HOCHSCHULE DER MEDIEN

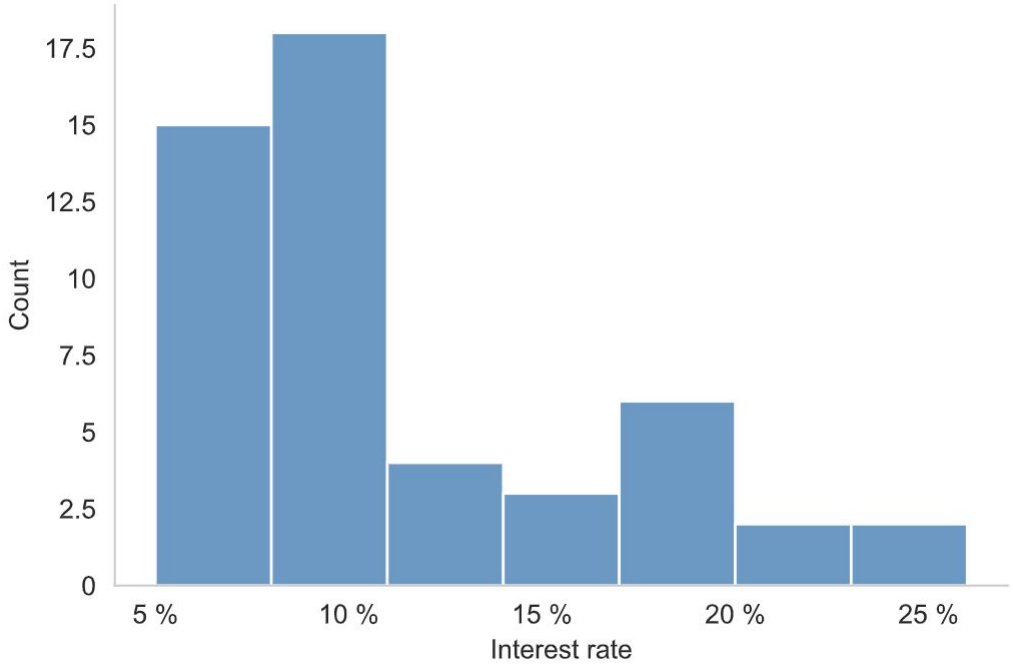
# Exploratory data analysis

## Exploring numerical data

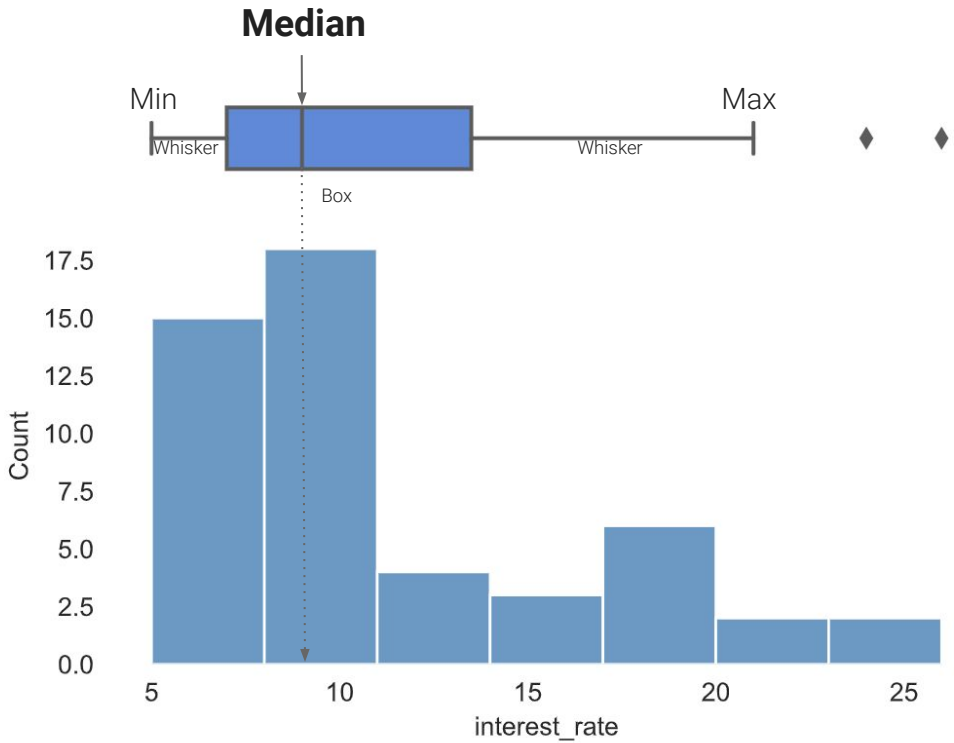
Prof. Dr. Jan Kirenz
HdM Stuttgart

# Box plots, quartiles, and the median

# **Histogram** for variable interest rate



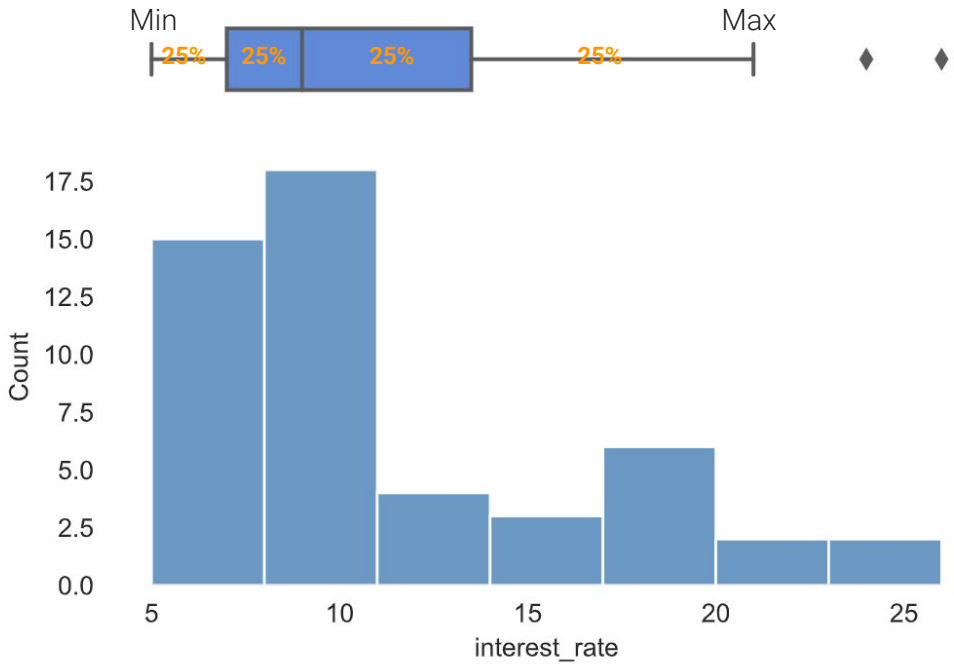
# **Boxplot** and histogram for interest rate



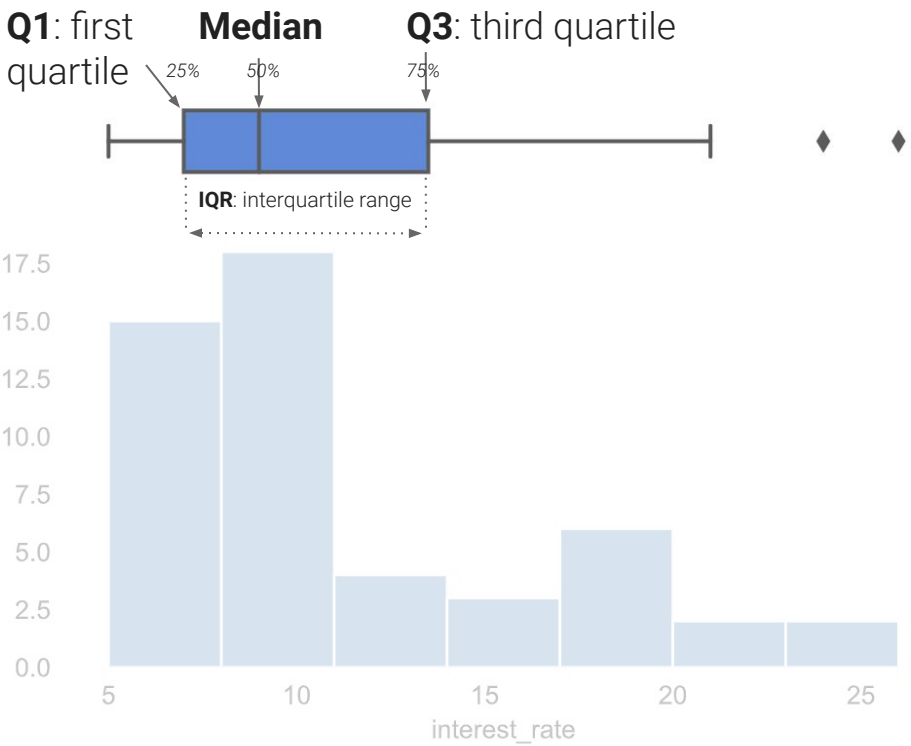

A box plot summarizes a dataset using five statistics while also identifying unusual observations

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

# **Boxplot** and quartiles



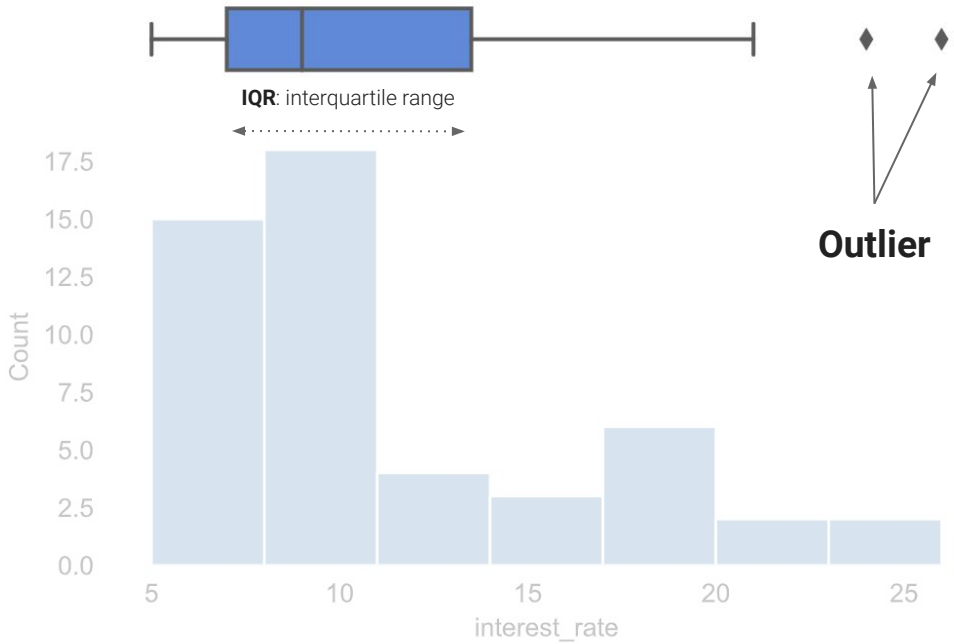
# **Boxplot**, first quartile, third quartile and IQR




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 × IQR** away from the first or the third quartile is considered an outlier.

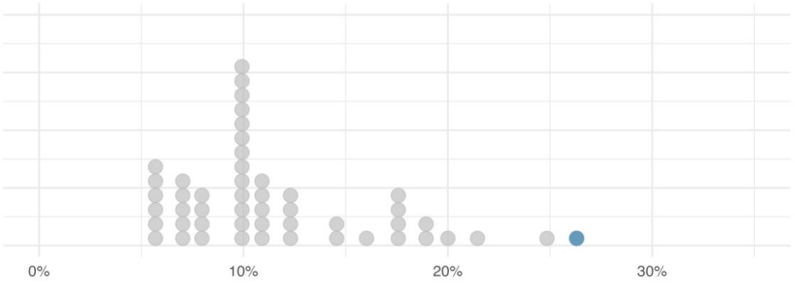
# Outliers are extreme

- 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.

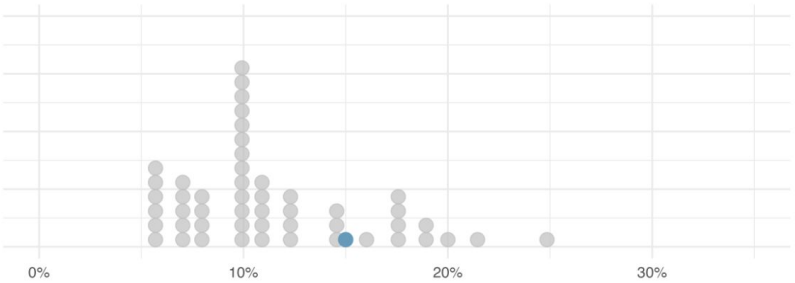
Keep in mind, however, that some datasets have a naturally long skew and outlying points do not represent any sort of problem in the dataset.

# Robust statistics

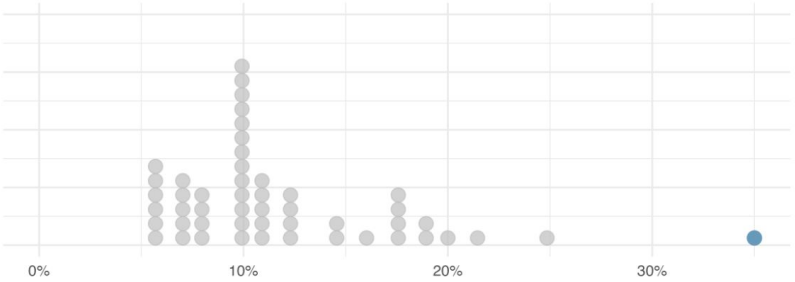
Original data




Move 26.3% to 15%




Move 26.3% to 35%

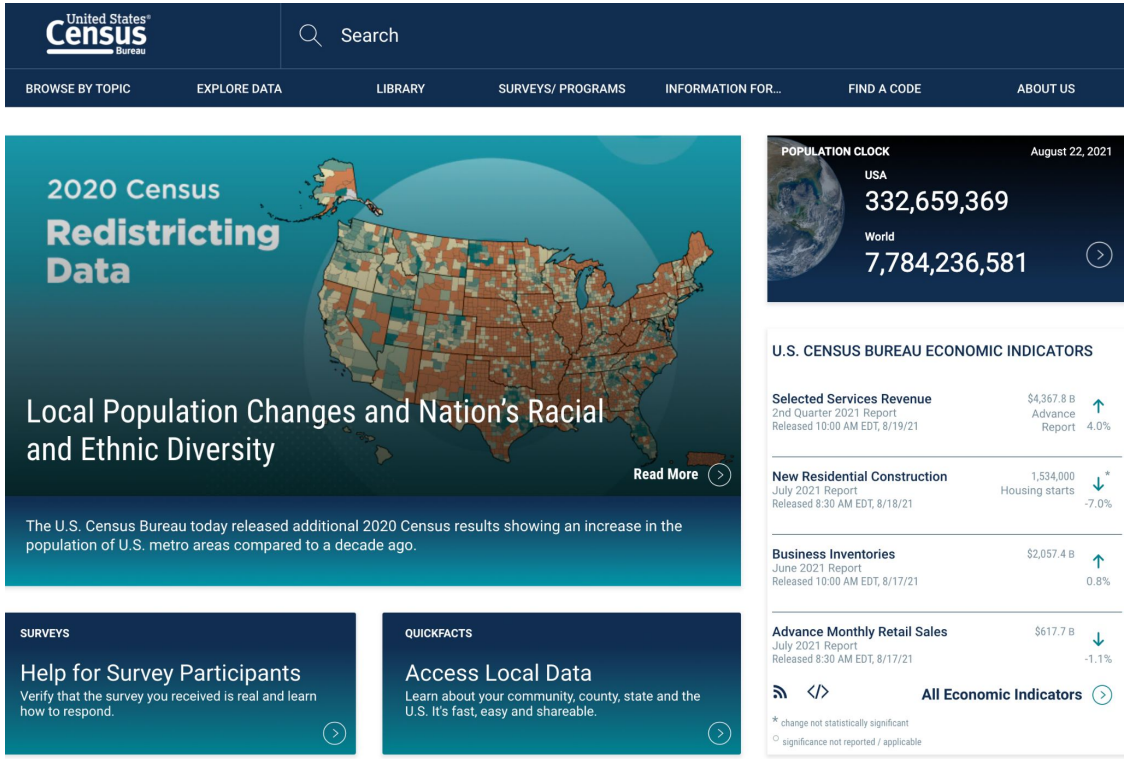



| | Robust | | Not robust | |
|---|---|---|---|---|
| **Scenario** | **Median** | **IQR** | **Mean** | **SD** |
| Original data | 9.93 | 5.75 | 11.6 | 5.05 |
| Move 26.3% to 15% | 9.93 | 5.75 | 11.3 | 4.61 |
| Move 26.3% to 35% | 9.93 | 5.75 | 11.7 | 5.68 |

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.

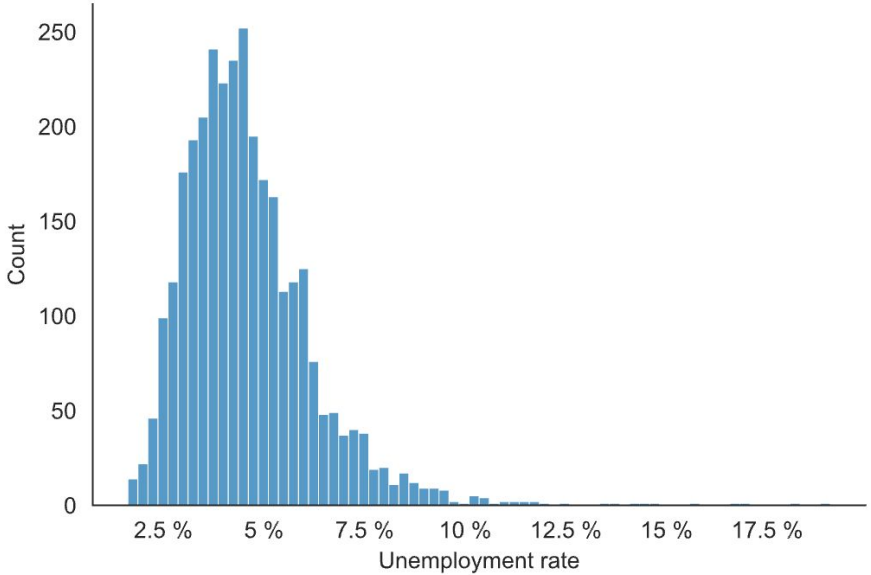
# Transforming data

# Data for 3142 counties in the United States

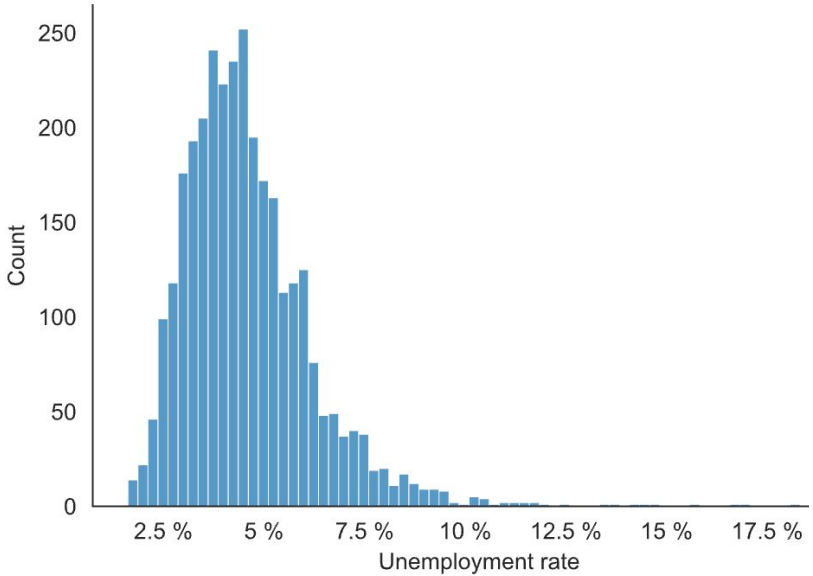


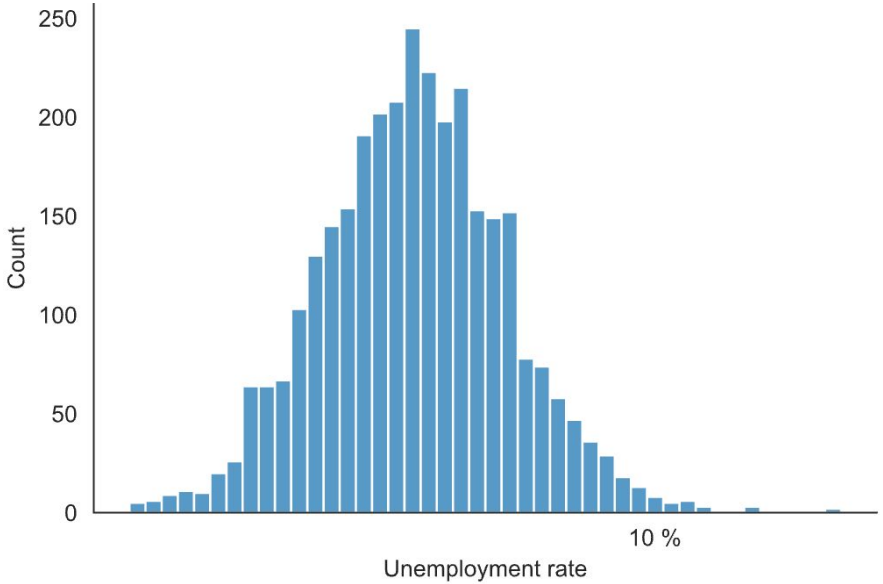

County dataset: https://www.openintro.org/data/index.php?data=county

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



# Skewed data and **log transformed** data

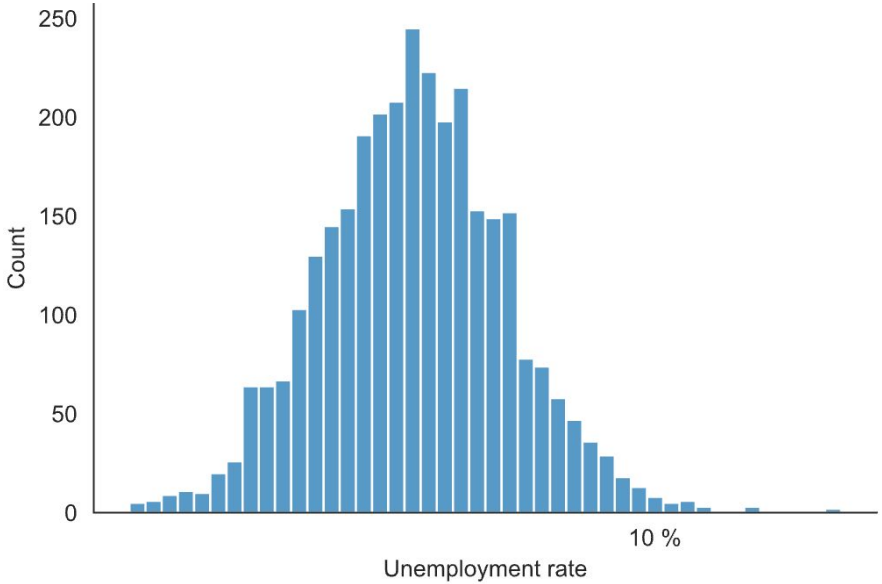


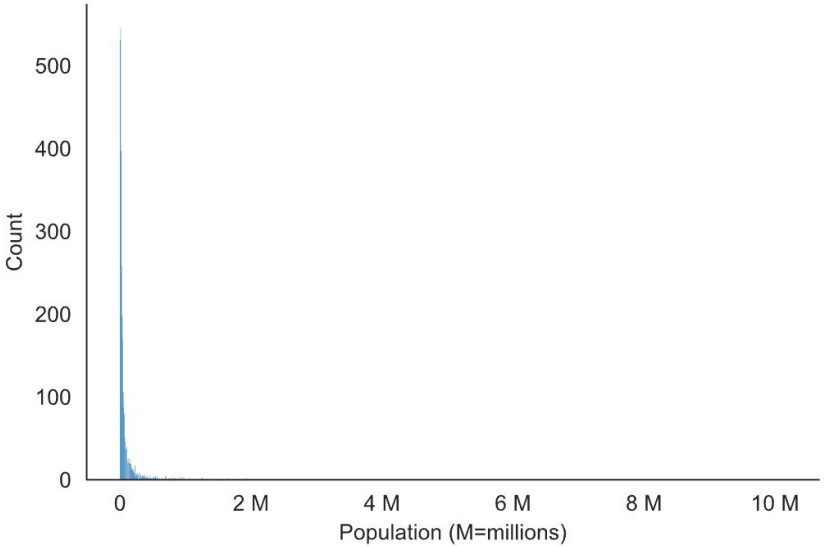
# Skewed data and **log transformed** data

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

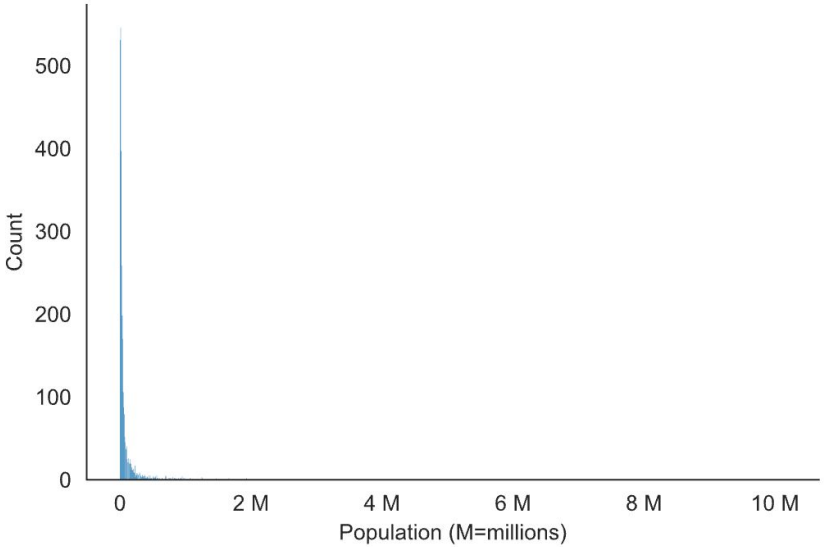
- 1 on the x-axis corresponds to $10^1 = 10$
- 5 on the x-axis corresponds to $10^5 = 100{,}000$.

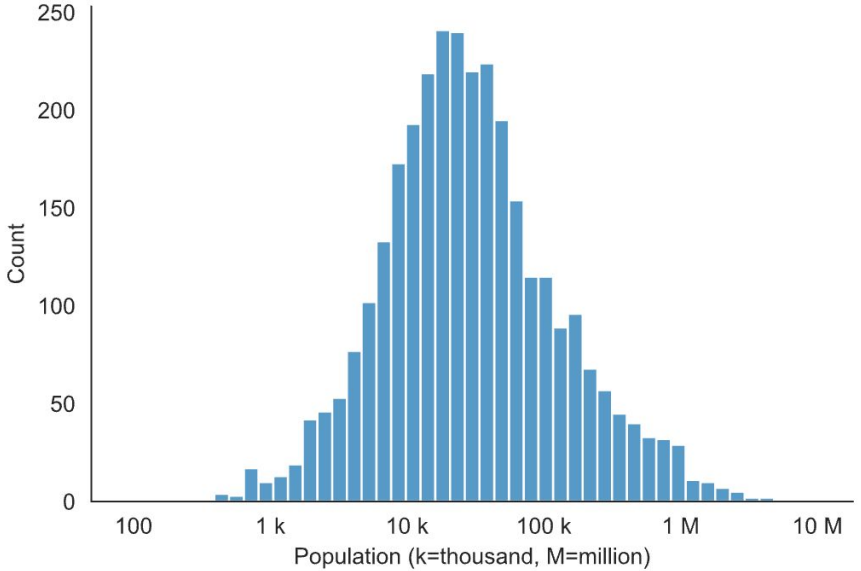


# A histogram of population in all US counties

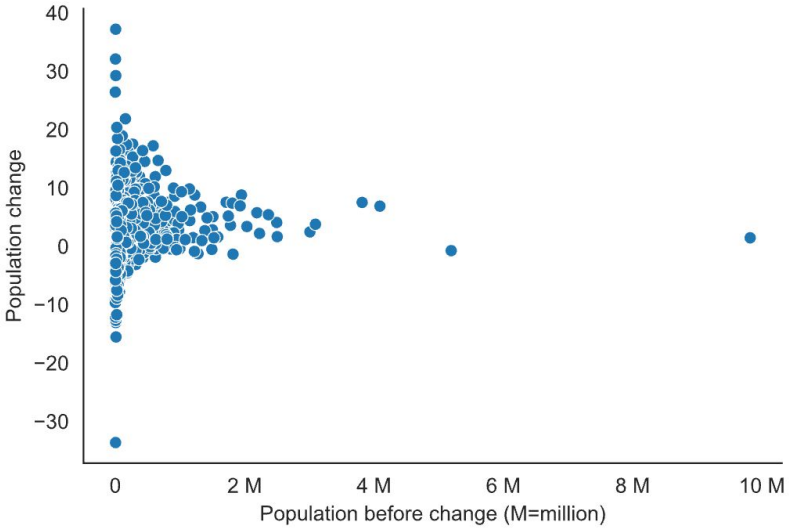


# Skewed and **log transformed** data

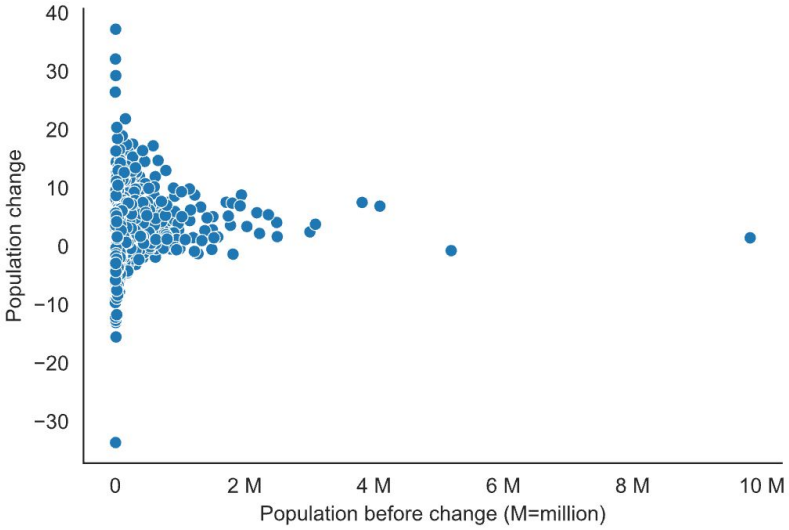


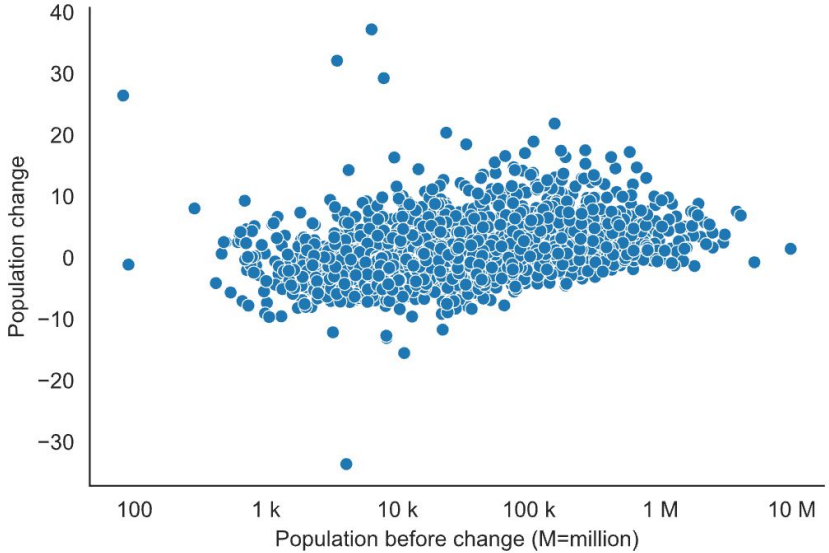



# Scatterplot of population change against the population before the change

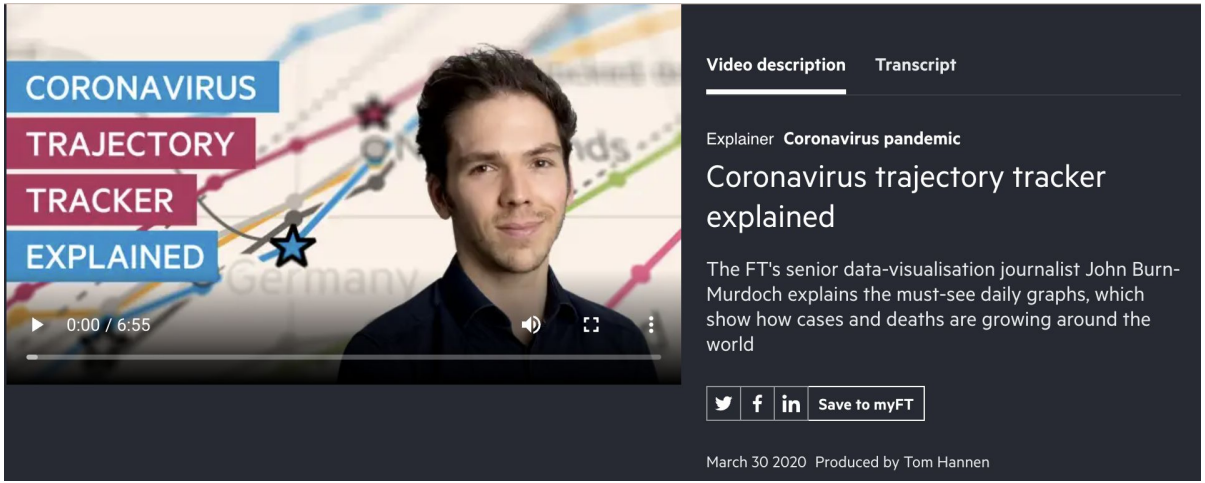


# Result of **log transformation**



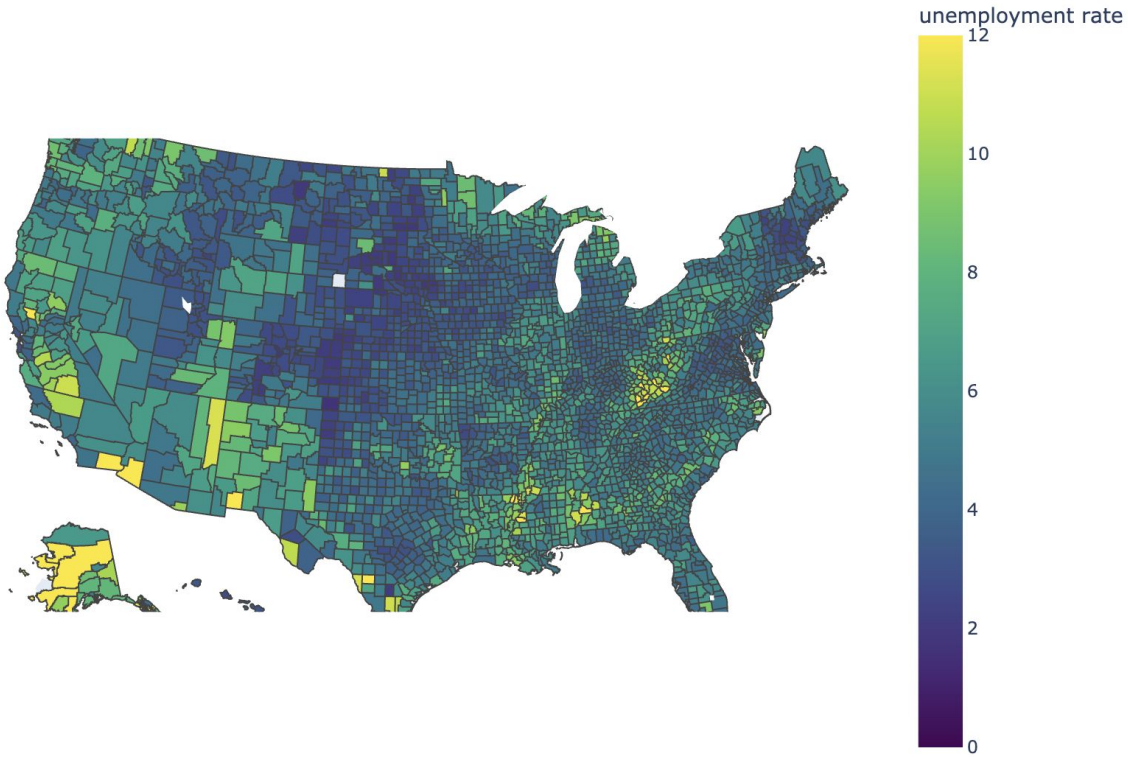
# Example for usage of log-scales




https://www.ft.com/video/9a72a9d4-8db1-4615-8333-4b73ae3ddff8

# Mapping

# **Intensity map** of the unemployment rate (percent)



# Terms you should know

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