Data Pre-Processing-I

(Introduction, Need, Data Cleaning)
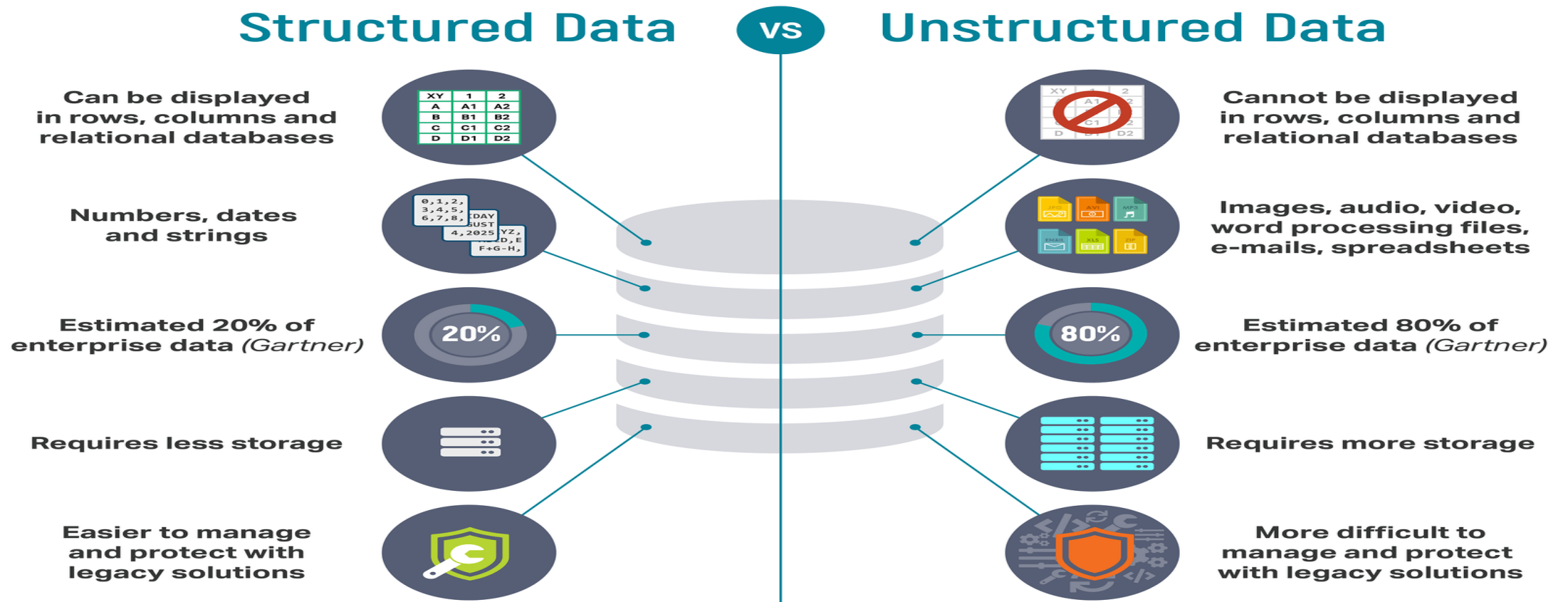
TIET, PATIALA

Data

- Data is a unprocessed fact, value, text, sound or picture that is not being interpreted and analyzed.

- Data is the most important part of all Data Analytics, Machine Learning, Artificial Intelligence.

- Big Enterprises are spending lots of money just to gather as much certain data as possible.

In 2021, Facebook acquire WhatsApp by paying a huge price of $19 billion

Structured vs. Unstructured Data in ML

Structured Data **vs** **Unstructured Data**

Can be displayed in rows, columns and relational databases

Numbers, dates and strings

Estimated 20% of enterprise data *(Gartner)*

Requires less storage

Easier to manage and protect with legacy solutions

Cannot be displayed in rows, columns and relational databases

Images, audio, video, word processing files, e-mails, spreadsheets

Estimated 80% of enterprise data *(Gartner)*

Requires more storage

More difficult to manage and protect with legacy solutions

Structured Data in ML

- Structured data in Machine Learning is stored in the form of rows and columns.

- Each instance of structured data represents a feature/attribute.

- For instance a well known ML dataset, *Iris,* has five features about species namely Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species.

Types of Features

Quantitative (Numerical)	Binary	Qualitative Nominal (Categorical)
• Quantitative data being measured. • Can be **Continuous** (infinite values- length, mass, weight) or **Discrete** (finite integer values- no. of workers absent)	• Which have two values (0/1, yes/no, true/false) • Examples- Marital Status, Permanent Employee, etc.	• Categorical data where order of categories is arbitrary • Example- account type (savings, current, fixed term, etc).

Types of Features Contd....

Qualitative Ordinal (Ranked)	Interval	Ratio
• Categorical data where there is some logical ordering of categories • Example: Size (S, M, L, XL, XXL, etc.), Likert Scale (Strongly Disagree, Disagree, Neutral, Agree, etc.)	• Has meaningful intervals between measurement. • No true starting point (zero) • Example- Temperature	• Have highest level of measurement. • Ratios between measurements and intervals are meaningful because there is true starting point (zero) • Example: weight, Age

Data Pre-Processing

▪ **Data Pre-processing:** It is that phase of any Machine Learning process, which transforms, or Encodes, the data to bring it to such a state where it can be easily interpreted by the learning algorithm.

"Data pre-processing is not a single standalone entity but a collection of multiple interrelated tasks"

"Collectively data pre-processing constitutes majority of the effort in machine learning process (approx. 90 %)"

Data Pre-Processing

- **Data Pre-processing:** It is that phase of any Machine Learning process, which transforms, or Encodes, the data to bring it to such a state where it can be easily interpreted by the learning algorithm.

"Data pre-processing is not a single standalone entity but a collection of multiple interrelated tasks"

"Collectively data pre-processing constitutes majority of the effort in machine learning process (approx. 90 %)"

Need of Data Pre-Processing

➢ Data in the real world is "quite messy"

- **incomplete**: missing feature values, absence of certain crucial feature, or containing only aggregate data.

  ▪ e.g. Height=" "

- **noisy**: containing errors or outliers

  ▪ e.g. Weight="5000" or "-60"

- **inconsistent**: containing discrepancies in feature values.

  ▪ e.g. Age="20" and dob="12 july 1990"

  ▪ e.g. contradictions between duplicate records

Need for data Pre-processing

➢ **Unstructured Data (Text)**
- Lower Case
- Normalization (remove punctuation, special symbols, urls)
- Stopwords Removal (of, and, the,....)
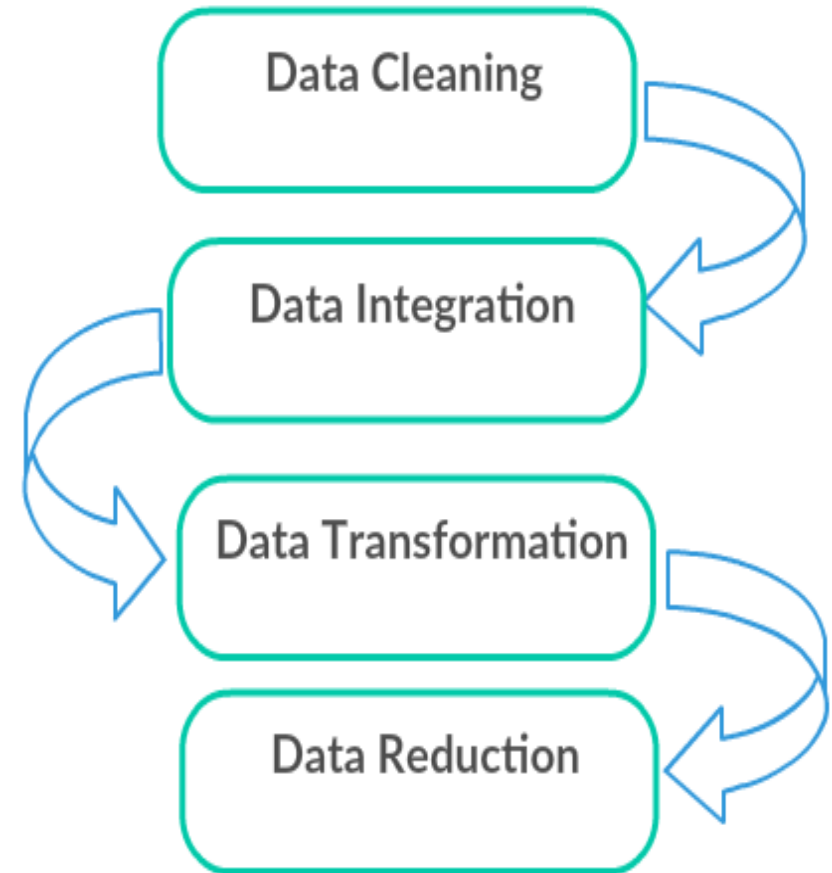- Stemming/Lemmatization (plays, playing, played → play)

➢ **Unstructured Data (Images)**
- Read image
- Resize image
- Remove noise(Denoise)
- Segmentation
- Morphology(smoothing edges)

Pre- Processing in Structured Data

➢ Major data pre-processing tasks

- Data cleaning

- Data integration

- Data transformation

- Data reduction

Data Cleaning

- **Data cleaning:** It is a procedure to "clean" the data by filling in missing values, smoothening noisy data, identifying or removing outliers, and resolving data inconsistencies.

- Data cleaning tasks
  - Fill missing values
  - Noise smoothening and outlier detection
  - Resolving inconsistencies

Data Cleaning- Missing Values

Missing values: data values are not available.

i.e. many data entities have no data values corresponding to a certain feature like BMI value missing for some persons in a diabetes dataset.

➤ Probable reasons for missing values:

- faulty measuring equipment
- reluctance of person to share certain detail
- negligence on part of data entry operator
- feature unimportance at time of data collection

Data Cleaning- Missing Values Contd...

➢ **Missing data handling techniques**

- Removing the data entity
- Manually filling the values
- Imputation (process used to determine and assign replacement values for missing, invalid, or inconsistent data)

"Technique selection is specific to user's preference, dataset or feature type or problem set"

Data Cleaning- Missing Values Contd...

➤Sample dataset related to forest fires

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4

Data Cleaning- Missing Values Contd…

Removing the data entity: Most easiest way directly to clean the data, but this is usually discouraged as it leads to loss of data, as you are removing the data entity or feature values that can add value to data set as well.

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4
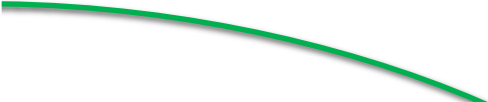
Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
mar	89.3	102.2	11.4	99	1.8
aug	91.5	608.2	8	86	2.2
sep	92.5	698.6	22.8	40	4

Data Cleaning- Missing Values Contd...

◦Manually filing up of values : This approach is time consuming, and not recommended for huge data sets.

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	17	33	0.8
mar	91.6	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	380	22.2	92	1.8
aug	90	495.6	24.1	27	2
aug	91.5	608.2	8	86	2.2
sep	91	692.6	22	63	5.4
sep	92.5	698.6	22.8	40	4

Data Cleaning- Missing Values Contd...

◦ Imputation : process used to determine and assign replacement values for missing, invalid, or inconsistent data. Various imputation methods include:

➢ Central Tendency Imputation

➢ Hot Deck Imputation

➢ Cold Deck Imputation

➢ Model Based Imputation

  ➢ Nearest Neighbor Imputation

  ➢ Tree-Based Imputation

Data Cleaning (Missing values)- Contd…

Central tendency Imputation : Replacing the missing value by central tendency (mean, median, mode) for a feature vector or belonging to same class of feature vector.

Mean

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Median

$$md = x_{\frac{(n-1)}{2}} \quad \text{for n is odd}$$

$$md = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) \quad \text{for n is even}$$

Mode : Mode is the most frequent value corresponding to a certain feature in a given data set

Data Cleaning- Missing Values Contd…

◦ Replacing Mean Value:

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	15.3	33	3.57
mar	90.5	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	458.3	22.2	58.7	3.57
aug	90.5	495.6	24.1	27	3.57
aug	91.5	608.2	8	86	2.2
sep	91	692.6	15.3	63	5.4
sep	92.5	698.6	22.8	40	4

"

Data Cleaning- Missing Values Contd…

◦ " Replacing by mean value: Not a suitable method if data set has many outliers"

◦For example: weighs of humans   67, 78, 900,-56,389,-1 etc.

Outlier

Mean is 229.5

o   Can be replaced with median in such cases.

o   "Mode is a good option for missing values in case of categorical variables"

Data Cleaning- Missing Values Contd...

Hot Deck Imputation

- Computes how many number of features (other than feature with missing data) have same values in the entire training examples and choose it for replacement.
- Used mostly in categorical data. (Clustering)

Cold Deck Imputation

- Similar to hot deck imputation.
- In it missing observations are replaced by values from a source unrelated to the data set under consideration. (Previous Study)

Data Cleaning- Missing Values Contd…

Nearest Neighbor-Based Imputation

- Rely on distance metrics
- evaluate the distance between recipients and donors.
- Used after converting all features to numerical (quantitative)

Data Cleaning- Noisy Data

- **Noise** is defined as a random variance in a measured variable.
- For numeric values, boxplots and scatter plots can be used to identify outliers.
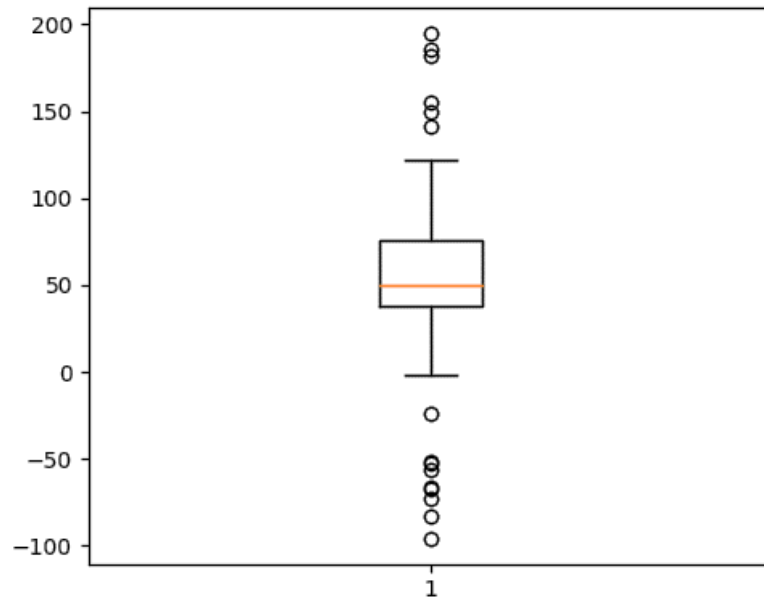
Boxplot
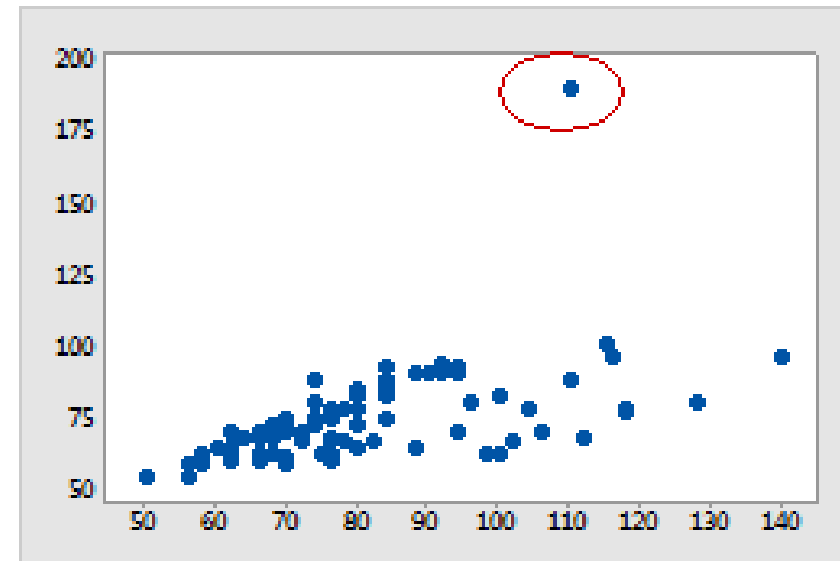
Scatter plot

Data Cleaning- Noisy Data

➤ Major reasons of random variations in data are:

- Malfunctioning of collection instruments.
- Data entry lags.
- Data transmission problems

To deal with these anomalous values, data smoothing techniques are applied, some of the popular ones are

- Binning method
- Regression
- Outlier analysis

Binning Method for Noisy Data

Binning method : performs the task of data smoothening.

Steps to be followed under binning method are:

Step 1: Sort the data into ascending order.

Step 2: Calculate the bin size (i.e. number of bins)

Step 3: Partition or distribute the data equally among the bins starting with first element of sorted data.

Step 4: perform data smoothening using bin means, bin boundaries, and bin median.

Last bin can have one less or more element!!

Binning Method for Noisy Data

Example : 9, 21, 29, 28, 4, 21, 8, 24, 26

Step1: sorted the data 4, 8, 9, 21, 21, 24, 26, 28, 29

Step 2 : Bin size calculation

$$\text{Bin size} = \frac{Max\ value - Min\ value}{data\ size}$$

$$= \frac{29\text{-}4}{9} = 2.777$$

But we need to take ceiling value, so bin size is 3 here

Binning Method for Noisy Data

- Step 3 : Bin partitioning (equi-size bins)

  Bin 1: 4, 8, 9
  Bin 2: 21, 21, 24
  Bin 3: 26, 28, 29

  Step 4 : data smoothening

- Using mean value : replace the bin values by bin average

  Bin 1: 7, 7, 7
  Bin 2: 22, 22, 22
  Bin 3: 27, 27, 27

Binning Method for Noisy Data

- ➢ Using boundary values : replace the bin value by a closest boundary value of the corresponding bin.

Bin 1: 4, 9, 9

Bin 2: 21, 21, 24

Bin 3: 26, 29, 29

"Boundary values remain unchanged in boundary method"

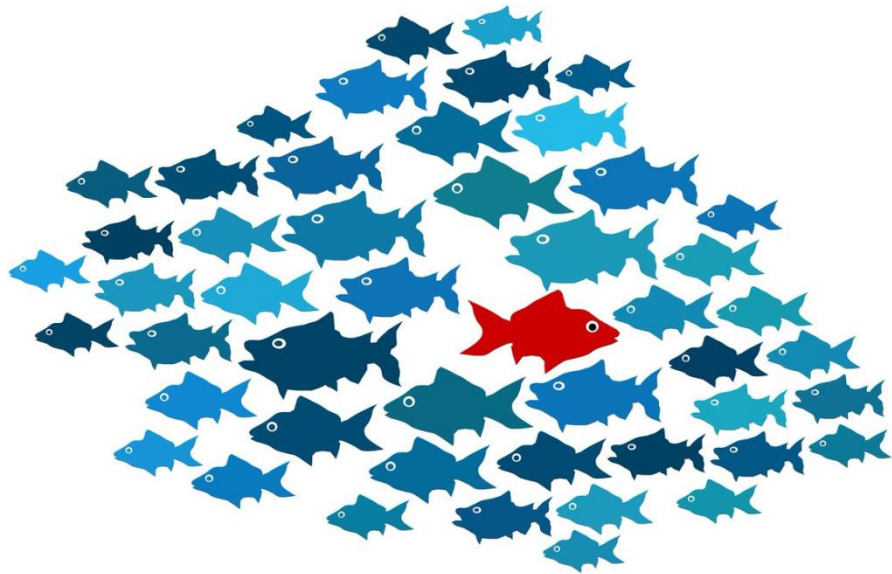- ➢ Using median values : replace the bin value by a bin median.

Bin 1: 8, 8, 8

Bin 2: 21, 21, 21

Bin 3: 28, 28, 28

Outlier Analysis

- An **outlier** is an object that deviates significantly from the rest of the objects.
- They can be caused by measurement or execution error.
- The analysis of outlier data is referred to as outlier analysis or outlier mining.
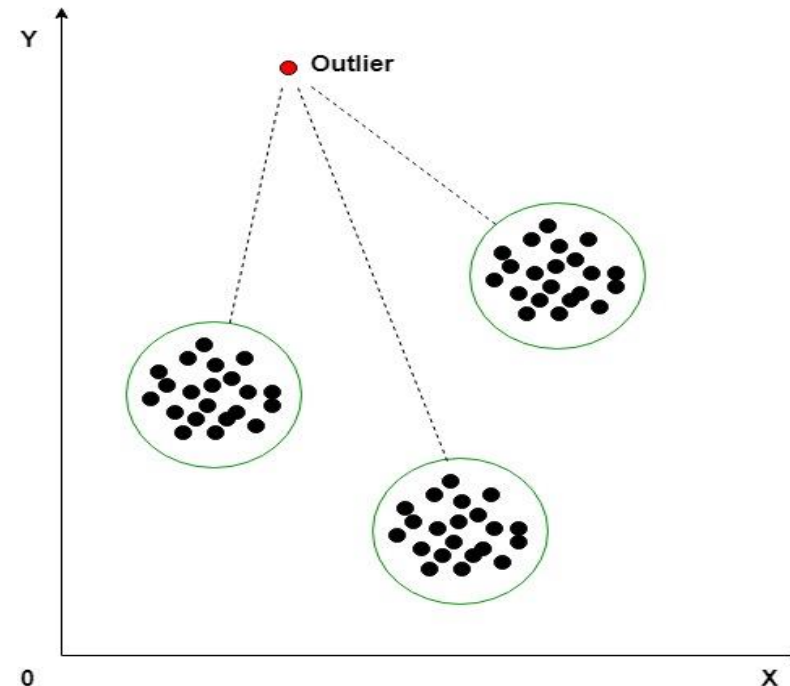
Outlier Analysis

- Types of Outliers:
  - ➤ Univariate: A univariate outlier is a data point that consists of an extreme value on one variable.
  - ➤ Multivariate: A multivariate outlier is a combination of unusual scores on at least two variables.

- Outlier Detection and Handling Methods
  - ➤ Extreme Value Analysis
  - ➤ Linear Models
  - ➤ Proximity-based Methods
  - ➤ Information Theoretic Methods

Extreme Value Analysis- Outlier Analysis

▪**Numeric Outlier**

▪This is the simplest, nonparametric outlier detection method in a one dimensional feature space.

▪Here outliers are calculated by means of the *IQR* (InterQuartile Range).

▪The first and the third quartile (*Q1, Q3*) are calculated.

▪An outlier is then a data point x_i that lies outside the interquartile range. That is:

$$x_i > Q_3 + k(IQR) \; and \; x_i < Q_1 - k(IQR)$$
where IQR=Q3-Q1 and k >=0

Assume the data 6, 2, 1, 5, 4, 3, 50. If these values represent the number of chapatis eaten in lunch, then 50 is clearly an outlier.

Sorted Values: 1, 2, 3, 4, 5, 6, 50

Q1 25 percentile of the given data is, 2

Q2 50 percentile of the given data is, 4.0

Q3 75 percentile of the given data is, 6

IQR= 6-2=4, k=1.5

Range: 6-1.5 X 4=0 and 6+1.5X4=12

50 is Outlier

Extreme Value Analysis- Outlier Analysis

▪**Z-score** is a parametric outlier detection method in a one or low dimensional feature space.

▪This technique assumes a Gaussian distribution of the data.

▪The outliers are the data points that are in the tails of the distribution and therefore far from the mean.

▪How far depends on a set threshold z_{thr} for the normalized data points z_i calculated with the formula:

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

where x_i is a data point, μ is the mean of all x_i and is the standard deviation of all x_i.

An outlier is then a normalized data point which has an absolute value greater than z_{thr}.

$$|z_i| > z_{thr}$$

Commonly used z_{thr} values are 2.5, 3.0 and 3.5.

Outlier Analysis

- **Linear Models**:
  - ➢Projection methods that model the data into lower dimensions using linear correlations.
  - ➢For example, principle component analysis and data with large residual errors may be outliers.

- **Proximity-based Models**:
  - ➢Data instances that are isolated from the mass of the data as determined by cluster, density or nearest neighbor analysis.

- **Information Theoretic Models**:
  - ➢Outliers are detected as data instances that increase the complexity (minimum code length) of the dataset.

Regression for Noisy Data

- Regression method : Linear regression and multiple linear regression can be used to smoothen the data, where the values are conformed to a function.
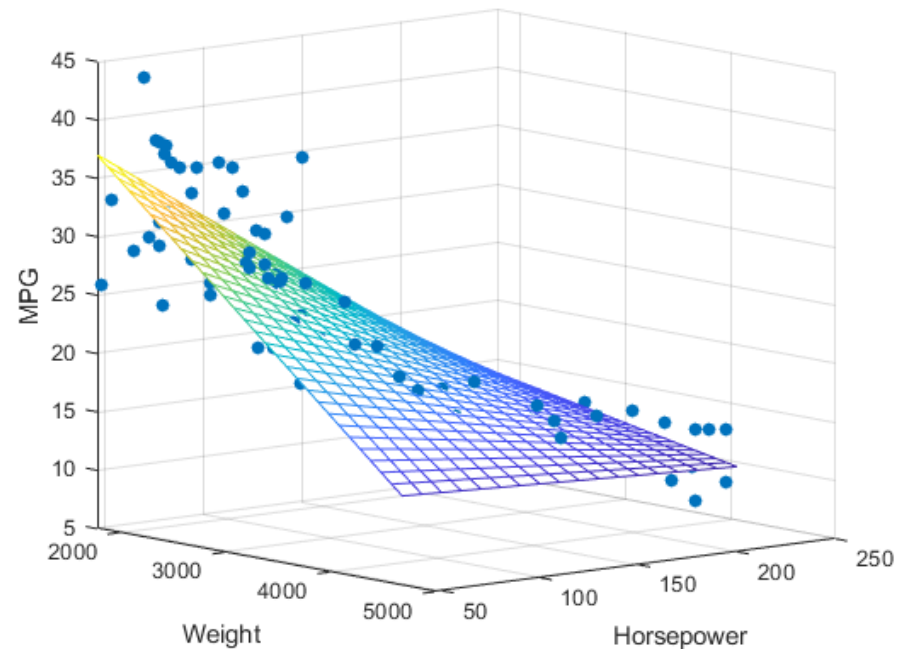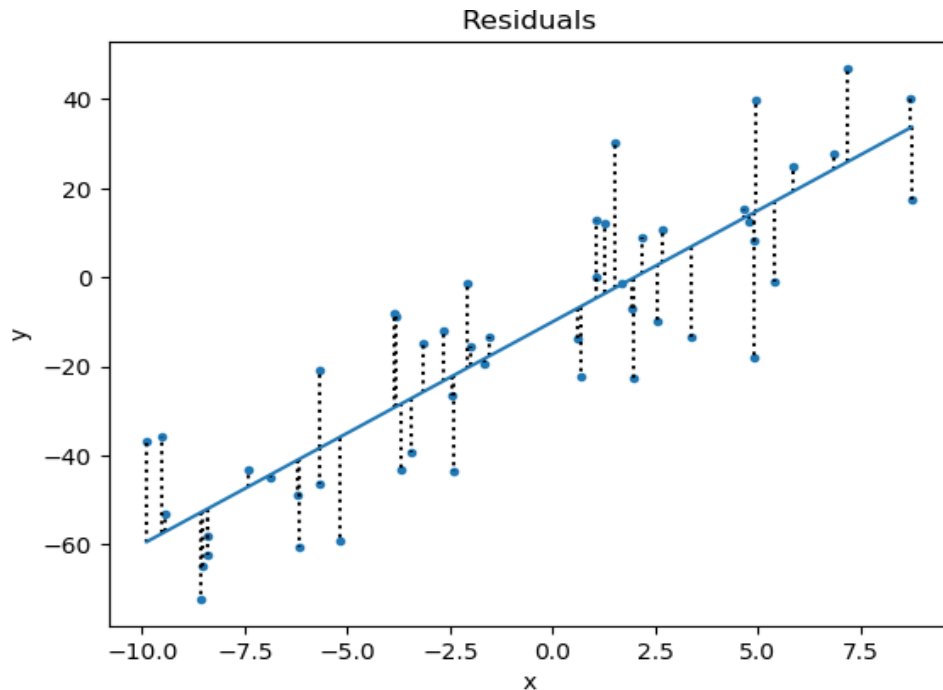
Data Cleaning – Inconsistent Data

Inconsistent Data: discrepancies between different data items.

e.g. the "Address" field contains the "Phone number"

To resolve inconsistencies

➢ Manual correction using external references
➢ Semi-automatic tools
  • To detect violation of known functional dependencies and data constraints
  • To correct redundant data

To avoid inconsistencies, perform data assessment like

knowing what the data type of the features should be and whether it is the same for all the data objects."

Data Cleaning-Summary