Statistical Analysis

Jafar Namdar

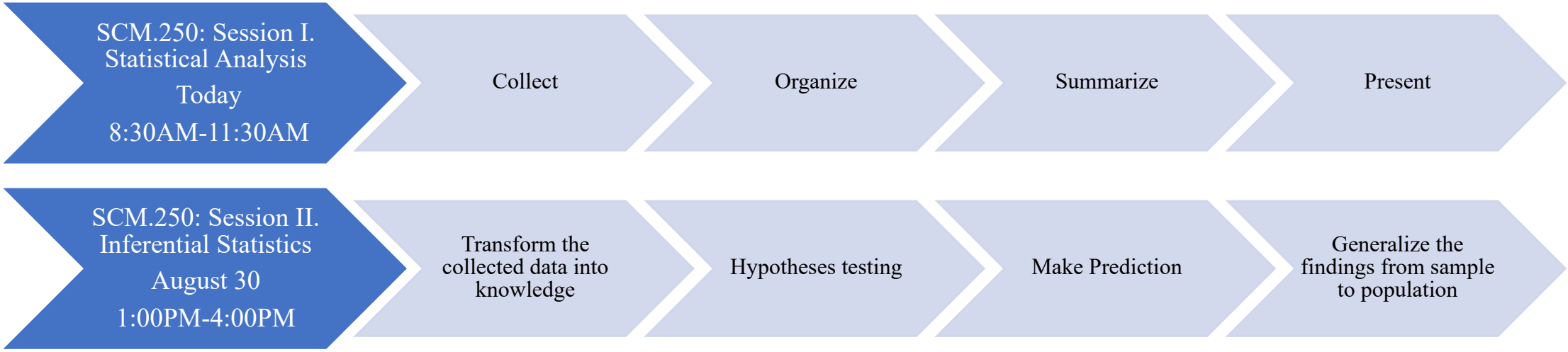
Postdoctoral Associate

MIT SCM & MIT Digital Supply Chain Transformation

Contact

- Email: Jnamdar@mit.edu
- My Office: E40-365

Two Sessions



What is statistics?

A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.*

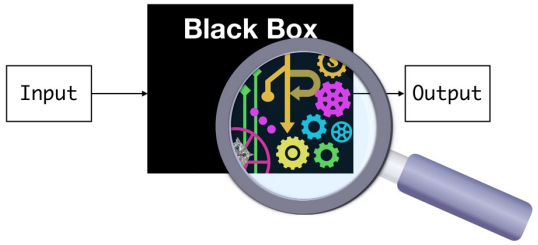
* *Merriam-Webster (retrieved from: https://www.merriam-webster.com/dictionary/statistics)*

What is statistics?

A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.*

Machine Learning Vs. Statistics

Machine learning models are designed to make the most accurate predictions possible.



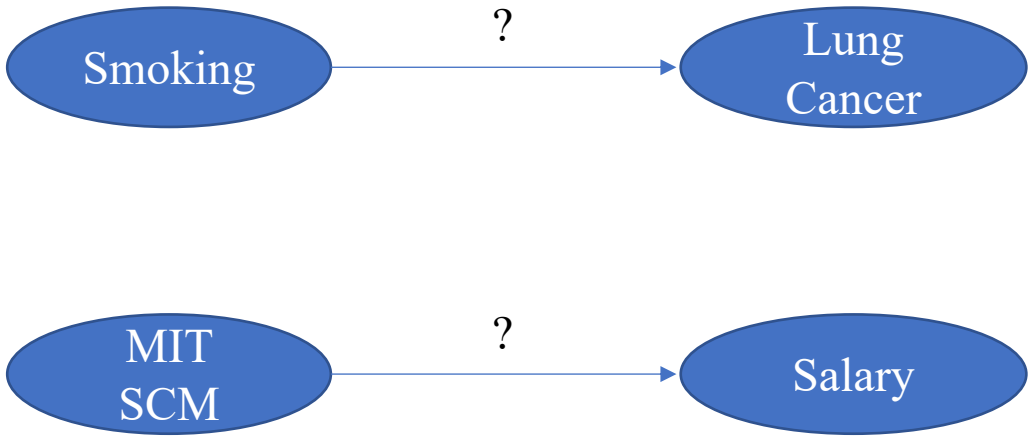
Statistical models are designed for inference about the relationships between variables.

Doctors
Scientist
Policy makers
Public Health Professional

https://www.kdnuggets.com/2016/11/machine-learning-vs-statistics.html
https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3
https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Statistical Thinking

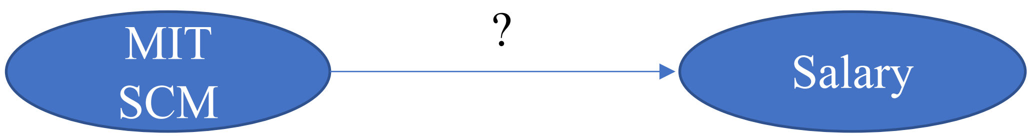
Statistical thinking involves the careful design of a study to collect meaningful data to answer a focused research question, detailed analysis of patterns in the data, and drawing conclusions that go beyond the observed data.

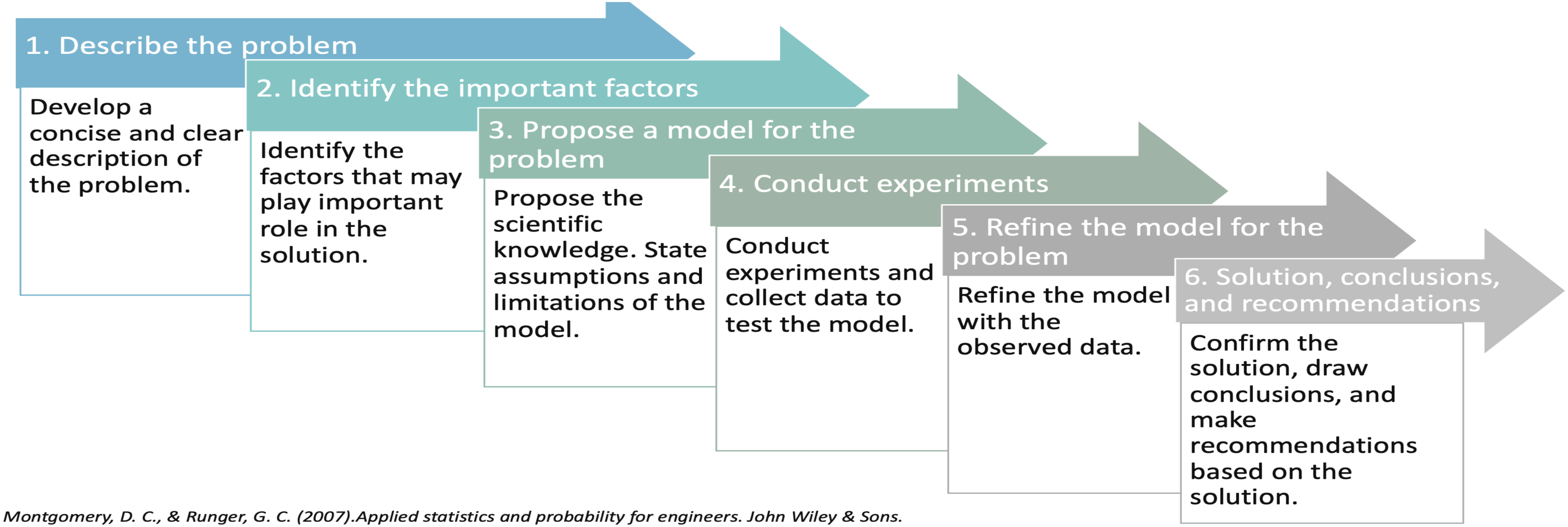


https://lardel.li/2019/02/data-science-is-not-a-black-box-managers-must-understand-it-too.html

Statistical Thinking

Statistical thinking involves the careful design of a study to collect meaningful data to answer a focused research question, detailed analysis of patterns in the data, and drawing conclusions that go beyond the observed data.





Montgomery, D. C., & Runger, G. C. (2007).Applied statistics and probability for engineers. John Wiley & Sons.

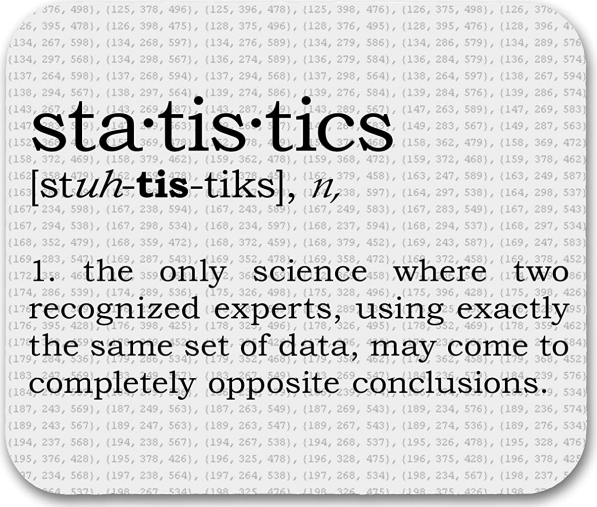
Statistical Thinking

Variability

Successive observations that do not produce exactly same results

How to incorporate variability into decision-making processes

Statistical Thinking



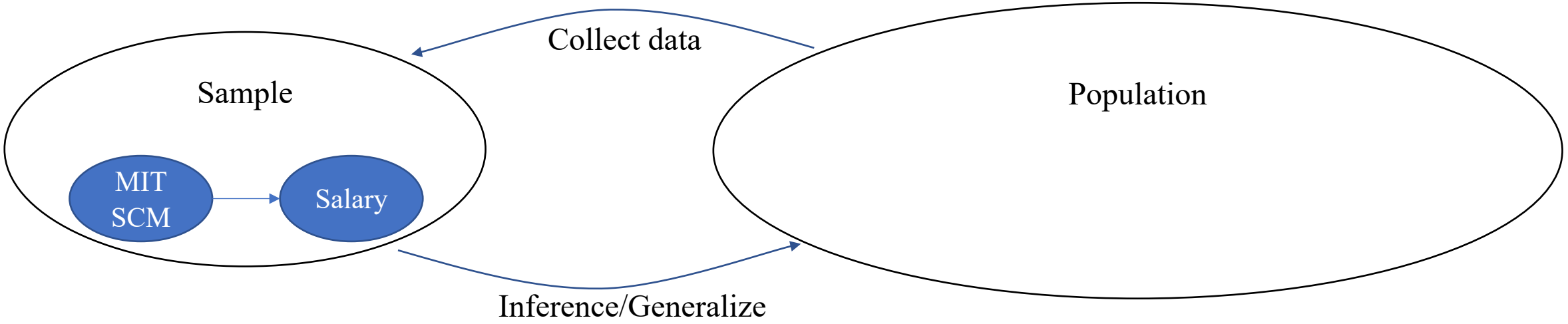
Example:

Before: Blood Pressure: 150/100
Doctor gives drugs
After: Blood Pressure: 148/90

Is the reduction in blood pressure related to drugs or variability? Statistics gives us an objective tool!

Descriptive Statistics

- We always work on sample data and try to generalize it to the entire population. Unfortunately, we never have access to the entire population.

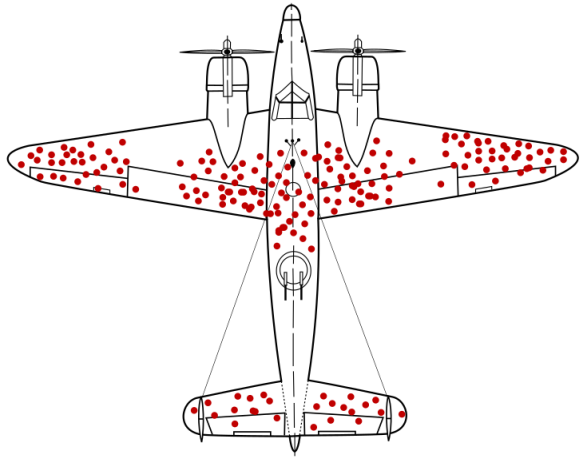


- Before making any inferences, you should "**get to know**" the data.

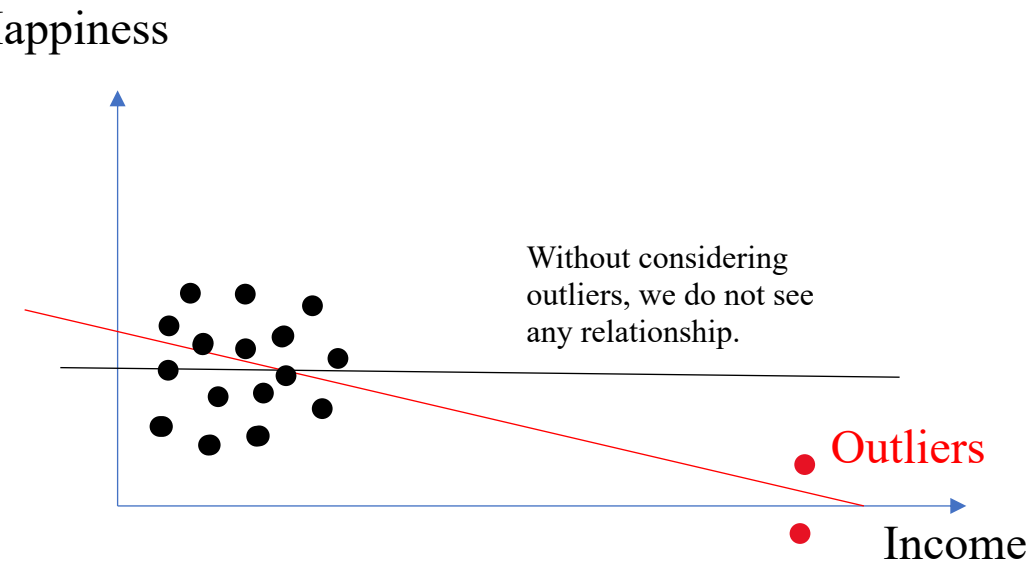
Descriptive Statistics

- Before making any inferences, you should "**get to know**" the data.

Abraham Wald's Analysis



Selection Bias

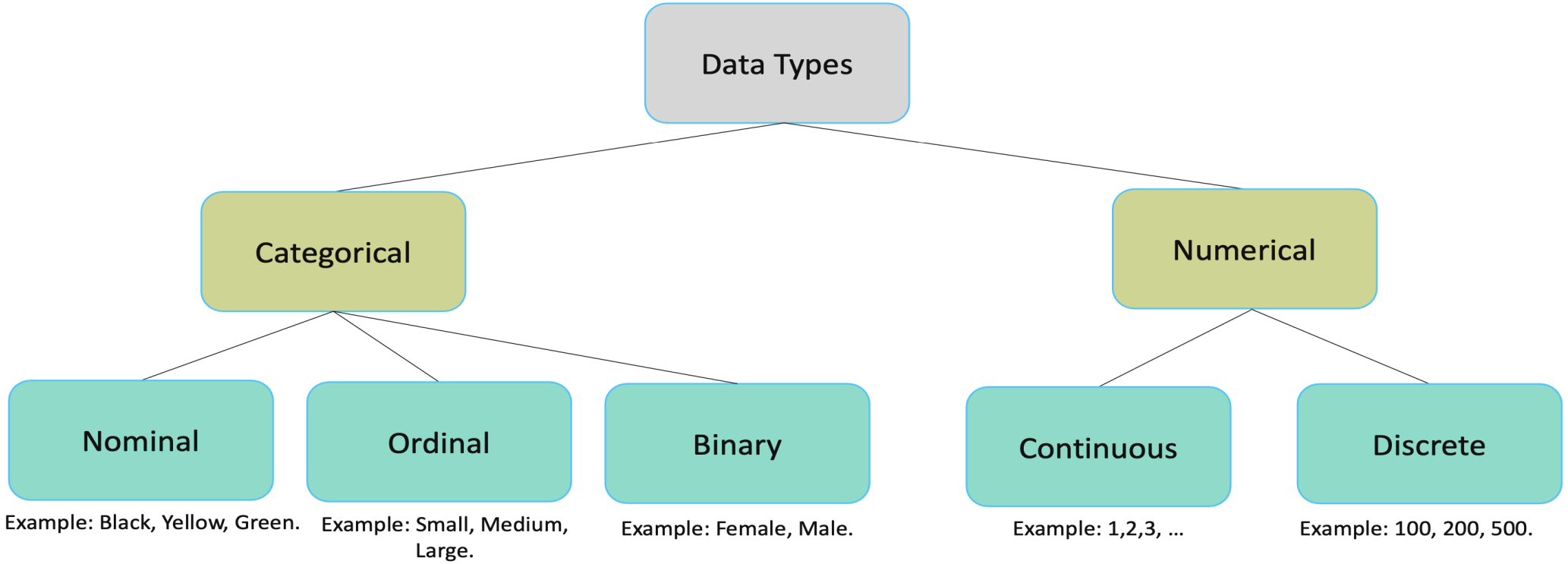


Failure to do so can lead to erroneous conclusions.

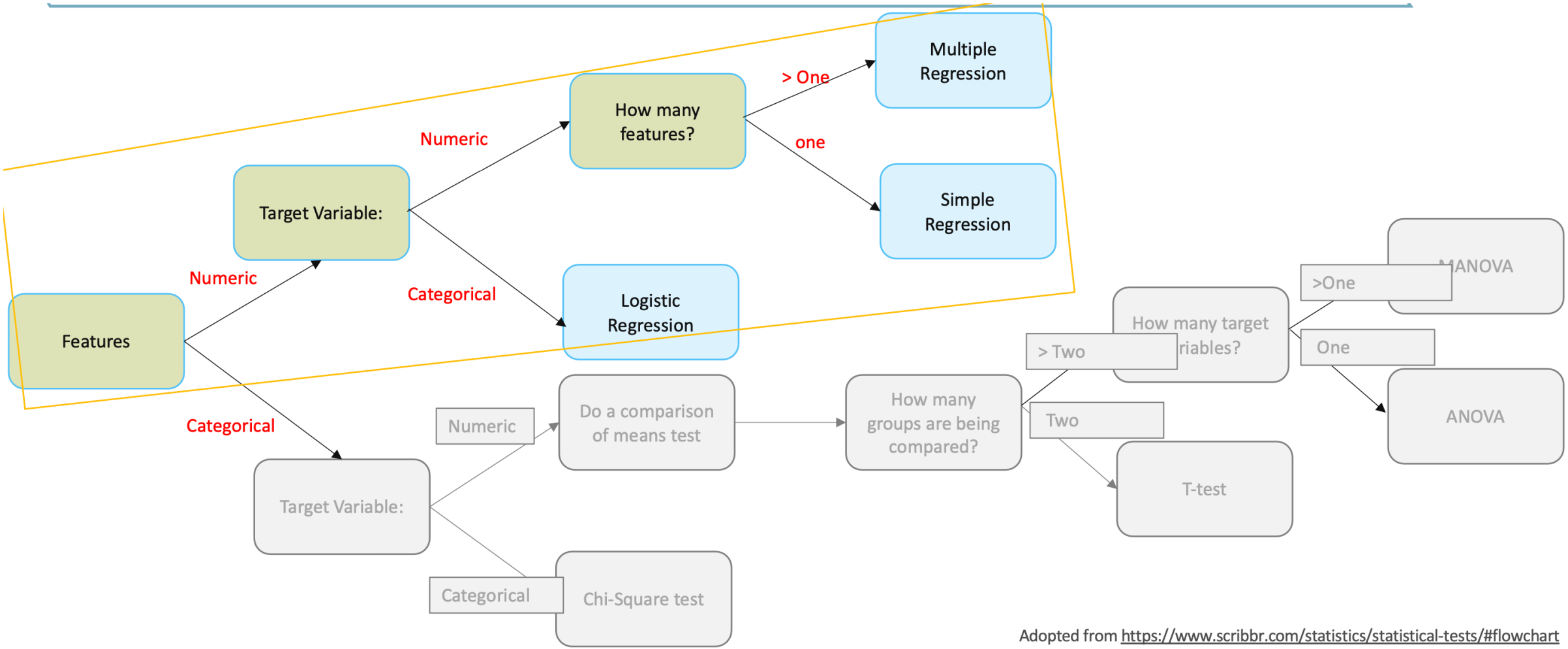
https://www.fharrell.com/post/introduction/#:~:text=An%20excellent%20example%20of%20statistical,which%20no%20damage%20was%20observed.
https://en.wikipedia.org/wiki/Abraham_Wald

Descriptive Statistics

- Data types influence the choice of your statistical models.



Data Types and Statistical Tests

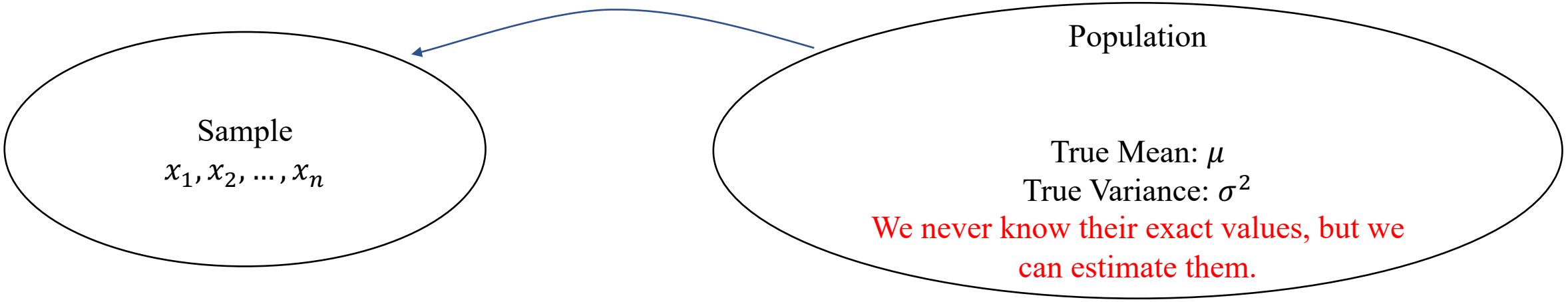


Adopted from https://www.scribbr.com/statistics/statistical-tests/#flowchart

Numerical Summaries of Data

Let's recall sample and population:

- Sample observations are part of larger population of observations.



$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad \text{Estimate for} \quad \mu$$

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} \quad \text{Estimate for} \quad \sigma^2$$

Numerical Summaries of Data

- Sometimes mean can be misleading!

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Salary	\$60k	\$135k	\$160k	\$140k	\$155k	\$120k	\$160k	\$155k	\$170k	\$3.8M

$\overline{x}$ =\$505.5k

$\overline{x}$ =\$139.4k after removing outlier

- What is the solution here? We should look at other measures to get a good picture of data.

Numerical Summaries of Data

- Measure of Dispersion

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Salary	\$60k	\$135k	\$160k	\$140k	\$155k	\$120k	\$160k	\$155k	\$170k	\$3.8M

Mean

$\bar{x}$ =\$505.5k

$\bar{x}$ =\$139.4k after removing outlier

Mean gives a good picture regarding the center of data

Variance

S^2 = 1206867.25

S^2 =996.91 after removing outlier

Standard Deviation

$S = \sqrt{1206867.25} = 1098.57$

S = 31.57 after removing outlier

Variance and Standard Deviation are the measure of dispersion

Numerical Summaries of Data

- In addition to Variance and Standard Deviation, Range is also used as a measure for dispersion in the data.
- Range: *Overall dispersion of values in the data.*

$$R = \max(x_i) - \min(x_i)$$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Salary	\$60k	\$135k	\$160k	\$140k	\$155k	\$120k	\$160k	\$155k	\$170k	\$3.8M

Mean

$\overline{x}$ =\$505.5k

$\overline{x}$ =\$139.4k after removing outlier

Mean gives a good picture regarding the center of data

Variance

S^2 = 1206867.25

S^2 =996.91 after removing outlier

Standard Deviation

$S = \sqrt{120686725} = 1098.57$

S = 31.57 after removing outlier

Range

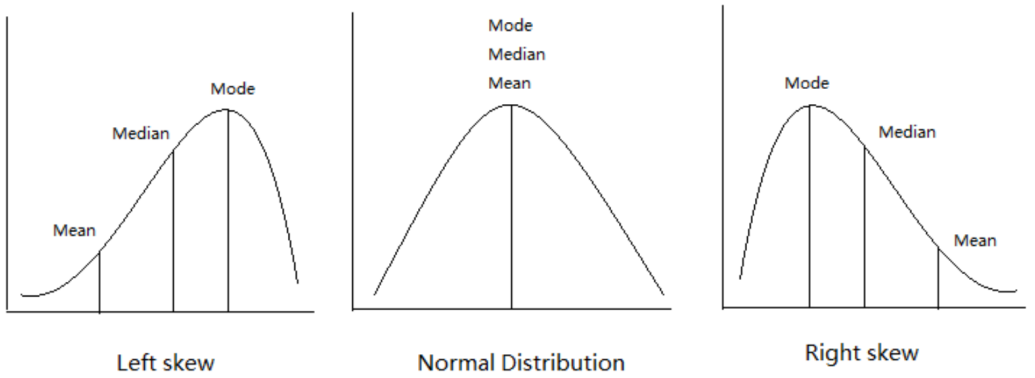
R=\$3.8M-\$60k=\$3.74M

R =110 after removing outlier

Range, Variance and Standard Deviation are the measure of dispersion

Numerical Summaries of Data

- Measure of Location
  - ➢ Mean: *Average value of observations (Center)*
  - ➢ Median: *Halfway between two central values.*
  - ➢ Mode: *Most frequently occurring data value.*



Source: https://medium.com/@nhan.tran/mean-median-an-mode-in-statistics-3359d3774b0b

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Salary	$60k	$135k	$160k	$140k	$155k	$120k	$160k	$155k	$170k	$3.8M

Calculate the median and mode of the salary data?

Median=?

Median=? After removing outlier

Mode=?

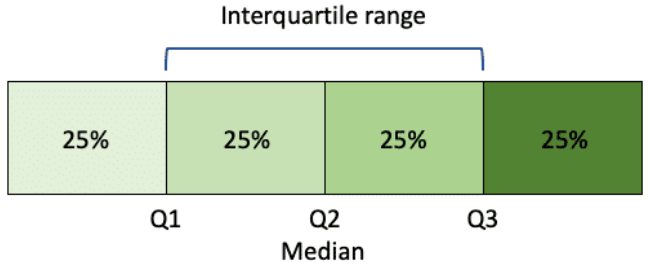
Mode=? After removing outlier

Numerical Summaries of Data

- Interquartile range: *The value between the third (75%) and first(25%) quartiles.*

$$IQR = q_3 - q_1$$

- Q1: 25%
- Q2: 50% or median
- Q3: 75%



Note: IQR is a Measure of Dispersion.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Salary	$60k	$135k	$160k	$140k	$155k	$120k	$160k	$155k	$170k	$3.8M

Calculate the IQR of the salary data?

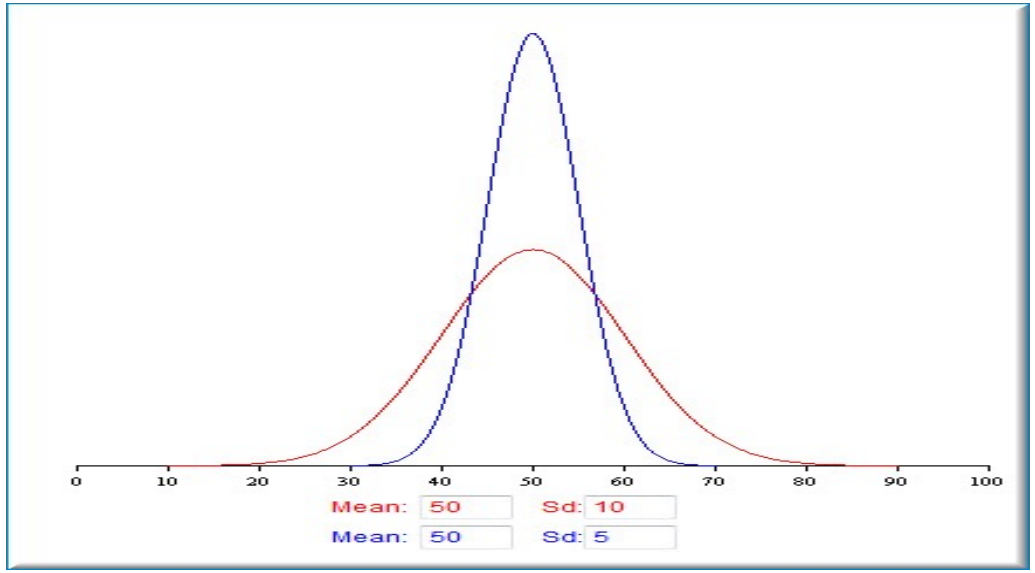
IQR=?

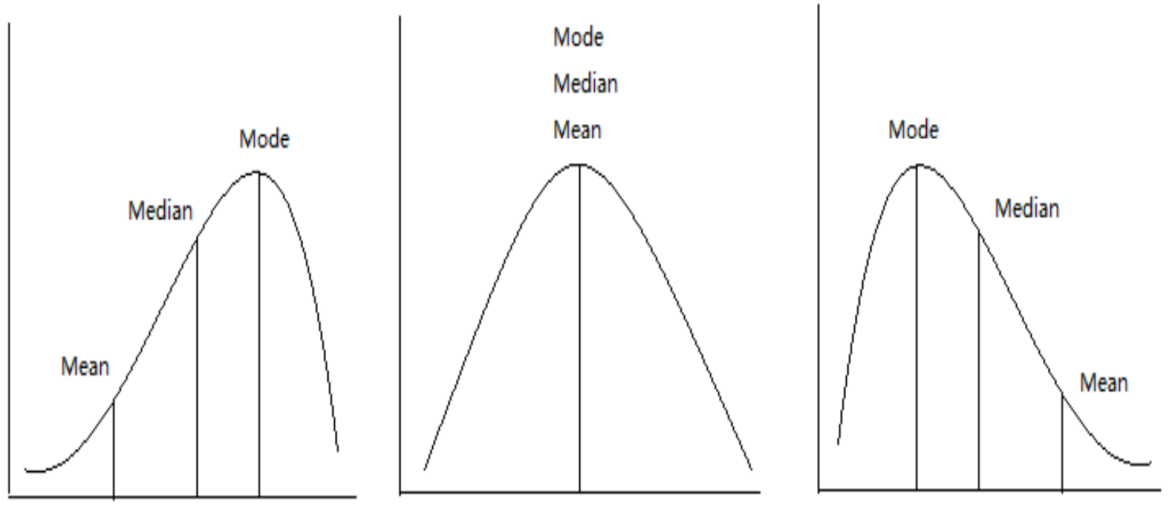
IQR=? After removing outlier

https://www.scribbr.com/statistics/interquartile-range/

Numerical Summaries of Data

- Plotting the distribution gives a good idea about data!





Source: https://medium.com/@nhan.tran/mean-median-an-mode-in-statistics-3359d3774b0b

https://onlinestatbook.com/2/summarizing_distributions/spread_sim_instructions.html

Limitations of Descriptive Statistics

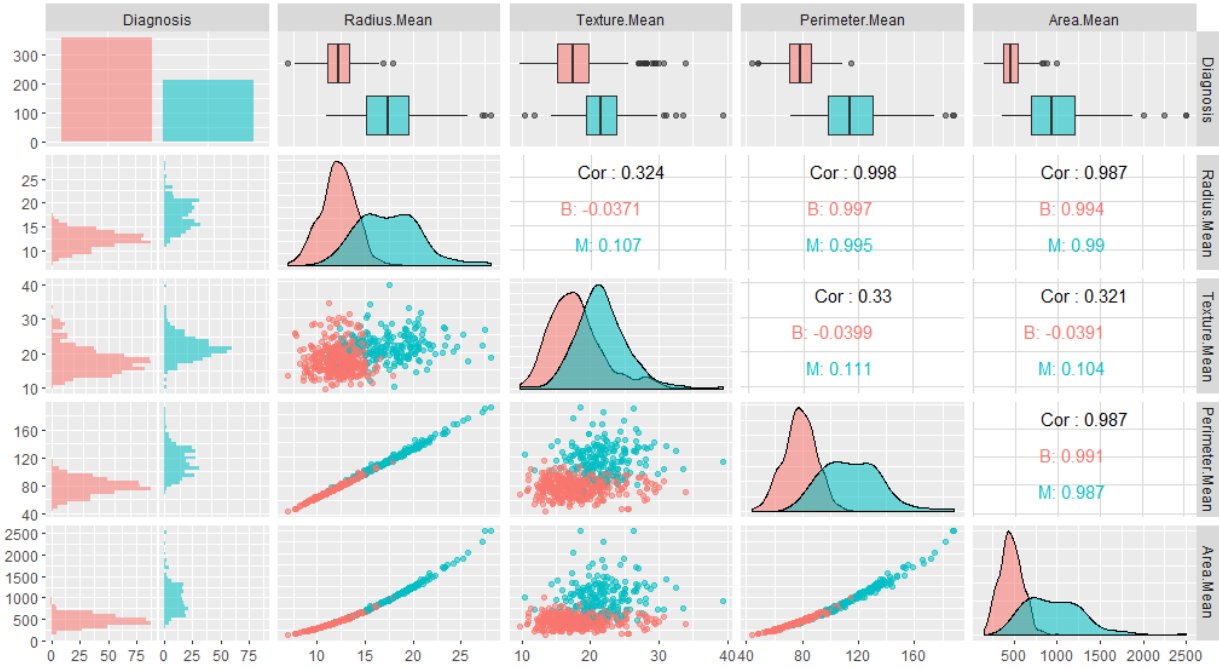
- Descriptive statistics is unable to:
  - Find causality of relationships between variables
  - Reach conclusion and generalize the results from sample to population
  - Correlate variables

- Using one descriptive statistics measure leads to lose important information.
- A good analysis of data should start with plotting the data.

- By measures of descriptive statistics (e.g., mean, median, etc.), you get a good picture of data and its structure which is essential for further analysis (inference).

Exploratory Data Analysis (EDA)

- EDA helps:
  - Understand characteristics of datasets
  - Identify outliers
  - Find interesting relationship among variables
  - Extract important features
  - Test underlying assumptions
  - Often employs data visualization methods



https://training.galaxyproject.org/training-material/topics/statistics/tutorials/intro-to-ml-with-r/tutorial.html

Exploratory Data Analysis (EDA)

- Some EDA tools:
  - Visualization (e.g., plotting distribution, scatter plot, etc.)
  - Correlation matrix, (helps to find the pair-wise correlated variables)
  - The measure of statistics such as mean, median, standard deviation, IQR, etc.
  - Dimensionality reduction (LASSO, PCA, etc.)
  - K-means clustering

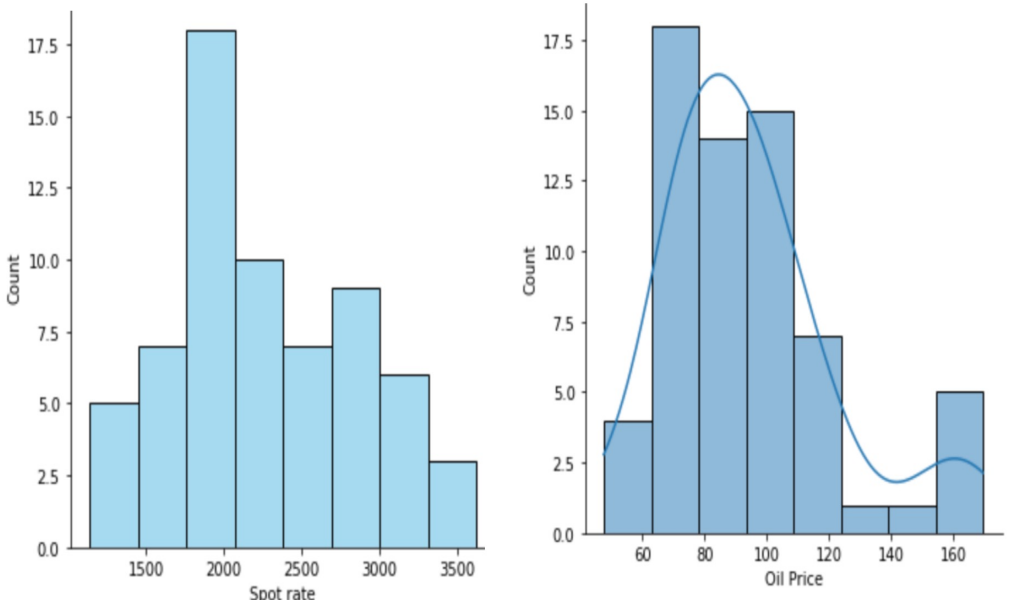
Dimensionality of Datasets

- **Univariate**
  - Measurement consists of one variable only.
  - Graphs: Stem-and-leaf diagram, histograms, boxplots.
- **Bivariate**
  - Measurement consists of two variables.
- **Multivariate**
  - Measurement consists of two or more variables.
    Graphs: Scatter plots, heat map, line graphs, bubble charts.

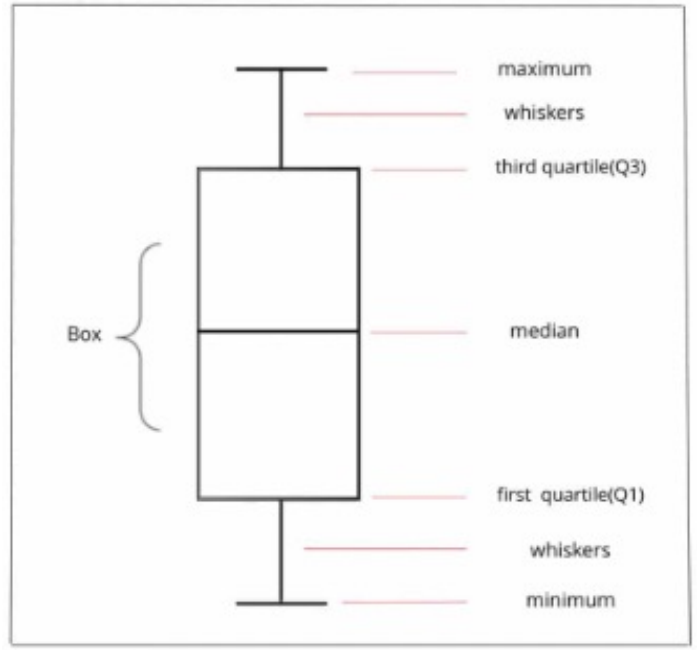
Univariate

Stem	Leaf
0	4
1	0, 7, 8
2	3, 3, 4, 7, 8
3	2, 2, 2, 3, 5, 7, 7
4	0, 0, 1, 1, 3
5	6, 7

Stem-and-leaf diagram

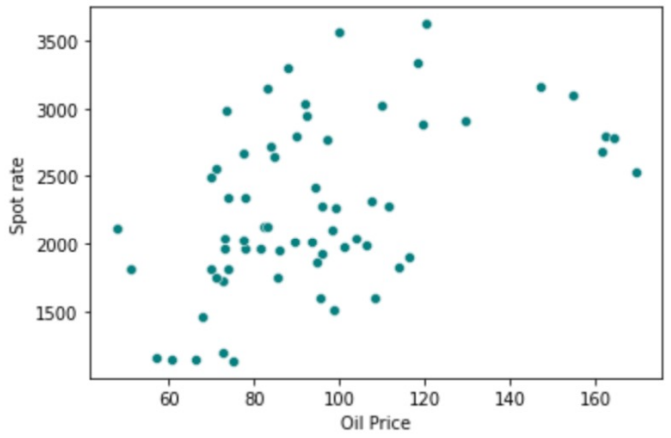


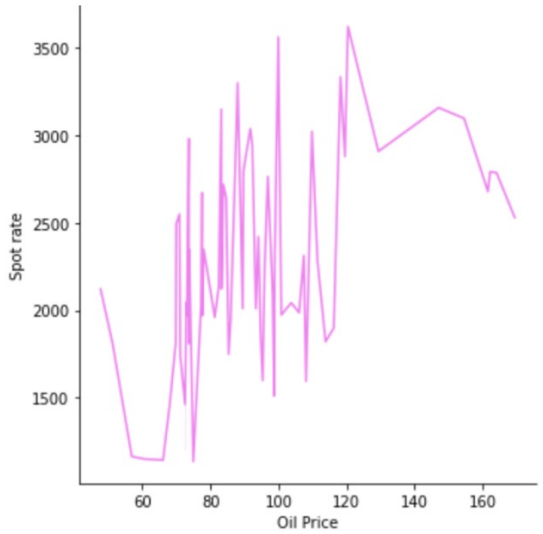
Histograms



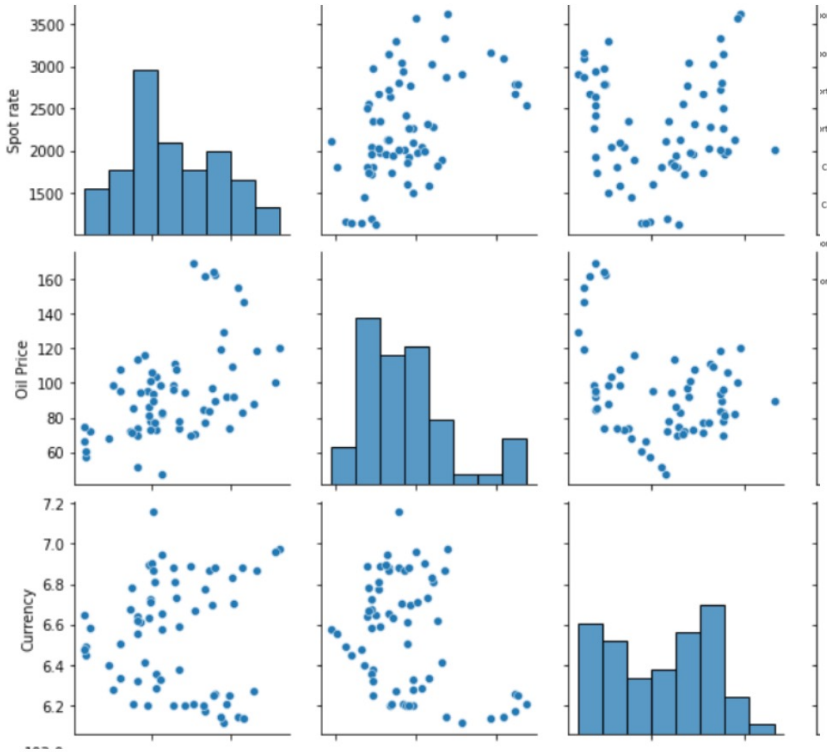
Boxplots

Multivariate Visualization



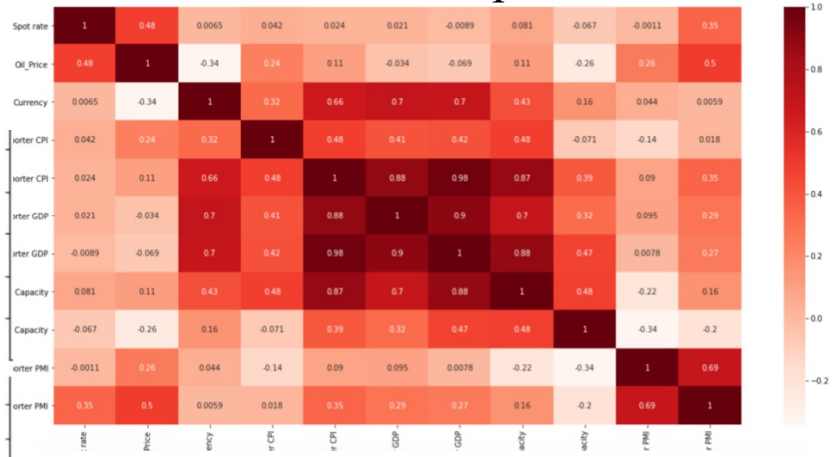


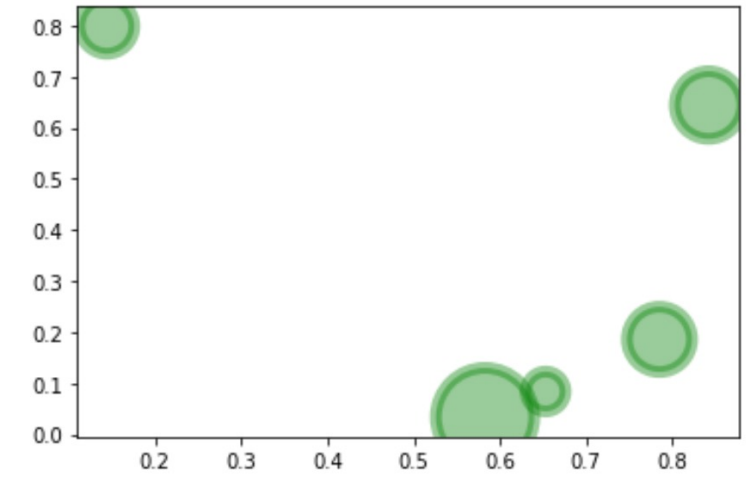
Line graphs



Scatter plots

Heat map





Bubble charts

EDA-Outliers

- Outliers are as samples that are exceptionally far from the mainstream of the data.

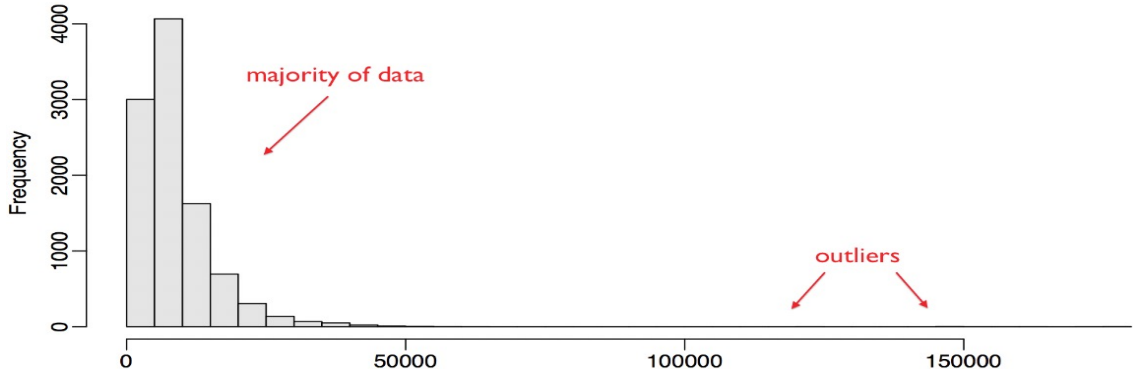
- Outliers in the datasets might be either:
  - Measurement or input error
  - True outlier observation (e.g., Jeff Bezos wealth)

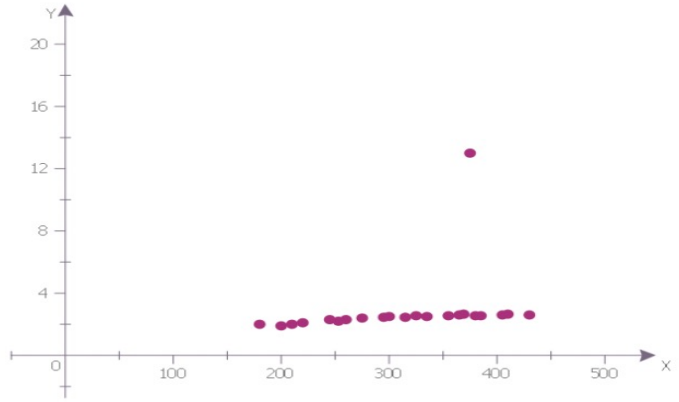
How can we detect outliers?

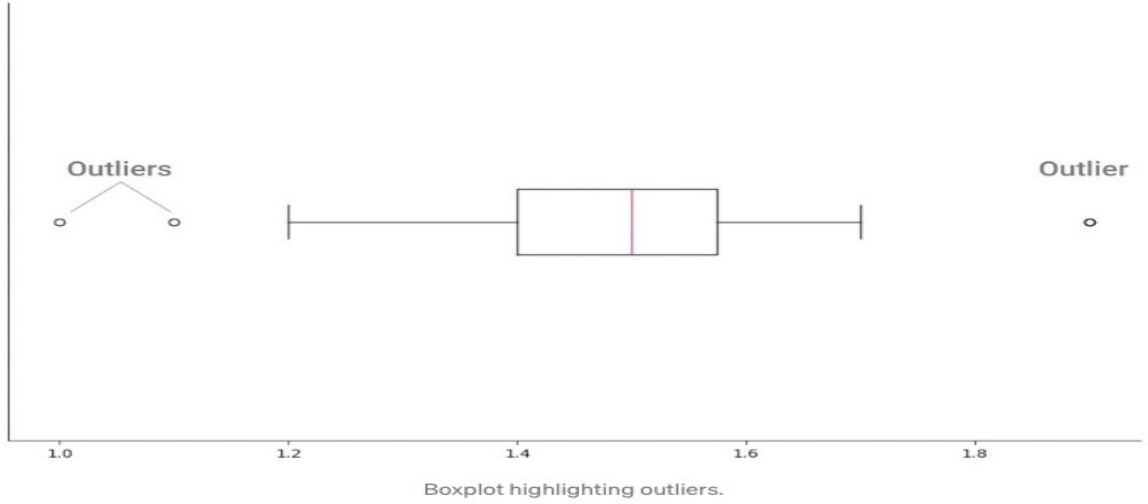
EDA-Outliers

➢ How can we detect outliers?

1. Plotting: Histograms, Scatterplots, Box plots.







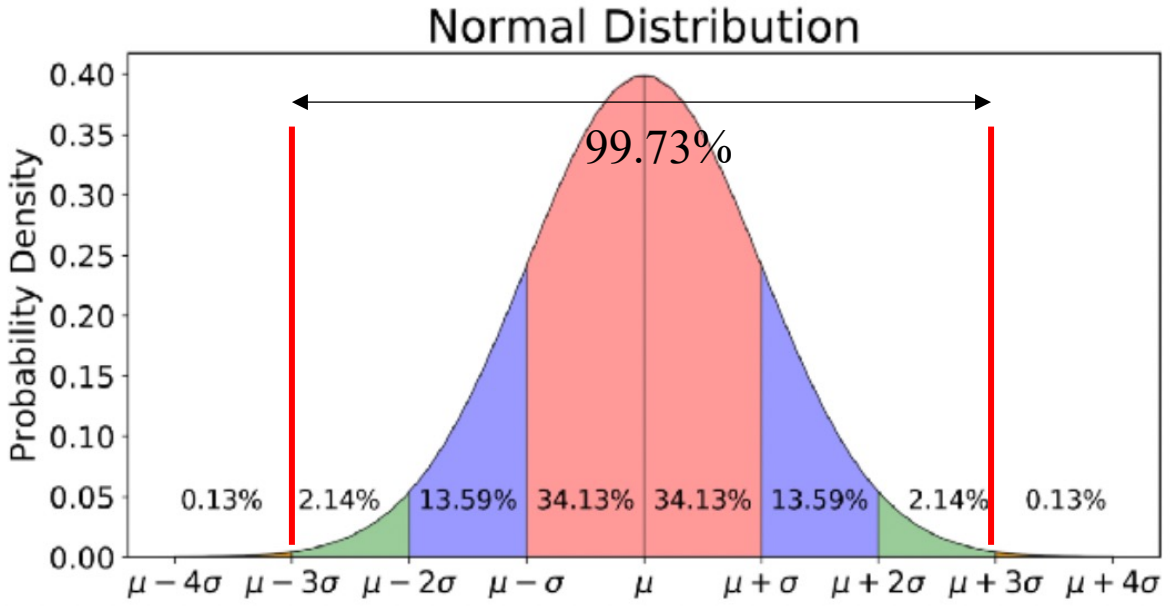
https://www.brendangregg.com/FrequencyTrails/intro.html
https://www.conceptdraw.com/examples/outliers-on-a-scattergraph
https://towardsdatascience.com/create-and-customize-boxplots-with-pythons-matplotlib-to-get-lots-of-insights-from-your-data-d561c9883643

EDA-Outliers

➢ How can we detect outliers?

2. Standardizing (Z-score)

$$z_i = \frac{x_i - \mu}{\sigma}$$ If $|z_i| > 3$, it raises flag



https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2

EDA-Outliers

➢ How can we detect outliers?

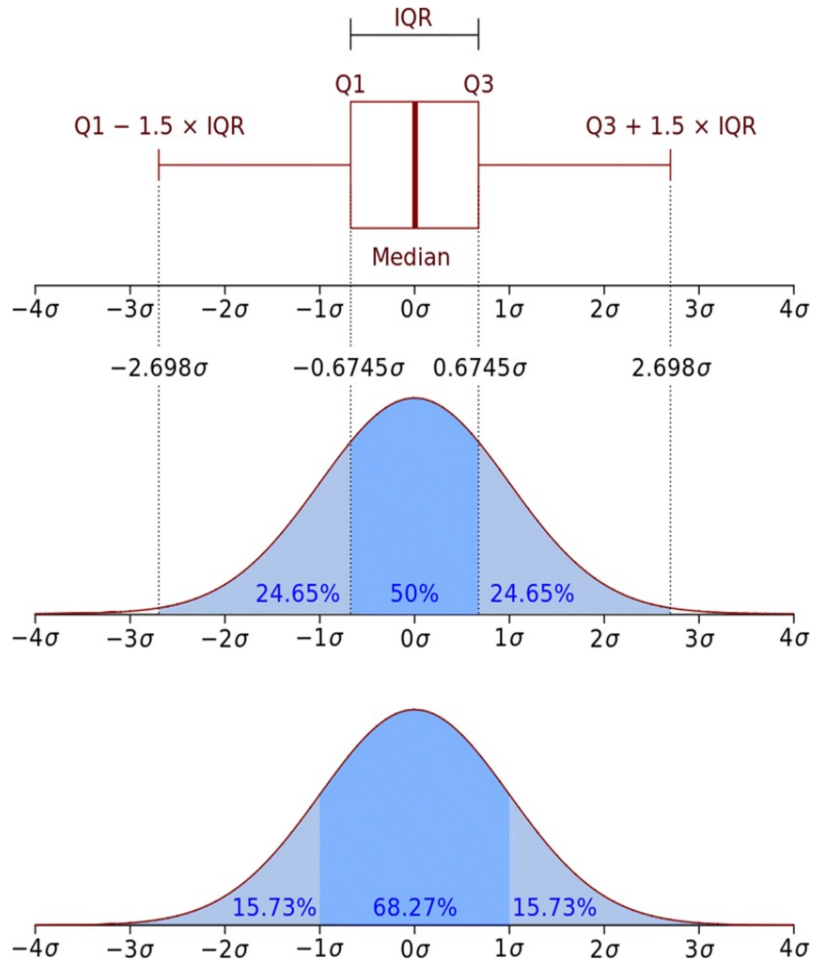
3. IQR method

Step 1: Calculate the interquartile range for the data ($q_3 - q_1$)

Step 2: Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers)

Step 3: Add $1.5 \times IQR$ to third quartile. Any number greater than this is a suspected outlier.

Step 4: Subtract $1.5 \times IQR$ from the first quartile. Any number less than this is a suspected outlier.



Ibrahim, E., Shouman, M. A., Torkey, H., & El-Sayed, A. (2021). Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system. *Multimedia Tools and Applications*, *80*(13), 20125-20147.

EDA-Rescaling Data

- ✓ Statistical variables may contain different scales and units such as volume, mile, pound, dollars, and more.
- ✓ The models can be more effective in the same scale.
- ✓ Two very common rescaling techniques:

Normalization (Min-Max scaling): rescales data to have the range between 0 and 1.

$$X_{i,new} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Standardizing: transforms data to have a mean of zero and a standard deviation of 1.

$$z_i = \frac{x_i - \mu}{\sigma}$$

Simple Linear Regression

- A regression model is used to model and explore relationships between variables that are related nondeterministic manner.
- Simple Linear Regression: Only one independent variable x (regressor or predictor) and one dependent variable Y (response).

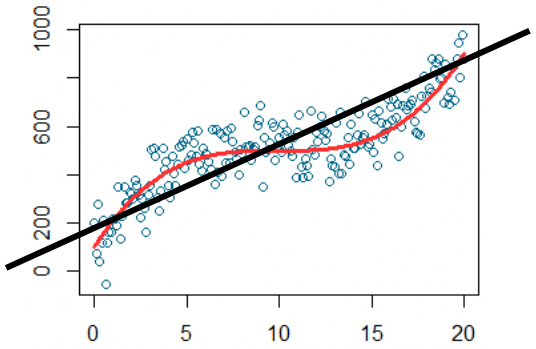
$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

- The mean of Y is a linear function of x. However, the actual observed value y does not have exact linear relationship.

- **The fitted or estimated regression line:**

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x \quad \text{Estimated}$$

$$y = \beta + \beta_1 x + \epsilon \quad \text{True relationship}$$



Multiple Linear Regression

- If the regression includes more than one predictor (aka independent variables, features, or x), then it is called multiple linear regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$ True relationship

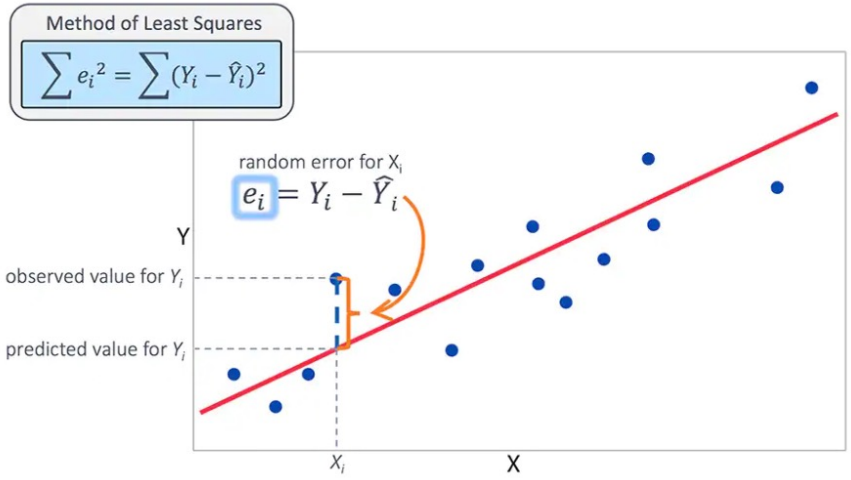
$$y = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_p} x_p + \epsilon$$

y : Dependent variable (also called: target, response)
x_i: Independent variable (also called: predictors, explanatory, features)

β_i: Regression coefficient

ϵ: Random error (also called residuals, noise) $e = y - \hat{y}$

e is used for predicting ϵ



https://www.jmp.com/en_no/statistics-knowledge-portal/what-is-multiple-regression/fitting-multiple-regression-model.html

EDA-Multi-collinearity issue

- It refers to situation when two or more independent variables are correlated.
- In this situation, the model gets unstable and confused, meaning it cannot separate out the effect of variables.
- For example:

$$BMI = \beta_0 + \beta_1 Weight_{lb} + \beta_2 Weight_{kg} + \beta_3 Height + \epsilon$$

$$Revenue = \beta_0 + \beta_1 TV_{Ad} + \beta_2 Radio_{Ad} + \epsilon$$

EDA-Multi-collinearity issue

- How to detect multi-collinearity issues?
  1. Correlation Matrix
     - If the magnitude of correlation is greater than 0.8, then you need to be careful.
  2. Variance Inflation Factor (VIF)
     - It should be less than 5 or 10. The smaller is better but it should not be greater than 10.

$$BMI = \beta_0 + \beta_1 Weight_{lb} + \beta_2 Weight_{kg} + \beta_3 Height + \epsilon$$

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$