3.1 Exploratory Data Analysis

Applied Data Analysis (ADA)

Oxford DH Summer School - 2023

Exploratory data analysis

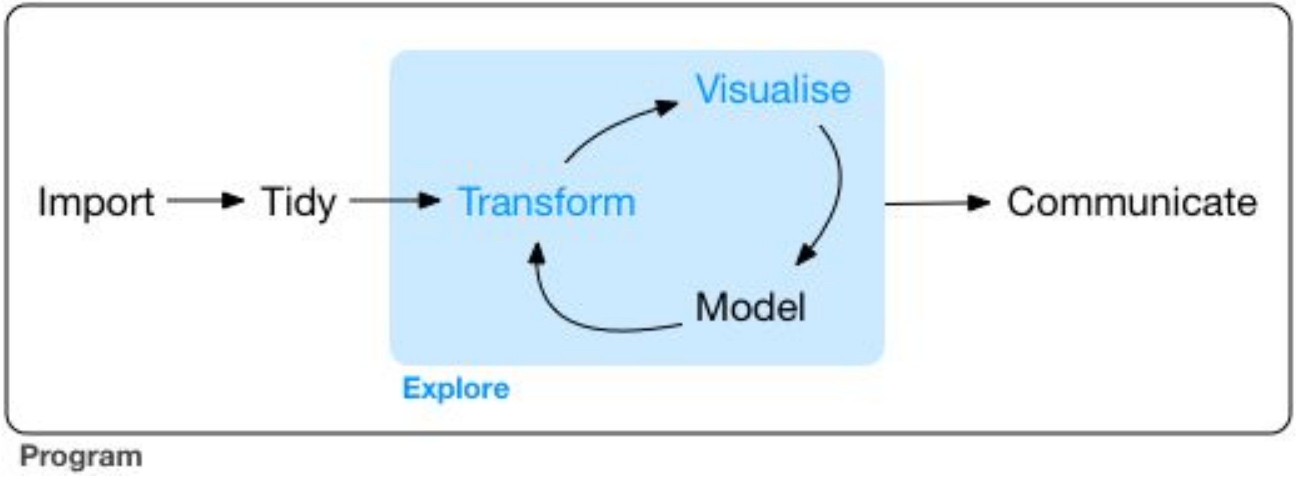
We want to use visualisation and transformation to explore your data in a systematic way, a task that statisticians call **exploratory data analysis**. Exploratory data analysis is the **iterative process** of:

1. Generating questions about your data.
2. Searching for answers by visualising, transforming, and modelling your data.
3. Using what you learn to refine your questions and/or generate new questions.

Exploratory data analysis is an important part of any data analysis, even if the questions are handed to you on a platter, because you always need to investigate the quality of your data. **Data cleaning** is just one application of it: you ask questions about whether your data meets your expectations or not.

https://r4ds.had.co.nz/exploratory-data-analysis.html

Exploratory data analysis



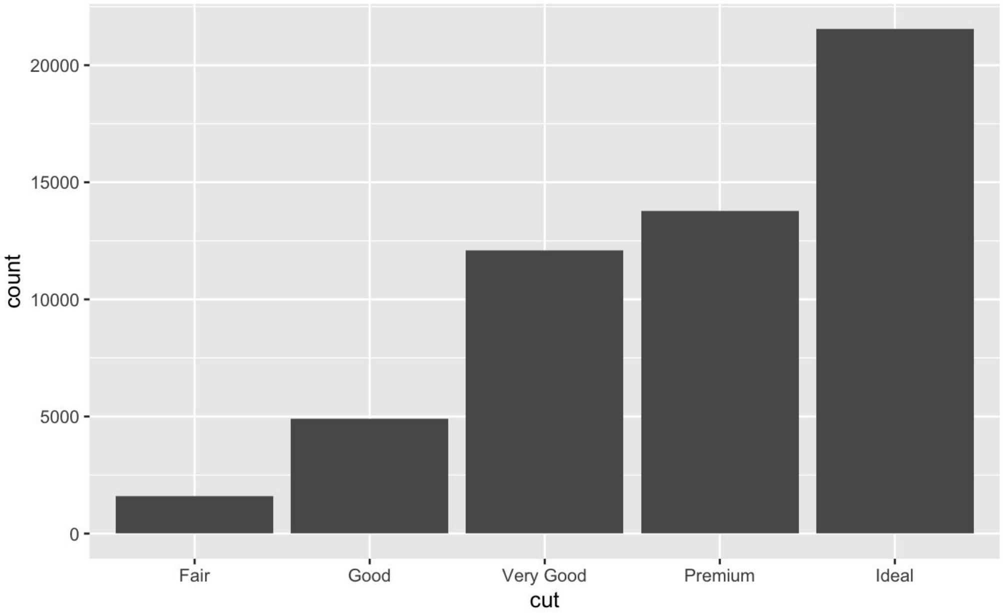
https://r4ds.had.co.nz/tidy-data.html

Today

- Basic plots: histograms, scatter plots, bar plots and box plots
- Two important distributions: normal and long-tail
- Descriptive statistics
- Outliers
- Measuring change
- Measuring co-variation

Visualizing variation in data

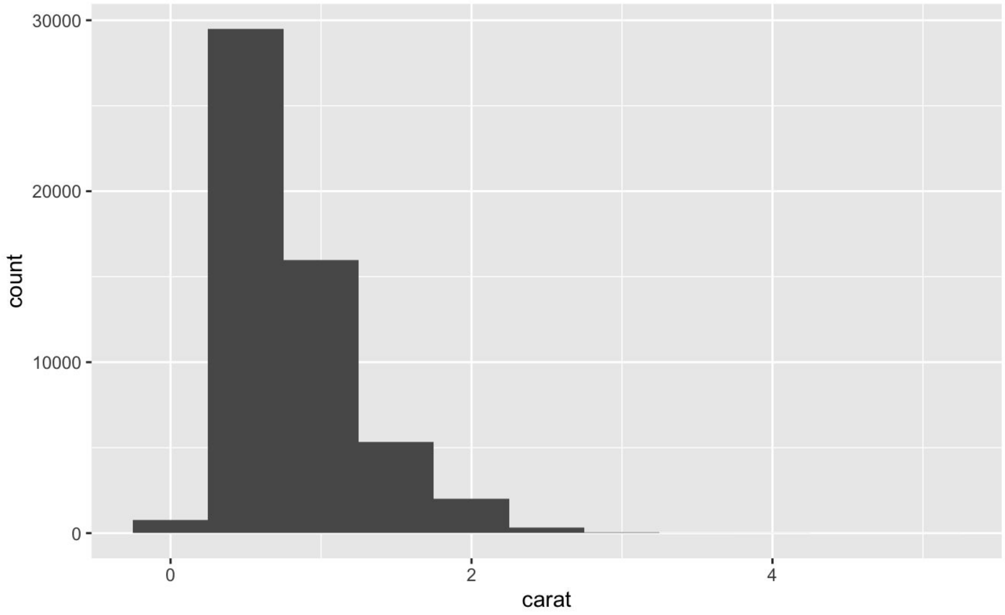
Categorical variables: bar plots



https://r4ds.had.co.nz/exploratory-data-analysis.html

Visualizing variation in data

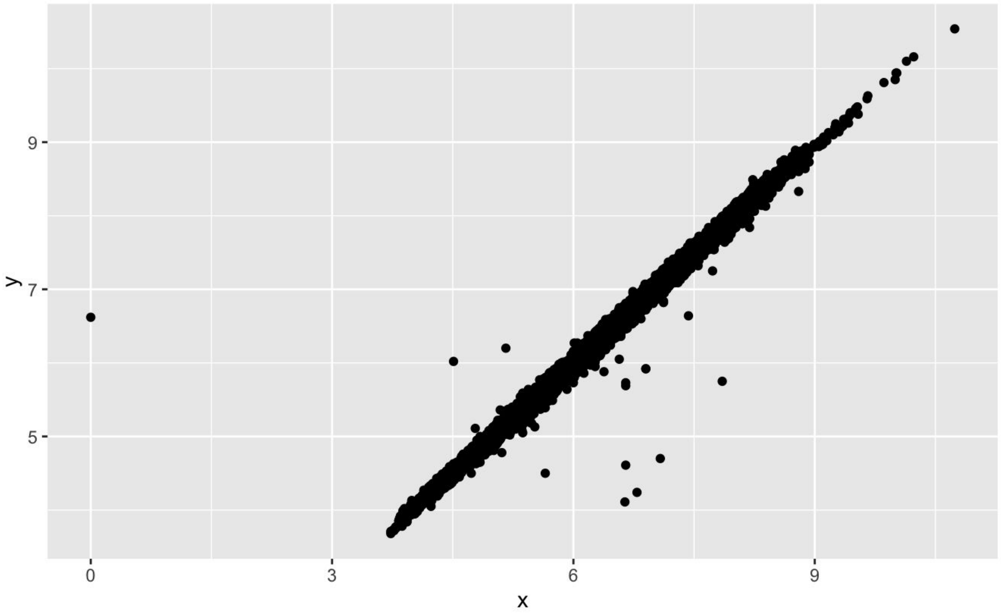
Continuous variables: histograms



https://r4ds.had.co.nz/exploratory-data-analysis.html

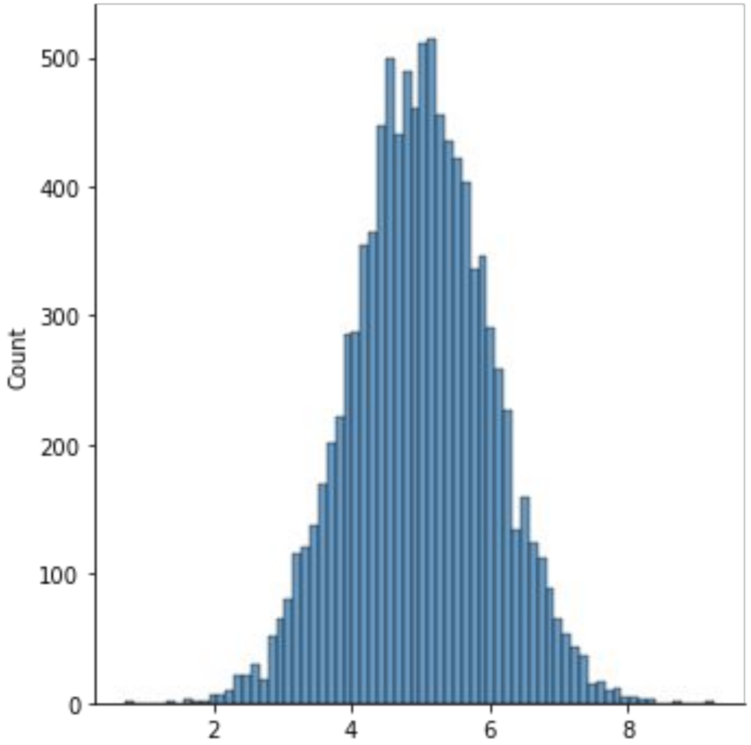
Visualizing variation in data

Comparing two continuous variables: scatter plots



https://r4ds.had.co.nz/exploratory-data-analysis.html

The normal distribution (bell curve, Gaussian curve)



Properties of the normal distribution / bell curve

The distribution that occurs naturally in many situations.

E.g., **the bulk of students will score the average (C)**, while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell.

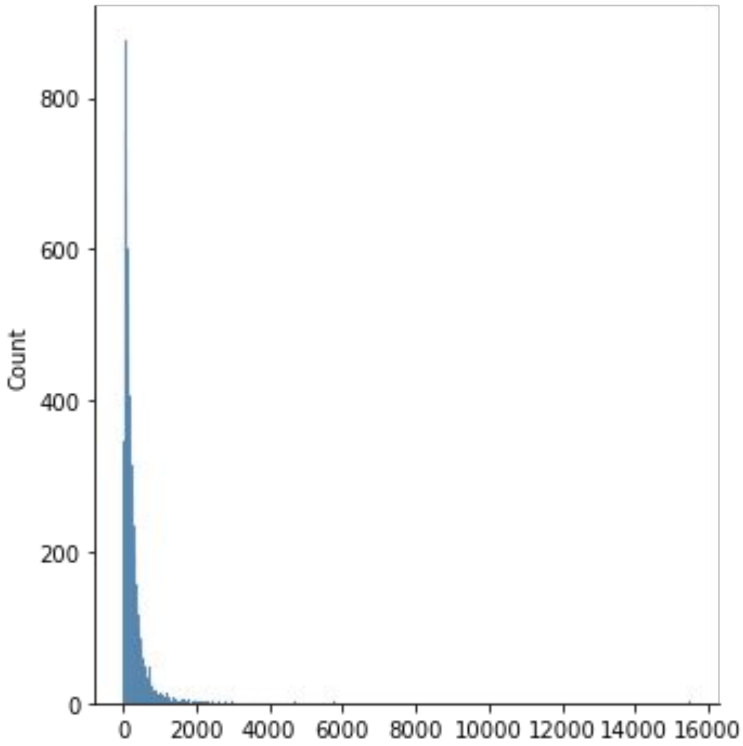
Other examples:

- Heights of people.
- Measurement errors.
- Blood pressure.
- Points on a test.
- IQ scores.

Properties of a normal distribution

- The curve is symmetric at the center (i.e., around the mean).
- Half of the values are to the left of center and half the values are to the right.
- The mean, mode and median are all equal or very close.

Long-tail distributions / Zipf distribution



Examples of long tail distributions

Many socioeconomic and cultural phenomena take long-tailed distributions:

- city population sizes
- word frequencies
- occurrences of natural resources (e.g., size of reserves in a certain geological region)
- stock price fluctuations
- size of companies

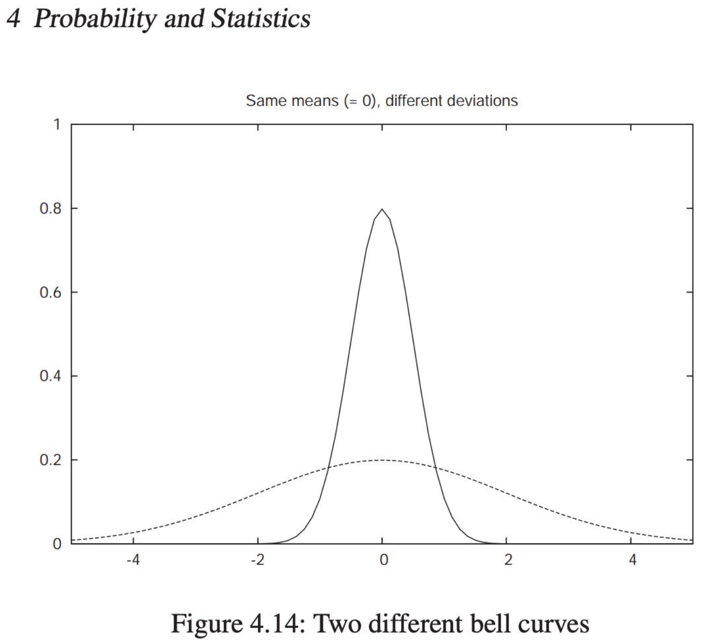
Descriptive statistics

- **Mean**: average value
- **Mode**: most frequent value

- **Median**: value such that 50% of data points are below and 50% above it
- **Quartile**:
  - 1st) value such that 25% of data points are below and 75% above it
  - 2nd) the median
  - 3rd) value such that 75% of data points are below and 25% above it

- Other useful stats: minimum, maximum, **standard deviation** (a measure of spread)

Descriptive statistics

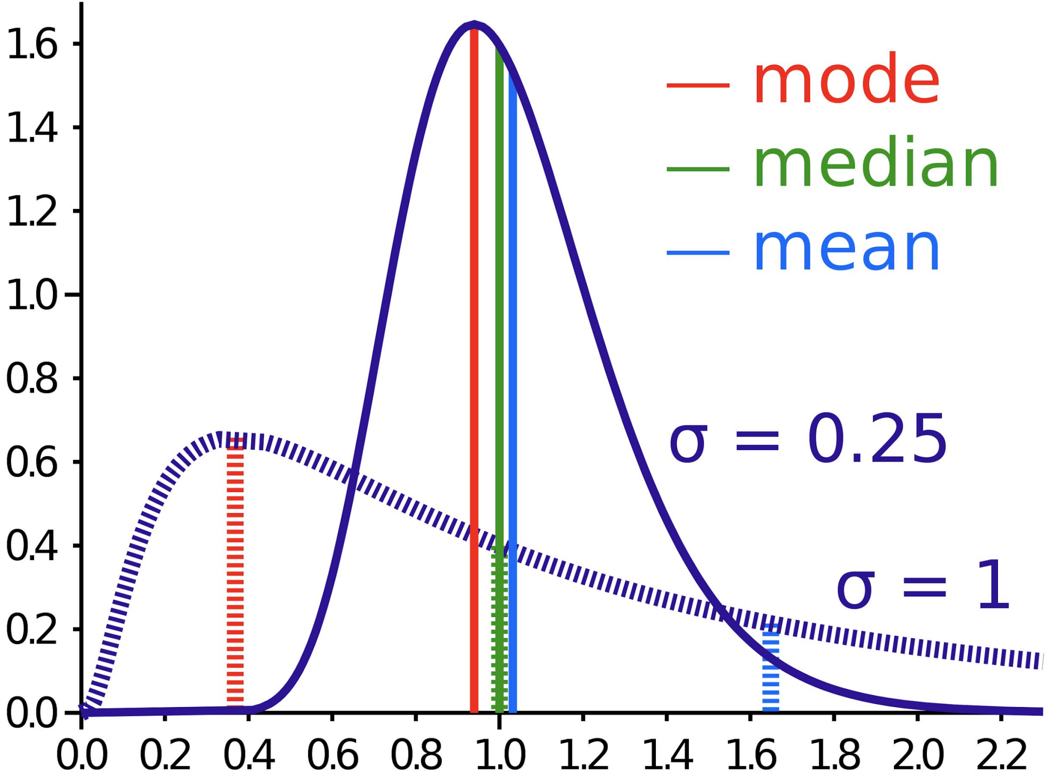


4 Probability and Statistics

Figure 4.14: Two different bell curves

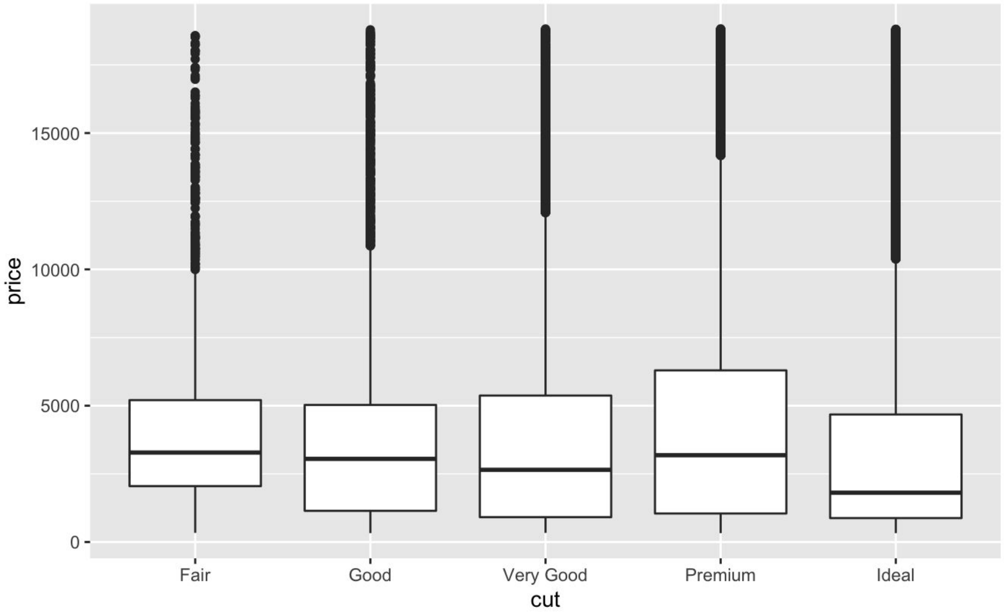
P. Juola and S. Ramsay, *Probability and Statistics of the book Six Septembers: Mathematics for the Humanist*.

Descriptive statistics



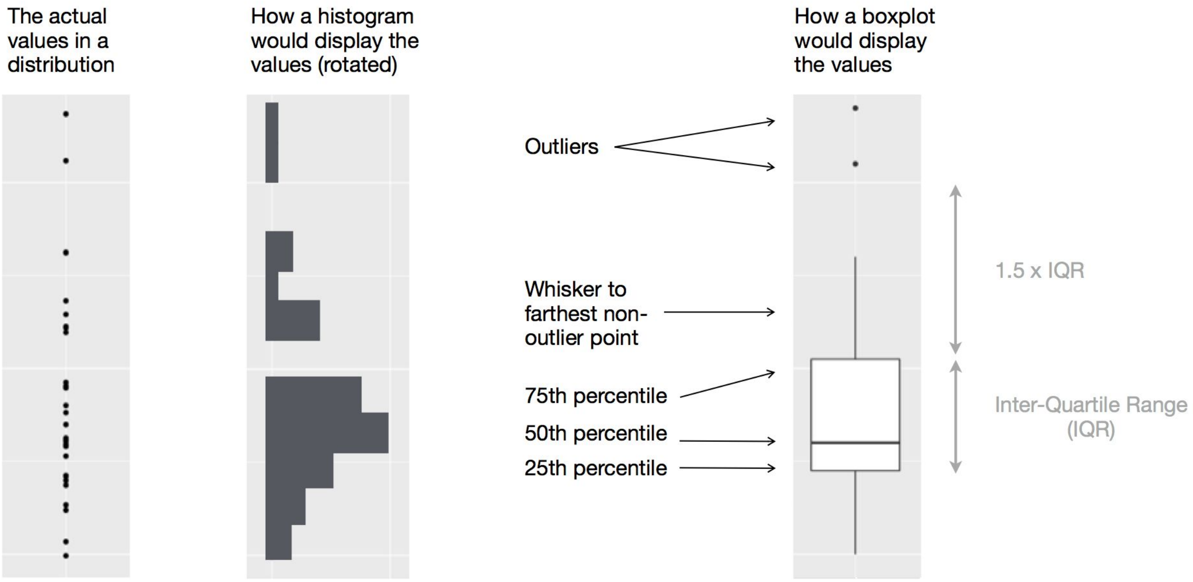
Visualizing variation in data

Exploring the distribution of a continuous variable and expose outliers: box plots



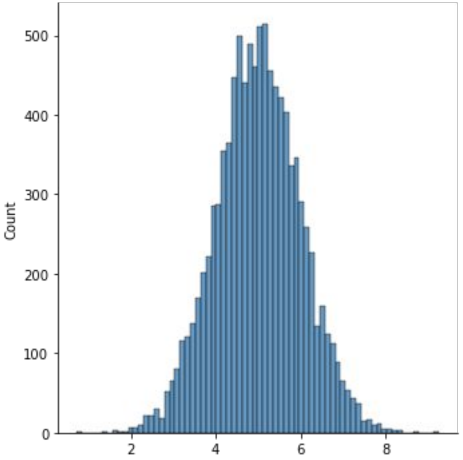
https://r4ds.had.co.nz/exploratory-data-analysis.html

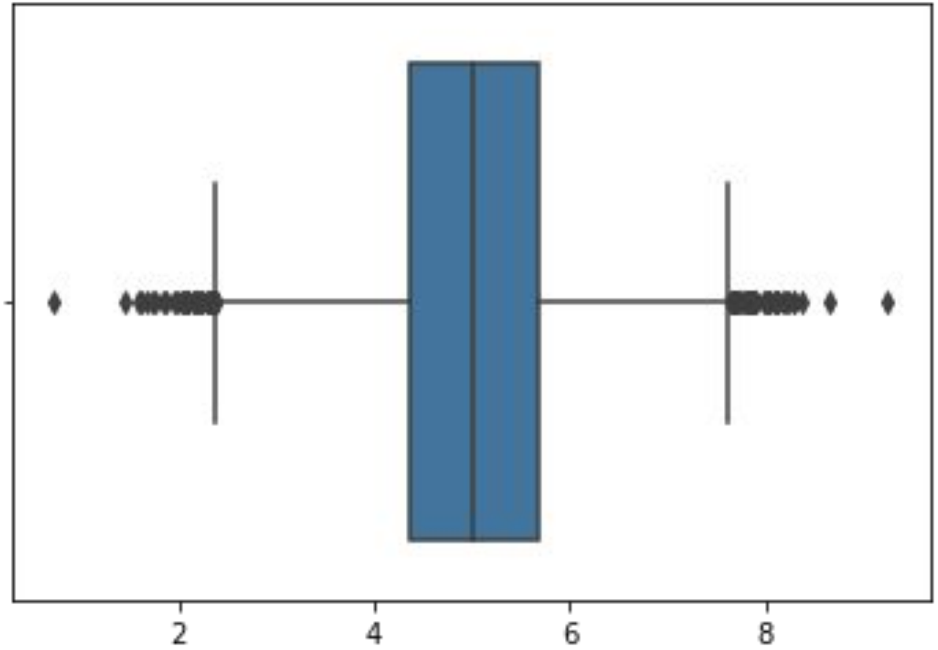
Visualizing variation in data


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)

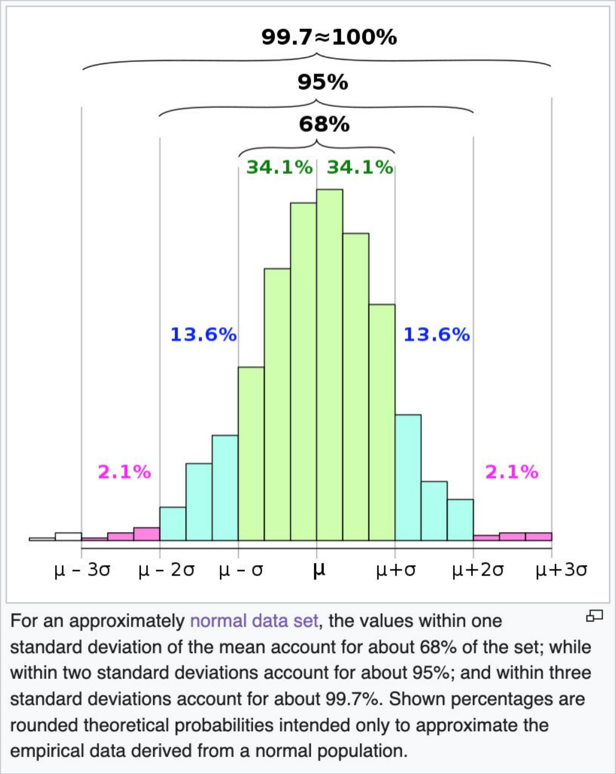
https://r4ds.had.co.nz/exploratory-data-analysis.html

The normal distribution



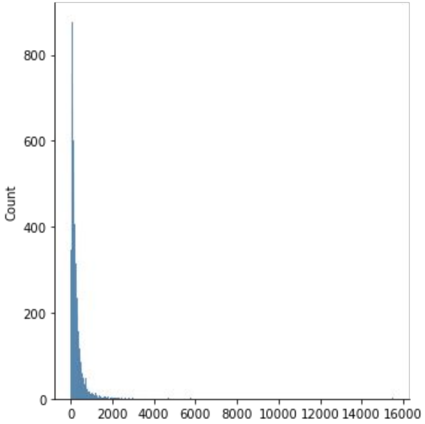


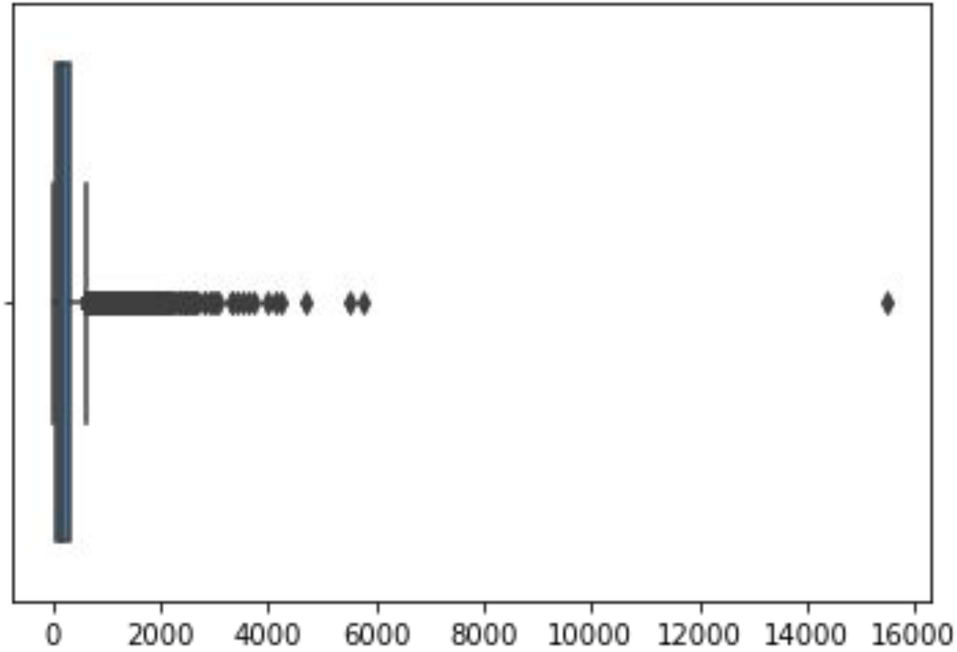
The normal distribution



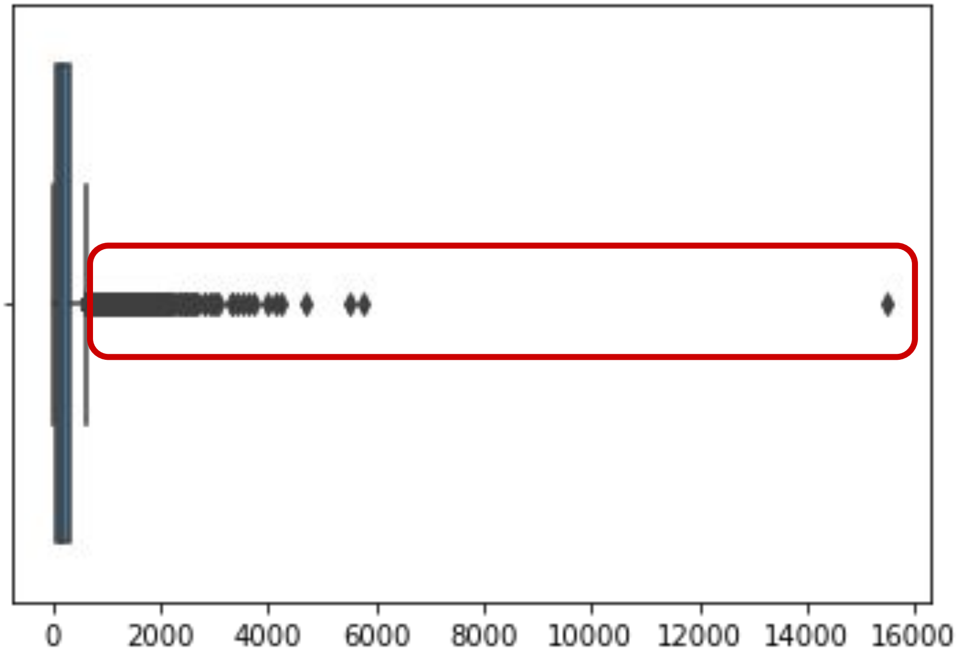
For an approximately normal data set, the values within one standard deviation of the mean account for about 68% of the set; while within two standard deviations account for about 95%; and within three standard deviations account for about 99.7%. Shown percentages are rounded theoretical probabilities intended only to approximate the empirical data derived from a normal population.

Long-tail distributions





Outliers



Measuring change

An effective measure of change are % increases (or decreases) over a certain time.

The calculation is very simple: ((new_value - old_value) / old_value) * 100

Example:

- house price index 2019: 500
- house price index 2020: 550
- house price index 2021: 490

Change 2019 to 2020: +10%; change 2020 to 2021: -10.9%

Measuring co-variation

How do two variables change together, when considering the same observations?

Covariance is a linear measure of such variation:

$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right], \quad \textbf{(\textit{Eq.1})}$$

Q&A