Big Data and Analytics

Prepared By: Prof. Preeti Sharma
Ph.D. Pursuing CE/IT (LDCE, GTU, Ahmedabad)
M.E. IT (LDCE, Ahmedabad)
B.E. CE (VGEC, Chandkheda, Ahmedabad)
D.E. CE (GPG, Ahmedabad)

Syllabus Covered:

1. Introduction to Big Data
2. Handling of Data
    a. Types of Data
    b. Knowledge Discovery Data (KDD)
        i. Data Cleaning
        ii. Data Integration
        iii. Data Selection
        iv. Data Transformation
        v. Data Mining
            1. Association Rule Mining
                a. Market Basket Analysis
        vi. Pattern Evaluation
        vii. Data Summarization Techniques using Plots

1. Introduction to Big Data

1.1 The Three V's of Big Data:

- The 3 V's (volume, velocity and variety) are three defining properties or dimensions of big data. Volume refers to the amount of data, velocity refers to the speed of data processing, and variety refers to the number of types of data.

Volume:

- Volume can be in Terabytes or Petabytes or Zettabytes.

Velocity:

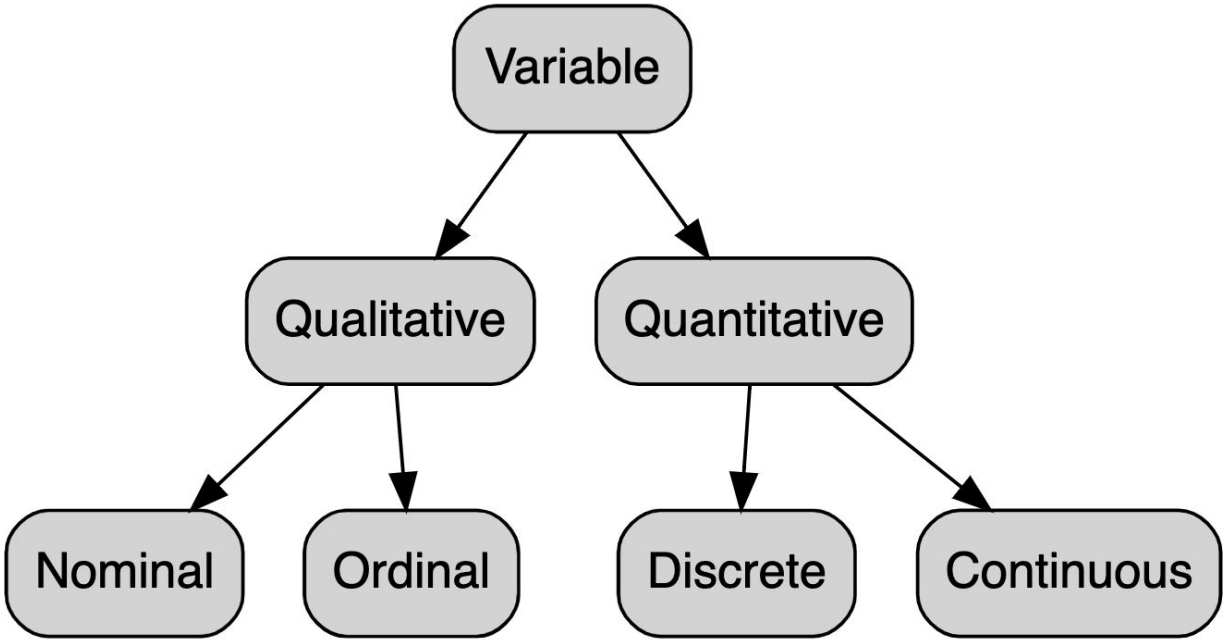
- Velocity essentially refers to the speed at which data is being created in real- time. We have moved from simple desktop applications to real- time processing applications.

Variety:

- Data can be structured data, semi-structured data and unstructured data. Data stored in a database is an example of structured data.HTML data, XML data, email data, Data in CSV files are the examples of semi-structured data. Powerpoint presentation, images, videos, researches, white papers, body of email etc are the examples of unstructured data.

2. Handling of Data

2.1 Types of Data:


Variable
Qualitative
Quantitative
Nominal
Ordinal
Discrete
Continuous

Qualitative or Categorical Data

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.

The Qualitative data are further classified into two parts :

Nominal Data

Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.

Examples of Nominal Data :

- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)
- Gender (Male, Female, Others)
- Eye Color (Black, Brown, etc.)

Ordinal Data

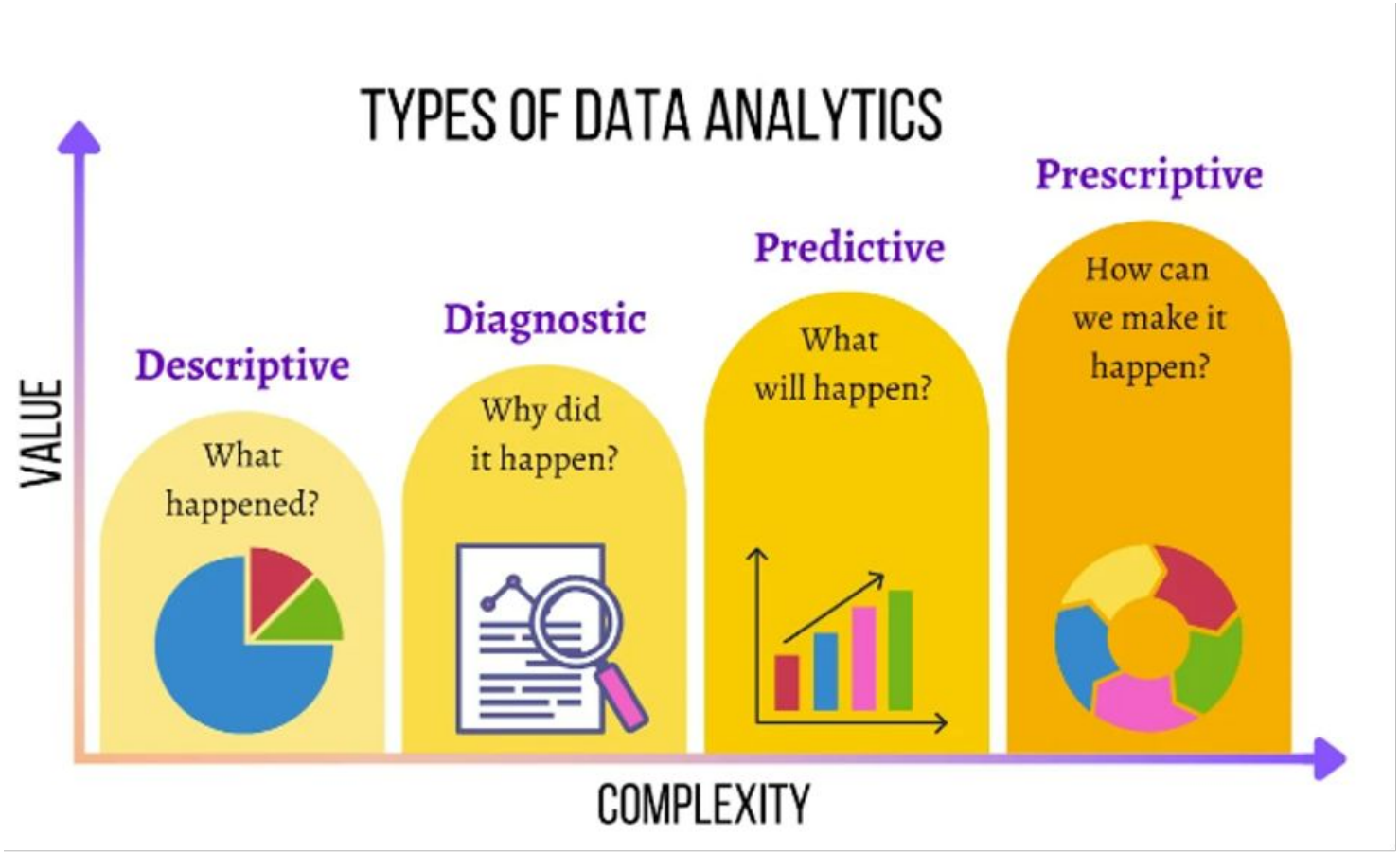
Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetic tasks on them.

Examples of Ordinal Data :

- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of people in a competition (First, Second, Third, etc.)
- Economic Status (High, Medium, and Low)
- Education Level (Higher, Secondary, Primary)

2.2 Types of Analytics:

2.2 Types of Analytics:



- The data analytics objective is to find insights and patterns in data that can help with operations, corporate decisions, and scientific research.

2.2.1 Descriptive Analysis:

- Descriptive Analysis is a method used to summarize and describe a data set's main features.
- "What happened" is answered by descriptive analytics.
- Tools that are commonly used in descriptive analysis include Excel, R, and Python.
- These tools have built-in functions that can be used to calculate summary statistics and create visualizations.
- The expensiveness of conducting a descriptive analysis depends on the data collection method.
- If the data is already available and accessible, the cost of conducting a descriptive analysis is usually minimal.

Examples of tasks that need descriptive analytics:

- Financial reports
- Survey reports
- Social media initiative

2.2.2 Diagnostic Analysis:

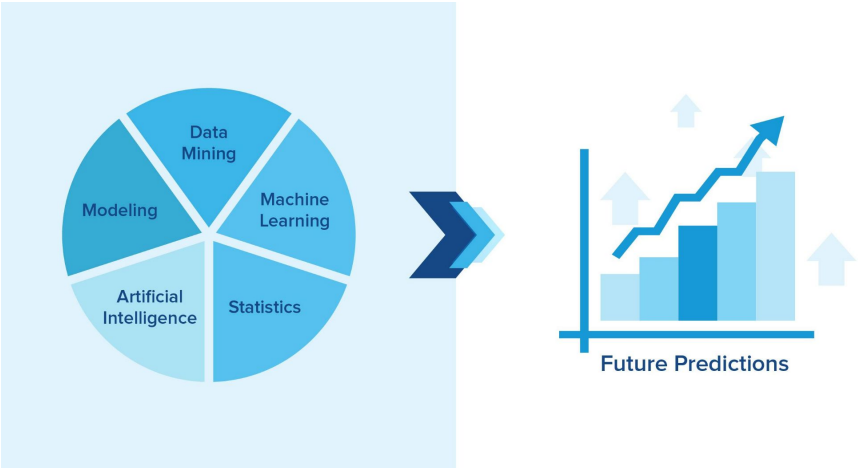
- Diagnostic analysis is a technique used to look into and pinpoint the root of a particular issue or problem. The subsequent logical inquiry, "**Why did this happen**?" is answered by diagnostic analytics.

- Tools that are commonly used in diagnostic analysis include Excel, R, and Python for data analysis, and specialized software like Minitab or JMP for statistical analysis.

Examples of tasks that need diagnostic analytics:

- Searching for patterns in the data groups
- Filtering of the data sets
- Probability theory
- Regression analysis

2.2.3 Predictive Analysis:



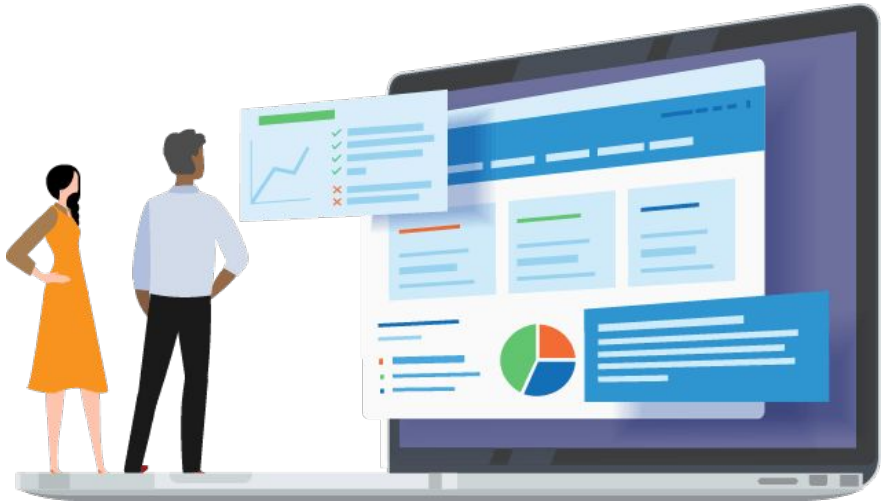
- In order to predict future trends or events or to provide a response to the question “What might occur in the future,” predictive analytics is used.

- Along with the commonly used SQL and Python, H2O Driverless AI, Microsoft Azure Machine learning , and IBM Watson Studio are some of the most used tools for model selection and semantic data analysis used in predictive analysis.

Examples of tasks that need predictive analytics:

- Predict the customer’s demands
- Handle shipping schedules
- Remain on top of inventory needs

2.2.4 Prescriptive Analysis:



- Prescriptive analytics is a form of business analytics which suggests decision options for how to take advantage of a future opportunity or mitigate a future risk, and shows the implication of each decision option.

- Prescriptive analytics finally responds to the query, “What should we do next?”. It recommends the course of action to be taken.

- Prescriptive analytics is the use of advanced processes and tools to analyze data and content to recommend the optimal course of action or strategy moving forward.

Examples of tasks that need prescriptive analytics:

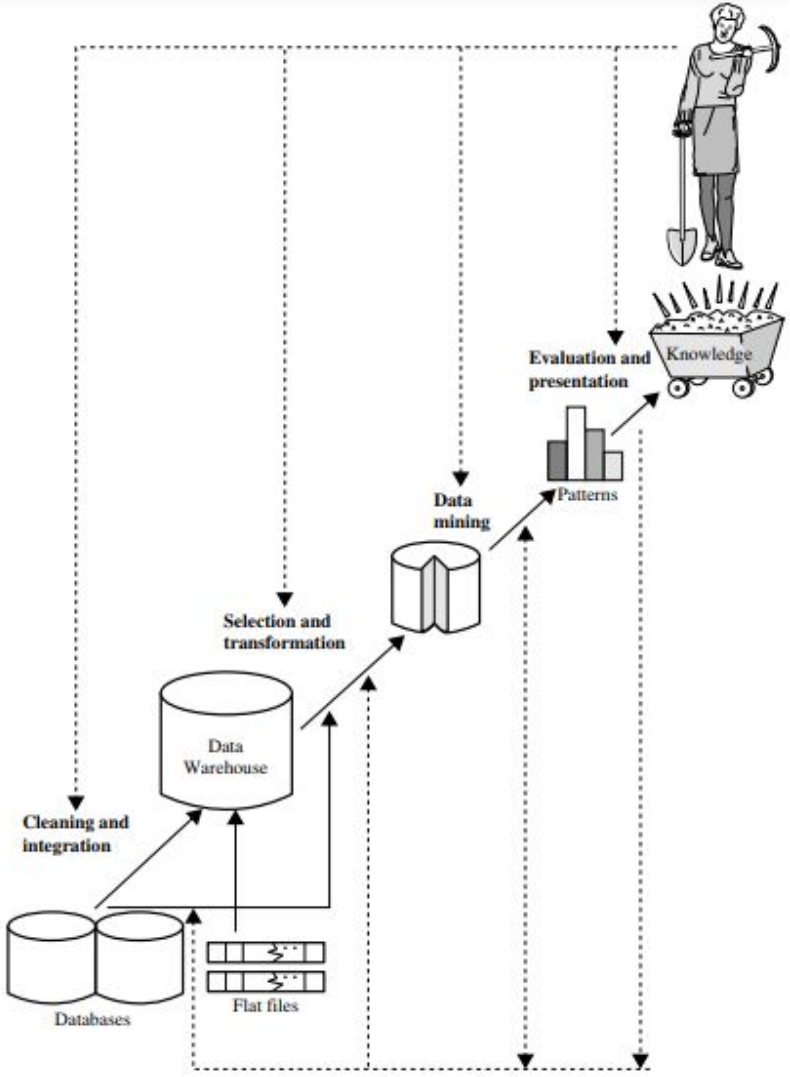
- Enhance processes
- Enable campaigns
- Steer production
- Facilitate customer services

2.3 Knowledge Discovery Data (KDD)

2.3 Knowledge Discovery Data (KDD):

- KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets.

- The knowledge discovery process is an iterative sequence of the following steps:
  1. **Data cleaning** (to remove noise and inconsistent data)
  2. **Data integration** (where multiple data sources may be combined)
  3. **Data selection** (where data relevant to the analysis task are retrieved from the database)
  4. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
  5. **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
  6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures.
  7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)


Evaluation and presentation
Knowledge
Patterns
Data mining
Selection and transformation
Data Warehouse
Cleaning and integration
Databases
Flat files

Data Preprocessing:

2.3.1 Data Preprocessing: Data Cleaning, Integration, Transformation, and Reduction

What is Data Preprocessing?

- Data preprocessing is the process of transforming raw data into an understandable format.
- It is also an important step in data mining as we cannot work with raw data.
- The quality of the data should be checked before applying machine learning or data mining algorithms.

Why to preprocess the Data?

- **Data Quality:** There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.
  - **Accuracy:** There are many possible reasons for inaccurate data (i.e., having incorrect attribute values).

    The data collection instruments used may be faulty.

    There may have been human or computer errors occurring at data entry.

    Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value “January 1” displayed for birthday). This is known as disguised missing data.

    Errors in data transmission can also occur.

    There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption.

    Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date).

    Duplicate tuples also require data cleaning.

Completeness: Incomplete data can occur for a number of reasons.

- Attributes of interest may not always be available, such as customer information for sales transaction data. Other data may not be included simply because they were not considered important at the time of entry.
- Relevant data may not be recorded due to a misunderstanding or because of equipment malfunctions.
- Data that were inconsistent with other recorded data may have been deleted.
- Furthermore, the recording of the data history or modifications may have been overlooked.
- Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

Timeliness also affects data quality.

- Suppose that you are overseeing the distribution of monthly sales bonuses to the top sales representatives at AllElectronics.
- Several sales representatives, however, fail to submit their sales records on time at the end of the month.
- There are also a number of corrections and adjustments that flow in after the month's end.
- For a period of time following each month, the data stored in the database are incomplete.
- However, once all of the data are received, it is correct.

- Two other factors affecting data quality are believability and interpretability.
- **Believability** reflects how much the data are trusted by users, while **interpretability** reflects how easy the data are understood.
- Suppose that a database, at one point, had several errors, all of which have since been corrected.

- The past errors, however, had caused many problems for sales department users, and so they no longer trust the data.
- The data also use many accounting codes, which the sales department does not know how to interpret.
- Even though the database is now accurate, complete, consistent, and timely, sales department users may regard it as of low quality due to poor believability and interpretability

Major Tasks in Data Preprocessing:

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction

2.3.1 Data Cleaning:

2.3.1 Data Cleaning:

Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

1. Cleaning in case of **Missing values**.
2. Cleaning **noisy** data, where noise is a random or variance error.
3. Cleaning with **Data discrepancy detection.**

1. Handling missing values

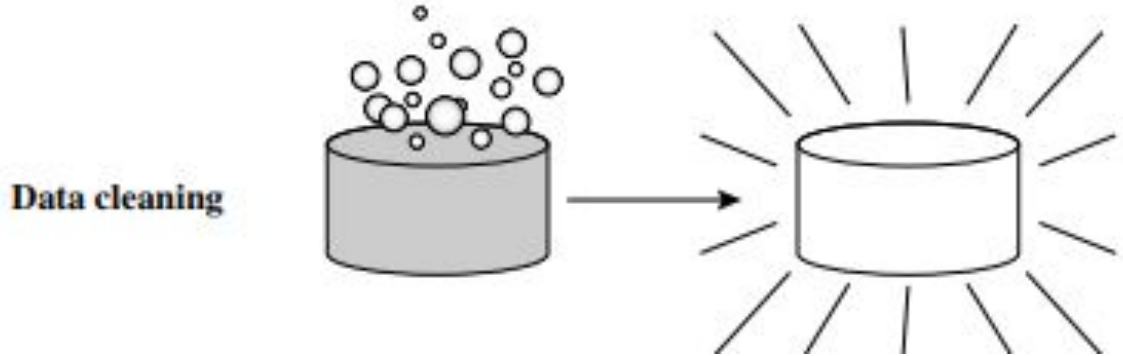
Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification).

This method is not very effective, unless the tuple contains several attributes with missing values.

It is especially poor when the percentage of missing values per attribute varies considerably.

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.



- **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.
- **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” or NA, Interpolate, Dropna

	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/4/2017	NaN	9.0	Sunny
2	1/5/2017	28.0	NaN	Snow
3	1/6/2017	NaN	7.0	NaN
4	1/7/2017	32.0	NaN	Rain
5	1/8/2017	NaN	NaN	Sunny
6	1/9/2017	NaN	NaN	NaN
7	1/10/2017	34.0	8.0	Cloudy
8	1/11/2017	40.0	12.0	Sunny

	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/4/2017	32.0	9.0	Sunny
2	1/5/2017	28.0	9.0	Snow
3	1/6/2017	28.0	7.0	Snow
4	1/7/2017	32.0	7.0	Rain
5	1/8/2017	32.0	7.0	Sunny
6	1/9/2017	32.0	7.0	Sunny
7	1/10/2017	34.0	8.0	Cloudy
8	1/11/2017	40.0	12.0	Sunny

Replacing with previous value – Forward fill

Replacing with next value – Backward fill

In backward fill, the missing value is imputed using the next value. It is mostly used in time series data.

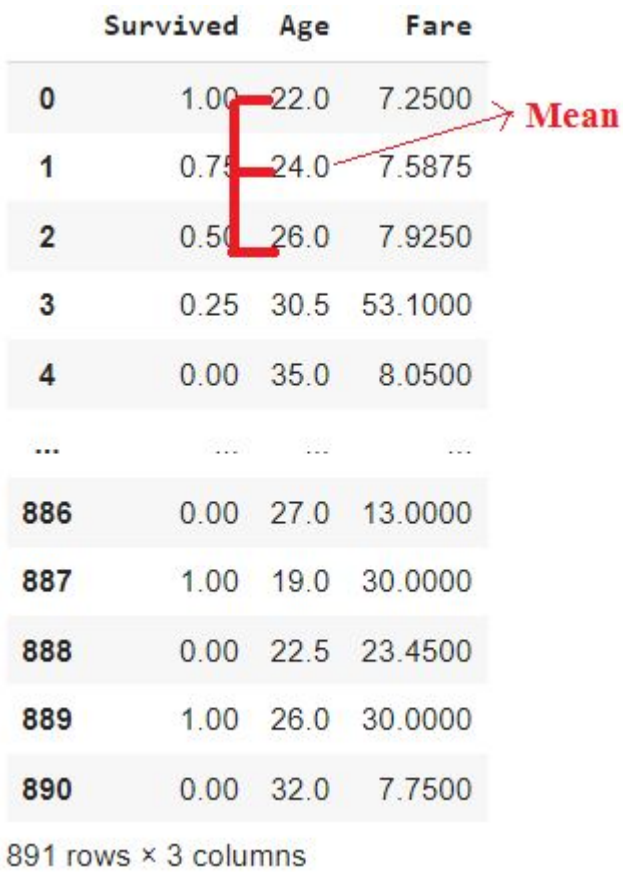
	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/4/2017	NaN	9.0	Sunny
2	1/5/2017	28.0	NaN	Snow
3	1/6/2017	NaN	7.0	NaN
4	1/7/2017	32.0	NaN	Rain
5	1/8/2017	NaN	NaN	Sunny
6	1/9/2017	NaN	NaN	NaN
7	1/10/2017	34.0	8.0	Cloudy
8	1/11/2017	40.0	12.0	Sunny

	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/4/2017	28.0	9.0	Sunny
2	1/5/2017	28.0	7.0	Snow
3	1/6/2017	32.0	7.0	Rain
4	1/7/2017	32.0	8.0	Rain
5	1/8/2017	34.0	8.0	Sunny
6	1/9/2017	34.0	8.0	Cloudy
7	1/10/2017	34.0	8.0	Cloudy
8	1/11/2017	40.0	12.0	Sunny

- **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.

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

- **Use the most probable value to fill in the missing value**



	Survived	Age	Fare
0	1.00	22.0	7.2500
1	0.75	24.0	7.5875
2	0.50	26.0	7.9250
3	0.25	30.5	53.1000
4	0.00	35.0	8.0500
...	...	...	...
886	0.00	27.0	13.0000
887	1.00	19.0	30.0000
888	0.00	22.5	23.4500
889	1.00	26.0	30.0000
890	0.00	32.0	7.7500

891 rows × 3 columns

2. Handling noisy data

Noisy generally means random error or containing unnecessary data points.

Handling noisy data is one of the most important steps as it leads to the optimization of the model we are using Here are some of the methods to handle noisy data.

- **Binning:** This method is to smooth or handle noisy data.
  Data binning, **bucketing** is a data pre-processing method used to minimize the effects of small observation errors.
  First, the data is sorted then, and then the sorted values are separated and stored in the form of bins.
  The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin.
  This has a smoothing effect on the input data and may also reduce the chances of overfitting in the case of small datasets.
  There are three methods for smoothing data in the bin.
  Smoothing by bin mean method: In this method, the values in the bin are replaced by the mean value of the bin;
  Smoothing by bin median: In this method, the values in the bin are replaced by the median value;
  Smoothing by bin boundary: In this method, the using minimum and maximum values of the bin values are taken, and the closest boundary value replaces the values.

Smoothing by Equal Width Binning Method:

Range Of Each Bin :

[min + W]

[min + 2W]

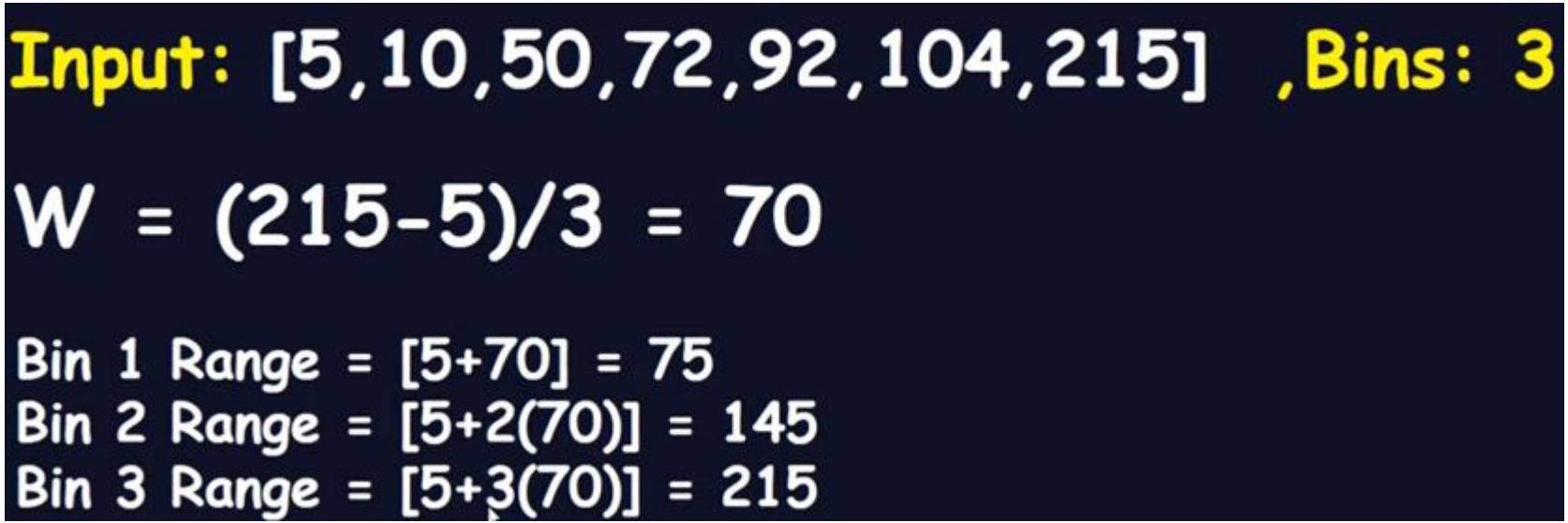
.....

[min + nW]

n : Number Of Bins

W = (max-min)/Number Of Bins

First sort the data, then distribute them in bins.



Unsorted data for price in dollars

Before sorting: 8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

First of all, sort the data

After Sorting: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

Smoothing the data by Equal Frequency:

Bin 1: 8, 9, 15, 16

Bin 2: 21, 21, 24, 26,

Bin 3: 27, 30, 30, 34

Smoothing the data 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 the data by Bin Boundaries:

Bin 1: 8, 9, 15, 16

Bin 2: 21, 21, 24, 26,

Bin 3: 27, 30, 30, 34

Before bin Boundary: Bin 1: 8, 9, 15, 16

Here, 1 is the minimum value and 16 is the maximum value.9 is near to 8, so 9 will be treated as 8. 15 is more near to 16 and farther away from 8. So, 15 will be treated as 16.

After bin Boundary: Bin 1: 8, 8, 16, 16

Before bin Boundary: Bin 2: 21, 21, 24, 26,

After bin Boundary: Bin 2: 21, 21, 26, 26,

Before bin Boundary: Bin 3: 27, 30, 30, 34

After bin Boundary: Bin 3: 27, 27, 27, 34

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:

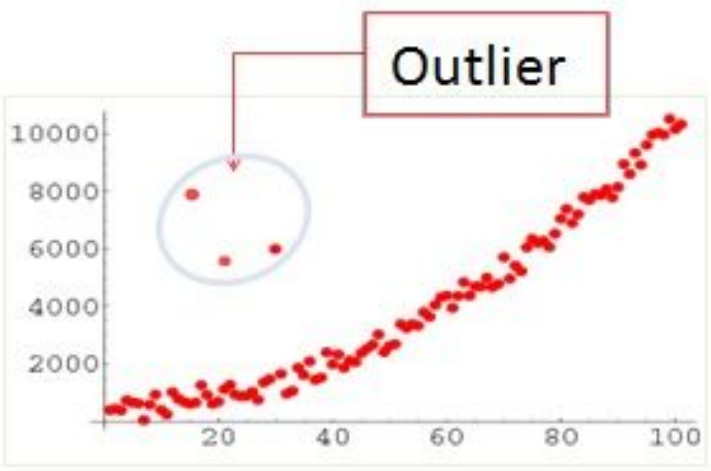
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

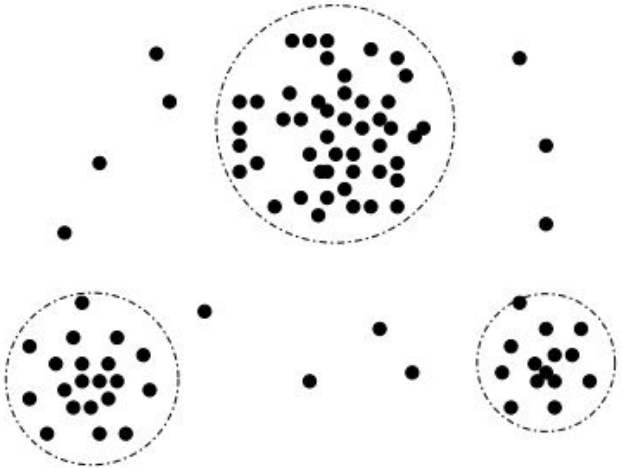
Smoothing by bin boundaries:

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

- **Outlier analysis:** 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.

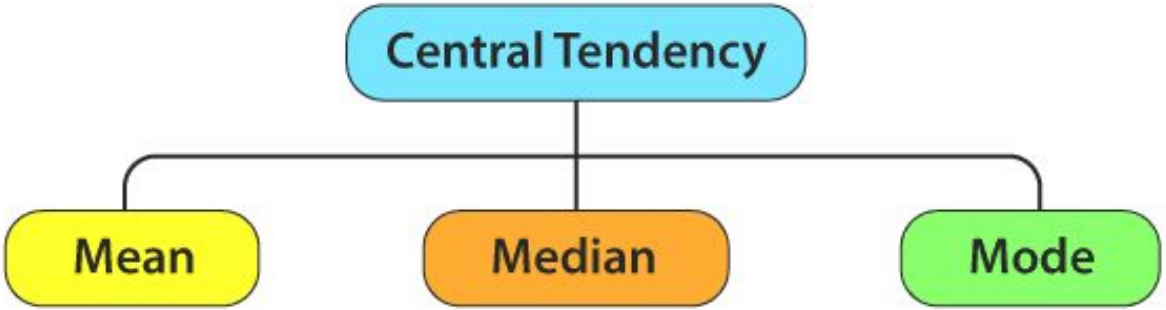
- **Clustering:** This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.





3. Cleaning with Data discrepancy detection:

- In statistics, the central tendency is the descriptive summary of a data set. Through the single value from the dataset, it reflects the centre of the data distribution.
- Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset.
- Generally, the central tendency of a dataset can be defined using some of the measures in statistics.



Measuring the Central Tendency: Mean, Median and Mode

Mean

- Mean: The most common and effective numeric measure of the "center" of a set of data is the (arithmetic) mean.
- The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values.
- In general, it is considered as the arithmetic mean.
- Let x1,x2,...,xN be a set of N values or observations, such as for some numeric attribute X, like salary. The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

Example: Mean. Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000.

Sometimes, each value xi in a set may be associated with a weight wi for i = 1,...,N. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

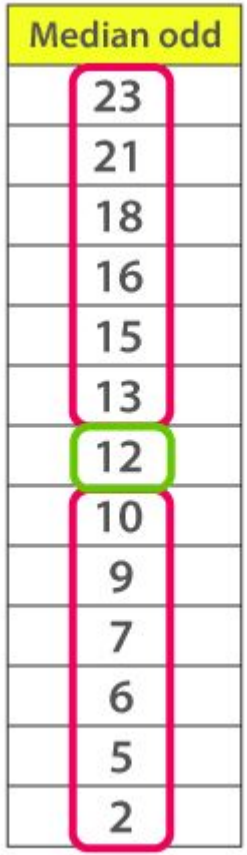
This is called the **weighted arithmetic mean** or the **weighted average.**

Median

- Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order.
- When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.
- The median is expensive to compute when we have a large number of observations. For numeric attributes, however, we can easily approximate the value.
- Assume that data are grouped in intervals according to their xi data values and that the frequency (i.e., number of data values) of each interval is known.
- Let the interval that contains the median frequency be the median interval. We can approximate the median of the entire data set (e.g., the median salary) by interpolation using the formula

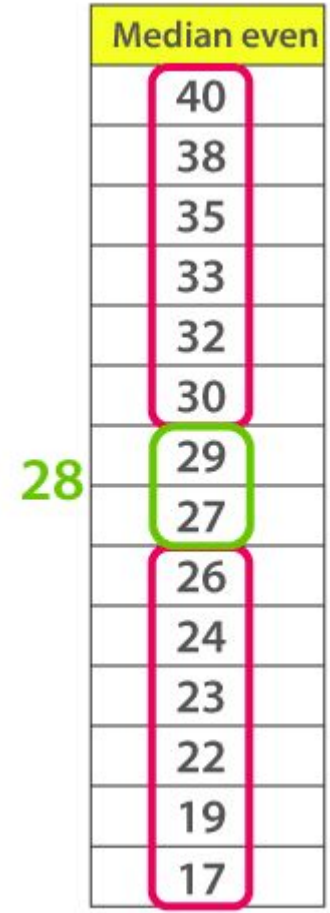
$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

Example: Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2.



Here 12 is the middle or median number that has 6 values above it and 6 values below it.

Now, consider another example with an even number of observations that are arranged in descending order – 40, 38, 35, 33, 32, 30, 29, 27, 26, 24, 23, 22, 19, and 17



- When you look at the given dataset, the two middle values obtained are 27 and 29.
- Now, find out the mean value for these two numbers.
- i.e.,(27+29)/2 =28
- Therefore, the median for the given data distribution is 28.

Mode

- The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.
- Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal. In general, a data set with two or more modes is multimodal.
- At the other extreme, if each data value occurs only once, then there is no mode

Example: Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5.

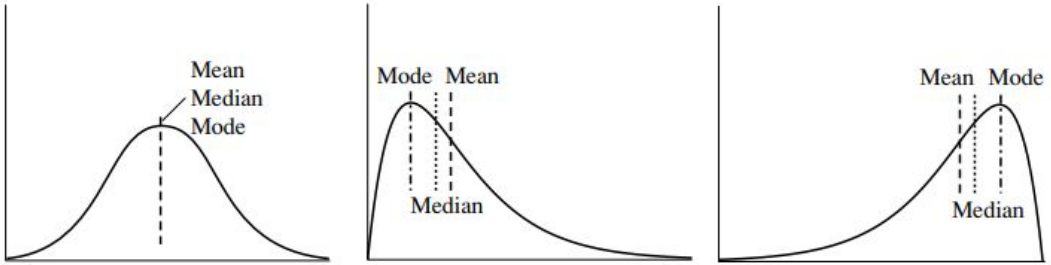
Mode
5
5
5
4
4
3
2
2
1

- Since the mode represents the most common value. Hence, the most frequently repeated value in the given dataset is 5.

- Data in most real applications are not symmetric. They may instead be either positively skewed, where the mode occurs at a value that is smaller than the median, or negatively skewed, where the mode occurs at a value greater than the median.
-

Based on the properties of the data, the measures of central tendency are selected.

- If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset.
- If you have skewed distribution, the best measure of finding the central tendency is the median.
- If you have the original data, then both the median and mode are the best choice of measuring the central tendency.
- If you have categorical data, the mode is the best choice to find the central tendency.



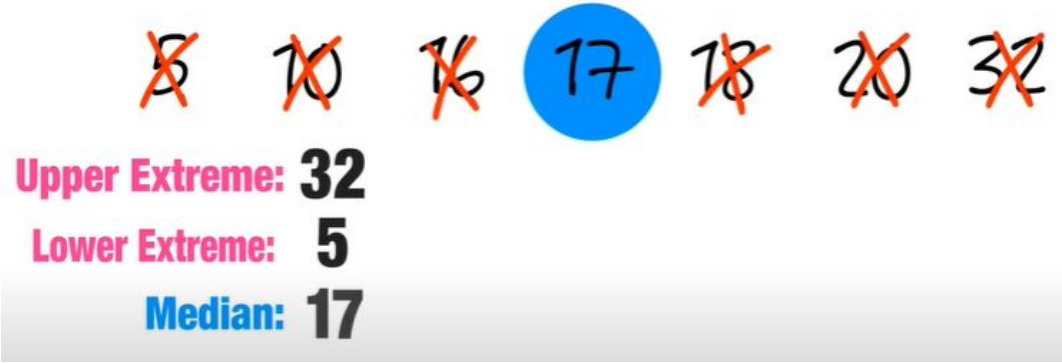
(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Measuring the Dispersion of Data: Range, Quartile, Interquartile Range, Variance, Standard Deviation

- Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.
- In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogeneous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

Range, Quartile, Interquartile / FIVE Number Summary: Visualization using Box & Whisker Plots:

- Boxplots are a popular way of visualizing a distribution.



Upper Extreme: 32

Lower Extreme: 5

Median: 17

Upper Quartile: 20

Lower Quartile: 10

Upper Extreme: 32

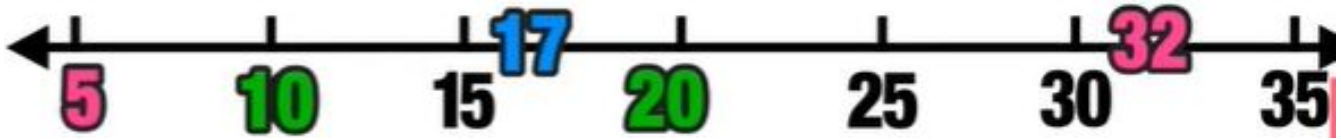
Lower Extreme: 5

Median: 17

Upper Quartile: 20

Lower Quartile: 10

-Points Scored Per Game-



Upper Extreme: 32

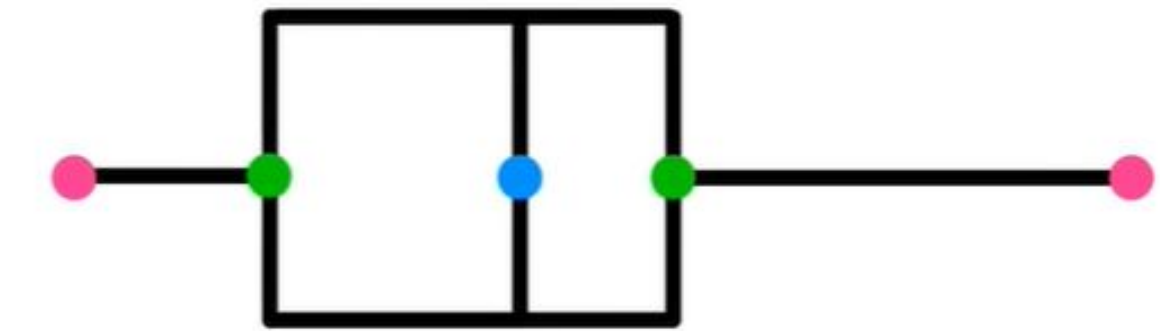
Lower Extreme: 5

Median: 17

Upper Quartile: 20

Lower Quartile: 10

-Points Scored Per Game-



Variance and Standard Deviation:

- Variance and standard deviation are measures of data dispersion.
- They indicate how spread out a data distribution is.
- A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.
- **Standard deviation** is the degree of dispersion or the scatter of the data points relative to its mean, in descriptive statistics. It tells how the values are spread across the data sample and it is the measure of the variation of the data points from the mean. The standard deviation of a data set, sample, statistical population, random variable, or probability distribution is the square root of its variance.
- When we have n number of observations and the observations are X1, X2,...Xn, then the mean deviation of the value from the mean is determined.
- However, the sum of squares of deviations from the mean doesn't seem to be a proper measure of dispersion. If the average of the squared differences from the mean is small, it indicates that the observations.
- are close to the mean.. This is a lower degree of dispersion. If this sum is large, it indicates that there is a higher degree of dispersion of the observations from the mean.

Standard Deviation Formula:

Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\bar{x}$ - The Sample Mean n - Total Number of Observations

There are two types of data sets: populations and samples. A population is an entire group that we are interested in studying, while a sample is a smaller group of individuals that is taken from the population. The formulas to calculate the standard deviations of population and sample differ a little.

Variance:

- Variance is a measure of dispersion. A measure of dispersion is a quantity that is used to check the variability of data about an average value.
- Data can be of two types - grouped and ungrouped. When data is expressed in the form of class intervals it is known as grouped data. On the other hand, if data consists of individual data points, it is called ungrouped data.
- The sample and population variance can be determined for both kinds of data

.

Population Variance - All the members of a group are known as the population. When we want to find how each data point in a given population varies or is spread out then we use the population variance. It is used to give the squared distance of each data point from the population mean.

Sample Variance - If the size of the population is too large then it is difficult to take each data point into consideration. In such a case, a select number of data points are picked up from the population to form the sample that can describe the entire group. Thus, the sample variance can be defined as the average of the squared distances from the mean. The variance is always calculated with respect to the sample mean.

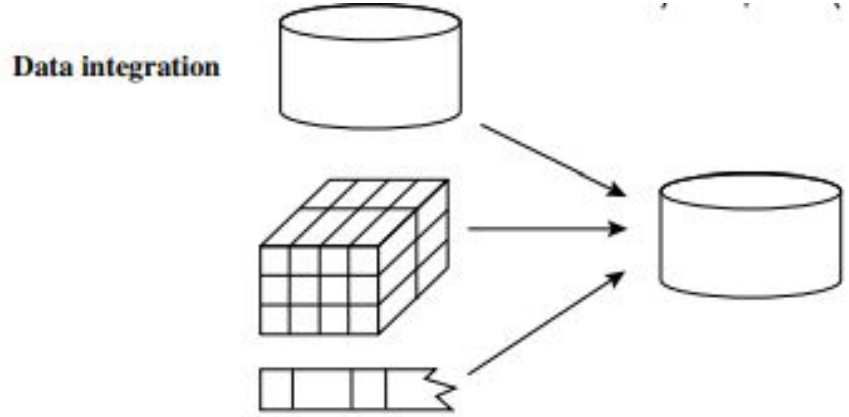
- A general definition of variance is that it is the expected value of the squared differences from the mean.

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

2.3.2 Data Integration:

2.3.2 Data Integration:

- Data integration is defined as heterogeneous data from multiple sources combined in a common source(Data Warehouse).
- Detecting and resolving data value concepts: The data taken from different databases while merging may differ. The attribute values from one database may differ from another database. For example, the date format may differ, like “MM/DD/YYYY” or “DD/MM/YYYY”.



Data Integration using ETL Process:

- ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse.

The process of ETL can be broken down into the following three stages:

1.

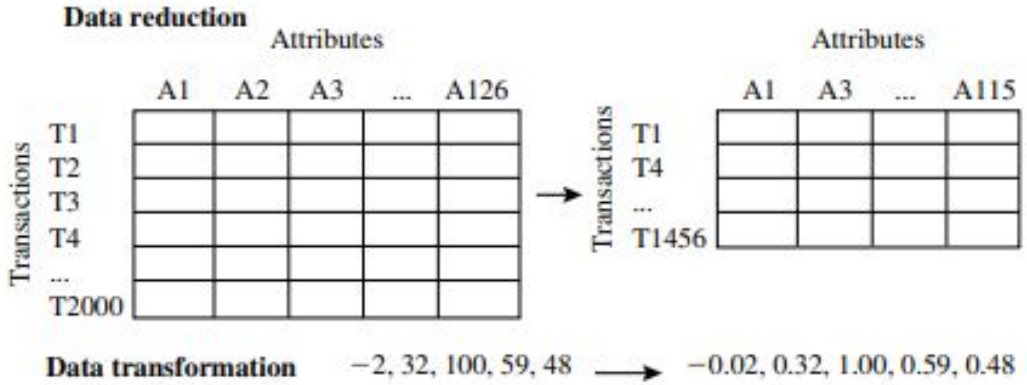
1. **Extract**: The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area.
2. **Transform**: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.
3. **Load**: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.

2.3.3 Data Selection:

2.3.3 Data Reduction:

This process helps in the reduction of the volume of the data, which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space. Some of the data reduction techniques are dimensionality reduction, numerosity reduction, and data compression.

- **Dimensionality reduction**: In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced.
  Combining and merging the attributes of the data without losing its original characteristics.
  This also helps in the reduction of storage space, and computation time is reduced.



- **Numerosity Reduction**: In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.

- **Data compression:** The compressed form of data is called data compression.
  This compression can be lossless or lossy.
  When there is no loss of information during compression, it is called lossless compression; Whereas lossy compression reduces information, but it removes only the unnecessary information.

2.3.4 Data Transformation:

2.3.4 Data Transformation:

The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods for data transformation.

- **Smoothing:** With the help of algorithms, we can remove noise from the dataset, which helps in knowing the important features of the dataset. By smoothing, we can find even a simple change that helps in prediction.

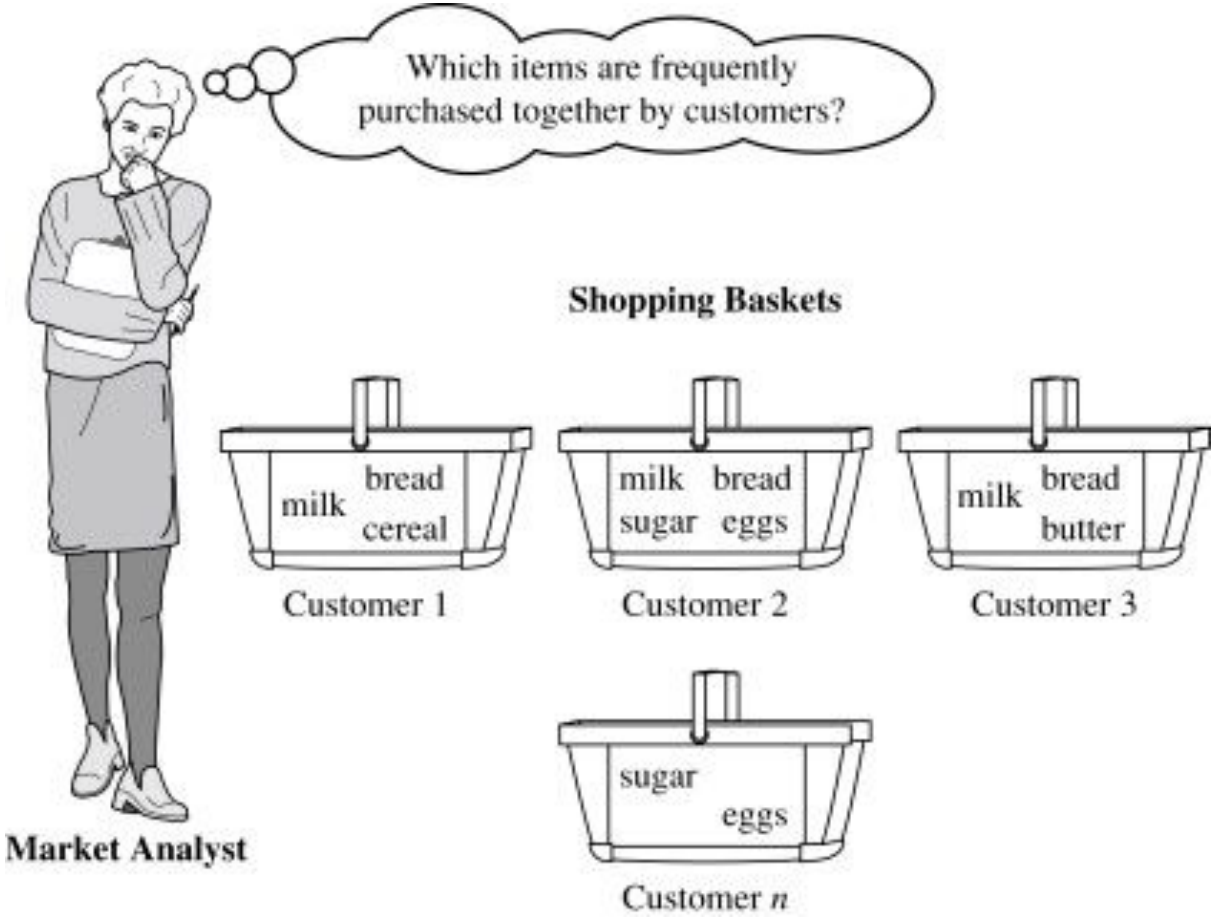
- **Aggregation:** In this method, the data is stored and presented in the form of a summary.
  The data set, which is from multiple sources, is integrated into with data analysis description.
  This is an important step since the accuracy of the data depends on the quantity and quality of the data.
  When the quality and the quantity of the data are good, the results are more relevant.

- **Discretization:** The continuous data here is split into intervals. Discretization reduces the data size. For example, rather than specifying the class time, we can set an interval like (3 pm-5 pm, or 6 pm-8 pm).

- **Normalization:** It is the method of scaling the data so that it can be represented in a smaller range.
  Example ranging from -1.0 to 1.0.

2.3.5 Data Mining:

2.3.5 Market Basket Analysis:



- In market basket analysis (also called association analysis or frequent itemset mining), you analyze **purchases that commonly happen together**. For example, people who buy bread and peanut butter also buy jelly. Or people who buy shampoo might also buy conditioner.
- What relationships there are between items is the target of the analysis. Knowing what your customers tend to buy together can help with marketing efforts and store/website layout.

- **Market basket analysis** is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.
- This can be done using **Association Rule Mining.**
- If can tell you what items do customers frequently buy together by generating a set of rules called **Association Rules**. In simple words, it gives you output as rules in form **if this then that**.

When you apply Association Rule Mining on a given set of transactions T your goal will be to find all rules with:

1. Support greater than or equal to min_support
2. Confidence greater than or equal to min_confidence

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

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

Example - 1:

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 confidence threshold (c = 60%)

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

Item set	Sup-count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

Item set	Sup-count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

Item set	Sup-count
Hot Dogs, Buns	2
Hot Dogs, Ketchup	1
Hot Dogs, Coke	2
Hot Dogs, Chips	2
Buns, Ketchup	1
Buns, Coke	0
Buns, Chips	0
Ketchup, Coke	0
Ketchup, Chips	1
Coke, Chips	3

Item set	Sup-count
Hot Dogs, Buns	2
Hot Dogs, Coke	2
Hot Dogs, Chips	2
Coke, Chips	3

Item set	Sup-count
Hot Dogs, Buns, Coke	0
Hot Dogs, Buns, Chips	0
Hot Dogs, Coke, Chips	2

Item set	Sup-count
Hot Dogs, Coke, Chips	2

There is only one itemset with minimum support 2.

So only one itemset is frequent.

Frequent Itemset (I) = {Hot Dogs, Coke, Chips}

Association rules,

[Hot Dogs^Coke]=>[Chips] //confidence = sup(Hot Dogs^Coke^Chips)/sup(Hot Dogs^Coke)

= 2/2*100=100% //**Selected**

[Hot Dogs^Chips]=>[Coke] //confidence = sup(Hot Dogs^Coke^Chips)/sup(Hot Dogs^Chips)

= 2/2*100=100% //**Selected**

[Coke^Chips]=>[Hot Dogs] //confidence = sup(Hot Dogs^Coke^Chips)/sup(Coke^Chips)

= 2/3*100=66.67% //**Selected**

[Hot Dogs]=>[Coke^Chips] //confidence = sup(Hot Dogs^Coke^Chips)/sup(Hot Dogs)

= 2/4*100=50% //**Rejected**

[Coke]=>[Hot Dogs^Chips] //confidence = sup(Hot Dogs^Coke^Chips)/sup(Coke)

= 2/3*100=66.67% //**Selected**

[Chips]=>[Hot Dogs^Coke] //confidence = sup(Hot Dogs^Coke^Chips)/sup(Chips)

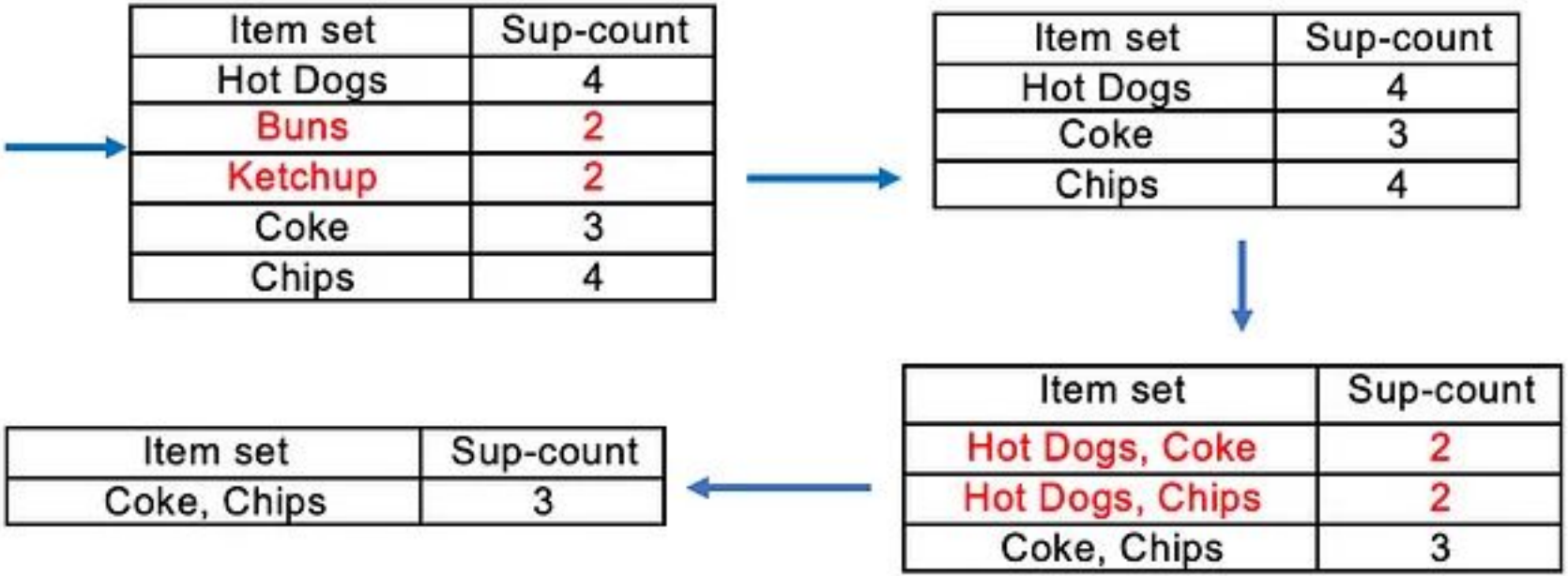
= 2/4*100=50% //**Rejected**

There are four strong results (minimum confidence greater than 60%)

Example - 2:

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** on this. Assume that minimum support (s = 3)



There is only one itemset with minimum support 3. So only one itemset is frequent.

Frequent Itemset (I) = {Coke, Chips}

Example - 3:

Transaction ID	Items
100	I1,I2
200	I2,I3,I4,I5
300	I2,I3
400	I1
500	I1,I2,I3

Market Basket Transactions.
Items labeled as I1,I2 and so on

Given are minimum support count and minimum confidence threshold
min_sup=2
min_confidence=50%

C1

Itemset	Support Count
I1	3
I2	4
I3	3
I4	1
I5	1

1. First you will start with all individual items called candidates and calculate their support counts. This is called **candidate list generation.**

L1

Itemset	Support Count
I1	3
I2	4
I3	3

2. Remove candidates that fail min_sup count. **I4** and **I5** fail min_sup=2. The list is now called **L1** containing the **frequent item sets.**
Here we have used **APRIORI principle:**
No supersets of infrequent itemset must be generated and tested.

C2= all possible 2-itemset combinations

Itemset	Support Count
I1,I2	2
I1,I3	1
I2,I3	3
I3,I1	1

3. Generate second Candidate list by L1 cross join L1. And note support counts. {I1,I2} appear in 2 transactions together.

L2

Itemset	Support Count
I1,I2	2
I2,I3	3

4. Remove candidates that fail min_sup count.

C3= all possible 2-itemset combinations

Itemset	Support Count
I1,I2,I3	1

5. Generate third Candidate list by L2 cross join L2. And note support counts. {I1,I2,I3} appear in only 1 transactions together.

6. L3 is null. L3={} since Support count for {I1,I2,I3} fails min_sup. Here First step of Association rule mining is completed and there will be no C4 candidate list

Transaction ID	Items
100	I1,I2
200	I2,I3,I4,I5
300	I2,I3
400	I1
500	I1,I2,I3

L2

Itemset	Support Count
I1,I2	2
I2,I3	3

Non-empty Subsets for {I1,I2}
I2
I1

Association Rules	Confidence
{I2}=>I1	2/4=50% ✓
{I1}=>I2	2/3=66% ✓

Green ticked ✓ are all strong rules satisfying **min_confidence>=50%** and **min_sup count=2**

Non-empty Subsets for {I2,I3}
I2
I3

Association Rules	Confidence
{I2}=>I3	3/4=75% ✓
{I3}=>I2	3/3=100% ✓

Transaction ID	Items
100	I1,I2
200	I2,I3,I4,I5
300	I2,I3
400	I1
500	I1,I2,I3

Given Min_confidence=50%. Confidence is calculated by:

C(A=>B)=P(AUB)/P(A)=n(AUB)/n(A)

Confidence is number of times A and B are together in all transactions containing A

Attribute Oriented Induction:

- Attribute Oriented Induction (AOI) is a data mining algorithm used for extracting knowledge of relational data, taking into account expert knowledge.
- Attribute-Oriented Induction (AOI) is a descriptive database mining technique, which compresses the original set of data into a generalized relation, providing concise and summative information about the massive set of the original data.
- It is an online data analysis, query oriented and generalization based approach.
- In this approach, we perform generalization on basis of different values of each attributes within the relevant data set.
- After that same tuple are merged and their respective counts are accumulated in order to perform aggregation.

How it is done?

1. **Data Focusing:** Collect the task-relevant data(initial relation) using a relational database query.

2. **Attribute Removal:** Perform generalization by attribute removal or attribute generalization. Attribute removal is based on the following rule: If there is a large set of distinct values for an attribute of the initial working relation, but either (case 1) there is no generalization operator on the attribute, or (case 2) its higher-level concepts are expressed in terms of other attributes, then the attribute should be removed.

3. **Attribute Aggregation:** Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.

   a. Attribute Generalization Threshold control Technique

   b. Generalized Relation Threshold Control Technique

4. Reduces the size of the generalized data set.

5. Interactive presentation with users.

Example:

Let's say there is a University database that is to be characterized, for that its corresponding DMQL will be

use University_DB
mine characteristics as “Science_Students”
in relevance to name, gender, major, birth_place, birth_date, residence, phone_no, GPA
from student

STEP -1: Data Focusing: The data mining query presented above is transformed into the following relational query for the collection of the task-relevant data set.

Select name, gender, major, birth_place, birth_date, residence, phone_no, GPA
from student
where status in {“Msc”, “MBA”, “Ph.D.” }

The transformed query is executed against the relational database,. This table is called the **(task-relevant) initial working relation.**

STEP -2 : Attribute Removal:

From this table, we are querying task-relevant data.

From this table, we also removed a few attributes like name and phone_no, because they make no sense in concluding insights.

name: Since there are a large number of distinct values for name and there is no generalization operation defined on it, this attribute is removed.

Phone_no: This contains too many distinct values and should therefore be removed in generalization.

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Gender: Since there are only two distinct values for gender, this attribute is retained and no generalization is performed on it.

Birth_Place: This attribute has a large number of distinct values; therefore, we would like to generalize it. Suppose that a concept hierarchy exists for Birth_Place defined as “city < province_or_state < country”. If the number of distinct values for country in the initial working relation is greater than the attribute generalization threshold, then birth_place should be removed, because even though a generalization operator exists for it, the generalization threshold would not be satisfied.

Birth_date: Suppose that a hierarchy exists that can generalize birth_date to age and age to age_range and that the number of rage ranges (intervals) is small with respect to the attribute generalization threshold.

Residence: Suppose that residence is defined by the attributes number, street, residence_city, state, and country. Then the attribute number and street should be removed so that residence is then generalized to city, which contains fewer distinct values.

GPA: The concept hierarchy exists for gpa that groups values for grade point average into numeric intervals {6-7,7-8-8-9}, which in turn are grouped into descriptive values such as {“Excellent”,”Very Good”}. The attribute therefore can be generalized.

STEP -3 : Attribute Generalization:

The generalization process will result in groups of identical tuples.

For example, the tuples with the CS and Physics are merged together with one single field “Major”.

Such identical tuples are then merged into one, with their counts accumulated.

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

STEP -3 : Attribute Generalization:

Now, from the final result we can perform OLAP operations; we may view count() as a measure, and the remaining attributes as dimensions.

Birth_Region Gender	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

References:

[1] http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Thank You