Day-7

11 August 2024 22:48

1. Conceptual aspects
2. Explanation
3. Examples/ Use cases
4. Python code demo

#	Topic	Number of Use cases/ Examples discussed	Explained Y/N, comments if any
1	Recap on data and features	Duplicates Data quality Best practices shared	
2	Missing values	Why we need to handle missing values Nature of NA Python coding for simple missing values (CA housing)	
3	Outlier detection (col level)	IQR Z-score Box plots 1.5 factor with 3 sigma rule	
4.	Encoding methods	Find and replace Label encoding OHE Frequency (freq, target)	

Data table

In the context of machine learning, a data table, also known as a dataset or data matrix, refers to a structured arrangement of data

rows typically represent individual instances or observations,

columns represent attributes, features, or variables associated with those instances.

A data table is a fundamental building block in machine learning

columns represent attributes such as age, gender, height, weight

each row represents a person

ID	Age	Gender	Height	Weight	Class
1	25	Male	178	75	A
2	30	Female	162	58	B
3	22	Male	185	82	A
4	28	Female	170	63	B

class label indicating the group the person belongs to

names

Record

Record, samples, point, case, entity, instance, entry, Objects
Data points, Document, tuple, Transaction, feature vector

Columns

Attributes, Features, Variables
Field, Predictors, characteristics

Data science a study of 3 or 4 different disciplines. Hence there are a lot of vocab, often many for the same term!

FEATURE TYPES

- independent variable,
  - sometimes called an experimental or predictor variable,
  - is a variable that is being manipulated in an experiment in order to observe the effect …
- dependent variable, sometimes called an outcome/response/target variable.

Example

- Dependent Variable: Test Mark (measured from 0 to 100)
- Independent Variables:
  - Revision time (measured in hours),
  - Intelligence (measured using IQ score)

Sample dataset (temp forecasts)

year	month	day	week	Temp 2 days before	Temp 1 day before	Average temp	Actual temp on that day	Temp forecast by noaa	Temp Forecast by acc	Temp forecast by friend
2016	1	1	Fri	45	45	45.6	45	43	50	29
2016	1	2	Sat	44	45	45.7	44	41	50	61
2016	1	3	Sun	45	44	45.8	41	43	46	56
2016	1	4	Mon	44	41	45.9	40	44	48	53
2016	1	5	Tues	41	40	46	44	46	46	41
2016	1	6	Wed	40	44	46.1	51	43	49	40
2016	1	7	Thurs	44	51	46.2	45	45	49	38
2016	1	8	Fri	51	45	46.3	48	43	47	34
2016	1	9	Sat	45	48	46.4	50	46	50	47
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
2016	1	19	Tues	50	54	47.6	48	47	49	53
2016	1	20	Wed	54	48	47.7	52	44	52	61
2016	1	21	Thurs	48	52	47.8	52	43	51	57

Independent variables

Independent variables

Dependent variable

Feature types

Numerical Features: age, temperature, height, and income

Categorical Features: variables that represent different categories or labels.

Textual Features: such as documents, sentences, or paragraphs.

Date and Time Features represent temporal information.

Boolean Features: take on two possible values: True or False.

Geospatial Features: location-based data, such as latitude, longitude, postal codes, or addresses.

Image and Video Features: computer vision tasks.

Audio Features: speech recognition, music analysis, and more.

Derived Features: features are created by transforming or combining existing features.

Continuous variables - decimal
Discreate variables - countable

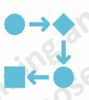
Categorical variables
- Nominal (no order,)
- Ordinal var (order, rank, but the intervals may not be the same across ranks)
- Binary (0,1)
  - Dichotomous

Basic sanity checks on the data
- Steps would vary from data type to data type (projects/ domain)
  - Numeric
  - Text Data
  - Image
  - Time Series data

- Scope of understanding (numeric)
  - General inspection
  - Handling duplicates
  - Missing data – imputation
  - Handling outliers/noise/novelty
  - Cardinality assessment
  - Encoding - dummy variables
  - Scaling , normalization

General inspection of data

conducting a preliminary assessment of your raw data to identify any initial issues, inconsistencies, or patterns that might affect your analysis.



gain an understanding of the data's quality, distribution, and potential challenges before you proceed with more detailed preprocessing steps.

General inspection serves as a starting point to determine the extent of data cleaning, transformation, and feature engineering required.

General inspection

General inspection

Data Overview

Data Size: Determine the number of rows and columns in your dataset.

Data Types: Identify the types of data in each column (numeric, categorical, text, etc.).

Basic Statistics: Calculate basic summary statistics like mean, median, and standard deviation for numerical columns.

Missing Data

Missing Values: Identify columns with missing data and assess the percentage of missing values for each column.

Data Distributions

Histograms: Create histograms to visualize the distribution of numerical variables.

Bar Plots: Plot bar charts to visualize the distribution of categorical variables.

General inspection

Data Quality

Domain Knowledge

Data Consistency: Look for inconsistent data formats or values that don't make sense.

Unique Values: Identify columns with a high number of unique values, which might indicate data quality issues.

Contextual Understanding: Apply domain knowledge to interpret any unusual patterns or inconsistencies in the data.

General inspection

Diligence	Comments
nominal or ordinal scale (where there are a fixed number of possible values)	inspect all possible values to uncover mistakes, duplications and inconsistencies. variable Company may include a number of different spellings for the same company such as "General Electric Company," "General Elec. Co.," "GE," "Gen. Electric Company," "General electric company," and "G.E. Company."
Numeric variables with inclusion of nonnumeric terms.	a variable generally consisting of numbers may include a value such as "above 50" or "out of range."

General inspection

Diligence	Comments
Timeliness of data	• how up-to-date the observations are and whether the quality is the same across different sources of data.
Data been collected over time	• changes related to the passing of time may no longer be relevant to the analysis. • *Cost of production field* - collected over many years, the rise in costs attributable to inflation may need to be considered for the analysis.

Subjective aspect.

- completeness of data.
- Timeliness ...
- specificity ... (project)
- relevance.

** imp + difficult.

Duplicate data

- **Duplicate** Records:
  - Duplicates can skew statistical analyses and lead to incorrect conclusions.
  - They can also result in **overestimations** or **underestimations** of certain patterns.

Handling duplicates

Customer Database	Clinical Trial Data	Social Media Engagement data	Sensor Readings
• contains information about individuals who have made purchases. Due to errors or system glitches, some customer records are duplicated.	• Data from a clinical trial includes duplicate entries for some patients, possibly due to data entry errors.	• contains duplicated entries for user engagement metrics, possibly due to tracking issues.	• includes duplicate readings, possibly due to communication errors.

Best practices on containing duplicates

Identify and Understand the Source of Duplicates

identify why duplicates exist. .. data entry errors, system glitches, or other reasons,

Deduplication During Data Collection

Implement deduplication mechanisms during the data collection phase.

Utilize Advanced Matching Techniques

... such as fuzzy matching or record linkage, to identify duplicates that may have variations in spelling or formatting.

Iterate and Refine:

Deduplication is often an iterative process.

17-08-2024

Next topic : Handling Missing values

What is a missing value?

A missing value refers to the absence of a specific piece of information or data in a particular variable in a dataset.

When a data point is missing for a variable in a given observation, it is denoted as a missing value. As NA or NaN

Meaning of missing values is broader ...

list of common representations for missing data

NULL or N/A
- Some datasets represent missing values with the term "NULL" or "N/A" (Not Available).

Underscore (_)
- In some datasets, an underscore character (_) might be used

Question Mark (?)
- Occasionally, a question mark may be used to indicate missing or unknown information

Empty String
- In text-based datasets, missing values may be represented by an empty string ("").

Placeholder Text
- Some datasets use placeholder text, such as "Unknown" or "Missing," to explicitly label missing values.

Custom Codes
- Datasets may use custom codes or values designated to represent missing data, such as "-9999" or "99999."

Other Special Characters
- Depending on the dataset and its conventions, other special characters may be used to denote missing values.

what is "Handling missing values" ... why it happens

"Handling missing values" refers to the process of addressing and managing data points that are not present or have undefined/wrongly defined values in a dataset.

Missing values can occur for various reasons and handling them is a crucial step in the data preprocessing phase of data science, machine learning, and statistical analysis.

Reasons for Missing Values

Data Entry Errors
- Human errors during data entry can lead to missing values.

Equipment Malfunction or Sensor Issues
- In datasets collected from sensors, instruments, or equipment, missing values may occur

Survey Non-Response:
- In survey-based data collection, respondents may choose not to answer certain questions

Privacy and Confidentiality Concerns:
- Sometimes, data may be intentionally missing or masked

Incomplete Data Collection:
- Data collection processes may be incomplete or not cover all relevant variables

Natural Causes:
- missing values may occur naturally.

Understand types of missing values

1. structurally missing
2. missing completely at random (MCAR)
3. missing at random (MAR)
4. Non-ignorable (also known as missing not at random).

** (ML) (approach)

structurally missing data

- refers to a specific pattern of missing values in a dataset
- is not random but systematic and related to the structure or design of the data collection process or the data model.
- structurally missing data follows a predictable and non-random pattern.

Examples

Surveys: If a survey asks respondents if they own a car and then asks follow-up questions about the car's make and model only to those who answered "yes," the data for those who answered "no" would be structurally missing.

Medical Studies: In a medical study, if patients who do not have a particular condition are not tested for related symptoms, the symptom data for these patients would be structurally missing.

Example

- Blood Pressure and Cholesterol Level are health metrics, but they are not applicable (N/A) or not measured for some patients.
- Vaccination Status is systematically incomplete for some patients, possibly due to ongoing vaccination schedules.
- Diabetes Medication is relevant only for patients diagnosed with diabetes, leading to missing values for patients without the condition.

Age	Gender	Blood Pressure	Cholesterol Level	Vaccination Status	Diabetes Medication
28	Female	120/80	180 mg/dL	Complete	No
45	Male	130/85	210 mg/dL	Incomplete	Yes
60	Female	N/A	N/A	N/A	Yes
35	Male	118/75	190 mg/dL	Complete	No
28	Female	N/A	N/A	N/A	No
50	Male	140/90	220 mg/dL	Incomplete	Yes
32	Female	125/82	200 mg/dL	Complete	N/A

Detecting structurally missing data

Descriptive Statistics

- Examine summary statistics for variables with missing data.
- Look for consistent patterns or trends in the descriptive statistics of complete cases versus cases with missing values.

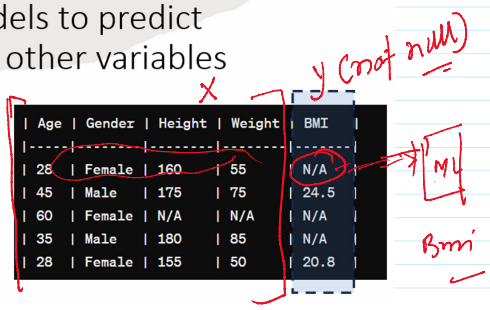
Train machine learning models

- predict missingness based on other variables.
- Feature importance analysis can highlight variables that are crucial in predicting the missingness patterns.

machine learning models to predict missingness based on other variables

- **Example: Predicting Missing BMI Values**
- Consider a dataset with information about individuals, including their age, gender, and other health-related variables.
- dataset has missing values for Body Mass Index (BMI), and you want to predict whether BMI is missing based on the available information.

Age	Gender	Height	Weight	BMI
25	Female	160	55	N/A
45	Male	175	75	24.5
60	Female	N/A	N/A	N/A
35	Male	180	85	N/A
28	Female	155	50	20.8



How to deal with structurally missing data

Understand the Study Design:	Exclude Irrelevant Variables:	Consider Time Points and Events:	Imputation with Care:	Consult Domain Experts:
• Gain a thorough understanding of the study design and the reasons behind structurally missing data.	• it might be appropriate to exclude variables from the analysis	• Exclude assessments or measurements that are not applicable at certain time points.	• If imputation is considered, use caution, especially when structurally missing data is systematic.	• Collaborate with domain experts or researchers familiar with the study area to gain insights.

MCAR (missing completely at random).

Missing Completely at Random (MCAR)

Definition

- refers to a situation where the probability of missing data on a variable is independent of both observed and unobserved data.

Characteristics of MCAR

- **Independence**: The likelihood of data being missing is unrelated to the values of other variables or to the missing values themselves.
- **Randomness**: The missing data is essentially random and does not show any systematic pattern.

Handling MCAR

Student_ID	Hours_of_Sleep	GPA	Extracurriculars
1	7	3.6	Yes
2	6	3.2	No
3	8	3.8	Yes
4	5	2.9	Yes
5	7	3.4	No
6	6	3.1	N/A (MCAR)
7	8	3.7	Yes
8	5	2.8	N/A (MCAR)
9	7	3.6	Yes
10	6	3.0	No

- **Hours_of_Sleep** represents the number of hours of sleep each student gets per night.
- **GPA** represents the student's grade point average.
- **Extracurriculars** indicates whether the data on extracurricular activities is present ("Yes") or missing ("N/A") for some students.

- Assume that the missingness in the **Extracurriculars** variable is completely at random (MCAR).

- means that the probability of missing values in the **Extracurriculars** column is unrelated to both observed and unobserved data.

random.

Ways to handle MCAR

- **Approach**: Remove observations with missing values.

Original Data:

Hours_of_Sleep	GPA	Extracurriculars
7	3.5	Yes
6	3.2	No
8	3.8	Yes
5	2.9	Yes
7	3.4	No
6	3.1	N/A (MCAR)
8	3.7	Yes
5	2.8	N/A (MCAR)
7	3.6	Yes
6	3.0	No

After Listwise Deletion:

Hours_of_Sleep	GPA	Extracurriculars
7	3.5	Yes
6	3.2	No
8	3.8	Yes
5	2.9	Yes
7	3.4	No
8	3.7	Yes
6	3.0	No

Ways to handle MCAR

- **Approach:** Replace missing values with the mean/median/mode of the observed values for that variable.

Original Data:

Hours_of_Sleep	GPA	Extracurriculars
7	3.5	Yes
6	3.2	No
8	3.8	Yes
5	2.9	Yes
7	3.4	No
6	3.1	N/A (MCAR)
8	3.7	Yes
5	2.8	N/A (MCAR)
7	3.6	Yes
6	3.0	No

After Mean Imputation:

Hours_of_Sleep	GPA	Extracurriculars
7	3.5	Yes
6	3.2	No
8	3.8	Yes
5	2.9	Yes
7	3.4	No
6	3.1	Yes (imputed)
8	3.7	Yes
5	2.8	Yes (imputed)
7	3.6	Yes
6	3.0	No

multiple types of imputation → mean (vanilla) → mode → median

Analysis

- Perform the desired analysis (e.g., regression) on each imputed dataset separately.
- Combine the results using appropriate rules (e.g., Rubin's rules) to incorporate the uncertainty from multiple imputations.

Imputed Dataset 3:

Hours_of_Sleep	GPA	Extracurriculars
7	3.5	Yes
6	3.2	No
8	3.5	Yes
5	2.9	Yes
7	3.2	No
6	3.1	Yes
8	3.7	Yes
5	3.5	No
7	3.6	Yes
6	3.0	No

Imputed Dataset 2:

Hours_of_Sleep	GPA	Extracurriculars
7	3.5	Yes
6	3.2	No
8	3.6	Yes
5	2.9	Yes
7	3.4	No
6	3.1	Yes
8	3.7	Yes
5	3.6	No
7	3.6	Yes
6	3.0	No

Imputed Dataset 1:

Hours_of_Sleep	GPA	Extracurriculars
7	3.5	Yes
6	3.2	No
8	3.4	Yes
5	2.9	Yes
7	3.3	No
6	3.1	Yes
8	3.7	Yes
5	3.3	No
7	3.6	Yes
6	3.0	No

1. Statistical method.; age sleep gpa
   - t-test
   - Anova
   - Chi
   - non-parametric
2. ML.

linkage → pattern (non random)

no linkage → pattern (random)

Imputation by interpolation method

- Imagine you have a dataset representing the speed of a car at different time points.
- However, due to some measurement issues, you have missing values for certain time points.

Time	Speed
0	10
1	15
2	20
3	25
4	?
5	35
6	40
7	?
8	55
9	60

Key points about linear interpolation

Applicability:	It is applicable to any ordered dataset where there is a meaningful linear relationship between consecutive data points.
Linear Relationship:	Assumes a linear relationship between the known data points. This is based on the idea that the change between consecutive points is constant.
Usage in Temporal Data:	Commonly used in time series data to estimate values at specific time points. For example, filling in missing values in a time-stamped dataset.

time chronological.

MCAR and MAR

Difference between MCAR and MAR

- **Example Dataset**
  - Consider a dataset of survey responses on job satisfaction, where each row represents an individual's response. The columns are:
  - **Age**: Age of the respondent
  - **Salary**: Annual salary of the respondent
  - **Job Satisfaction**: Self-reported job satisfaction (1 to 10 scale)
  - **Hours Worked**: Number of hours worked per week

Age	Salary	Job Satisfaction	Hours Worked
25	50000	8	40
30	60000	7	35
40	70000	NA	45
22	40000	6	NA
35	NA	9	50
28	55000	NA	38
50	80000	5	42
29	53000	7	44

- hypothesis testing → next module.
- ML method → (ML tq)

MCAR Example

Scenario:	Suppose the missing data for **Job Satisfaction** and **Hours Worked** is due to a survey design flaw, such as a printing issue that caused some responses to be skipped randomly.
MCAR Characteristics:	The missingness of **Job Satisfaction** and **Hours Worked** is unrelated to any other values in the dataset, including **Age** or **Salary**. For example, the missing values are not systematically higher or lower based on any specific characteristic of the respondent.

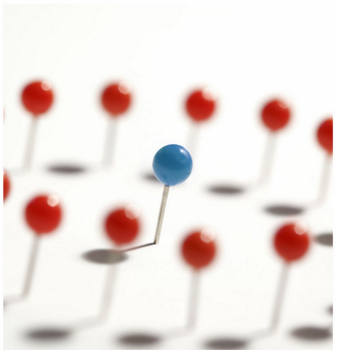
MAR Example

Scenario:	Suppose the missing data for **Job Satisfaction** and **Hours Worked** is related to **Salary**. Specifically, the data is missing for respondents with higher salaries, perhaps because higher-paid employees were less likely to complete the survey or chose not to disclose their satisfaction levels.
MAR Characteristics:	The missingness of **Job Satisfaction** and **Hours Worked** is related to the observed variable **Salary**. Higher salary is associated with missing values in these fields. However, the missingness is not related to the actual values of **Job Satisfaction** or **Hours Worked** themselves.

Illustration of Differences

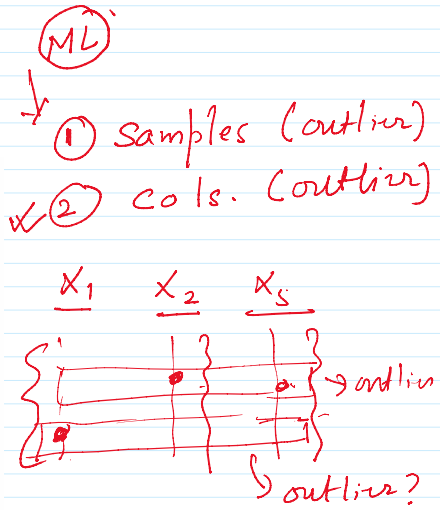
MCAR	MAR
• **Analysis Impact**: If the missing data is MCAR, removing or imputing missing values will not bias the results because the missingness does not depend on any observed or unobserved variables. • **Simple Imputation**: You can use methods like mean imputation or listwise deletion without worrying about introducing bias.	• **Analysis Impact**: If the missing data is MAR, the missingness is related to observed variables (e.g., **Salary**), so you must use methods that account for this relationship, such as multiple imputation or maximum likelihood estimation. • **Advanced Handling**: Techniques that model the relationship between missing data and observed data are required to obtain unbiased results.

Next topic : Outlier assessment



Outliers

- instances too distinct when compared to the other examples of the data set.
- Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature



handling outliers

Identify Outliers:

Common techniques include visualizations (box plots, scatter plots), statistical methods (z-scores, IQR).

Understand Context:

Consider the context of your data and the domain in which it was collected.

Outliers that are valid and meaningful in one context might be errors in another.

Domain Knowledge:

Consult domain experts to understand the potential impact of outliers on your analysis.

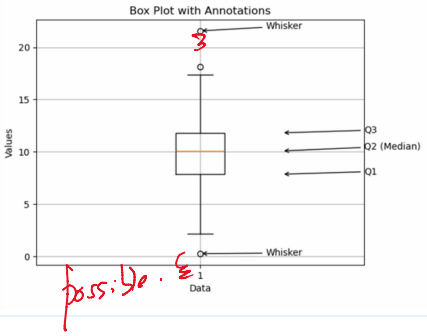
Ways to handle outliers

Visual Inspection	Descriptive Statistics	Machine Learning Models
• **Scatter Plots, Box Plots:** Use box plots to highlight the distribution of the data and identify points outside the whiskers.	• **Z-Score:** Calculate the z-score for each data point, indicating how many standard deviations • **IQR (Interquartile Range):** Use the IQR to define a range within which most data points lie. Points outside this range are considered potential outliers.	• **Clustering Algorithms:** Apply clustering algorithms, such as K-Means or DBSCAN, and identify data points that do not belong to any cluster. • **Isolation Forest:** Utilize isolation forest algorithms, which are effective at isolating outliers by partitioning the data.

Visual inspection – box plots

univariate dataset or assessing outliers in a single variable,

consider alternative visualizations such as box plots or histograms.



Interpretation

- If the Box is Symmetric: data is evenly distributed.
- If the Box is Skewed: longer tail of the box indicates the direction of skewness.
- Outliers: Identified as individual points beyond the whiskers.

Components of a Box plot - quartiles

Quartiles divide a dataset into four equal parts, representing different segments of the data distribution.

First Quartile (Q1): 25th percentile, marking the lower boundary of the first 25% of the data.

Second Quartile (Q2): 50th percentile or the median, dividing the data into two halves.

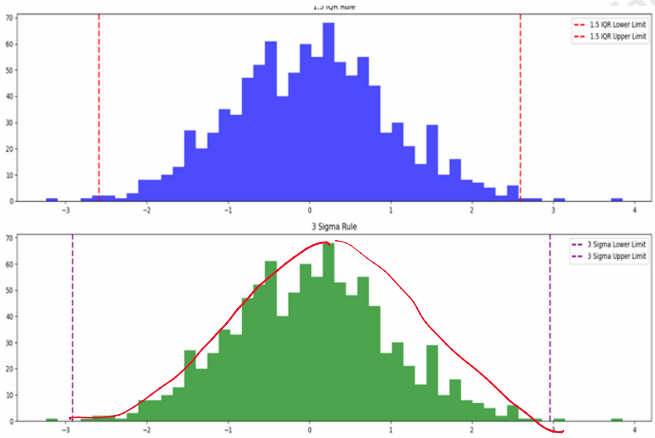
Third Quartile (Q3): The 75th percentile, marking the lower boundary of the upper 25% of the data.

Quartiles provide insights into the central tendencies and spread of the dataset.

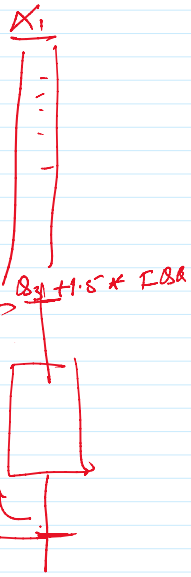
Components of a Box plot

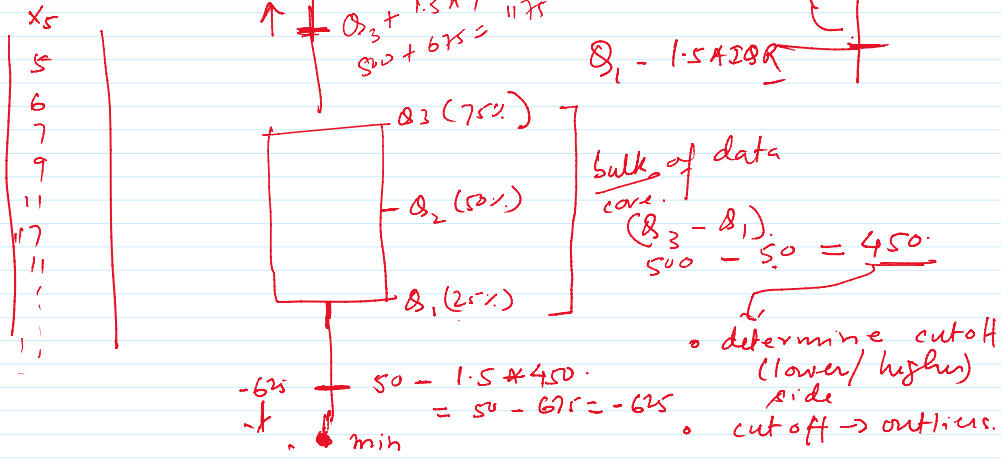
Box	Whiskers	Outliers
• box in a box plot represents the interquartile range (IQR), which is the range between Q1 and Q3.	• Whiskers extend from the edges of the box to the minimum and maximum values within a certain range. • The range is often set at 1.5 times the interquartile range (IQR).	• Outliers are individual data points beyond the whiskers.

IQR vs std dev



- two subplots showing a histogram of the data along with the 1.5 IQR limits and the 3 sigma limits.
- The first subplot focuses on the 1.5 IQR rule, and the second subplot focuses on the 3 sigma rule.
- You can observe how the 1.5 IQR rule aligns approximately with the 3 sigma rule in a normal distribution.

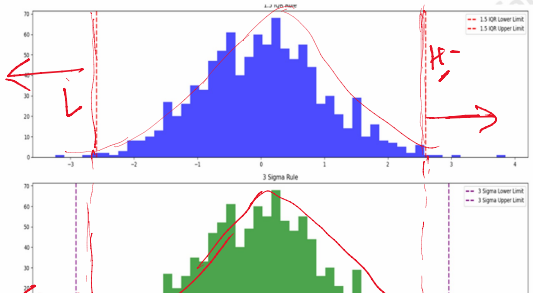




- determine cutoff (lower/ higher) side
- cutoff → outliers.

Why 1.5 ?

IQR vs std dev



IQR ①

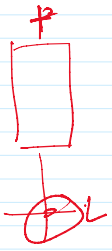
1.5 factor.

L

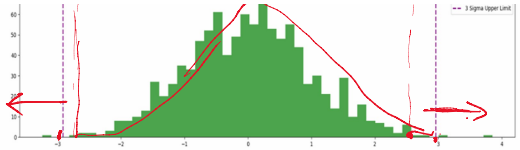
H

Z-score.

- two subplots showing a histogram of the data along with the 1.5 IQR limits and the 3 sigma limits.
- The first subplot focuses on the 1.5 IQR rule, and the second subplot focuses on the 3 sigma rule.
- You can observe how the 1.5 IQR rule aligns approximately with the 3 sigma rule in a



Z-score.



rule.

- You can observe how the 1.5 IQR rule aligns approximately with the 3 sigma rule in a normal distribution.

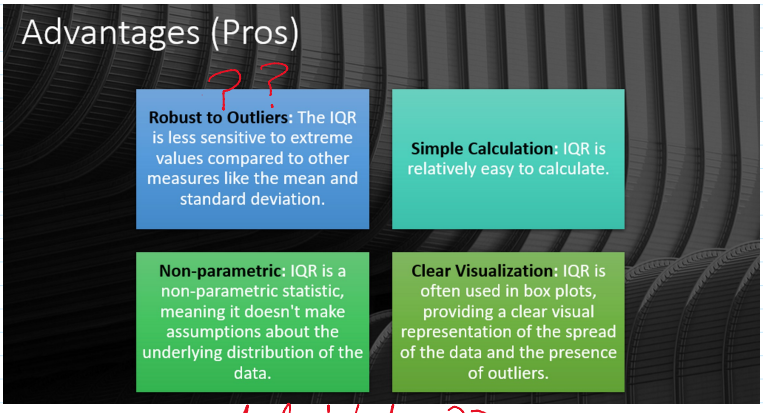
1.5 (high) ≈ +3σ
1.5 (low) ≈ -3σ
99.7% (empirically).

Outliers by IQR method

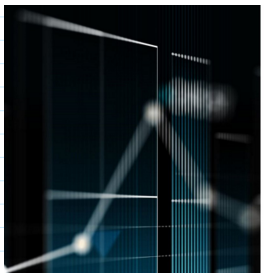
- **Dummy Dataset** : 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 500
- **Step 1**: Arrange the data in ascending order.
- **Step 2**: Calculate Q1 (25th percentile) = 15
- **Step 3**: Calculate Q3 (75th percentile) = 39
- **Step 4**: Calculate IQR (Interquartile Range) as Q3 - Q1 = 39 - 15 = 24

... next steps

- Define Lower and Upper Bounds:
  - Define lower bound as Q1 - 1.5 * IQR = 15 - 1.5 * 24 = -21
  - Define upper bound as Q3 + 1.5 * IQR = 39 + 1.5 * 24 = 75
- Identify Outliers:
  - Any data point outside the range [Lower Bound, Upper Bound] is considered an outlier.
  - Outliers: 500 (outside the upper bound)



assuming normal distribution ??
limitation: sample size.



Z-score method for outlier detection

n.d

- Z score, also known as the standard score, is a measure of how many standard deviations a data point is from the mean of a dataset.
- often used to standardize and compare data points from different distributions.
- formula for calculating the Z score of a data point (X) in a dataset with mean (μ) and standard deviation (σ) is given by:

$$z = \frac{(X-\mu)}{\sigma}$$

5
6
7
9
11
42

μ
σ
$\frac{(5-\mu)}{\sigma} = z$

distributions.

- formula for calculating the Z score of a data point (X) in a dataset with mean (μ) and standard deviation (σ) is given by:

$$Z = \frac{(X-\mu)}{\sigma}$$

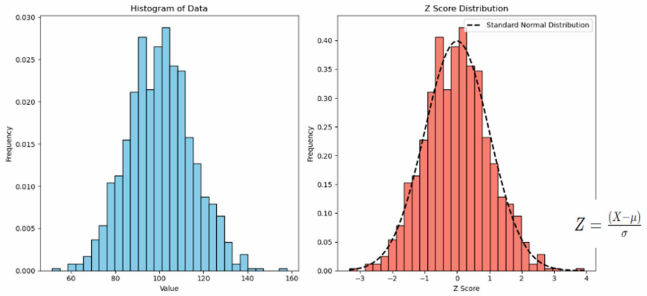
- Z score of
  - 0 indicates that the data point's value is exactly at the mean,
  - +1 indicates that it is one standard deviation above the mean
  - -1 indicates that it is one standard deviation below the mean.

$$\left(\frac{x-\mu}{\sigma}\right) = Z$$

Z-score.

Illustrate Z scores

- first subplot shows the histogram of the original data, and the second subplot shows the histogram of the corresponding Z scores.
- dashed line in the second subplot represents the standard normal distribution (mean=0, std_dev=1).
- You can see how the Z scores are distributed around the mean of 0, illustrating the standardization process

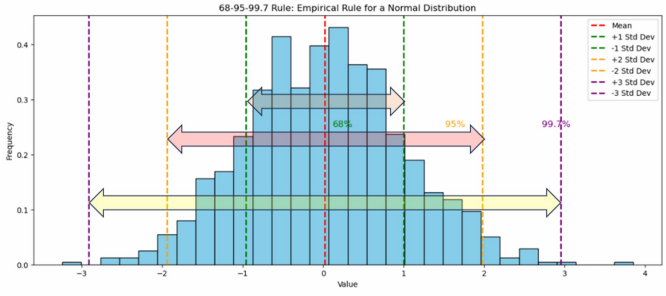


$$Z = \frac{(X-\mu)}{\sigma}$$

17-08-2024 @ Copyright – created by bhupen 26

68-95-99.7 Rule

- **Within 1 Standard Deviation (±1σ):** Approximately 68% of the data should fall within this range. If a data point is beyond this range, it might be considered unusual but not necessarily an outlier.
- **Within 2 Standard Deviations (±2σ):** Around 95% of the data should fall within this range. Data points beyond this range may be considered as potential outliers.
- **Within 3 Standard Deviations (±3σ):** About 99.7% of the data should fall within this range. Data points beyond this range are often classified as outliers.



Limitations

1. sample size
2. n.d. (skew will impact).
(μ, σ).

Next topic - Encoding of categorical data

(** very imp)

Define

01

Encoding categorical variables is a way to convert string/object data to numeric format

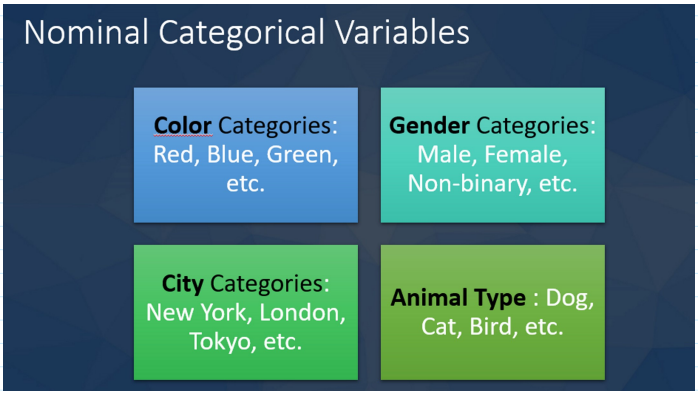
02

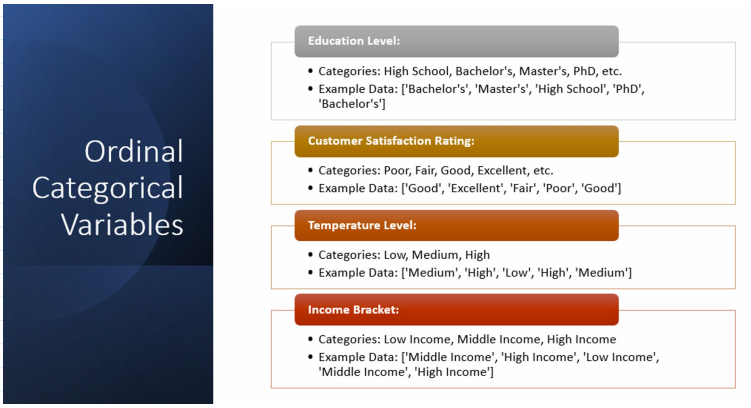
providing a way to include categorical data in machine learning models that require numerical input.

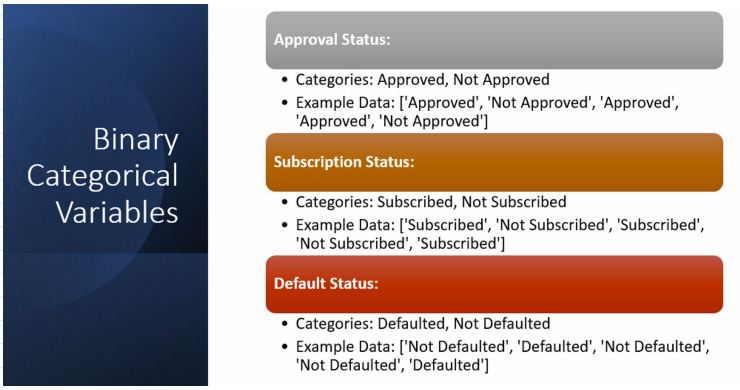
03

E.g. creating binary columns for each category and using 0s and 1s to indicate the presence or absence of each category.

17-08-2024 @ Copyright – created by bhupen 6







Example 1: One-Hot Encoding (Dummy Variables)

Consider a dataset with a 'Color' column containing categorical data: 3 categories

ID	Color
1	Red
2	Blue
3	Green

One-hot encoding would create binary columns for each color

ID	Color_Red	Color_Blue	Color_Green
1	1	0	0
2	0	1	0
3	0	0	1

Example 2: Label Encoding

Consider a 'Size' column with categorical data

ID	Size
1	Small
2	Medium
3	Large

Label encoding assigns numerical labels to categories

D	Size
1	0
2	1
3	2

very very popular.

Example: Frequency Encoding

Consider a dataset with a 'City' column:

ID	City
1	New York
2	Tokyo
3	London
4	Tokyo
5	New York

Frequency encoding assigns values based on the frequency of each category in the dataset:

ID	City_Frequency
1	0.4
2	0.4
3	0.2
4	0.4
5	0.4

'New York' and 'Tokyo' both appear twice in the dataset, resulting in a frequency of 0.4 for each. 'London' appears once, resulting in a frequency of 0.2.

??

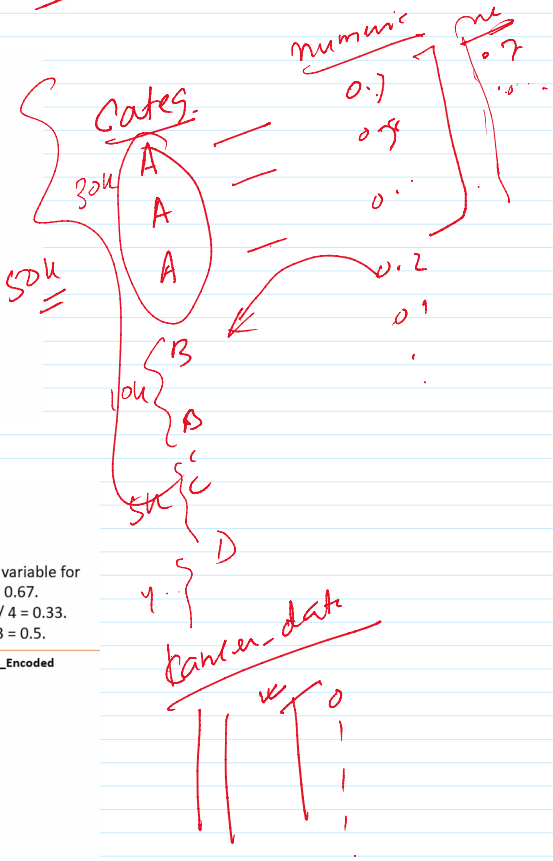
Interpretation

Higher Value:

A higher frequency-encoded value indicates that the category appears more frequently in the dataset.

Lower Value:

A lower value still indicates that the category is less common.



Target Encoding (Mean Encoding)

- Target encoding assigns values based on the mean of the target variable for each category

City	Target
New York	1
Tokyo	0
London	1
Tokyo	1
New York	0
London	0
Tokyo	1
New York	1
London	0
Tokyo	0

available to us

- For 'New York,' the mean of the target variable for rows with 'New York' is (1 + 0 + 1) / 3 = 0.67.
- For 'Tokyo,' the mean is (0 + 1 + 0 + 1) / 4 = 0.33.
- For 'London,' the mean is (1 + 0 + 0) / 3 = 0.5.

City	Target	City_Target_Encoded
New York	1	0.67
Tokyo	0	0.33
London	1	0.5
Tokyo	1	0.33
New York	0	0.67
London	0	0.5
Tokyo	1	0.33
New York	1	0.67
London	0	0.5
Tokyo	0	0.33

age Sale grade - - - - - - status (target)
R - 0
A - 1
H - 2

encoding → "embeddings" (LLMs)
→ "vectorization"