Data Curation Techniques

Siddharth R

Data Quality Challenges

- Incomplete data
- Inaccurate / Noisy data
- Inconsistent data
- Believability / Interpretability
- Perspective of data quality - Marketing Manager vs Sales Manager view of incorrect address

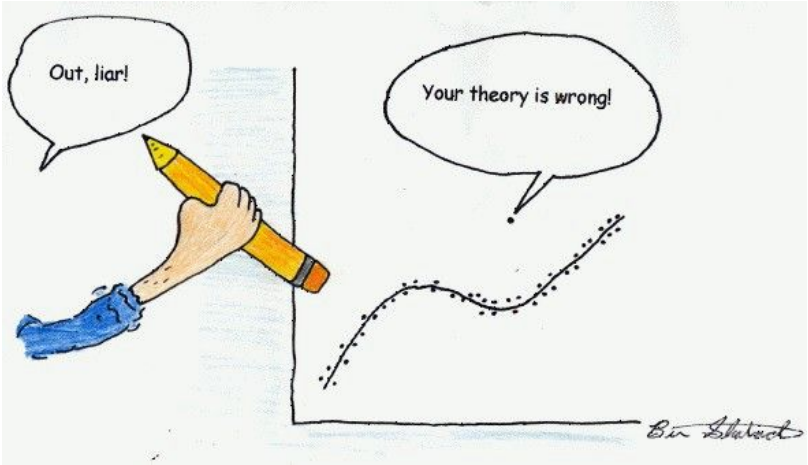
Data Cleaning as a Process

- Discrepancy detection
  - Poorly designed data entry forms
  - Human error in data entry
  - Deliberate errors
  - Data decay
  - Inconsistency in data representation and integration
  -
- Meta-data : data about data
- Check for Unique Rule, Consecutive Rule and Null Rule

Data Cleaning: Handling Missing Data

How do you handle missing values?

Handling Outliers / Noise

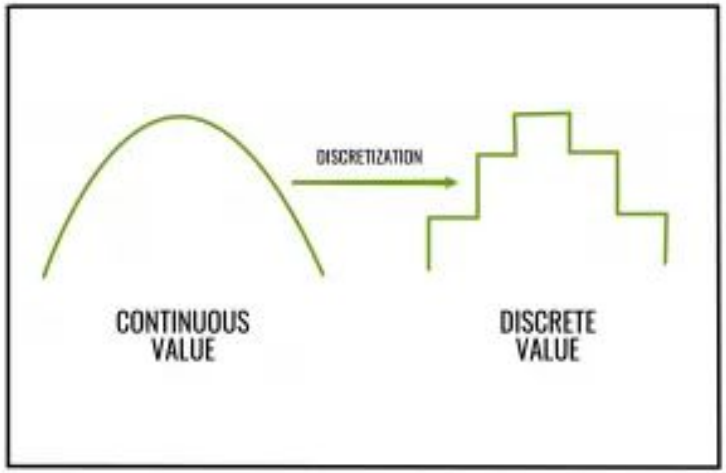


- An outlier is a data point that significantly differs from other observations in a dataset.
- What could be the causes of an outlier?
- Types of outliers:
  - Univariate vs Multivariate
  - Natural vs Artificial
  - Global vs Contextual vs Collective
- Detecting outliers:
  - Data discretization / Binning
  - Z-Score
  - IQR
  - Visualization

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

Data Discretization

- Also called as “Binning”
- The process of grouping continuous numerical data into discrete intervals or "bins."
- Advantages of binning
  - Smooths out noise
  - Better interpretability
  - Handling outliers
- Types of binning :
  - Equal width binning
  - Equal frequency binning



Equal Width vs Equal Frequency

Equal-Width Binning

- Each bin has an equal range (or width) of values.
- The range of values in each bin is calculated as:

$$\text{Bin Width} = (\text{Max Value} - \text{Min Value}) / \text{Number of Bins}$$

Use Case: When you want uniform bins irrespective of the data distribution.

Example - Equal Width Binning

Dataset: `[5, 12, 15, 18, 22, 26, 30, 35, 40]`

Number of Bins: 3

Step 1: Sort the data

Step 2: Find bin width -> bin width = (40-5) / 3 = 11.667 (Approximately 12)

Step 3: Create Bins : Bin 1 : [5 - 16] = [5,12,15]

Bin 2: [17-28] = [18,22,26]

Bin 3: [29-40] = [30,35,40]

Try for dataset with

Skewed distribution : [5, 10, 15, 20, 25, 30, 35, 40, 50, 100] , Number of bins = 3

Extreme outlier: [10, 15, 20, 25, 30, 1000] , Number of bins = 3

Equal Frequency Binning

- Equal frequency binning divides the data into intervals that contain approximately the same number of data points.

Use Case: When you want balanced bins, especially for skewed distributions.

Steps:

1. Sort the data
2. Determine the Number of Data Points per Bin
3. Start binning

Example: Equal Frequency Binning

Data in sorted order : 5,10,15,20,25,30,35,40,50,100

Number of bins : 3

Data per bins : 10 / 3 = 3.333 (2 bins with 3 data points and 1 bin with 4 data points)

- Bin 1: [5, 10, 15]
- Bin 2: [20, 25, 30]
- Bin 3: [35, 40, 50, 100]

Smoothing using binning

Data : 4, 7, 13, 16, 20, 24, 27, 29, 31, 33, 38, 42.

Equal Frequency :
BIN 1: 4, 7, 13, 16
BIN 2: 20, 24, 27, 29
BIN 3: 31, 33, 38, 42

Smoothing by Bin Mean
BIN 1 : 10, 10, 10, 10 (mean: 10)
BIN 2 : 25, 25, 25, 25 (mean: 25)
BIN 3 : 36, 36, 36, 36 (mean: 36)

Smoothing by Bin Boundaries

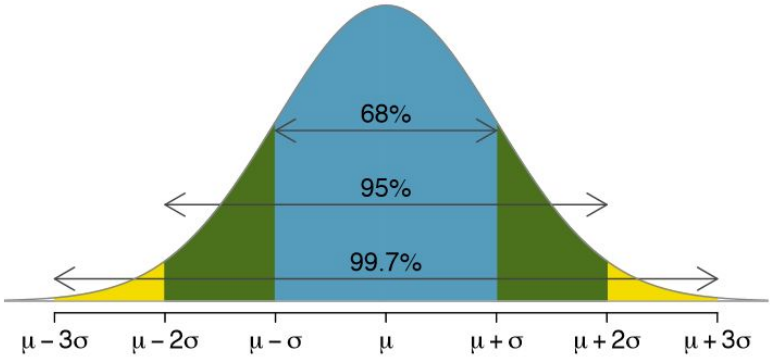
BIN 1: 4, 4, 16, 16 (min: 4 and max: 16)
BIN 2: 20, 20, 29, 29 (min: 20 and max: 29)
BIN 3: 31, 31, 42, 42 (min: 31 and max: 42)

Z-Score

- The Z-score is a statistical measure that indicates how many standard deviations a data point is from the mean of the dataset.

  Z-Score = (x - mean) / standard deviation

- If the z score of a data point is more than ± 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.



Example

Data = [10,12,13,14,15,16,17,18,19,20,100]

- Mean = ?
- Standard deviation = ?
- Z - score of all values
  - For 10, it is (10 - Mean) / Standard Deviation = ?
  - For 100, - ?

IQR

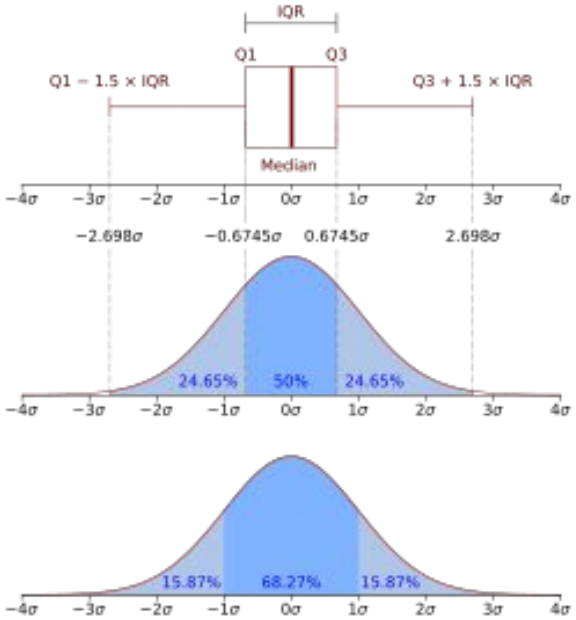
- IQR measures the middle 50% of the data.
- Outliers are points that fall below the first quartile Q1 or above the third quartile Q3 by 1.5 times the IQR.

  IQR = Q3 - Q1

  Lower bound = Q1 - 1.5 * IQR

  Upper bound = Q3 + 1.5 * IQR

- Percentage vs Percentile?



Visualization

Box Plot:

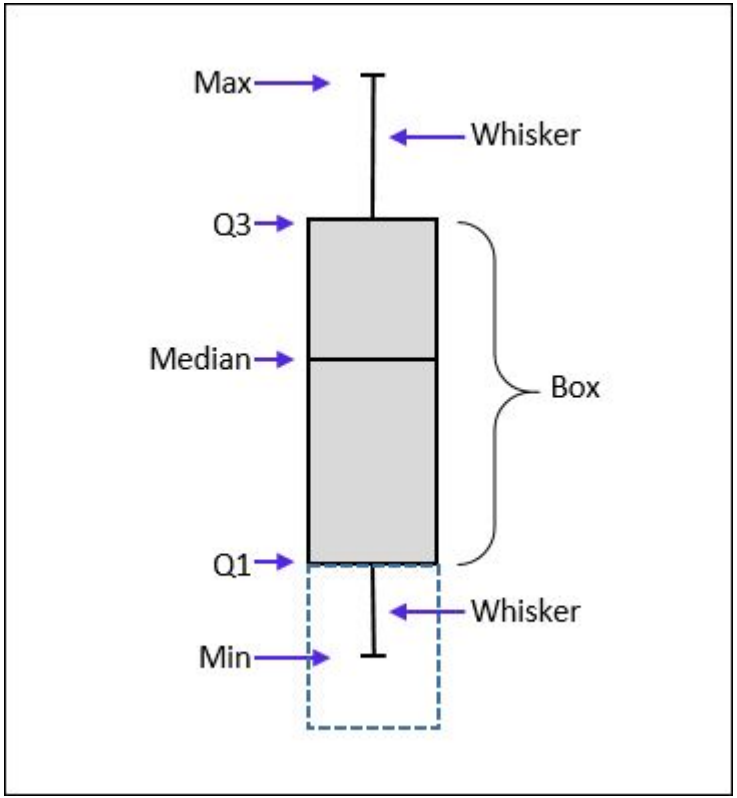
4 Components:

1. Median (Q2, 50th percentile)
2. IQR Box (Q1 to Q3)
3. Whiskers (Min & Max within 1.5×IQR range)
4. Outliers (Dots beyond whiskers)

- Histogram and Scatter plot

height = [150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 130]

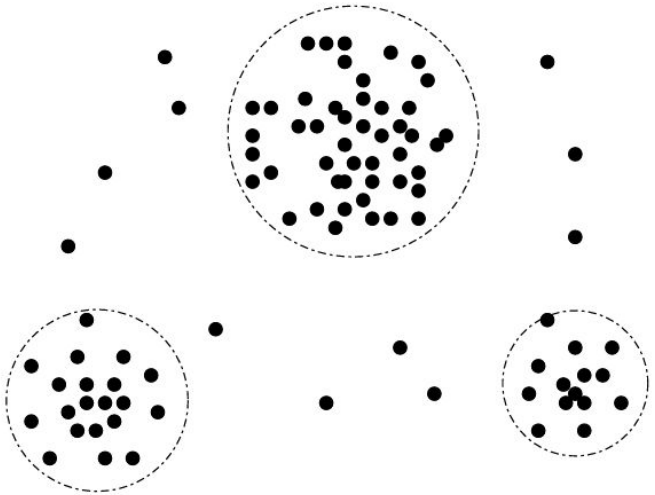
weight = [50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 120]



Clustering Approach

Clustering:

- Similar values are organized into groups, or “clusters.”
- Values that fall outside of the set of clusters may be considered outliers



Handling Duplicates

- Identifying Duplicates - duplicated()
  - sum()
  - subset
  - keep
- Removing Duplicates - drop_duplicates()
  - subset
  - keep
  - inplace
- Replacing Duplicates
  - replace()

Thank you