CADT
Cambodia Academy of Digital Technology

Data Preprocessing

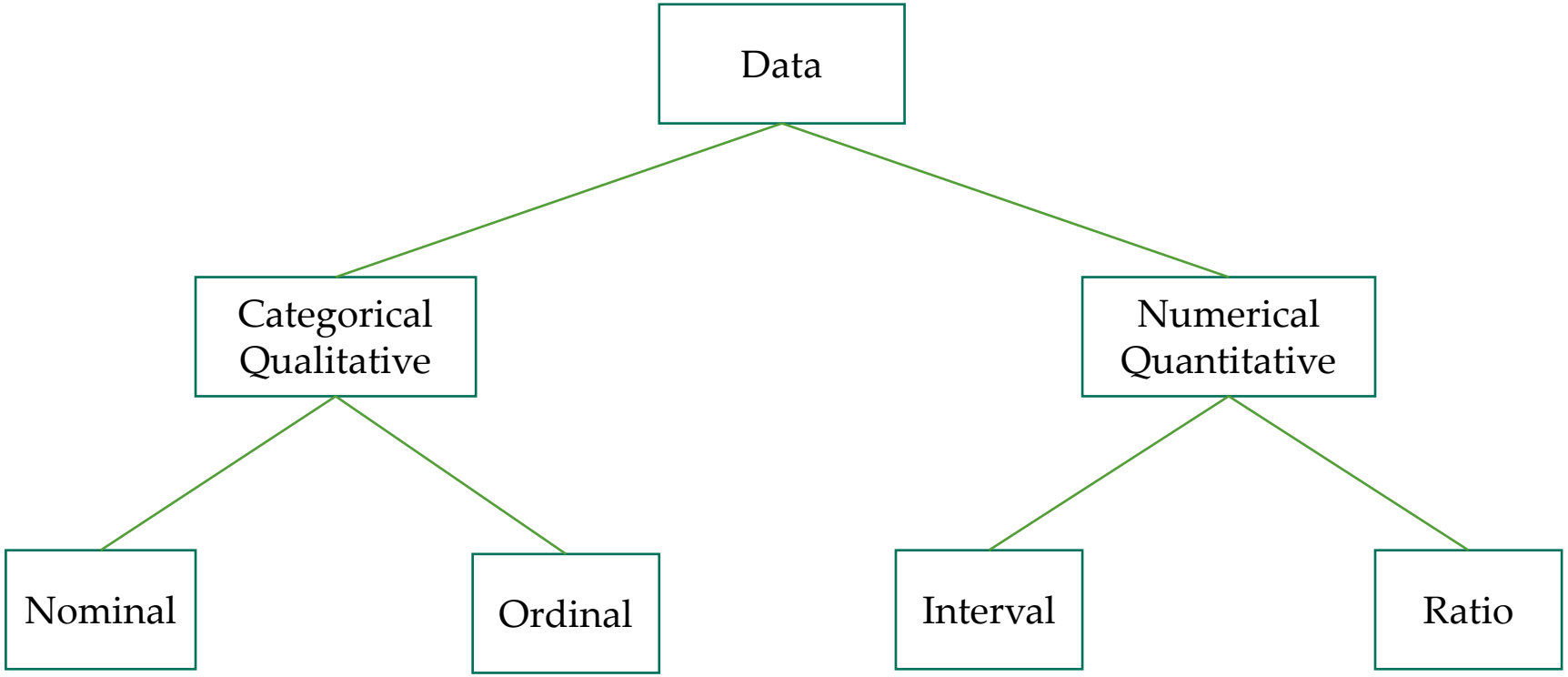
Lecturer:
Ms. Sokchovy MONIRATH
Mr. Soklay HENG

Academic Year: 2026-2027

Outline

1. Types of Data
2. Types of Variable
3. Data Cleaning
4. Feature Scaling
5. Feature Encoding
6. Feature Correlation

1. Types of Data



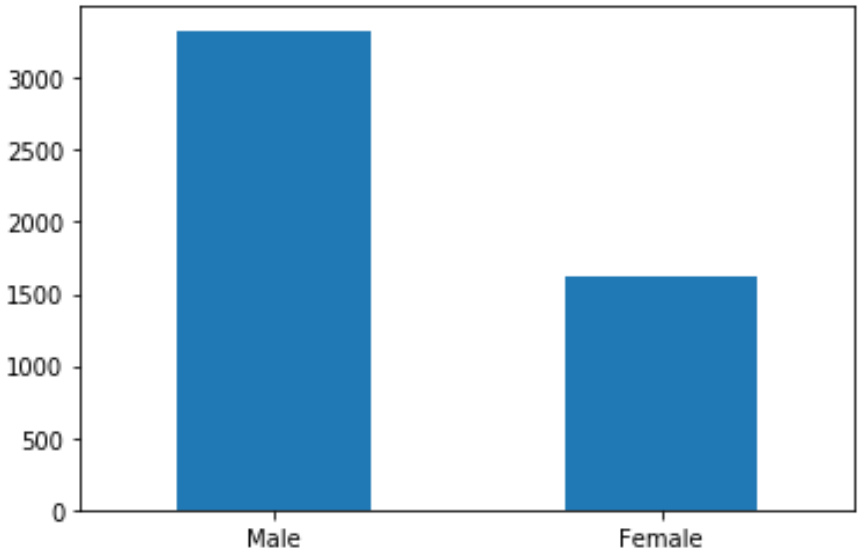
1. Categorical Data

Categorical Data refers to a collection of information divided into groups to represents the **characteristic** or the **quality** of someone or something. It does not have **mathematical meaning** and is classified into two groups:

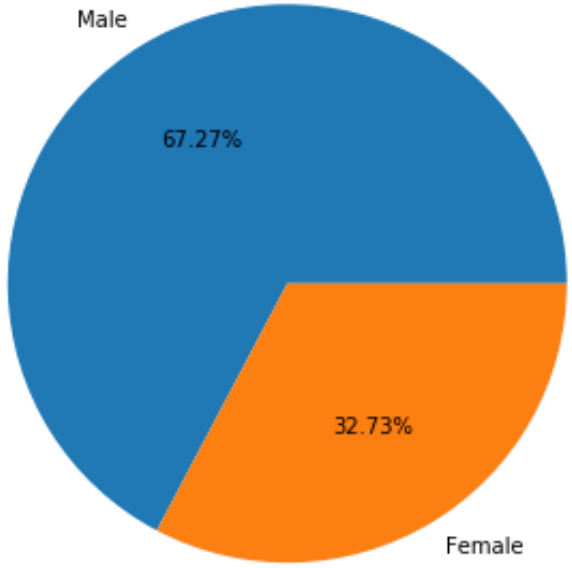
- Nominal (Label)
  - It **does not have order** (the values remain the same even changing the order).
  - E.g. gender, religion, color, language
- Ordinal
  - It **has order/rank**.
  - E.g. study grade, racing rank, language level, customer satisfaction

1. Categorical Data

- Graphical Representation: Sex



Bar Chart



Pie Chart

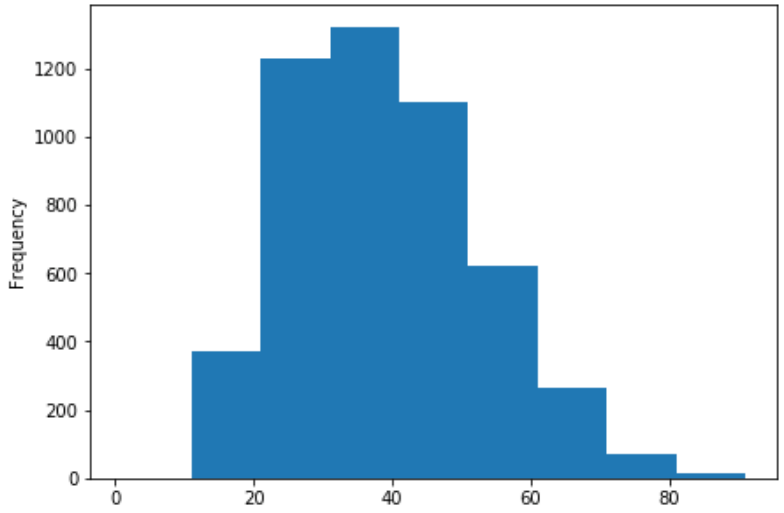
1. Numerical Data

Numerical Data refers to **discrete** or **continuous** values representing the **quantity** or the **measurement** of something. It does have **mathematical meaning** and is classified into two groups:

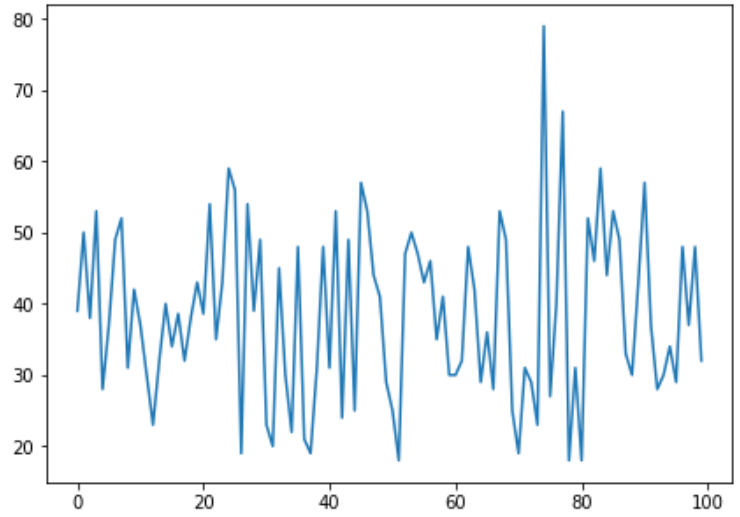
- Interval
  - The distances between values are equal, but the **zero point does not have the absolute value** (no true/natural zero).
  - E.g. temperature (°C, °F), time (12 hours), IQ score
- Ratio
  - It is for the interval data who **has true zero**.
  - E.g. length, height, weight, age

1. Numerical Data

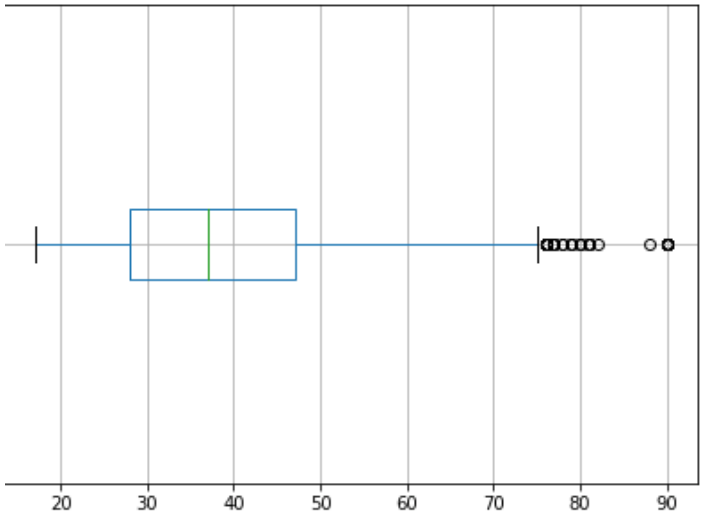
- Graphical Representation: Age



Histogram



Line Graph

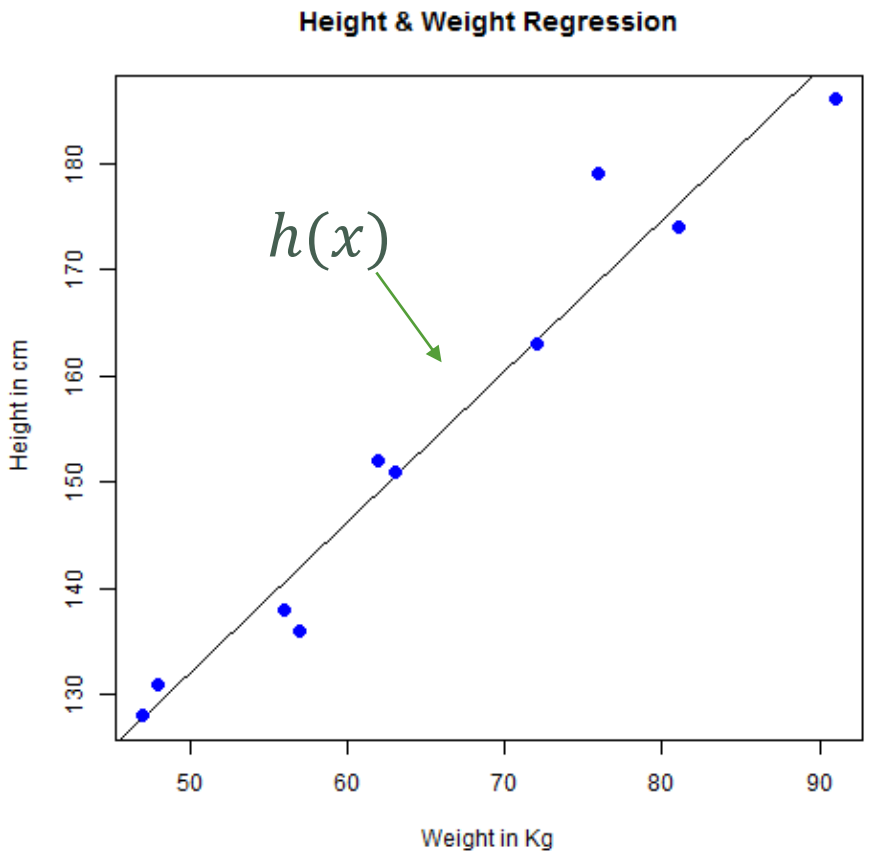


Boxplot

1. Detail Properties

Properties:	Nominal	Ordinal	Interval	Ratio
Order is known		✓	✓	✓
Count/Frequency	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median			✓	✓
Mean			✓	✓
Can add and subtract value			✓	✓
Can multiple and divide value				✓
Has ”true zero” representing the absence of the property being measured				✓

2. Types of Variable



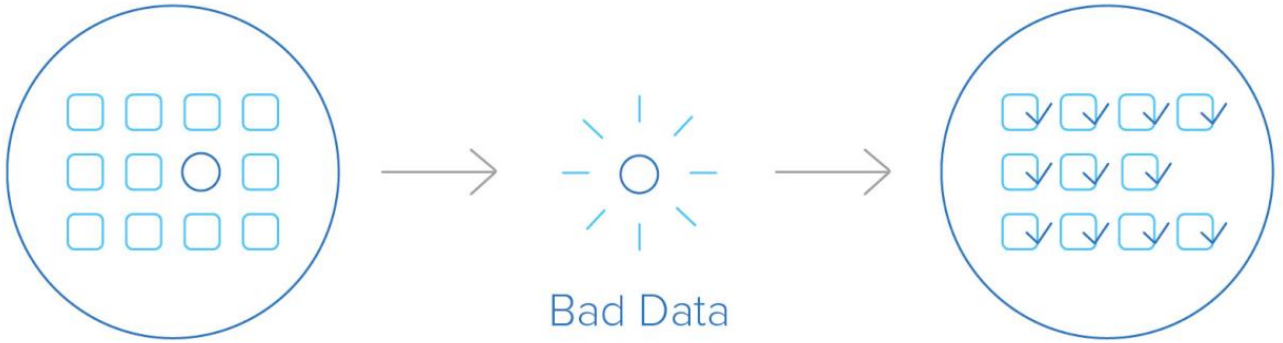
- **Independent variables** are regarded as the inputs to a system and may take on different values freely. They are also called as **predictor** or **explanatory variable** denoted by x.
- **Dependent variables** are those values that change as consequence of changes in other values in the system. They are also called as **response variable** denoted by y.

$x \rightarrow \boxed{h(x)?} \rightarrow y$

Weight ⟶ Height

3. Data Cleaning

- **Data Cleaning** is a process of identifying the corrupt or inaccurate parts from the data and correcting those parts by replacing, modifying or deleting the dirty or coarse data.
- The data have to be cleaned based on the context or problem so that it can be further analyzed and processed, especially before applying any Machine Learning algorithms.

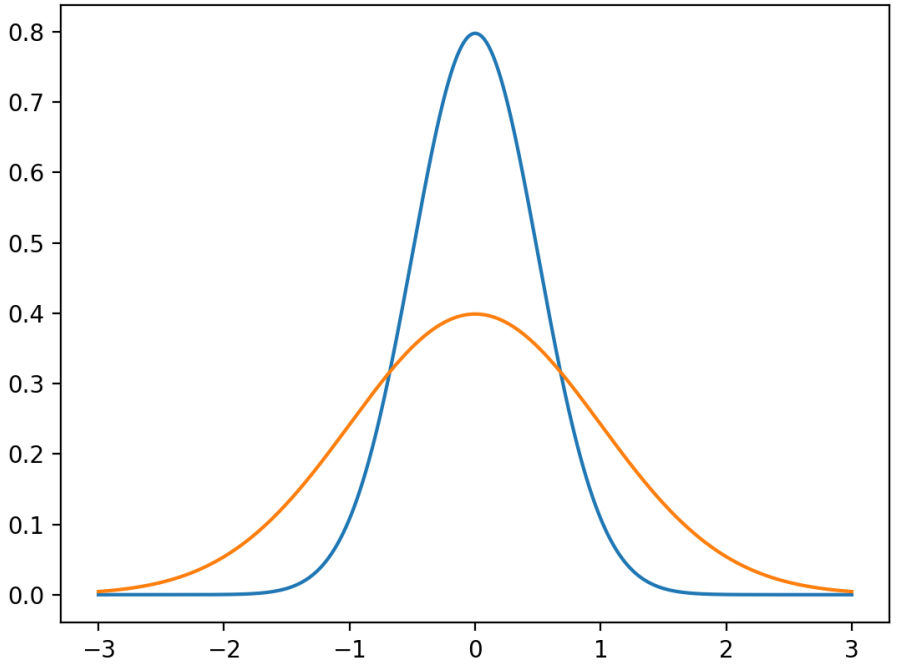


3.1. Low Variation

For variables with very low variation, we should either:

- ✓ Standardize all variables or use the standard deviation to account for variables with difference scales.
- ✓ Drop variables with **zero variance** (when the variable containing only one value) due to less predictive power.

- $var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - mean(x))^2$
- $std(x) = \sqrt{var(x)}$



Different Variances

3.2. Missing Value

Missing Values could be NA, n/a, empty string, ?,-1,-99,-999, null, etc. They should carefully be treated based on given context, because they would effect quality of the model during training.

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

3.2. Missing Value

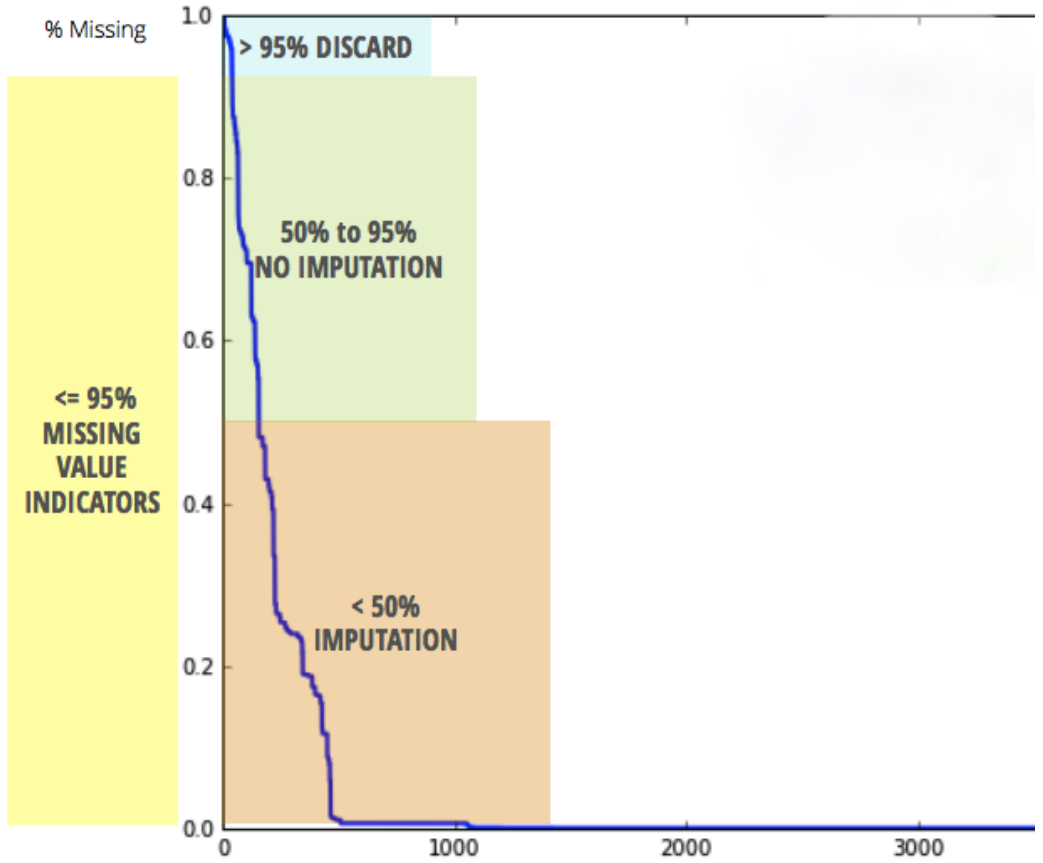
Possible ways to deal with missing values:

- ✓ Delete those records – not efficient when the dataset is small
- ✓ Apply **imputation**:
  - + Numerical variable: replace the missing values with either mean, median or mode
  - + Categorical variable: replace the missing values with the highest frequency value (mode)

- $mean(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $median(x)$ = middle value (when the data are arranged in order) of x
- $mode(x)$ = most common value of x

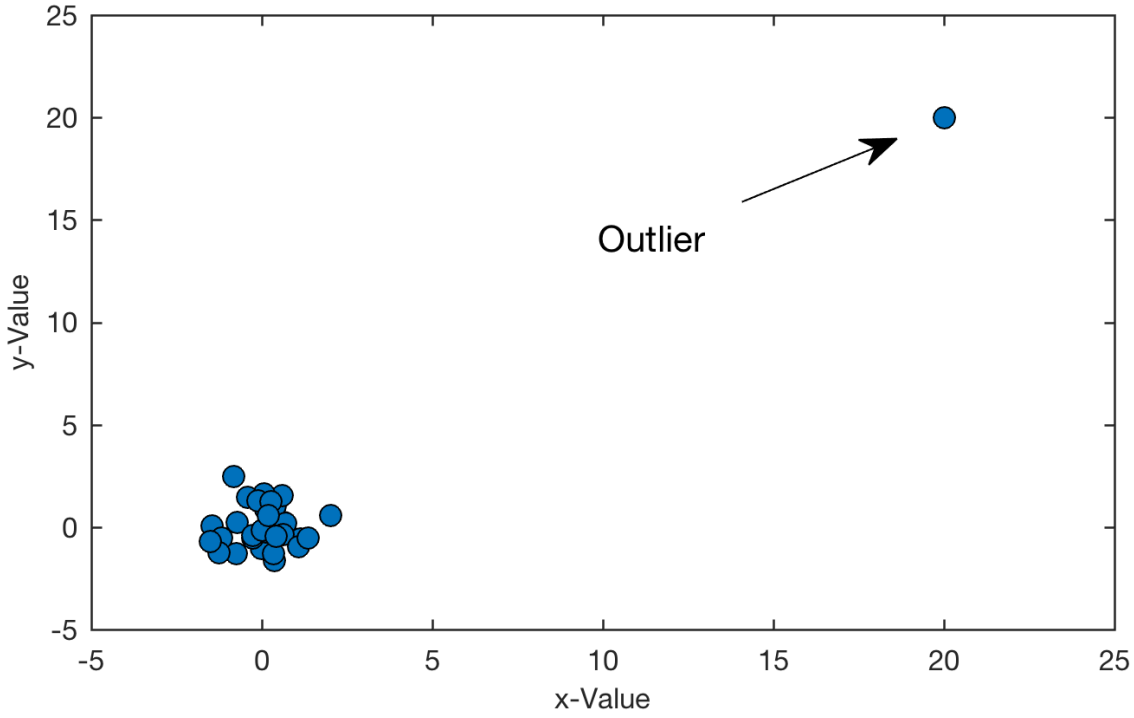
3.2. Missing Value

- Based on Vishal Patel, he suggests to apply imputation only when the amount of missing values of a feature is less than 50%.
- A feature whose missing values is over 95% should be filtered out.



3.3. Outlier

- In statistic, an **Outlier** is an observation that lies an abnormal distance from the other observations in the dataset. In some cases, it is considered as the corrupt or incorrect data.



3.3. Outlier

- As there is no precise way to define what are the outliers in general, we or a domain expert, must interpret the raw observations and decide whether a value is an outlier.
- The outliers deserve special attention or should be completely ignored for certain algorithms (i.e. Linear Regression, K-Means) because of their sensitivity.
- Common causes of outliers:
  - Data entry errors (human errors)
  - Measurement errors (instrument errors)
  - Sampling errors (extracting or mixing data from wrong or various sources)
  - Natural (not an error, **novelties** in data)

3.3. Outlier

The data point x is considered as an outlier when:

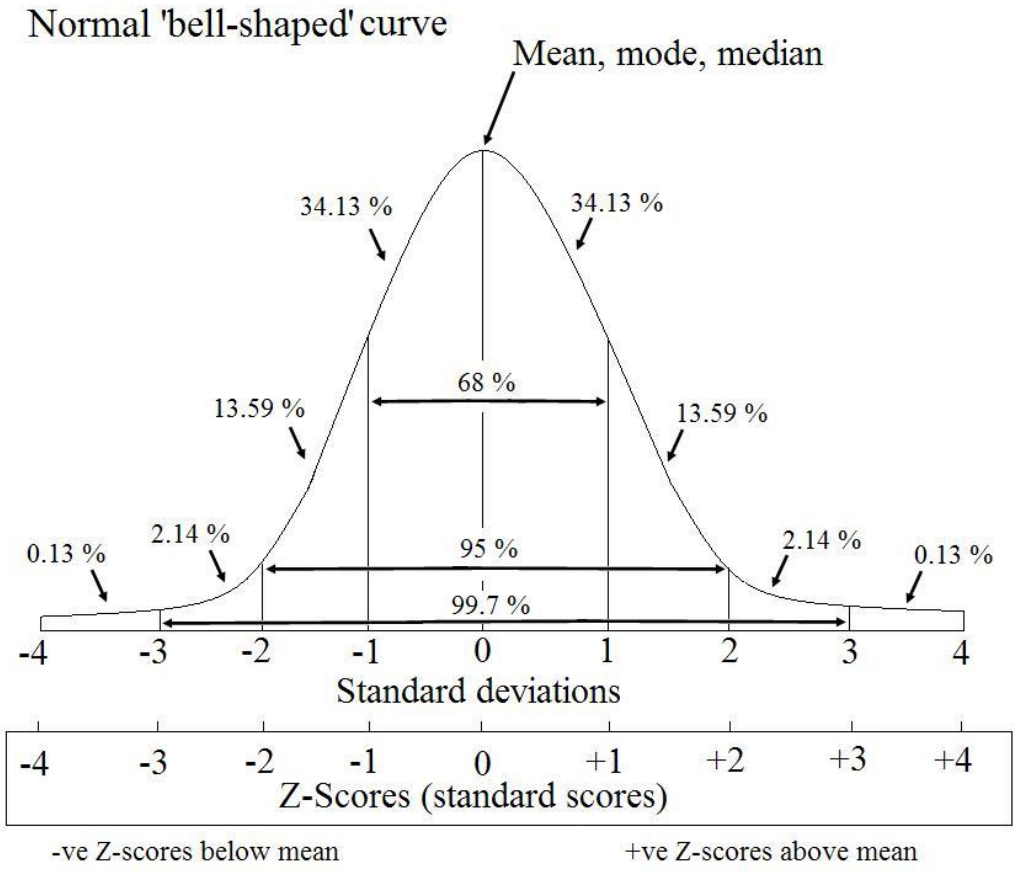
- It lies outside the mean plus or minus 3 times the standard deviation of the variable if the data distribution is Gaussian.

$$-3 < \textit{Z-score}(x) < 3$$

- It falls outside of 1.5 times of interquartile range above the 3rd quartile and below the 1st quartile.

$$Q_1 - 1.5 \times IQR > x > Q_3 + 1.5 \times IQR$$

3.3. Z-score



- Z-score measures how many standard deviations an observation is away from the mean.
- In Gaussian/Normal distribution, nearly 99.7 percent of the data are found between -3 and +3 of the Z-score.

$$Z\text{-}score(x) = \frac{x - mean(x)}{std(x)}$$

3.3. Z-score

Example: Find the outliers of the following data having mean=10 and standard deviation=30.

x
-120
84
66
-9
200
45
-1

3.3. Z-score

x	$Z\text{-}score(x)$
-120	-4.33
84	2.46
66	1.86
-9	-0.63
200	6.33
45	1.16
-1	-0.36

$$mean(x) = 10$$

$$std(x) = 30$$

$$Z\text{-}score(x) = \frac{x - mean(x)}{std(x)}$$

$$Z\text{-}score(x_1) = \frac{-120 - 10}{30} = -4.33$$

$$\vdots$$

$$Z\text{-}score(x_7) = \frac{-1 - 10}{30} = -0.36$$

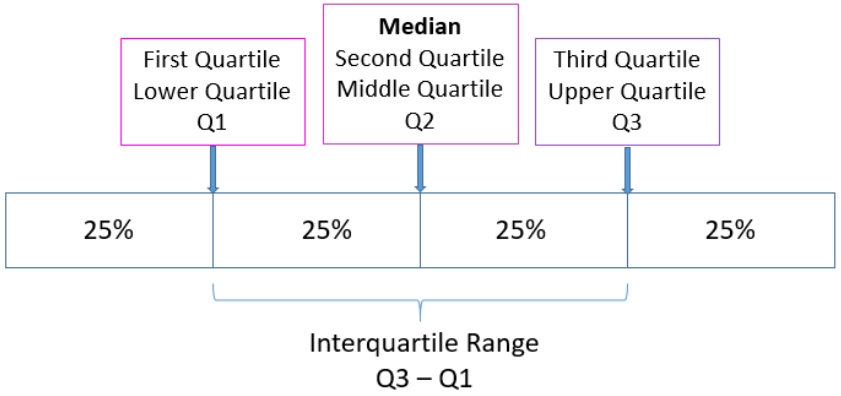
3.3. Z-score

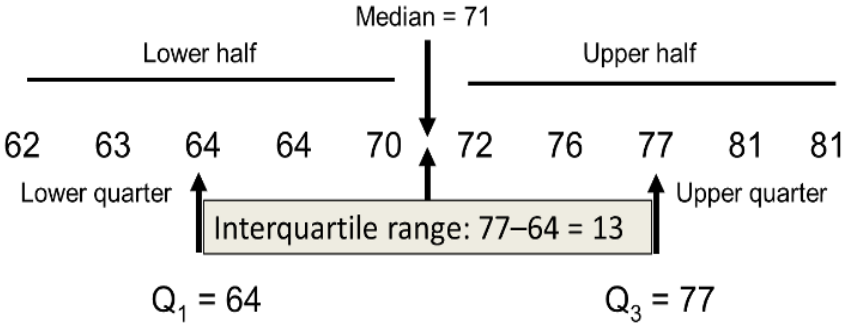
x	$Z\text{-score}(x)$
-120	-4.33
84	2.46
66	1.86
-9	-0.63
200	6.33
45	1.16
-1	-0.36

According to the values of the z-score in the table, "**-120**" and "**200**" are the outliers as their z-score is either less than -3 or greater than 3.

3.3. Interquartile Range (IQR)

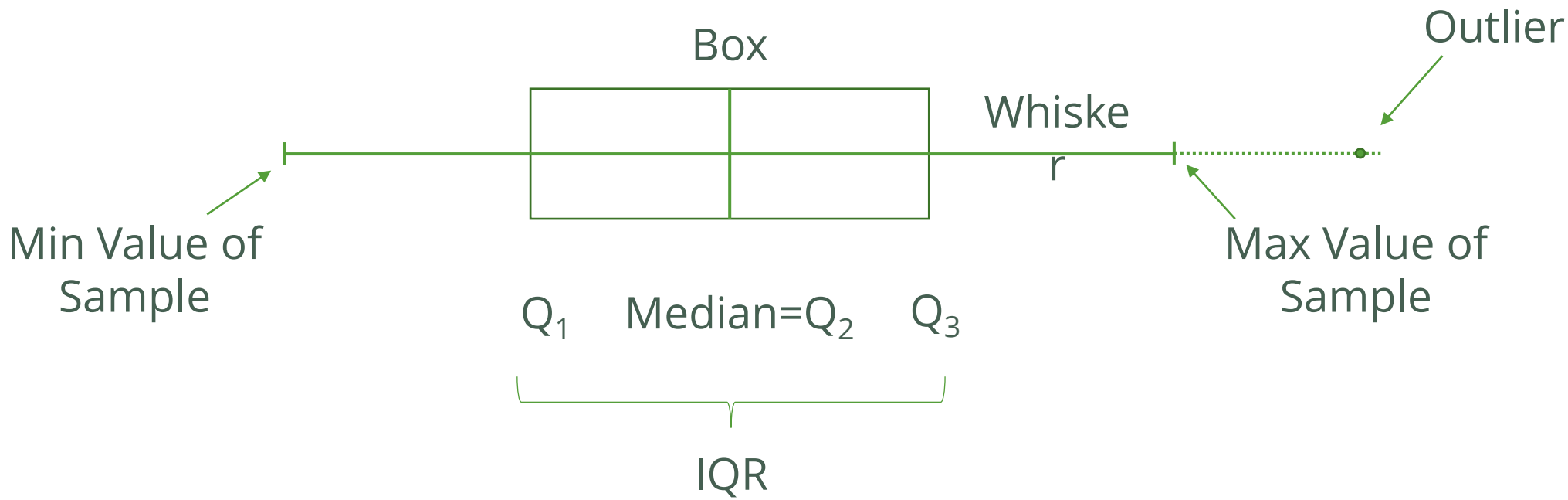
- The **interquartile range** (IQR) is a measure of variability, based on dividing a data set into quartiles.
- Quartiles divide a rank-ordered data set into **4 equal parts**. The values dividing each part are called the first, second, and third quartiles denoted by Q1, Q2, and Q3, respectively.
- IQR is the difference between Q3 – Q1.





3.3. Interquartile Range (IQR)

- Interquartile Range is frequently used with **Box and Whisker Plot** (boxplot) to visualize basic characteristic of the distribution and to identify the outliers.



3.3. Interquartile Range (IQR)

Example: Find the outliers among the following data using IQR.

x
15
10
39
20
44
55
48
109

3.3. Interquartile Range (IQR)

Step 1: Order the data in ascendant order

10, 15, 20, 39, 44, 48, 55, 109

Step 2: Find the median of data, which is the Q_2

$$Q_2 = median(x) = \frac{39+44}{2} = 41.5$$

Step 3: Split data into two parts and find Q_1 and Q_3

10, 15, 20, 39, 44, 48, 55, 109

17.5 41.5 51.5

$$Q_1 = \frac{15+20}{2} = 17.5 \text{ and } Q_3 = \frac{48+55}{2} = 51.5$$

3.3. Interquartile Range (IQR)

Step 4: Calculate IQR

$$IQR = Q_3 - Q_1 = 51.5 - 17.5 = 34$$

Step 5: Find the range of data such that $[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$

$$Q_1 - 1.5 \times IQR = 17.5 - 1.5 \times 34 = \text{-}33.5$$

$$Q_3 + 1.5 \times IQR = 51.5 + 1.5 \times 34 = 102.5$$

$\rightarrow$ Boundary range $= [-33.5, 102.5]$

3.3. Interquartile Range (IQR)

Step 6: Compare each value of x with boundary range $= [-33.5, 102.5]$

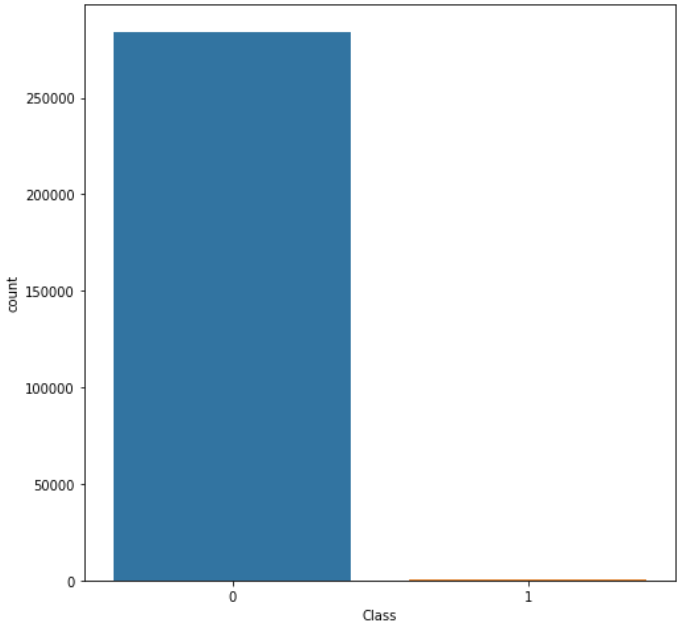
It shows that "**109**" lays outside the boundary range. Thus it is the outlier.



Q_1=-33.5 Q_2=41.5 Q_3=51.5

3.4. Imbalanced Dataset

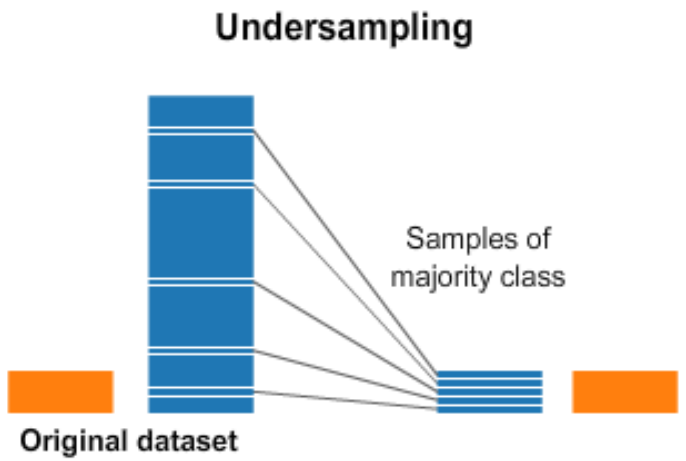
- Imbalanced dataset typically refers to the classification tasks where the classes are not represented equally.
- Machine learning algorithms such as K-Nearest Neighbor, Logistic Regression, etc. might be overfitting when they are trained with highly imbalanced dataset as they heavily focuses on the majority class.

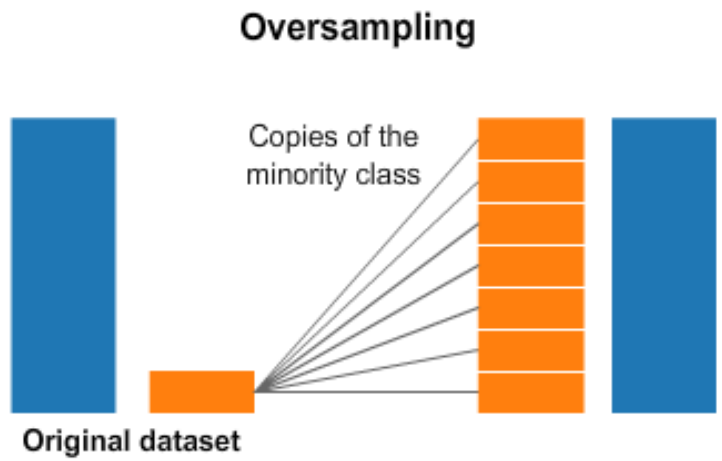


Imbalance dataset

3.4. Undersampling vs. Oversampling

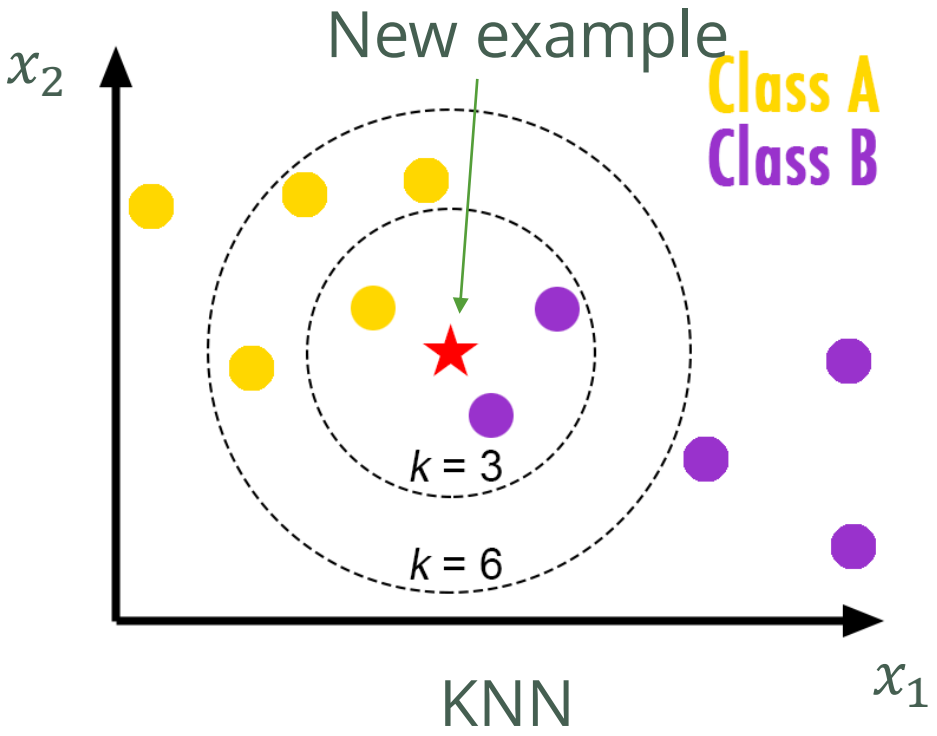
- **Undersampling (or downsampling)** drops the observations of the majority by randomly down sampling them to be of the same size as the smaller class.
- **Oversampling (or upsampling)** duplicates the observations of the minority class by repeatedly sampling them over and over until it reaches the same size as the majority class.



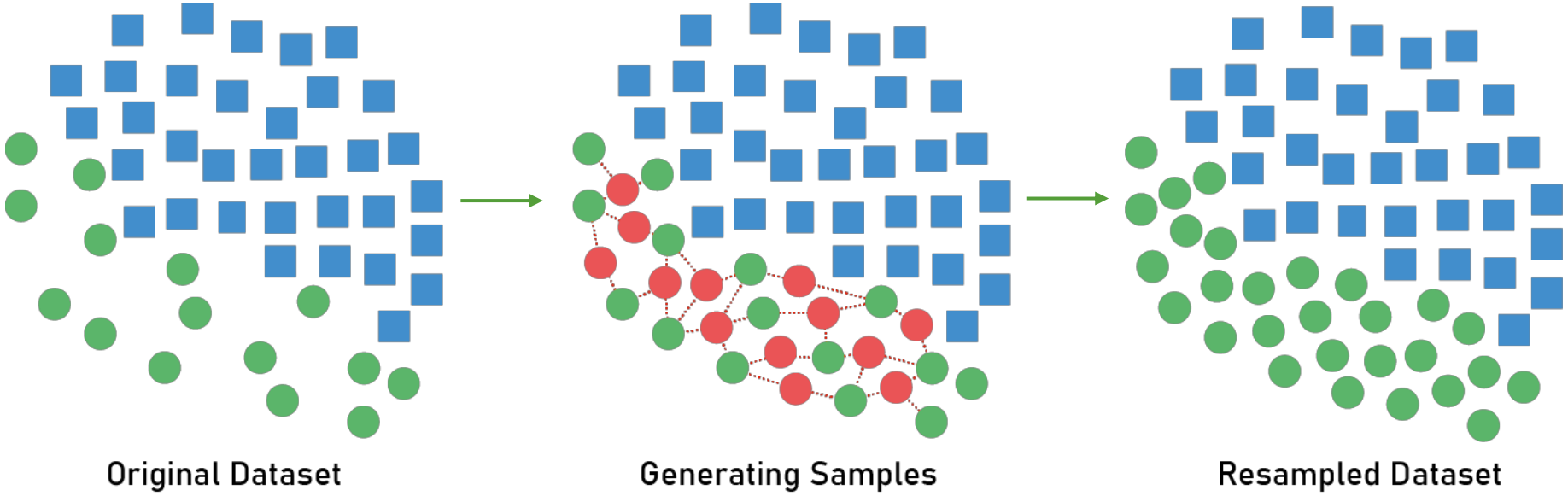


3.4. SMOTE

SMOTE stands for **S**ynthetic **M**inority **O**versampling **Te**chnique — it consists of creating **artificial samples** from the minority class by using K-Nearest Neighbors (KNN). SMOTE is used to avoid model overfitting.



3.4. SMOTE



4. Feature Scaling

- Most often, the dataset contains features highly vary in magnitudes and units. As learning algorithms only take in features magnitude neglecting their unit, the features with higher magnitude might dominate the one with low magnitudes.
- Feature scaling is essential for algorithms (i.e. KNN, K-Means, PCA) who use **Euclidean Distance** as feature with higher range will weigh a lot than the others in distance calculation.
- Another reason why feature scaling is applied is that training algorithms which use **Gradient Descent** tends to converge much faster.

$$a = \begin{pmatrix} a_{x_1} \\ a_{x_2} \end{pmatrix},\ b = \begin{pmatrix} b_{x_1} \\ b_{x_2} \end{pmatrix} \Big\} \ d_{euc}(a,b) = \sqrt{\left(a_{x_1} - b_{x_1}\right)^2 + \left(a_{x_2} - b_{x_2}\right)^2}$$

$$x_1 \in \{0.1, 0.01, \dots\},\quad x_2 \in \{10^3, 10^4, \dots\}$$

4. Feature Scaling

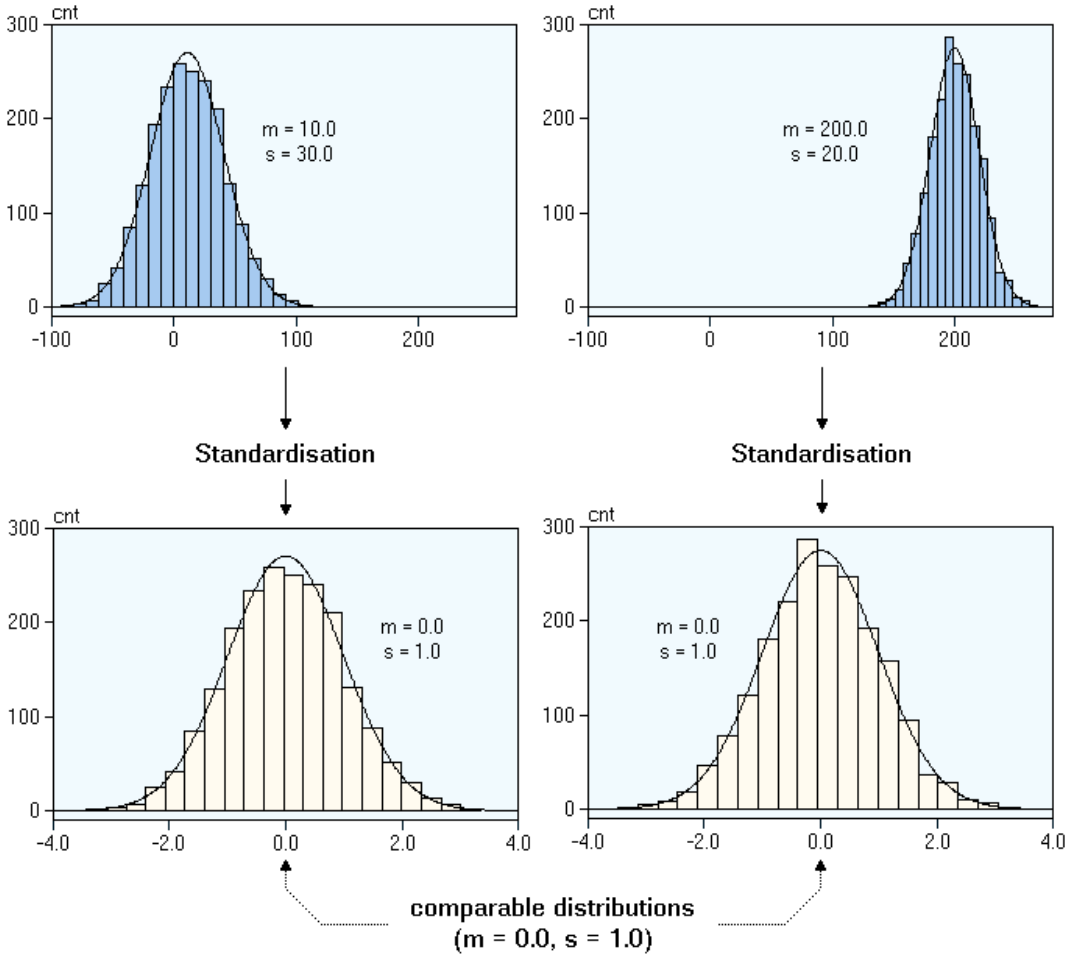
Example: After applying PCA, the dataset are more separable when features are scaled.



Without Scaling

With Scaling

4. Standardization



- Standardizing means subtracting a measure of location (mean) and dividing by a measure of scale (standard deviation). If a random variable follows Gaussian distribution, we obtain "Standard Normal" random variable with mean **0** and standard deviation 1.
- Standardizing replaces the values of a random variable by their z-score:

$$x_{new} = \frac{x - mean(x)}{std(x)}$$

4. Standardization

Example: Apply Standization with the following two Gaussian variables.

x_1	x_2
39	192
8	229
-12	187
1	187
0	223
-38	219
-37	174

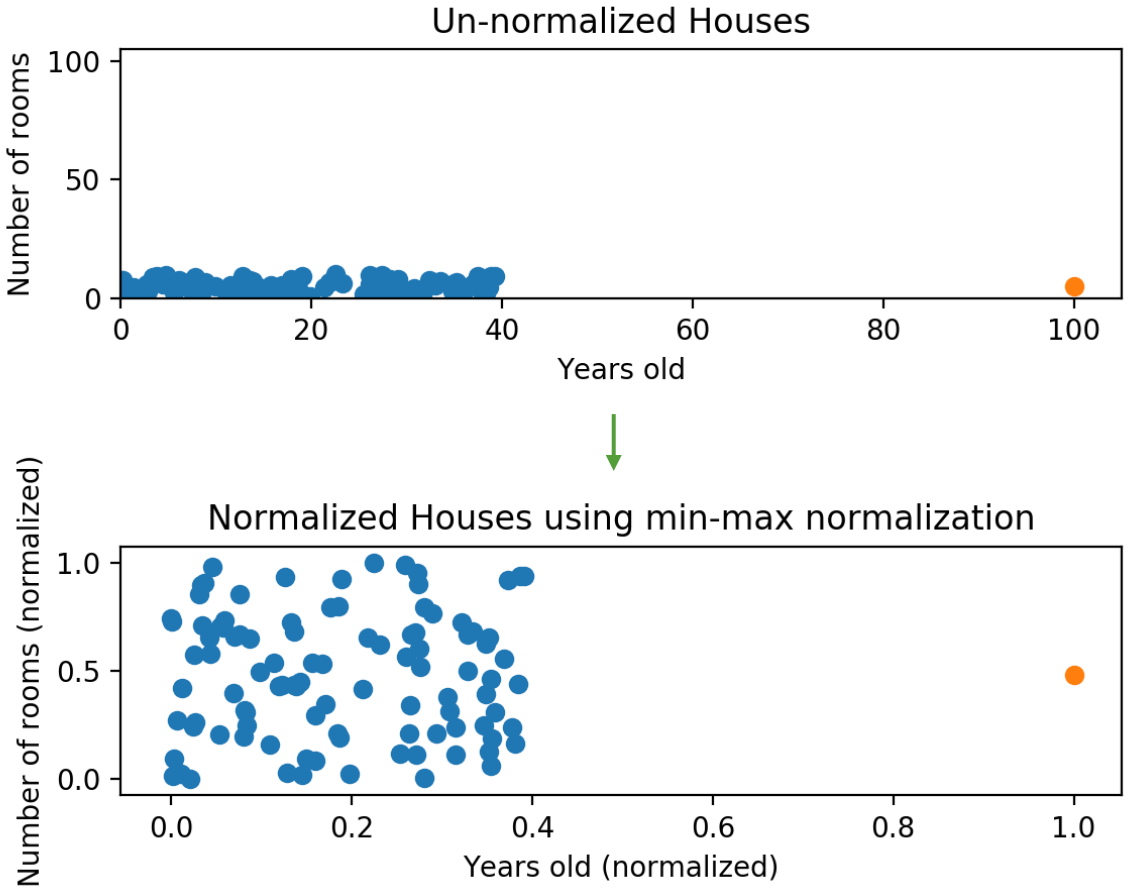
- $mean(x_1) = 10,\ std(x_1) = 30$
- $mean(x_2) = 200,\ std(x_2) = 20$

4. Standardization

- Compute Z-score of each variable

x_1	x_2	$Z\text{-}score(x_1)$	$Z\text{-}score(x_2)$
39	192	1.79	-0.47
8	229	0.54	1.37
-12	187	-0.25	-0.72
1	187	0.26	-0.72
0	223	0.22	1.07
-38	219	-1.30	0.87
-37	174	-1.26	-1.38

4. Min-Max Normalization



- A type of normalization which scales a variable by its minimum and range, to make all the elements lie between **0** and **1**, thus bringing all the values of numeric variables to a common scale.
- It is done by replacing all the values of x by:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4. Min-Max Normalization

Example: Apply Min-Max normalization with the following data.

x_1	x_2
39	192
8	229
-12	187
1	187
0	223
-38	219
-37	174

4. Min-Max Normalization

- **Step 1**: Find min and max value of each data

x_1	x_2
39	192
8	229
-12	187
1	187
0	223
-38	219
-37	174

- $min(x_1) = -38,\ max(x_1) = 39$
- $min(x_2) = 174,\ max(x_2) = 229$

4. Min-Max Normalization

- **Step 2**: Compute new values of each variable by using min-max formula

x_1	x_2	x_{new}^1	x_{new}^2
39	192	1	0.32
8	229	0.59	1
-12	187	0.33	0.23
1	187	0.50	0.23
0	223	0.49	0.89
-38	219	0	0.81
-37	174	0.01	0

5. Feature Encoding

- It is very common to see categorical features in a dataset. However most of learning algorithms can only deal with numerical values. Thus, it is essential to encode categorical features into numerical values.
- There are many techniques to encode categorical variable based on its characteristics:
  - Ordinal/Label Encoding
  - One-hot Encoding
  - Binary Encoding
  - Hashing Encoding
  - Etc.

5. Label Encoding

- **Label/Ordinal Encoding** is used only with categorical data that have a natural ordered relationship between each value (e.g. ‘Size’ with values such as ‘small’, ‘medium’, and ‘large’).
- The encoding is done by assigning the discrete number to the values based on their order.

E.g. X = [”small”, ”medium”, ”large”] → [0, 1, 2]

5. One-Hot Encoding

- A common way when there is no natural ordering between the values of a categorical variable.
- For each value of a categorical variable, a **dummy variable** is created where the value is **1** if for that instance the original value takes that value and **0** otherwise.

E.g. X = [”Cambodia”, ”Thai”, ”Lao”]

Dummy Variables

City	City_Cambodia	City_Thai	City_Lao	Encoding
Cambodia	1	0	0	[1, 0, 0]
Thai	0	1	0	[0, 1, 0]
Lao	0	0	1	[0, 0, 1]

6. Feature Correlation

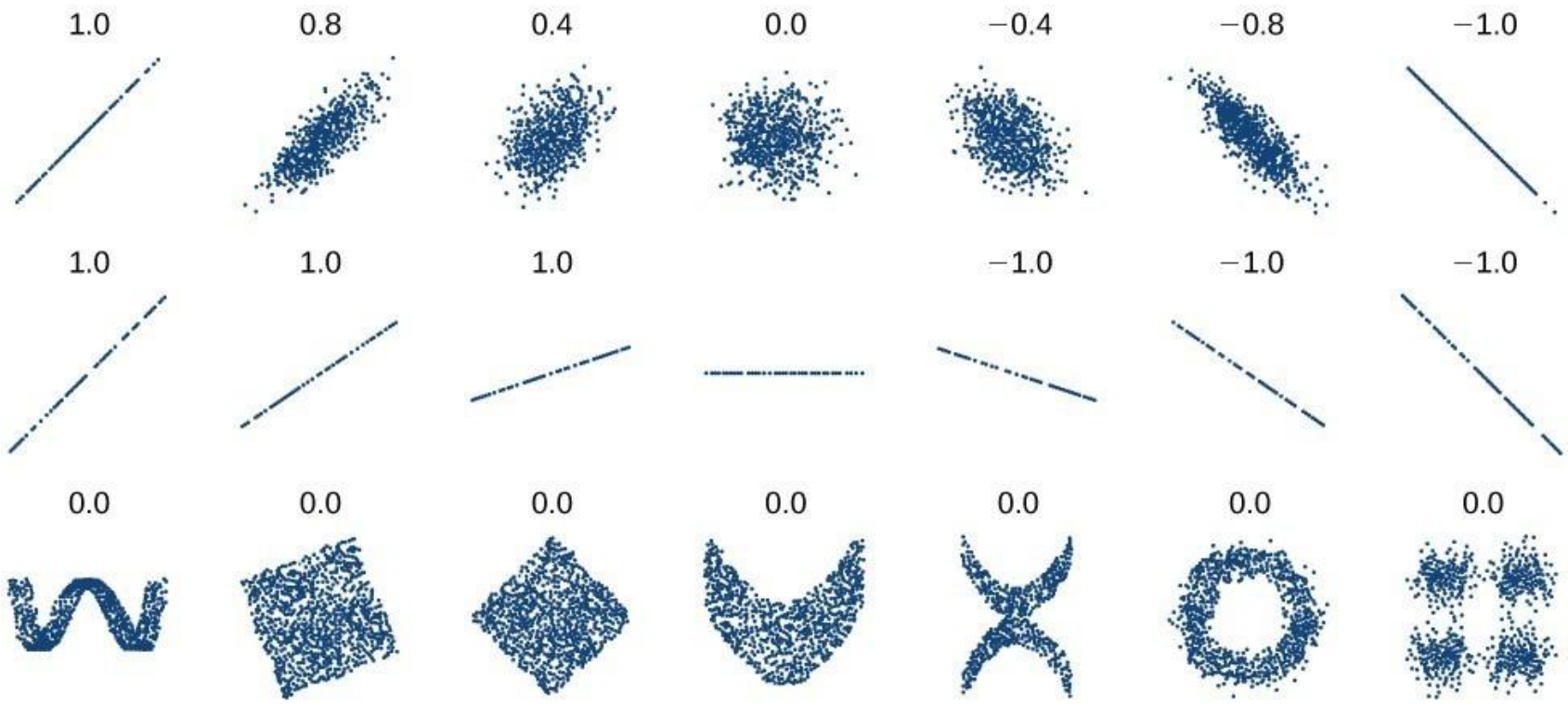
- Correlation is a statistical term which in common usage refers to how close two variables are to having a **linear relationship/dependency** with each other.
- It is used in "**Feature Selection**" to identify which input features are qualified in contribution to a learning algorithm in predicting the output. E.g. Input features with high correlation are more likely to have the same effect on the output. So, we can drop one of them.
- Common feature correlation methods:
  - Pearson Correlation, Spearman Correlation
  - Chi-Square Test
  - Anova Test

6. Pearson Correlation

- In statistics, Pearson correlation coefficient is a measure of linear correlation between two numerical variables x and y.
- According to **Cauchy-Schwarz**, it has a value between **+1** and **−1**, where 1 is perfect positive linear correlation, 0 is no linear correlation, and −1 is perfect negative linear correlation.
- Given n paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, Pearson correlation coefficient r_{xy} is defined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

6. Pearson Correlation



6. Pearson Correlation

Example: Given the records of a shop about daily temperature and sale as following, determine the correlation between temperature and sale using Pearson correlation coefficient.

Temp C	Sales
14.2	$215
16.4	$325
11.9	$185
15.2	$332
18.5	$406

6. Pearson Correlation

Denote:

x : Temperature of the day in degree Celsius

y : Amount of sale in dollar each day

6. Pearson Correlation

Step 1: Find the mean of x and y

x	y
14.2	215
16.4	325
11.9	185
15.2	332
18.5	406
$\bar{x}$ =15.24	$\bar{y}$ = 292.6

6. Pearson Correlation

Step 2: Compute $x - \bar{x},\ y - \bar{y}$

x	y	$x - \bar{x}$	$y - \bar{y}$
14.2	215	-1.04	-77.6
16.4	325	16.4	325
11.9	185	11.9	185
15.2	332	15.2	332
18.5	406	18.5	406
$\bar{x}$ =15.24	$\bar{y}$ = 292.6		

6. Pearson Correlation

Step 3: Compute $(x-\bar{x})(y-\bar{y})$, $(x-\bar{x})^2$, $(y-\bar{y})^2$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
14.2	215	-1.04	-77.6	80.704	1.0816	6021.76
16.4	325	1.16	32.4	37.584	1.3456	1049.76
11.9	185	-3.34	-107.6	359.384	11.1556	11577.76
15.2	332	-0.04	39.4	-1.576	0.0016	1552.36
18.5	406	3.26	113.4	369.684	10.6276	12859.56
$\bar{x}$ =15.24	$\bar{y}$ = 292.6					

6. Pearson Correlation

Step 4: Compute $\sum(x-\bar{x})(y-\bar{y})$, $\sum(x-\bar{x})^2$, $\sum(y-\bar{y})^2$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
14.2	215	-1.04	-77.6	80.704	1.0816	6021.76
16.4	325	1.16	32.4	37.584	1.3456	1049.76
11.9	185	-3.34	-107.6	359.384	11.1556	11577.76
15.2	332	-0.04	39.4	-1.576	0.0016	1552.36
18.5	406	3.26	113.4	369.684	10.6276	12859.56
$\bar{x}$ =15.24	$\bar{y}$ = 292.6			845.78	24.212	33061.2

6. Pearson Correlation

Step 5: Compute correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{845.78}{\sqrt{24.212} \times \sqrt{33061.2}} = 0.9453$$

$$r_{xy} = 0.9453$$

Based on the value $r_{xy} = 0.9453$, the two variables, temperature and sale, have strong positive linear relationship.