Data Preprocessing: An Overview

- **Data Quality: Why Preprocess the Data?**

- Data quality is a measure of the condition of data based on factors such as accuracy, completeness, consistency, reliability and whether it's up to date. Measuring data quality levels can help organizations identify data errors that need to be resolved and assess whether the data in their IT systems is fit to serve its intended purpose.

- Five Factors of High Quality Data

  1. **Completeness**
  2. **Consistency**
  3. **Accuracy**
  4. **Validity**
  5. **Timeliness**

1. Data Completeness

- Data completeness refers to whether there are any gaps in the data from what was expected to be collected, and what was actually collected.

- **Example**: An inspection is done on a vehicle and the inspector accidentally does not indicate the current hour meter reading on the vehicle, which is a required field for that inspection. This has rendered the inspection incomplete and less valuable because important information is left out.

2. Data Consistency

- Data consistency is the measure that indicates that data is not conflicting each other and not conflicting business rules. If data is replicated in multiple places, it needs to be consistent across all instances.

- Example: For a department store, you might hold data on a particular customer through a loyalty program, mailing list, online accounts payment system and order fulfillment system. In that tangled mess of systems there may be misspelled names, old addresses and conflicting status flags. This could cause problems in processes that read data

3. Data Accuracy

- Accuracy is the measure that indicates how well and how correctly is data represented in the data base, comparing its value to the real world or to a reference data. Accuracy is a crucial data quality characteristic because inaccurate information can cause significant problems with severe consequences.

- **Example**: on the same inspection example, if the operator records the mileage at 40,000 miles instead of 60,000 miles, this is inaccurate data, resulting in misinformation and related issues.

4. Data Validity

- Fixing invalid data often means that there is an issue with a process rather than a result. Validity of data is determined by whether the data measure that which it is intended to measure.

- Example: When new information is needed but forms don't get changed, the data is no longer valid because it does not properly measure what it is supposed to.

5. Data Timeliness

- Data timeliness refers to the expectation of when data should be received in order for the information to be used effectively.

- **Example:** At the end of the month, several sales representatives fail to file their sales record on time. These are also several corrections & adjustments which flow into after the end of the month. Data stored in the database are incomplete for a time after each month.

- **Data Cleaning**

- Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database.

- Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

- After cleaning, a dataset should be uniform with other related datasets in the operation.

  - **Missing Values**

- Missing data is a deceptively tricky issue in machine learning. We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important.

- Following are some methods

1. **Ignore the tuple**:

- This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values.

- By ignoring the tuple, we do not make use of the remaining attributes' values in the tuple. Such data could have been useful to the task at hand.

2. **Fill in the missing value manually:**

- In general, this approach is time consuming and may not be feasible given a large data set with many missing values.

3. **Use a global constant to fill in the missing value:**

- Replace all missing attribute values by the same constant such as a label like "Unknown". If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not foolproof.

4. **Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value:**

- Measures of central tendency, which indicate the "middle" value of a data distribution. For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median.

- For example, suppose that the data distribution regarding the income of AllElectronics customers is symmetric and that the mean income is $56,000. Use this value to replace the missing value for income.

5. **Use the attribute mean or median for all samples belonging to the same class as the given tuple:**

- For example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.

6. **Use the most probable value to fill in the missing value**:

- This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree.

- Induction for example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

- Methods 3 through 6 bias the data—the filled-in value may not be correct. Method 6, however, is a popular strategy. In comparison to the other methods, it uses the most information from the present data to predict missing values.

- **Noisy Data**

- Noise is a random error or variance in a measured variable. Noisy Data may be due to faulty data collection instruments, data entry problems and technology limitation.

- How to Handle Noisy Data? Following are data smoothing techniques.

1. Binning

- Binning methods sorted data value by consulting its "neighbor-hood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins.

- **For example**

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Methods of binning

1) smoothing by bin means

- In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

- The data for price are first sorted and then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values). For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

2) Smoothing by bin boundaries

- In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

- Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In general, the larger the width, the greater the effect of the smoothing.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

3) smoothing by bin median

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Smoothing By Bin Median

Bin 1:8,8,8
Bin 2:21,21,21
Bin 3:28,28,28

Advantages (Pros) of data smoothing

- Data smoothing clears the understandability of different important hidden patterns in the data set.
- Data smoothing can be used to help predict trends. Prediction is very helpful for getting the right decisions at the right time.
- Data smoothing helps in getting accurate results from the data.

Cons of data smoothing

- Data smoothing doesn't always provide a clear explanation of the patterns among the data.

- It is possible that certain data points being ignored by focusing the other data points.

8 ,16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

Bin 1: 8, 9, 15, 16

Bin 2: 21, 21, 24, 26,

Bin 3: 27, 30, 30, 34

Smoothing by bin means

For Bin 1:

(8+ 9 + 15 +16 / 4) = 12

(4 indicating the total values like 8, 9 , 15, 16)

Bin 1 = 12, 12, 12, 12

For Bin 2:

(21 + 21 + 24 + 26 / 4) = 23

Bin 2 = 23, 23, 23, 23

For Bin 3:

(27 + 30 + 30 + 34 / 4) = 30

Bin 3 = 30, 30, 30, 30

Smoothing by bin boundaries

Bin 1: 8, 9, 15, 16

Bin 2: 21, 21, 24, 26,

Bin 3: 27, 30, 30, 34

Answer

Bin 1: 8, 8, 16, 16

Bin 2: 21,21,26,26

Bin 3:27,27,27,34

Smoothing by bin median

(9+15)/2=12

Bin 1: 12,12,12,12

Bin 2: 23,23,23,23

Bin 3: 30,30,30,30

2. Regression:

- Data smoothing can also be done by regression, a technique that conforms data values to a function. Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.

- Regression refers to a type of supervised machine learning technique that is used to predict any continuous-valued attribute. Regression helps any business organization to analyze the target variable and predictor variable relationships. It is a most significant tool to analyze the data that can be used for financial forecasting and time series modeling.

- Regression involves the technique of fitting a straight line or a curve on numerous data points. It happens in such a way that the distance between the data points and cure comes out to be the lowest.

- The most popular types of regression are linear and logistic regressions. Other than that, many other types of regression can be performed depending on their performance on an individual data set.

- Regression is divided into five different types

  1. Linear Regression
  2. Logistic Regression
  3. Lasso Regression
  4. Ridge Regression
  5. Polynomial Regression

Application of Regression

- Regression is a very popular technique, and it has wide applications in businesses and industries. The regression procedure involves the predictor variable and response variable. The major application of regression is given below.

- Environmental modeling
- Analyzing Business and marketing behavior
- Financial predictors or forecasting
- Analyzing the new trends and patterns.

3. Outlier analysis:

Outlier

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

- An outlier cannot be termed as a noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects.

- Outliers are of three types, namely –

1. Global (or Point) Outliers **(a data point strongly deviates from all the rest of the data points, it is known as a global outlier)**

2. Collective Outliers(**some of the data points, as a whole, deviate significantly from the rest of the dataset)**

3. Contextual (or Conditional) Outliers (**A data point may be an outlier due to a certain condition and may show normal behavior under another condition.)**

Outlier analysis

- **The process in which the behavior of the outliers is identified in a dataset is called outlier analysis. It is also known as "outlier mining",** the process is defined as a significant task of data mining.

- But it is still used in many applications like fraud detection, medical, etc. It is usually because the events that occur rarely can store much more significant information than the events that occur more regularly.

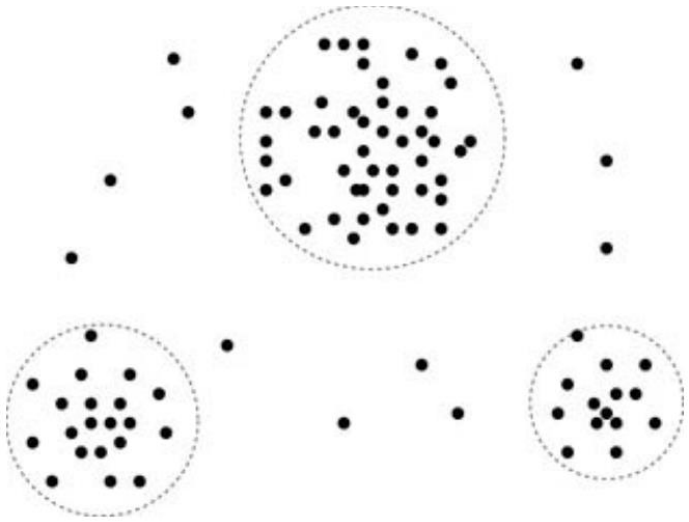
- Other applications where outlier detection plays a vital role are given below.

  - Fraud detection in the telecom industry
  - In market analysis, outlier analysis enables marketers to identify the customer's behaviors.
  - In the Medical analysis field.
  - Fraud detection in banking and finance such as credit cards, insurance sector, etc.

- Outliers may be detected **by clustering**, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers

Example

- A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.



- **Data Cleaning as a Process**

- The first step in data cleaning as a process is discrepancy detection. Discrepancies can be caused by several factors, including poorly designed data entry forms that have many optional fields, human error in data entry, deliberate errors (e.g., respondents not wanting to divulge information about themselves), and data decay (e.g., outdated addresses).

- Discrepancies may also arise from inconsistent data representations and inconsistent use of codes.

- **Data transformations.** That is, once we find discrepancies, we typically need to define and apply (a series of) transformations to correct them.

- Commercial tools can assist in the data transformation step. Data migration tools allow simple transformations to be

specified such as to replace the string "gender" by "sex." **ETL (extraction/transformation/loading) tools** allow users to specify transforms through a graphical user interface (GUI).

▪ Data Reduction

➢ Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

○ Overview of Data Reduction Strategies

➢ Data reduction strategies include **dimensionality reduction, numerosity reduction, and data compression.**

1. Dimensionality reduction

➢ **The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction.**

➢ A dataset contains a huge number of input features in various cases, which makes the predictive modeling task more complicated. Because it is very difficult to visualize or make predictions for the training dataset with a high number of features, for such cases, dimensionality reduction techniques are required to use.

➢ Dimensionality reduction technique can be defined as, "***It is a way of converting the higher dimensions dataset into lesser***

dimensions dataset ensuring that it provides similar information." These techniques are widely used in machine learning for obtaining a better fit predictive model while solving the classification and regression problems.

- It is commonly used in the fields that deal with high-dimensional data, such as **speech recognition, signal processing, bioinformatics, etc. It can also be used for data visualization, noise reduction, cluster analysis**, etc.

Benefits of applying Dimensionality Reduction

Some benefits of applying dimensionality reduction technique to the given dataset are given below:

- By reducing the dimensions of the features, the space required to store the dataset also gets reduced.
- Less Computation training time is required for reduced dimensions of features.
- Reduced dimensions of features of the dataset help in visualizing the data quickly.
- It removes the redundant features (if present) by taking care of multicollinearity.

Disadvantages of dimensionality Reduction

- There are also some disadvantages of applying the dimensionality reduction, which are given below:

- Some data may be lost due to dimensionality reduction.
- In the PCA (Principal Component Analysis) dimensionality reduction technique, sometimes the principal components required to consider are unknown.

2. Numerosity reduction

- **This technique replaces the original data volume by alternative, smaller forms of data representation.**

- These techniques may be **parametric** or **nonparametric**. For **parametric methods**, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.)
  Example- Regression and log-linear models.

- **Nonparametric methods** for storing reduced representations of the data include histograms, clustering, sampling, and data cube aggregation.

Difference between Dimensionality Reduction and Numerosity Reduction

Dimensionality Reduction	Numerosity Reduction
In dimensionality reduction, data encoding or transformation are applied to obtain a reduced or compressed representation of original data.	In numerosity reduction, data volume is reduced by choosing alternating, smaller forms of data representation.
It can be used for removing irrelevant and redundant	It is merely a representation technique of original data to a

attributes.	smaller form.
In this technique, some data can be lost which is inappropriate.	In this method, there is no loss of data but the whole data is represented in a smaller form.

3. Data compression

- Transformations are applied so as to obtain a reduced or "compressed" representation of the original data. **If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called "lossless"**. If, instead, **we can reconstruct only an approximation of the original data, then the data reduction is called "lossy"**.

Histograms

- Histograms use binning to approximate data distributions and are a popular form of data reduction.

- A histogram provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values (called "bins" or "bucket")

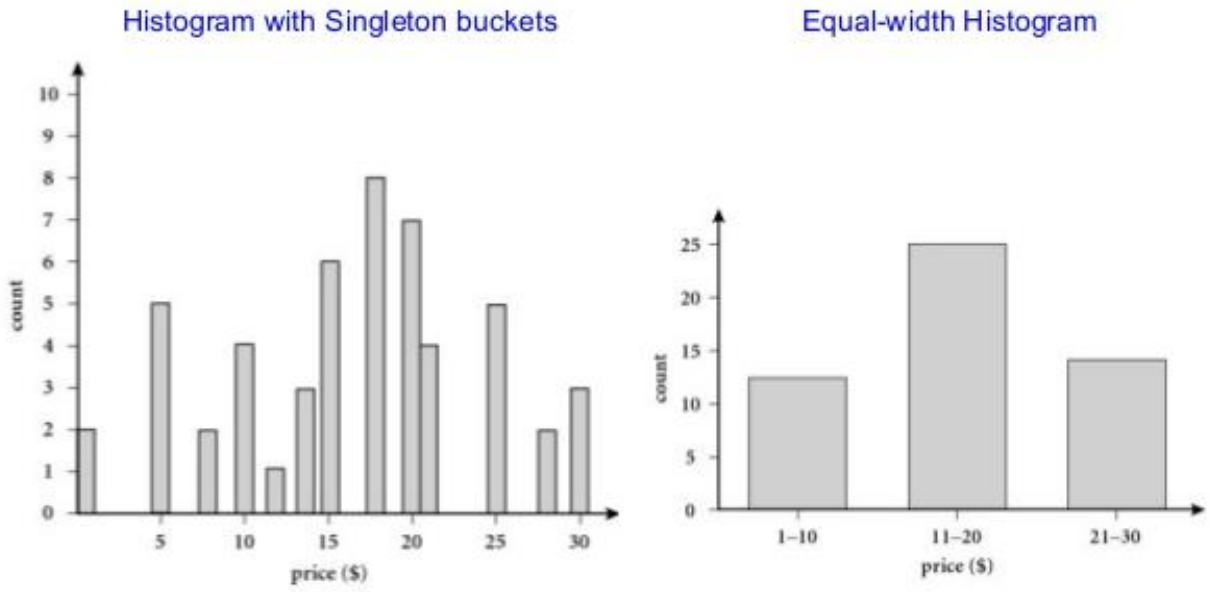
- **If each bucket represents only a single attribute–value/frequency pair, the buckets are called "singleton buckets"**.

- **Example**

The following data are a list of AllElectronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,30, 30.

Following diagram shows a histogram for the data using singleton buckets. To further reduce the data, it is common to have each bucket denote a continuous value range for the given attribute. In each bucket represents a different $10 range for price.

Histograms



List of prices: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30

Sampling

- Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random data sample (or subset).

- **Types of Sampling**

1. **Simple random sample without replacement (SRSWOR) of size**

- As each item is selected, it is removed from the population is a method of selection of n units out of the N units one by one such that at any stage of selection, anyone of the remaining units have same chance of being selected, i.e. 1/N .

2. **Simple random sample with replacement (SRSWR) of size**

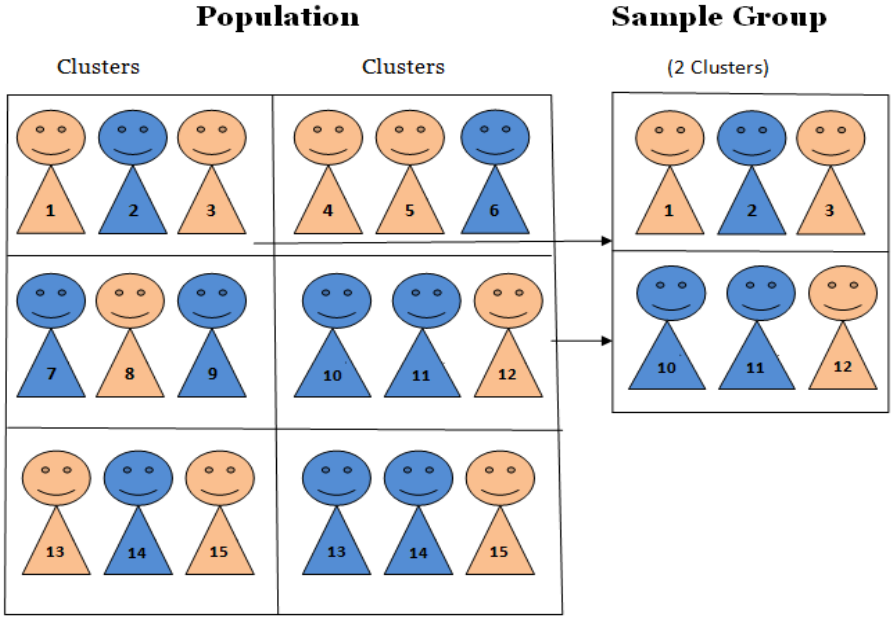
- is a method of selection of n units out of the N units one by one such that at each stage of selection each unit has equal chance of being selected, i.e., 1/ N .

- **Example**

- Suppose we have a bowl of 100 unique numbers from 0 to 99. We want to select a random sample of numbers from the bowl. After we pick a number from the bowl, we can put the number aside or we can put it back into the bowl. If we put the number back in the bowl, it may be selected more than once; if we put it aside, it can selected only one time.

- When a population element can be selected more than one time, we are **sampling with replacement**. When a population element can be selected only one time, we are **sampling without replacement**.

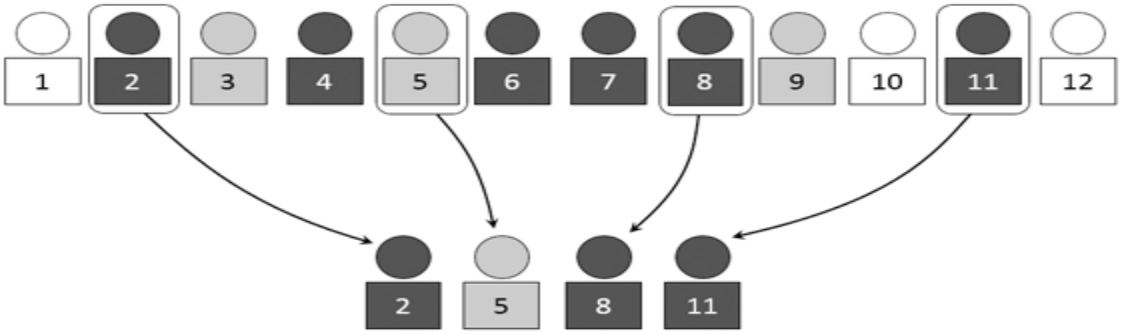
3. Cluster sample



- Cluster sampling refers to a type of sampling method. With cluster sampling, the researcher divides the population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. The researcher conducts his analysis on data from the sampled clusters.

- **For example**, let's consider a scenario where an organization is looking to survey the performance of smart phones across Germany. They can divide the entire country's population into cities (clusters) and further select cities with the highest population and also filter those using mobile devices. This multiple stage sampling is known as cluster sampling.

4. Stratified sample:

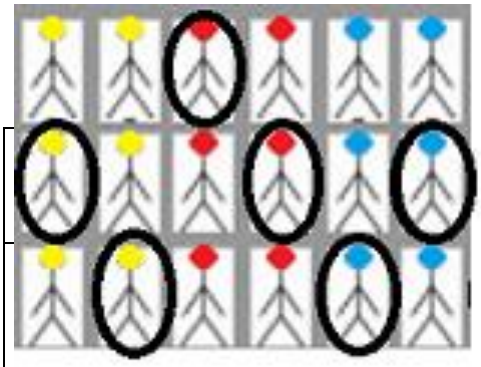


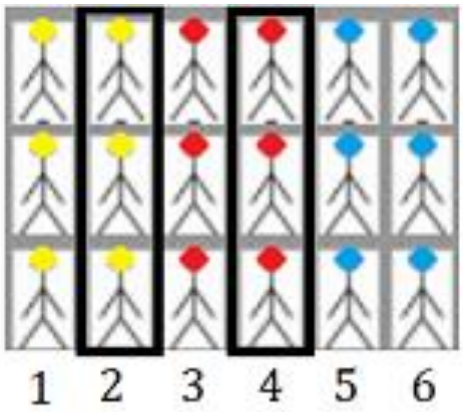
- Stratified sampling is a type of sampling method in which the total population is divided into smaller groups or **strata** to complete the sampling process.

- The strata are formed based on some common characteristics in the population data. After dividing the population into strata, the researcher randomly selects the sample proportionally.

- **For example**, a stratified sample may be obtained from customer data, where a stratum is created for each customer age group. In this way, the age group having the smallest number of customers will be sure to be represented.

- **Difference between cluster and stratified sampling in tabular form**



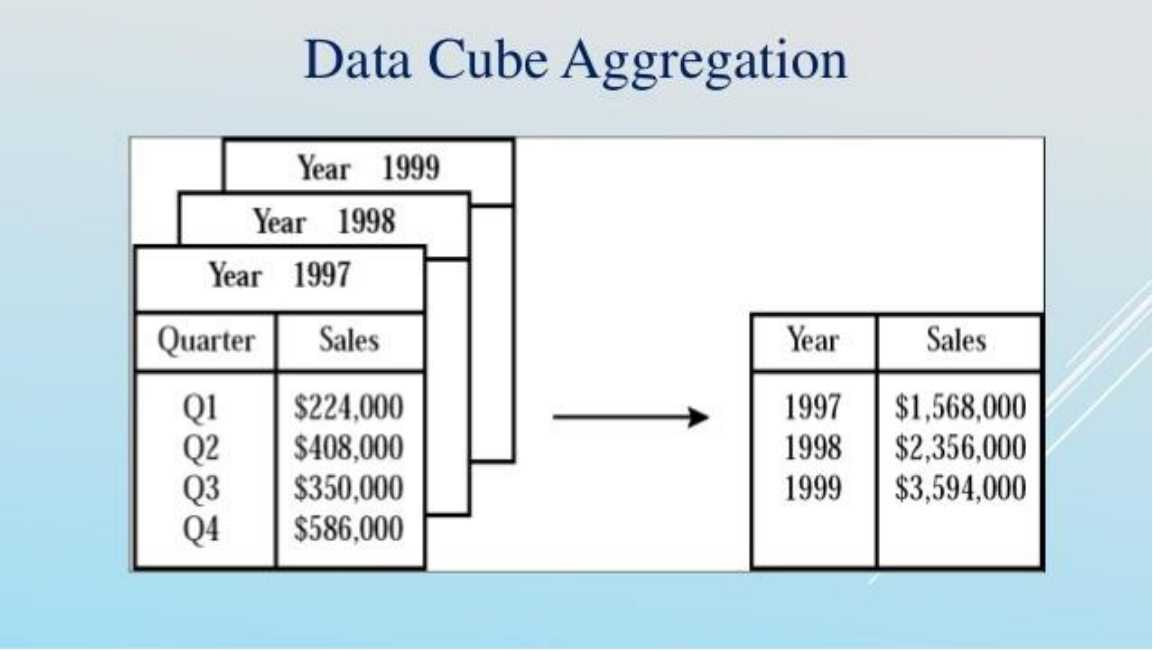


Stratified Sampling **Cluster Sampling**

	the segments.	occurring groups called 'cluster'.
Sample	Randomly selected individuals are taken from all the strata.	All the individuals are taken from randomly selected clusters.
Selection of population elements	Individually	Collectively
Homogeneity	Within group	Between groups
Heterogeneity	Between groups	Within group
Bifurcation	Imposed by the researcher	Naturally occurring groups
Objective	To increase precision and representation.	To reduce cost and improve efficiency.

Data Cube Aggregation

- Data cubes store multidimensional aggregated information



- **Association Rule Mining**

- **Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an item set occurs in a transaction.**

- Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

- Association rules are created by searching data for frequent if-then patterns and using the criteria

    - **Support** (Support indicates how frequently the if/then relationship appears in the database.)
    - **Confidence** (Confidence tells about the number of times these relationships have been found to be true.)

- **Example:** Support and Confidence can be represented by the following example

    Bread=> butter [support=2%, confidence-60%]

- The above statement is an example of an association rule. This means that there is a 2% transaction that bought bread and butter together and there are 60% of customers who bought bread as well as butter.

- **Support and Confidence for Itemset A and B are represented by formulas:**

$$\text{Support (A)} = \frac{\text{Number of transaction in which A appears}}{\text{Total number of transactions}}$$

$$\text{Confidence } (A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

- To identify the most important relationships. **A third metric, called "*lift*", can be used to compare confidence with expected confidence.**

- **Association rules are calculated from *itemsets***, which are made up of two or more items. If rules are built from analyzing all the possible itemsets.

- **Lift**

  The lift of a rule is defined as:

$$\text{lift}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) * \text{supp}(Y)}$$

- Association rule mining consists of 2 steps:

  1. Find all the frequent itemsets.
  2. Generate association rules from the above frequent itemsets.

- **The main applications of association rule mining:**

- **Basket data analysis (Market Basket analysis)**

- It is to analyze the association of purchased items in a single basket or single purchase.

- It is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently.

- **Example:** Data is collected using barcode scanners in most supermarkets. This database, known as the "market basket" database, consists of a large number of records on past transactions. A single record lists all the items bought by a customer in one sale.

- **Cross marketing**

- It is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.

- **Catalog design**

- The selections of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related.

- **Apriori Algorithm – Frequent Pattern Algorithms**

- It is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for Boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent item sets are used to find k+1 item sets.

The steps followed in the Apriori Algorithm of data mining are:

1. **Join Step**: This step generates (K+1) item set from K-item sets by joining each item with itself.

2. **Prune Step**: **This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed.** This step is performed to reduce the size of the candidate item sets.

Steps in Apriori

- Apriori algorithm is a sequence of steps to be followed to find the most frequent item set in the given database.

- This data mining technique follows the join and the prune steps iteratively until the most frequent item set is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

Example

- Consider the following dataset and we will find frequent item sets and generate association rules for them.

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 2
minimum confidence is 60%

Step-1: K=1

1) Create a table containing support count of each item present in dataset – Called **C1(candidate set)**

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

2) Compare candidate set item's support count with minimum support count (here min_support=2 if support_count of candidate set items is less than min_support then remove those items). **This gives us itemset L1**.

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Step-2: K=2

1)

- Generate candidate set C2 using L1 (this is called join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common.

- Check all subsets of an itemset are frequent or not and if not frequent remove that itemset.(Example subset of{I1, I2} are {I1}, {I2} they are frequent.Check for each itemset)

- Now find support count of these itemsets by searching in dataset.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

2) compare candidate (C2) support count with minimum support count(here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L2.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

Step-3:

1)

- Generate candidate set C3 using L2 (join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common. So here, for L2, first element should match.
  So itemset generated by joining L2 is {I1, I2, I3}{I1, I2, I5}{I1, I3, i5}{I2, I3, I4}{I2, I4, I5}{I2, I3, I5}

- Check if all subsets of these itemsets are frequent or not and if not, then remove that itemset.(Here subset of {I1, I2, I3} are {I1, I2},{I2, I3},{I1, I3} which are frequent. For {I2, I3, I4}, subset {I3, I4} is not frequent so remove it. Similarly check for every itemset)

- Find support count of these remaining itemset by searching in dataset.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

2) Compare candidate (C3) support count with minimum support count(here min_support=2 if support_count of candidate set

item is less than min_support then remove those items) this gives us itemset L3.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Step-4:

- Generate candidate set C4 using L3 (join step). Condition of joining L_{k-1} and L_{k-1} (K=4) is that, they should have (K-2) elements in common. So here, for L3, first 2 elements (items) should match.

- Check all subsets of these itemsets are frequent or not (Here itemset formed by joining L3 is {I1, I2, I3, I5} so its subset contains {I1, I3, I5}, which is not frequent). So no itemset in C4

- We stop here because no frequent itemsets are found further

Confidence

- A confidence of 60% means that 60% of the customers, who purchased milk and bread also bought butter.

  Confidence(A->B)=Support_count(A∪B)/Support_count(A)
- So here, by taking an example of any frequent itemset, we will show the rule generation.
  Itemset {I1, I2, I3} //from L3

- **So rules can be**

[I1^I2]=>[I3] //confidence = sup(I1^I2^I3)/sup(I1^I2) = 2/4*100=50%

[I1^I3]=>[I2] //confidence = sup(I1^I2^I3)/sup(I1^I3) = 2/4*100=50%

[I2^I3]=>[I1] //confidence = sup(I1^I2^I3)/sup(I2^I3) = 2/4*100=50%

[I1]=>[I2^I3] //confidence = sup(I1^I2^I3)/sup(I1) = 2/6*100=33%

[I2]=>[I1^I3] //confidence = sup(I1^I2^I3)/sup(I2) = 2/7*100=28%

[I3]=>[I1^I2] //confidence = sup(I1^I2^I3)/sup(I3) = 2/6*100=33%

So if minimum confidence is 50%, then first 3 rules can be considered as strong association rules.

Examples of apriori algorithm

1)

Transaction	List of items
T1	I1,I2,I3
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3,I4

Support threshold=50%, Confidence= 60%

Support threshold=50% => (50/100)*6= 3 => min_sup=3

Minimum support count =3
Confidence =60%

Transaction	List of items
T1	I1,I2,I3

Transaction	List of items
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3,I4

1. Count Of Each Item
TABLE-2

Item	Count
I1	4
I2	5
I3	4
I4	4
I5	2

L1

Item	Count
I1	4
I2	5
I3	4
I4	4

3. Join Step: Form 2-itemset. From **TABLE-1** find out the occurrences of 2-itemset.
TABLE-4

Item	Count

Item	Count
I1,I2	4
I1,I3	3
I1,I4	2
I2,I3	4
I2,I4	3
I3,I4	2

L2

4. Prune Step: TABLE -4 shows that item set {I1, I4} and {I3, I4} does not meet min_sup, thus it is deleted.

TABLE-5

Item	Count
I1,I2	4
I1,I3	3
I2,I3	4
I2,I4	3

5. Join and Prune Step: Form 3-itemset. From the **TABLE- 1** find out occurrences of 3-itemset. From **TABLE-5**, find out the 2-itemset subsets which support min_sup.

We can see for itemset {I1, I2, I3} subsets, {I1, I2}, {I1, I3}, {I2, I3} are occurring in **TABLE-5** thus {I1, I2, I3} is frequent.

We can see for itemset {I1, I2, I4} subsets, {I1, I2}, {I1, I4}, {I2, I4}, {I1, I4} is not frequent, as it is not occurring in **TABLE-5** thus {I1, I2, I4} is not frequent, hence it is deleted.

TABLE-6

Item	
I1,I2,I3	**3**
I1,I2,I4	
I1,I3,I4	
I2,I3,I4	

Only {I1, I2, I3} is frequent

2)

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Given: Minimum Support= 2, Minimum Confidence= 50%

3)

Transaction ID	Items
T1	Hot Dogs, Buns, Ketchup
T2	Hot Dogs, Buns
T3	Hot Dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hot Dogs, Coke, Chips

Find the **frequent itemsets** and generate **association rules** on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%)

Let's start,

$$\text{minimum support count} = \frac{33.33}{100} \times 6$$
$$= 2$$