlier	unimodal
first quartile	percentile	variability
histogram	point estimate	variance
intensity map	right skewed	weighted mean
interquartile range	robust statistics	whiskers