Feature Engineering

Papangkorn Inkeaw, Ph.D.

Feature Understanding

Chapter 2

Structure and Unstracture Data

Structured (organized) Data

- Can be broken down into obevations and characteristics
- Organized using a tabular method → Data tabel
- Such as:
  - Meteorological data
  - Address information
  - Billing information

Characteristics

Obevations

Name	Age	Gender	Address	Zip Code
Suwit	32	M	32/1	50200
Sitta	45	M	301 M.1	55000
…	…	…	…	…
Marisa	28	F		10500

Structure and Unstracture Data

Unstructured (unorganized) Data

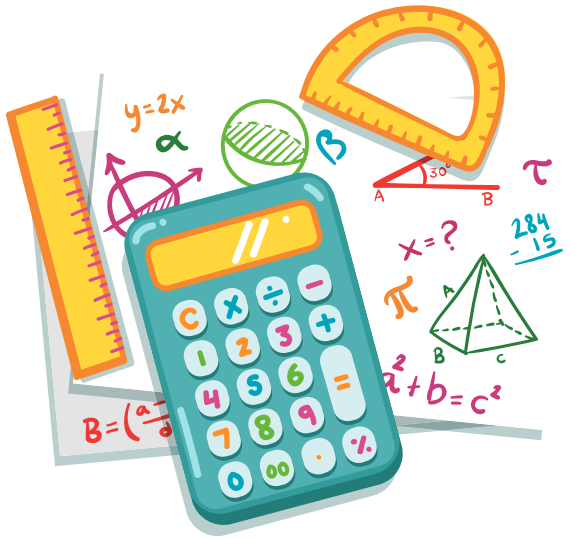
- Free-flowing entity
- Does not follow standard organizational hierarchy
- A single characteristic (column)
- Examples:
  - Free-text comments
  - Images
  - Sounds
  - Server log

Review	Sentiment
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	Positive
...	...
Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his par...	Negative

Quantitative and Qualitative Data

Quantitative Data

- Numerical in nature
- Basic mathematical procedures are possible on the set.
- Examples:
  - Temperature in Fahrenheit or Celsius
  - The amount of blood you donate
  - Weight and height



Discrete

- Take on a finite or countably infinite set
- Such as integer, grade and number of object

Continuous

- Take on any real value
- Such as height, weight and size

Quantitative and Qualitative Data

Qualitative Data

- Categorical and language in nature
- Cannot be described using numbers and basic mathematics.
- Examples:
  - Weather measured as cloudy or sunny
  - The name of a person
  - Reviews in social medias

Levels of Data

The four levels of data:

1. Nominal level
2. Ordinal level
3. Interval level
4. Ratio level

Each level comes with a varying level of control and mathematical possibilities.

Levels of Data

Categorical Data

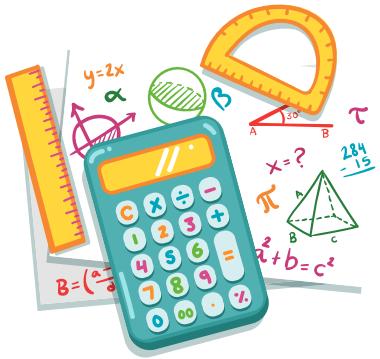
Nominal Level

- Attribute values in the domain are unordered.
- Can only equality (=) compare.
- Such as gender, type of hair, etc.

Ordinal Level

- Attribute values are ordered.
- Can both equality (=) and inequality (<, >) compare.
- Such as education, feel (unhappy, OK, happy), etc.

Levels of Data



Numeric Data

Interval Level

- Can compute only differences (addition or subtraction)
- For example, temperature measured in ˚C or ˚F.
  - If it is 20 ˚C on one day and 10 ˚C on previous day
  - We **can** talk about a temperature drop of 10˚C.
  - We **cannot** say that it is twice as cold as the previous day.

Ratio Level

- Can compute both differences and ratio between values,
- For example age.
  - If Jone is 20 years old and Jim is 10 years old.
  - We **can** say that Jone older than Jim with 10 years.
  - We **can** say that Jone is twice as old as Jim.

Levels of Data

Summary of data types and scale measures

Provides	Nominal	Ordinal	Interval-scaled	Ratio-scaled
The order of values is known		●	●	●
"Count," aka "Frequency of Distribution"	●	●	●	●
Mode	●	●	●	●
Median		●	●	●
Mean			●	●
Can quantify the difference between each values			●	●
Can add or subtract values			●	●
Can multiple and divide values				●
Has "true zero"				●

Source: https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

Describing Data

Descriptive Statistics

Mean

- A measure of a central or typical value for a probability distribution.
- The sum of all measurements divided by the number of observations in the data set.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Describing Data

Descriptive Statistics

Median

- Reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliners.
- The middle value that separates the higher half from the lower half of the data set.
- To compute the middle value, we need to arrange all the numbers from smallest to greatest.
- Then

$$\tilde{x} = \begin{cases} x_{\frac{(n+1)}{2}}, & if\ n\ is\ odd, \\ \frac{\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)}{2}, & if\ n\ is\ even, \end{cases}$$

Describing Data

Descriptive Statistics

Mode

- The most frequent value in the data set.

Describing Data

Example

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

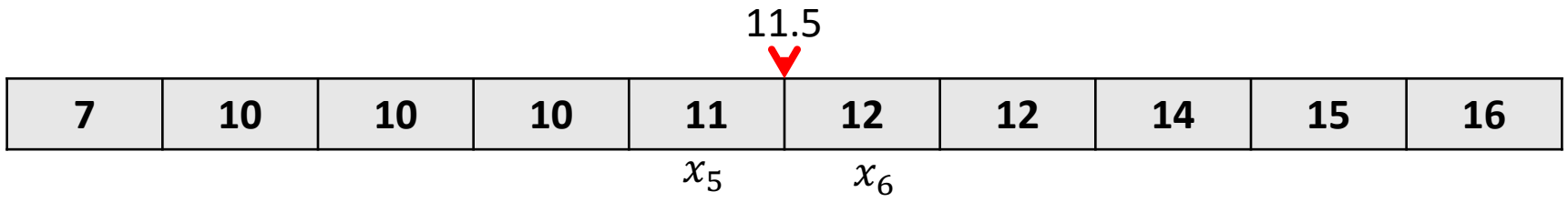
Mean:

$$\bar{x} = \frac{7+10+11+15+10+10+12+14+16+12}{10} = \frac{117}{10} = 11.7$$

Median:

First, sort the job performance. As the number of job performance data (n = 10) is even, the median of job performance is

$$\tilde{x} = \frac{x_5 + x_6}{2} = \frac{11 + 12}{2} = 11.5$$

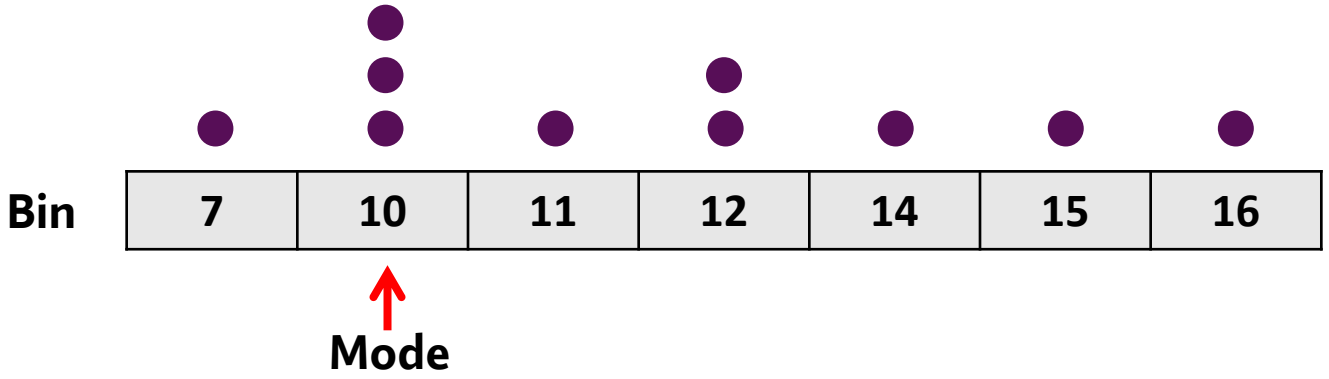


Describing Data

Example

Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12

Mode:

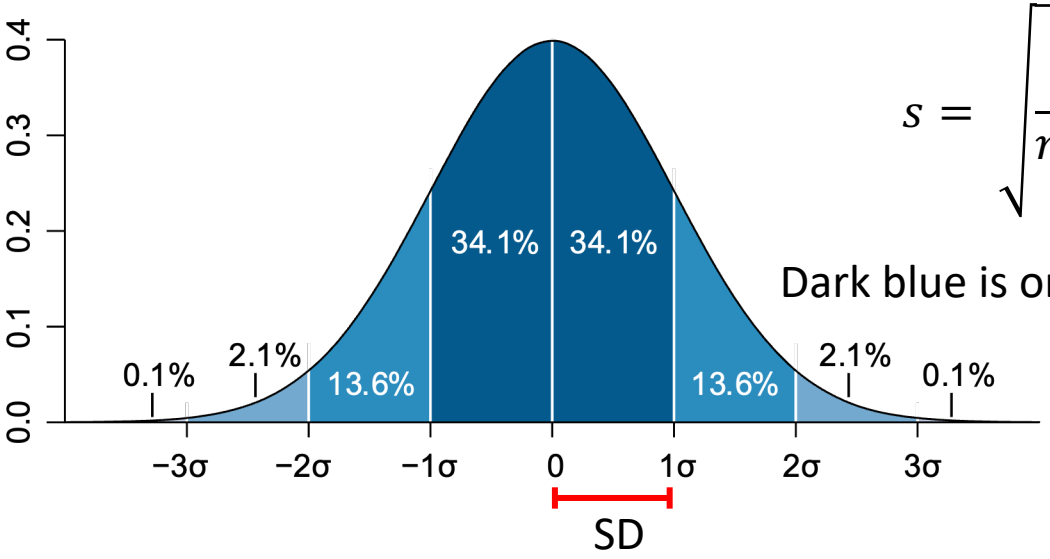


Describing Data

Descriptive Statistics

Standard Deviation

- A measure that is used to quantify the amount of variation or dispersion of a set of data values.
- A low standard deviation indicates that the data points tend to be close to the mean.
- A high standard deviation indicates that the data points are spread out over a wider range of values.
- The formula for the sample standard deviation is



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

Dark blue is one standard deviation on either side of the mean.

Source: https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

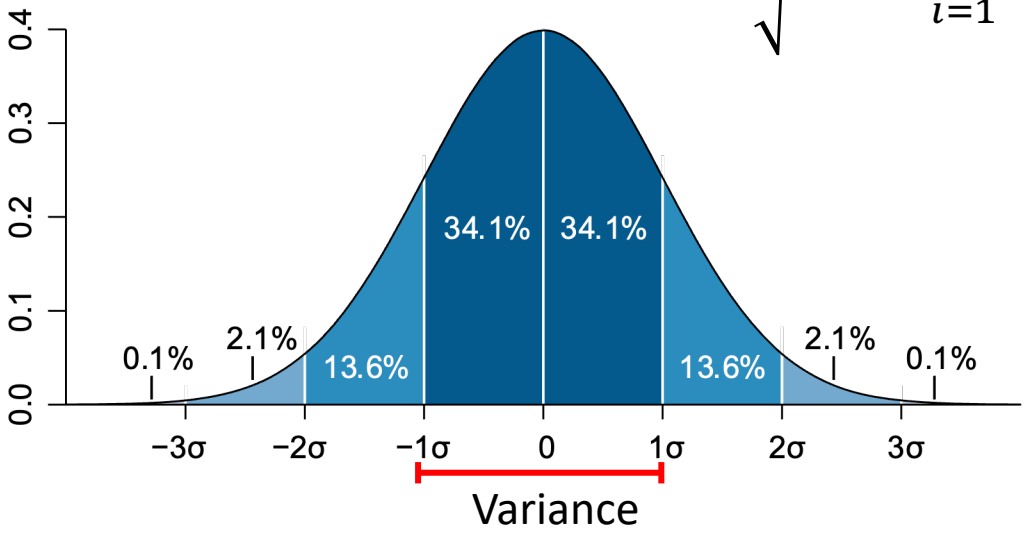
Describing Data

Descriptive Statistics

Variance

- How far a set of numbers are spread out from their average value.
- It is the square of the standard deviation

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



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Describing Data

Example

- Job performance; X={7, 10, 11, 15, 10, 10, 12, 14, 16, 12}
- Mean of job performance $\bar{x}$: 11.7
- Standard Deviation; $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 2.71$
- Variance; $\mathrm{var}(X) = SD^2 = 2.71^2 = 7.34$

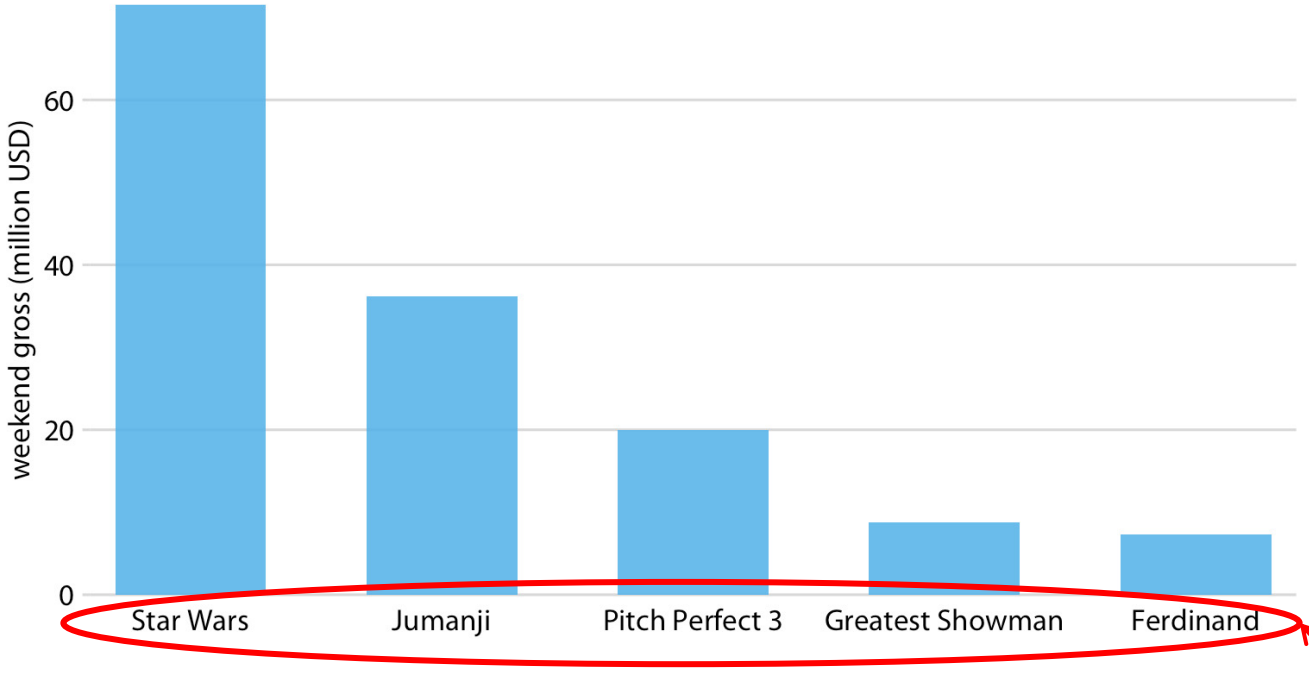
Job performance x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
7	-4.7	22.09
10	-1.7	2.89
11	-0.7	0.49
15	3.3	10.89
10	-1.7	2.89
10	-1.7	2.89
12	0.3	0.09
14	2.3	5.29
16	4.3	18.49
12	0.3	0.09
$\sum_{i=1}^{n}(x_i - \bar{x})^2$		66.1
$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$		2.71

Describing Data

Graphical Method

Bar Plot

- Each entity of the categoric variable is represented as a bar.
- The size of the bar represents its numeric value.



Highest-grossing movies for the weekend of December 22–24, 2017, displayed as a bar plot.

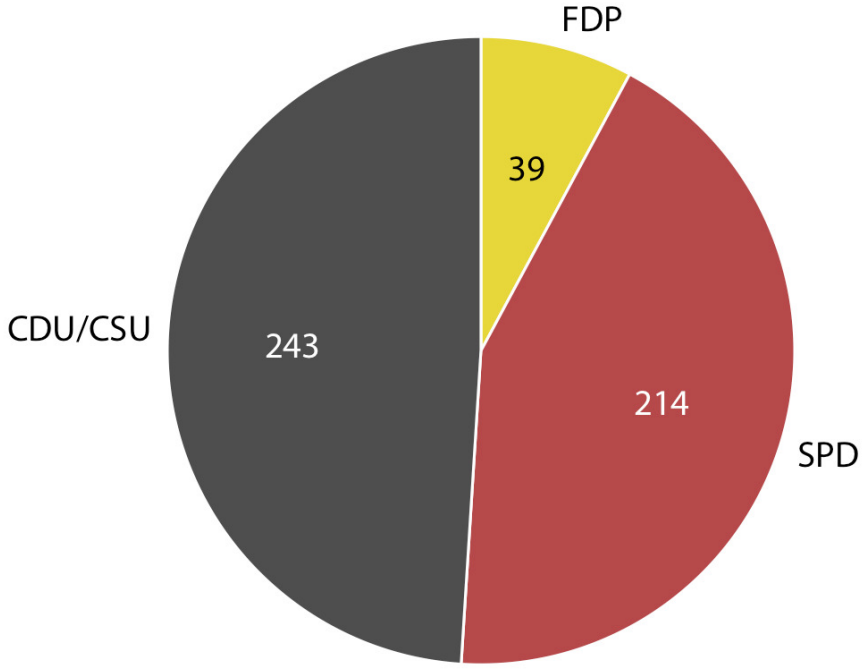
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Describing Data

Graphical Method

Pie

- A circle divided into sectors that each represent a proportion of the whole.
- It is often used to show proportion, where the sum of the sectors equal 100%.



Party composition of the eighth German Bundestag, 1976–1980, visualized as a pie chart. This visualization highlights that the ruling coalition of SPD and FDP had a small majority over the opposition CDU/CSU.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Describing Data

Graphical Method

Histogram

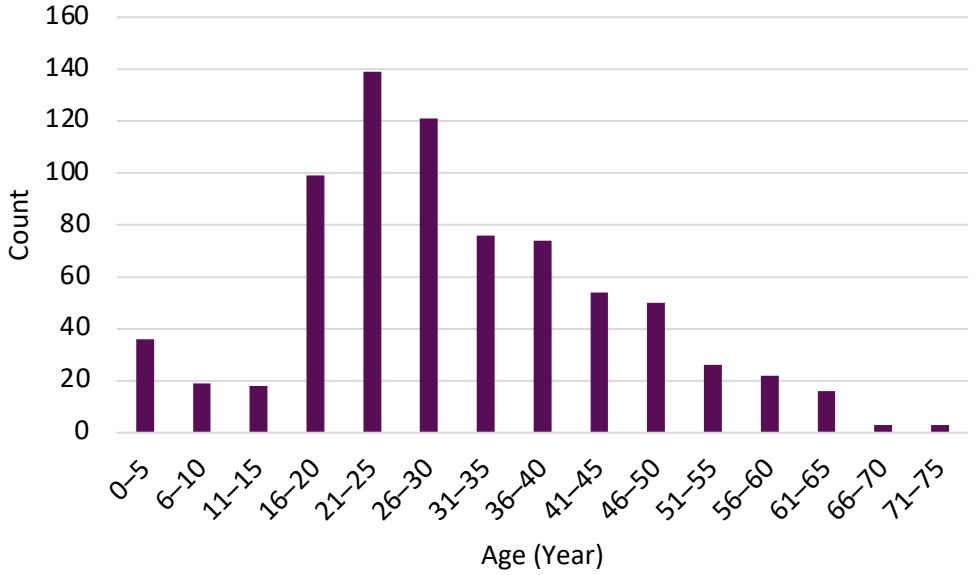
- A histogram takes as input a numeric variable only.
- The variable is cut into several bins
- The number of observation per bin is represented by the height of the bar.

	Age	…	…	…
x_1	35			
x_2	25			
…	…			
x_n	62			

$Age \in \{0,75\}$

Data

Bin of Age	Count
0-5	36
6-10	19
11-15	18
…	…
61-65	16
66-70	3
71-75	3



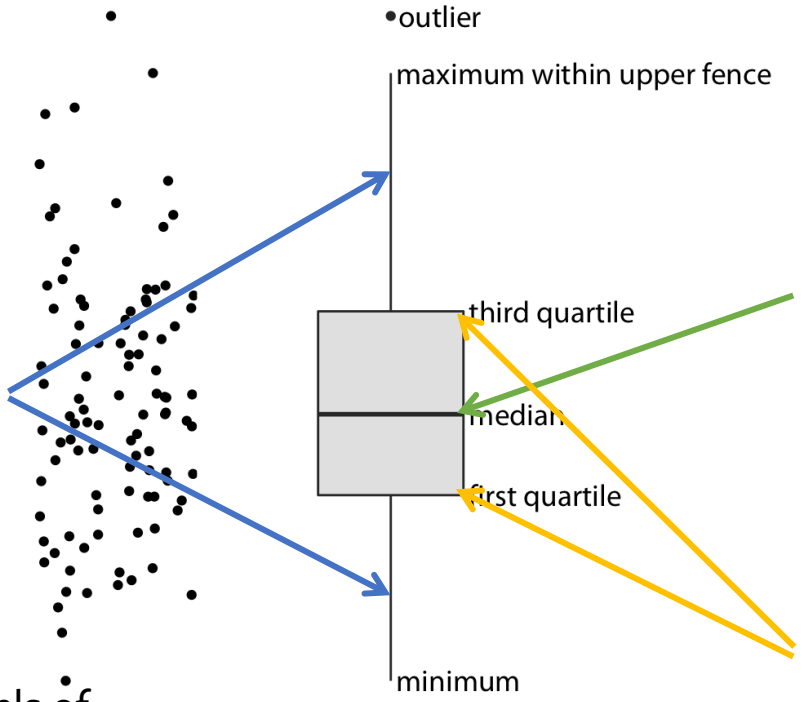
Histogram

Describing Data

Graphical Method

Boxplot

- Boxplot gives a nice summary of one or several distributions.



Extreme lines show the highest and lowest value excluding outliers

Line that divides the box into 2 parts represents the median of the data

The end of the box shows the third and first quartiles.

Anatomy of a boxplot

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

References & Study Resources

- Sinan Ozdemir and Divya Susarla. (2018). *Feature Engineering Made Easy*. Packt Publishing.
- Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.
- https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/
- https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg