BUSINESS INTELLIGENCE & ANALYTICS

P.RAJESH

CONTENTS ….

- ..\syllabus.docx

• Why is Business Intelligence

CHAPTER OBJECTIVES:

• Define Business Analytics.
• The Relationship Of Analytics And Business Intelligence To The Subject Of Business Analytics.
• The Three Steps Of The Business Analytics Process.
• Describe Four Data Classification Measurement Scales.
• Explain The Relationship Of The Business Analytics Process With The Organization Decision-making Process.

BIA TERMS ….

- What is it?
- Why is it important?
- How do you do it?

TERMINOLOGY

- **Business analytics** begins with a *data set* (a simple collection of data or a data file) or commonly with a *database* (a collection of data files that contain information on people, locations, and so on).
- As **databases** grow, they need to be stored somewhere. Technologies such as *computer clouds* (hardware and software used for data remote storage, retrieval, and computational functions)
- ***Data warehousing*** (a collection of databases used for reporting and data analysis) store data. Database storage areas have become so large that a new term was devised to describe them.
- ***Big data*** describes the collection of data sets that are so large and complex that software systems are hardly able to process them

- Define ***little data*** as anything that is not big data.
  - Little data describes the smaller data segments or files that help individual businesses keep track of customers. As a means of sorting through data to find useful information, the application of anal
- Three terms in business literature are often related to one another: analytics, business analytics, and business intelligence.
  - *Analytics* can be defined as a process that involves the use of statistical techniques (measures of central tendency, graphs, and so on), information system software (data mining, sorting routines), and operations research methodologies (linear programming) to explore, visualize, discover, and communicate patterns or trends in data.
  - Simply, analytics converts data into useful information. Analytics is an older term commonly applied to all disciplines, not just business. A typical example of the use of analytics is the weather measurements collected and converted into statistics, which in turn predict weather patterns.
  - There are many types of analytics, and there is a need to organize these types to understand their uses. We will adopt the three categories (*descriptive* , *predictive* , and *prescriptive*)

Type of Analytics	Definition
Descriptive	The application of simple statistical techniques that describe what is contained in a data set or database. Example: An age bar chart is used to depict retail shoppers for a department store that wants to target advertising to customers by age.
Predictive	An application of advanced statistical, information software, or operations research methods to identify predictive variables and build predictive models to identify trends and relationships not readily observed in a descriptive analysis. Example: Multiple regression is used to show the relationship (or lack of relationship) between age, weight, and exercise on diet food sales. Knowing that relationships exist helps explain why one set of independent variables influences dependent variables such as business performance.
Prescriptive	An application of decision science, management science, and operations research methodologies (applied mathematical techniques) to make best use of allocable resources. Example: A department store has a limited advertising budget to target customers. Linear programming models can be used to optimally allocate the budget to various advertising media.

Types of Data Analysis

Types of Data Analysis

Descriptive

- Aims to help uncover valuable insight from the data being analyzed
- Answers the question **"What happened?"**

Predictive

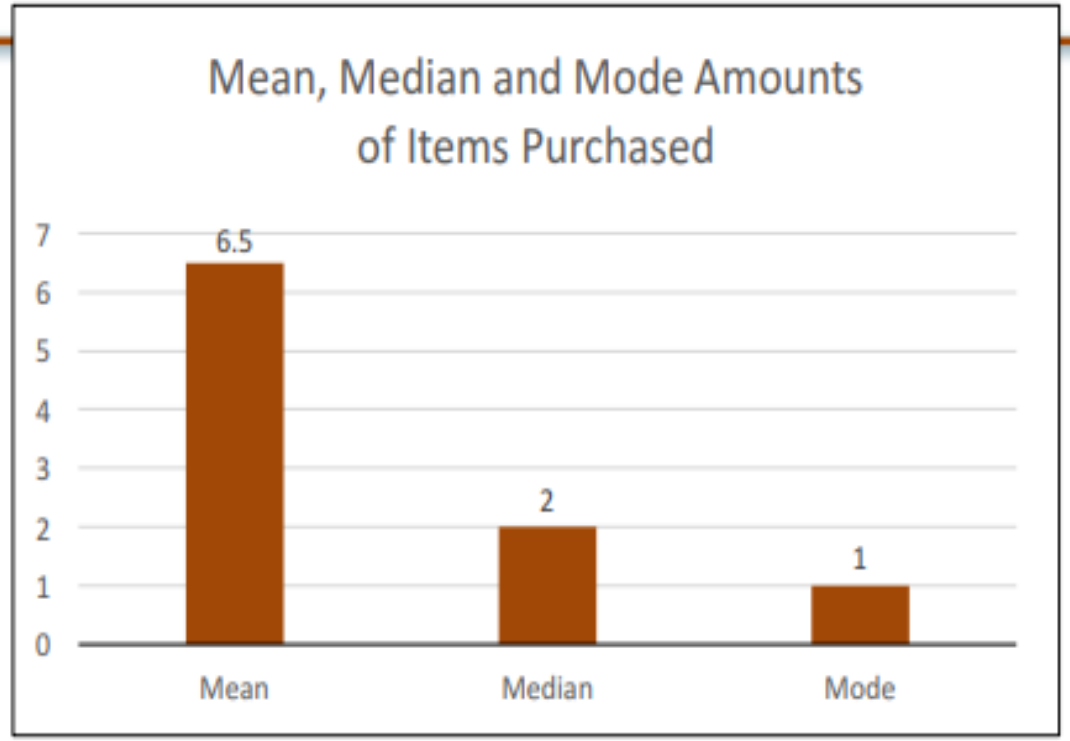
- Helps forecast behavior of people and markets
- Answers the question **"What could happen?"**

Prescriptive

- Suggests conclusions or actions that may be taken based on the analysis
- Answers the question **"What should be done?"**

Descriptive Data Analytics

- Though the most simple type, it is used most often.
- Two types of descriptive analysis:
  1. Measures of central tendency (tells us about the middle)
     - Mean – the average
     - Median – the midpoint of the responses
     - Mode – the response with the highest frequency
  2. Measures of dispersion
     - Range – the min, the max and the distance between the two
     - Variance – the average degree to which each of the points differ from the mean
     - Standard Deviation – the most common/standard way of expressing the spread of data



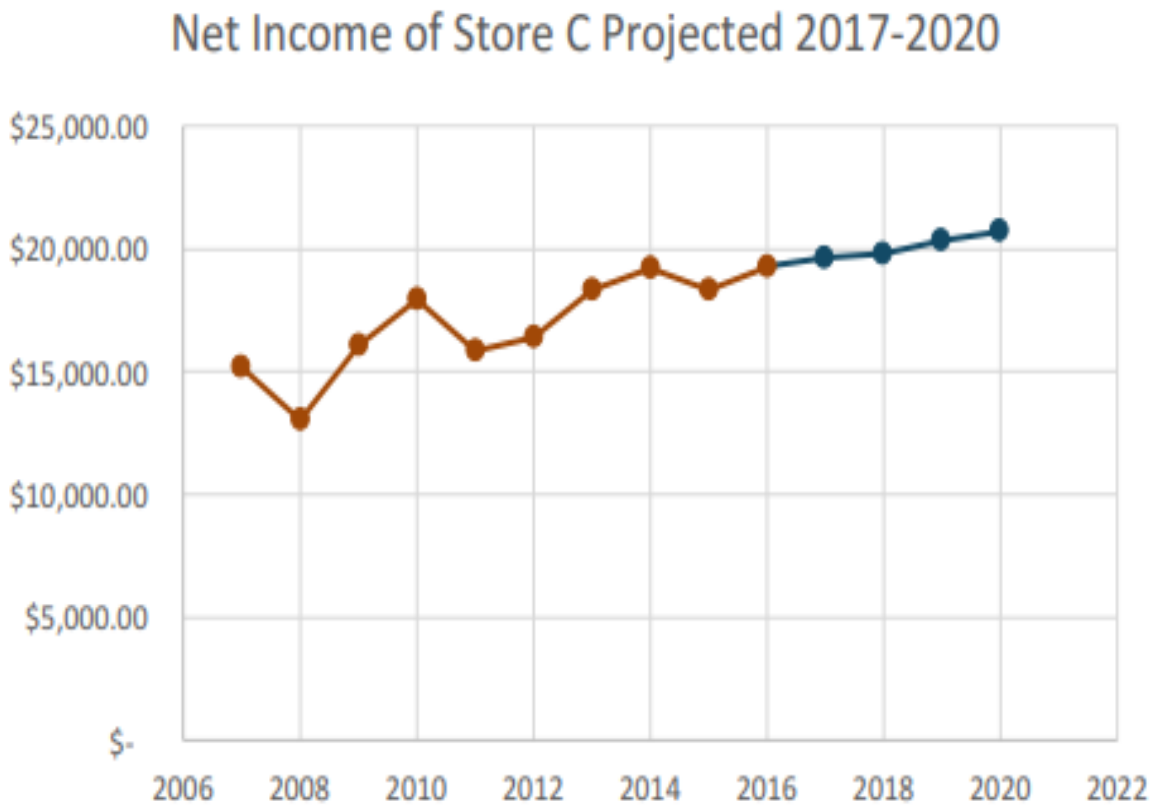
Customer_ID	Items Purchased	Amount Spent
29304	1	$ 1.09
28308	3	$ 44.43
19962	21	$ 218.58
30281	1	$ 73.02

Predictive Data Analytics

- Some mistake predictive analysis to have exclusive relevance to predicting *future* events.
  - However, in cases such as sentiment analysis, existing data (e.g., the text of a tweet) is used to predict non-existent data (whether the tweet is positive or negative).
- Several of the models that can be used for predictive analysis are:
  - Forecasting
  - Simulation
  - Regression
  - Classification
  - Clustering

Predictive Forecasting

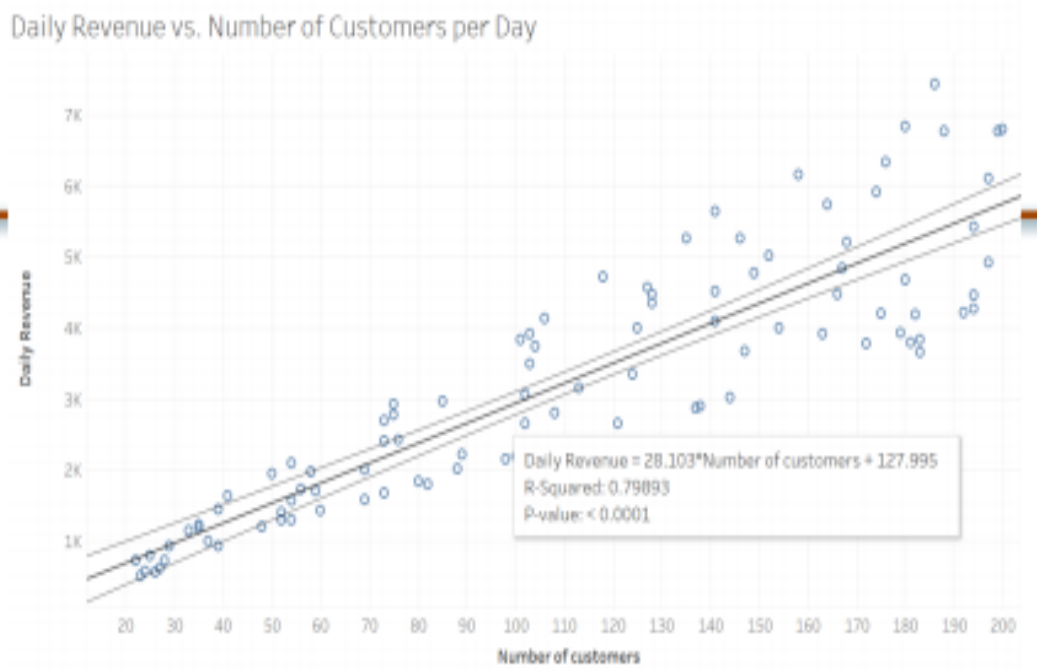
- Forecasting:
  - Moving average technique: use the mean of prior periods to predict the next
    - The mean of periods 1–4 = period 5
    - The mean of periods 2–5 = period 6
  - Exponential smoothing technique: similar, but more recent data points are weighted more heavily due to relevance
  - Regression techniques
- Use caution in forecasting – The larger the forecasted time period, the less accuracy there is in the projections.



Predictive Simulation

- Simulation
  - Queuing models: used to predict wait time and queue length
    - Results can be used to create staff schedules in a way that reduces inefficiencies, etc.
  - Discrete event model: used in special situations when queuing cannot be used
    - Results can be used to identify bottlenecks, etc.
  - Monte Carlo simulations: used to identify probable outcomes of a scenario based on many possible outcomes (uses random number generation and many iterations of the scenario).
    - Results can be used to predict the likelihood of profitability within the first two years, etc.

Predictive Regression

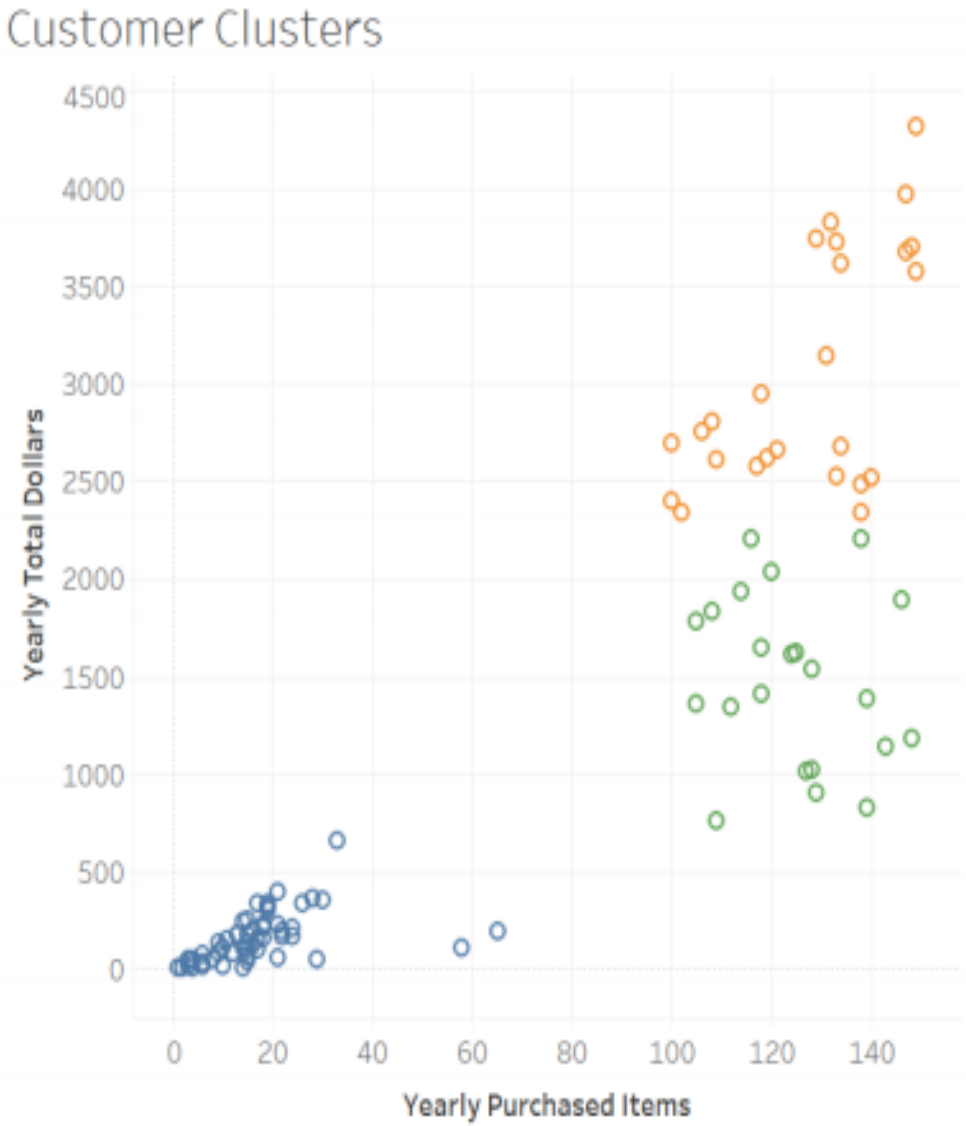


- Regression – generally speaking, used to understand the correlation of independent and dependent variables
- Types of regression models:
  - Logistic: used for categorical variables (i.e., will customers shop at your store or a competitor?)
  - Linear: used to identify a linear relationship between the dependent variable and at least one independent variables (i.e., daily store revenue predicted by the number of customers entering the store)
  - Step-wise: used to identify a relationship between dependent/independent variables. This is done by adding/removing variables based on how those variables impact the overall strength of the model.

Predictive

Classification & Clustering

- Classification: used to assign objects to one of several categories
  - Sentiment analysis of social media postings
- Clustering: another method of forming groups
  - Intragroup differences are minimized
  - Intergroup differences are maximized
  - Commonly used to create and better understand customer groups

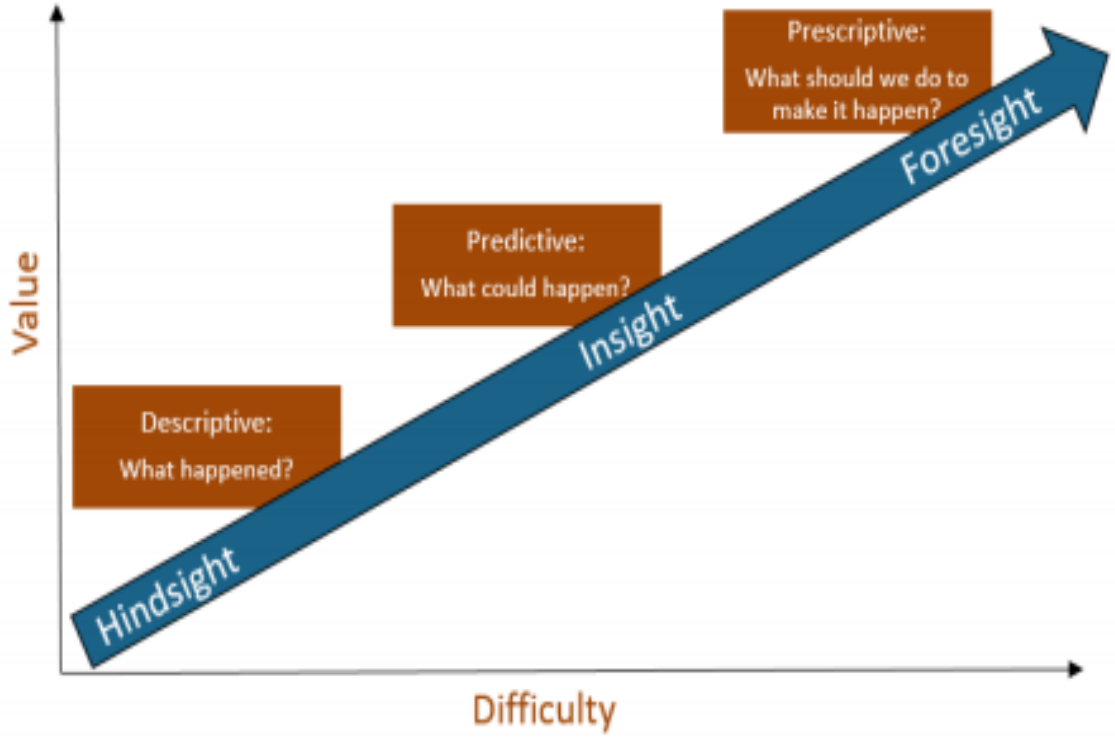


Prescriptive Data Analytics

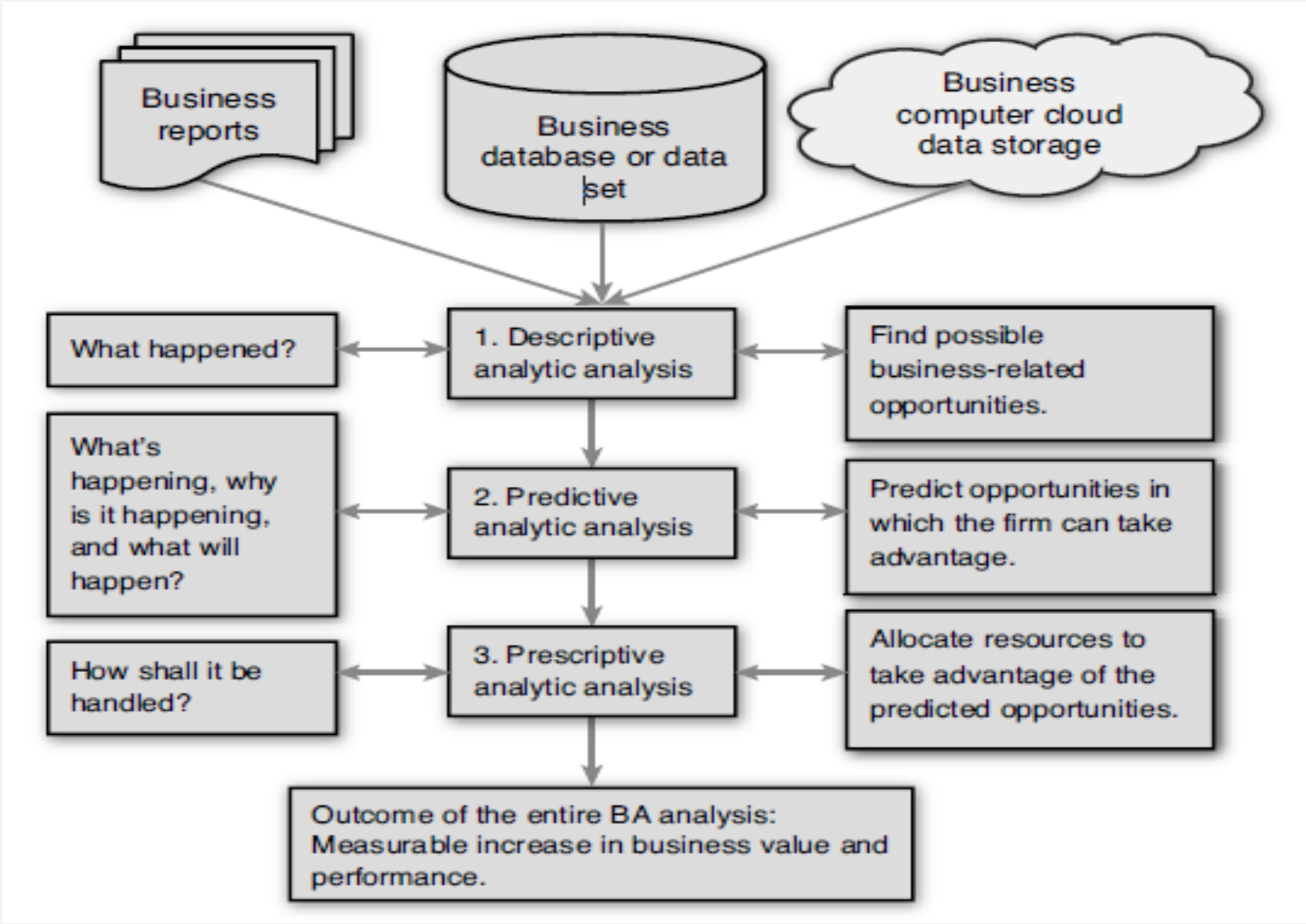
- Decisions can be formulated from descriptive and predictive analysis
  - If I need to cut a product and I know that product C is least preferred and least profitable, I will cut product C.
- However, prescriptive analytics explicitly tell you the decisions that should be made. This can be done using a variety of techniques:
  - Linear programming
  - Integer programming
  - Mixed integer programming
  - Nonlinear programming

Comparing the Three Types of Data Analytics

- Descriptive analysis is most common.
  - Best practice to perform descriptive analyses prior to prescriptive/predictive
    - Understand that distribution, variance, skew, etc., may exclude certain models
- How to know which type of analysis to pursue:
  - How much time do you have?
  - What resources are available to you?
  - How accurate is your data? How accurate do you need the model/analysis to be?
  - How popular/accepted is the model you are considering?
    - Don't subscribe to "that's how we've always done it," but remember to use a model that stakeholders will accept.



BUSINESS ANALYTICS PROCESS



TYPES OF DATA MEASUREMENT CLASSIFICATION SCALES

Type of Data Measurement Scale	Description
Categorical Data	Data that is grouped by one or more characteristics. Categorical data usually involves cardinal numbers counted or expressed as percentages. Example 1: Product markets that can be characterized by categories of "high-end" products or "low-income" products, based on dollar sales. It is common to use this term to apply to data sets that contain items identified by categories as well as observations summarized in cross-tabulations or contingency tables.
Ordinal Data	Data that is ranked or ordered to show relational preference. Example 1: Football team rankings not based on points scored but on wins. Example 2: Ranking of business firms based on product quality.
Interval Data	Data that is arranged along a scale, in which each value is equally distant from others. It is ordinal data. Example 1: A temperature gauge. Example 2: A survey instrument using a Likert scale (that is, 1, 2, 3, 4, 5, 6, 7), where 1 to 2 is perceived as equidistant to the interval from 2 to 3, and so on. Note: In ordinal data, the ranking of firms might vary greatly from first place to second, but in interval data, they would have to be relationally proportional.
Ratio Data	Data expressed as a ratio on a continuous scale. Example 1: The ratio of firms with green manufacturing programs is twice that of firms without such a program.

The *Descriptive Analytic analysis* some patterns or variables of business behavior should be identified representing targets of business opportunities and possible (but not yet defined) future trend behavior.

Additional effort (more mining) might be required, such as the generation of detailed statistical reports narrowly focused on the data related to targets of business opportunities to explain what is taking place in the data (what happened in the past).

- **There are many methods that can be used in this step of the BA process.**
  - A commonly used methodology is multiple regression. “Statistical Tools,” and “Forecasting,” for a discussion on multiple regression and ANOVA testing(analysis of variance, a statistical method in which the variation in a set of observations is divided into distinct components).
  - This methodology is ideal for establishing whether a statistical relationship exists between the predictive variables found in the descriptive analysis.
  - The relationship might be to show that a dependent variable is predictively associated with business value or performance of some kind.
  - Exploring a database using advanced statistical procedures to verify and confirm the best predictive variables is an important part of this step in the BA process.

RELATIONSHIP OF BA PROCESS AND ORGANIZATION DECISION-MAKING PROCESS

- The BA process can solve problems and identify opportunities to improve business performance.
- In the process, organizations may also determine strategies to guide operations and help achieve competitive advantages.
- Typically, solving problems and identifying strategic opportunities to follow are organization decision-making tasks.
- The latter, identifying opportunities, can be viewed as a problem of strategy choice requiring a solution.

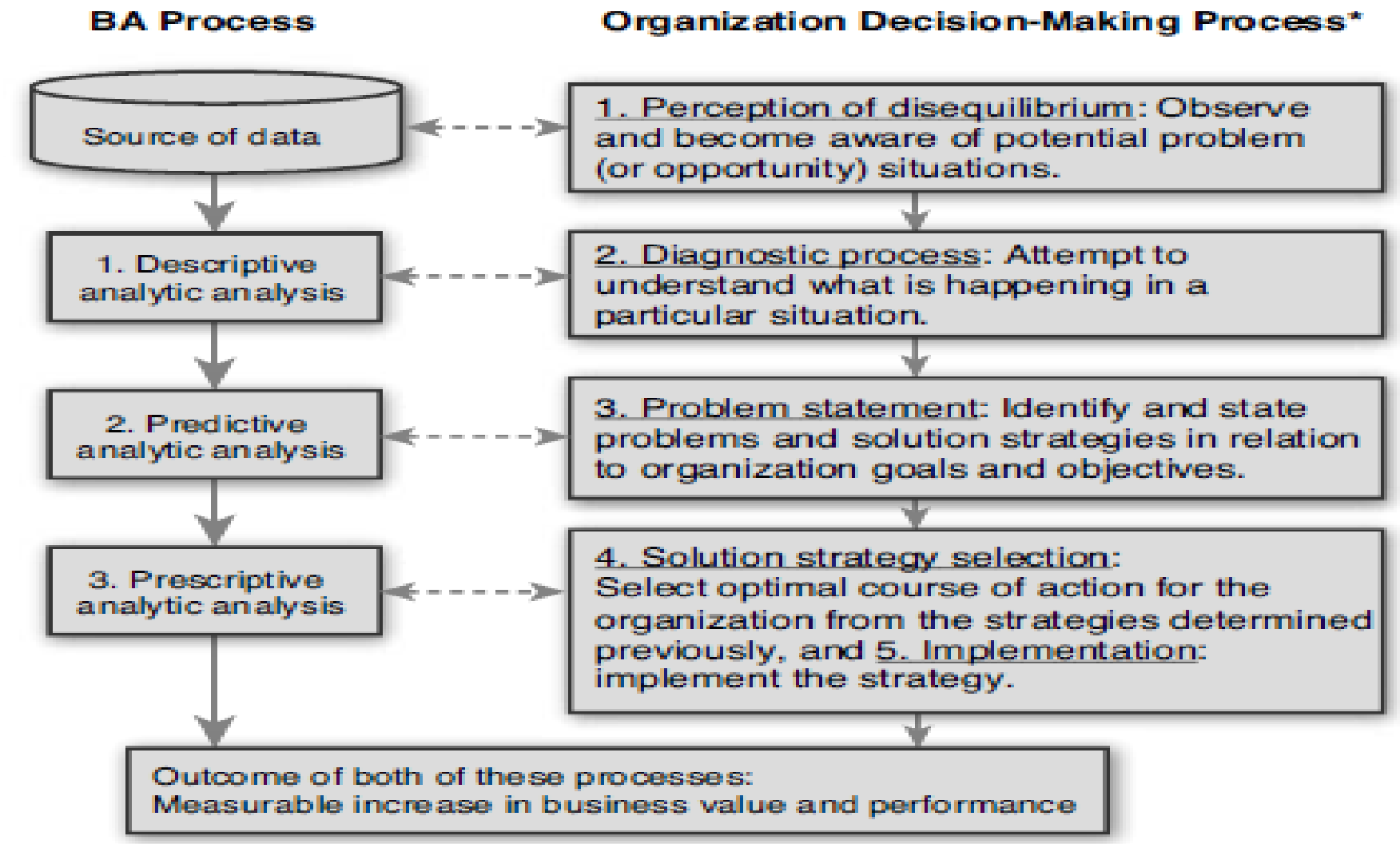


Figure 1.2 Comparison of business analytics and organization decision-making processes

**Source:* Adapted from Figure 1 in Elbing (1970), pp. 12–13.

- The *organization decision-making process* (ODMP) is focused on decision-making to solve problems but could also be applied to finding opportunities in data and deciding what is the best course of action to take advantage of them.
- The five-step ODMP begins with the perception of disequilibrium, or the awareness that a problem exists that needs a decision.
- Similarly, in the BA process, the first step is to recognize that databases may contain information that could both solve problems and find opportunities to improve business performance.
- Then in Step 2 of the ODMP, an exploration of the problem to determine its size, impact, and other factors is undertaken to diagnose what the problem is. Likewise, the BA descriptive analytic analysis explores factors that might prove useful in solving problems and offering opportunities.
- The ODMP problem statement step is similarly structured to the BA predictive analysis to find strategies, paths, or trends that clearly define a problem or opportunity for an organization to solve problems.
- Finally, the ODMP's last steps of strategy selection and implementation involve the same kinds of tasks that the BA process requires in the final prescriptive step(make an optimal selection of resource allocations that can be implemented for the betterment of the organization).
- The decision-making foundation that has served ODMP for many decades parallels the BA process.
- The same logic serves both processes and supports organization decision-making skills and capacities.

DATA COLLECTION,PROCESSING AND ANALYSIS

- **STEPS IN DATA COLLECTION**

❑ The first two steps relate to the collection of primary data while the third step relates to the collection of secondary data.

❑ The information/data collected by a person directly is known as primary data while records or data collected from offices/institutions is known as secondary data.

- **A. Steps in Primary Data Collection:**

1. Making oneself ready both mentally as well as physically for collecting primary data from field situations.

2. Keeping a field book/record book or diary for writing relevant information, doing field sketching or writing records of the occurrence of phenomenon at specific time intervals.

3. Administering questionnaire schedule to the target groups of area people across sampled sites.

4. Verifying the facts through cross checks in the answers and ground realties.

5. Integrating the observations, responses and recorded facts in a systematic and logical framework.

- **B. Steps in Secondary Data Collection:**

1. Knowledge about the offices/institutes etc. keeping the record of relevant data is of prime importance to obtain the secondary data/information.

2. Get an official letter containing your requirements of data and purpose of data collection from your Principal/Head of the Institute? Your identity card is also an essential requirement to get an entry in the offices.

3. Keep a note book/record file to transfer data for the purpose. It could also be done with the help of photo copying systems.

4. The secondary data, thus, collected forms the basis for tabulation and processing as per need.

TOOLS AND TECHNIQUES OF DATA COLLECTION

- For data collection we make use of certain tools and follow specific techniques. The tools that help in data collection are as under:
- ☐ Observing the phenomenon and recording the details,
- ☐ Inquiring about the facts through questionnaires/schedules
- ☐ Making measurements.
- ☐ Conducting tests.
- ☐ Recording the events.

- A. Questionnaires:
- (a) Contents of Questionnaire:
- (b) Form of Questionnaire:
- (c) The Interview

- B. The Schedules

- C. Rating Scales

Temperature Conditions:

Very Cold	Cold	Cool	Moderately Warm	Hot	Very Hot
0	1	2	3	4	5

Development Level:

Under Developed	VeryLow Level	Low Level	Medium Level	High Level	Very High Level
0	1	2	3	4	5

- D. Field Sketches
- E. Photographs

- F. Methods of Administering the Questionnaires and Survey Schedules
- G. Collection of Information
- H. Precautions in Collecting the Information
- I. Selection of Samples and Sample Size
  1) Identification of Samples:
  2) Sampling Techniques
     - Systematic Sampling
     - Random Sampling
       - Simple Random Sampling:
       - Stratified Random Sampling
  3) Sample Size:

- **PROCESSING OF DATA**

- **Processing of primary data:**
- **Editing of data:**
- **The coding of data:**
- **Organization of Data:**

Details / Households	Population			Functions				Facilities		
	P	M	F	Agri	Ind	Trade	Service	T.V.	Phone	Vehicle
01	20	12	08	5	-	1	12	1	1	1 Scooter
02	17	09	08	6	-	1	1	1	1	1 Scooter
03	9	04	05	-	-	2	1	1	2	1 Car and 1 Scooter
04	12	06	06		1		2	1	1	1 Scooter
05	13	07	06	2	-	-	2	1	-	1 Scooter

- **Classification of data:**

- **Presentation of data:**

- **Tabular Presentation**
- **Statistical Presentation of data:**
  - (a) Arithmetic mean or average
  - (b) Median
  - (c) Mode

Arithmetic Mean

Mean: The "average" number; found by adding all data points and dividing by the number of data points.
For example, the production of rice per acre in five districts is 10, 8, 12,9 and 6 quintals. The average production of rice for these districts is :

$$\frac{10+8+12+9+6}{5}=\frac{45}{5}=9 \text{ quintals per acre}$$

Example: The mean of 4, 1, and 7 is $(4+1+7)/3 = 12/3 = 4$.

The arithmetic mean is the sum of all of the data points divided by the number of data points.

$$\text{mean} = \frac{\text{sum of data}}{\#\text{ of data points}}$$

Here's the same formula written more formally:

$$\text{mean} = \frac{\sum x_i}{n}$$

- **Median:** The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Example: The median of 4, 1, and 7 is 4 because when the numbers are put in order $(1, 4, 7)$, the number 4 is in the middle.

Example

Find the mean of this data:

$1, 2, 4, 5$

Start by adding the data:

$1 + 2 + 4 + 5 = 12$

There are 4 data points.

$$\text{mean} = \frac{12}{4} = 3$$

The mean is 3.

- **Mode:** The most frequent number—that is, the number that occurs the highest number of times.

Example: The mode of $\{4, 2, 4, 3, 2, 2\}$ is 2 because it occurs three times, which is more than any other number.

Example 1

Ms. Norris asked students in her class how many siblings they each had.

Find the mode of the data:

$0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 5$

Look for the value that occurs the most:

$0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 5$

The mode is 1 sibling.

DATA EXPLORATION

1. Steps of Data Exploration and Preparation

Remember the quality of your inputs decide the quality of your output.

Example:

So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here. With estimate, data exploration, cleaning and preparation can take up to 70% of your total project time.

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment

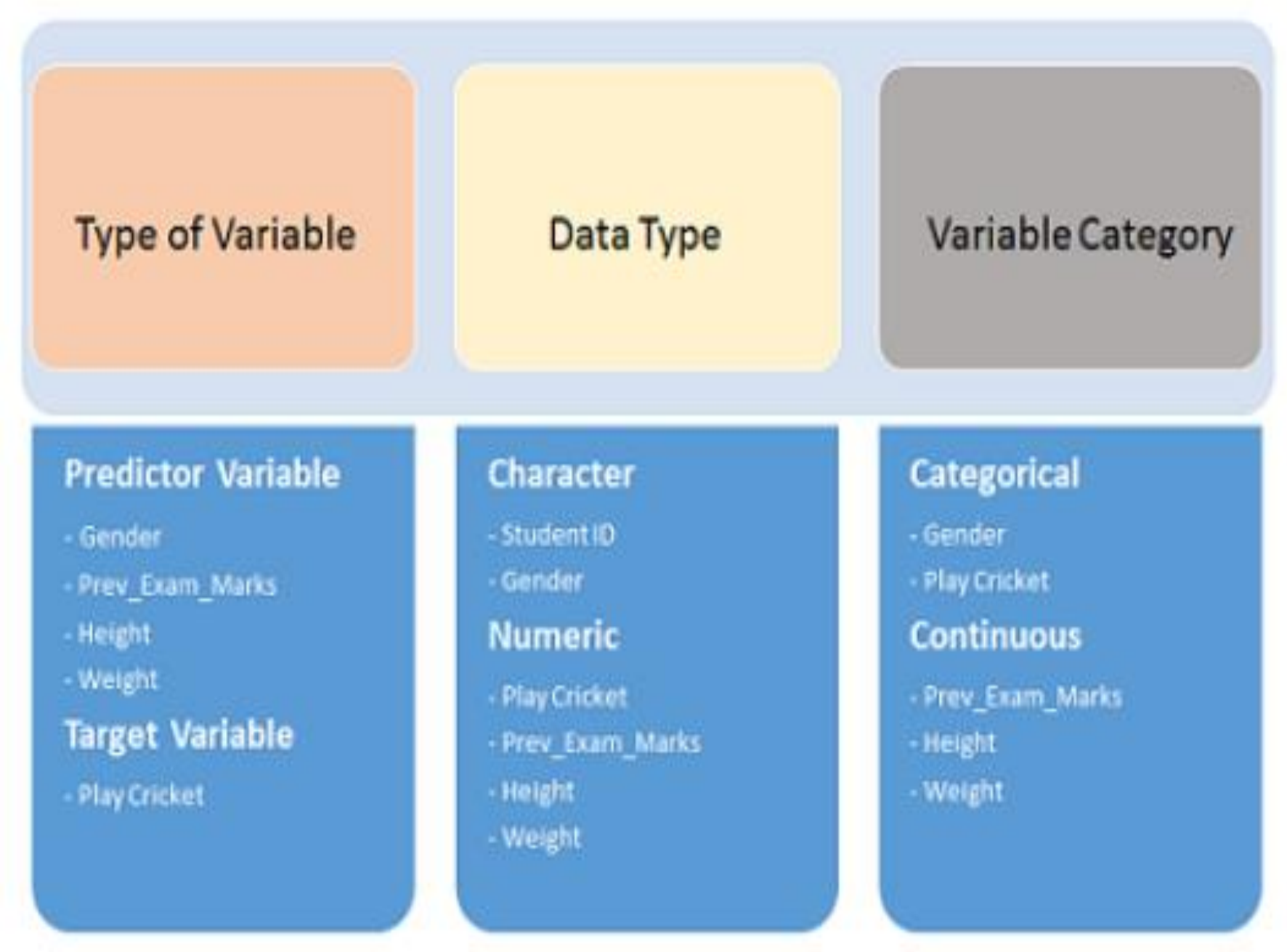
1.Variable Identification

First, identify **Predictor (Input) and Target (output) variables.**

Next, identify the data type and category of the variables.

- Example:- Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables.

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0


Type of Variable
Data Type
Variable Category
Predictor Variable
- Gender
- Prev_Exam_Marks
- Height
- Weight
Target Variable
- Play Cricket
Character
- Student ID
- Gender
Numeric
- Play Cricket
- Prev_Exam_Marks
- Height
- Weight
Categorical
- Gender
- Play Cricket
Continuous
- Prev_Exam_Marks
- Height
- Weight

2.UNIVARIATE ANALYSIS

At this stage, we explore variables one by one.

Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous.

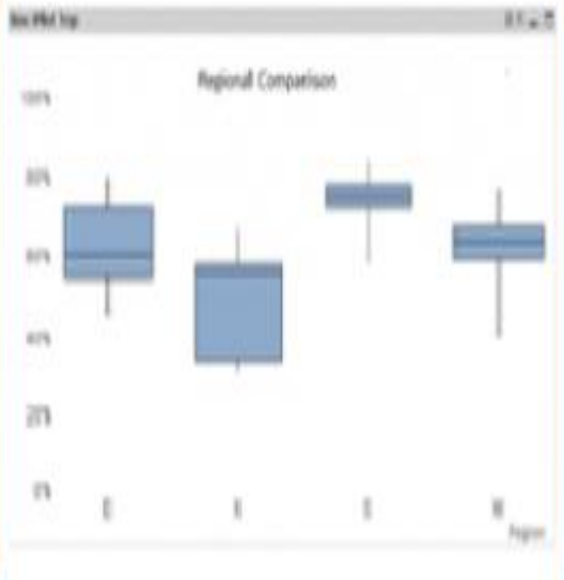
Let's look at these methods and statistical measures for categorical and continuous variables individually:

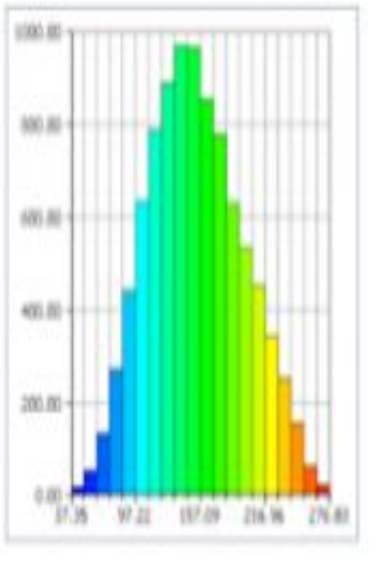
Continuous Variables:-

In case of continuous variables, we need to understand the central tendency and spread of the variable.

These are measured using various statistical metrics visualization methods as shown below:

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	





Categorical Variables:-

For categorical variables, we'll use frequency table to understand distribution of each category.

We can also read as percentage of values under each category.

- It can be be measured using two metrics, **Count and Count% against each category.**

Bar chart can be used as visualization.

3.Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables.

Here, we look for association and disassociation between variables at a pre-defined significance level.

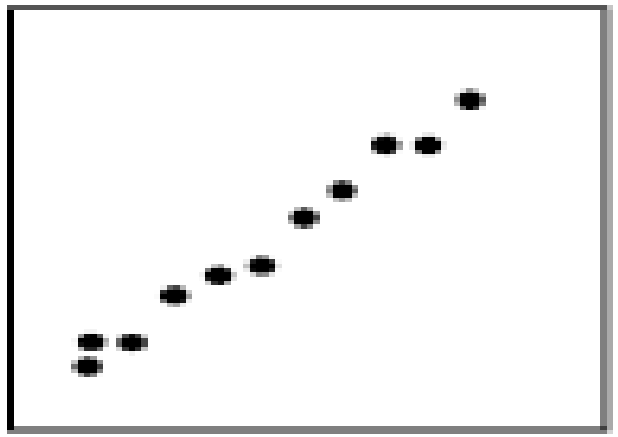
- We can perform bi-variate analysis for any combination of categorical and continuous variables.
- The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous.
- Different methods are used to tackle these combinations during analysis process.

- **Continuous & Continuous:**

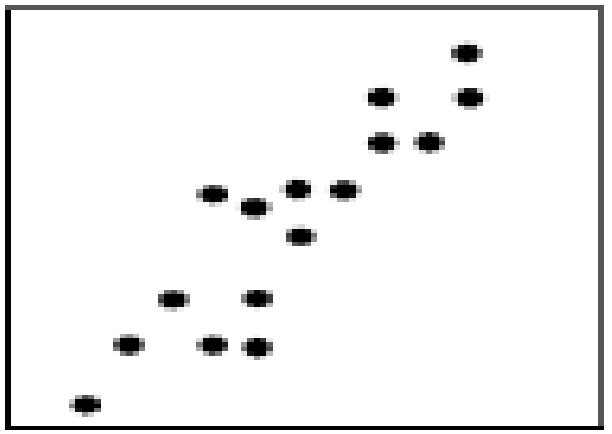
  While doing bi-variate analysis between two continuous variables, we should look at scatter plot.

  It is a nifty way to find out the relationship between two variables.

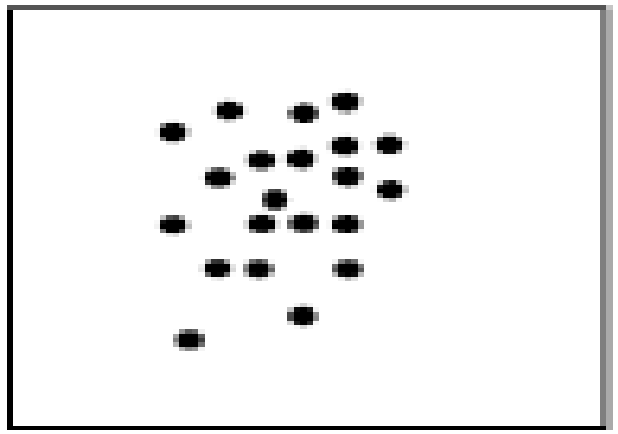
  The pattern of scatter plot indicates the relationship between variables.
- The relationship can be linear or non-linear.



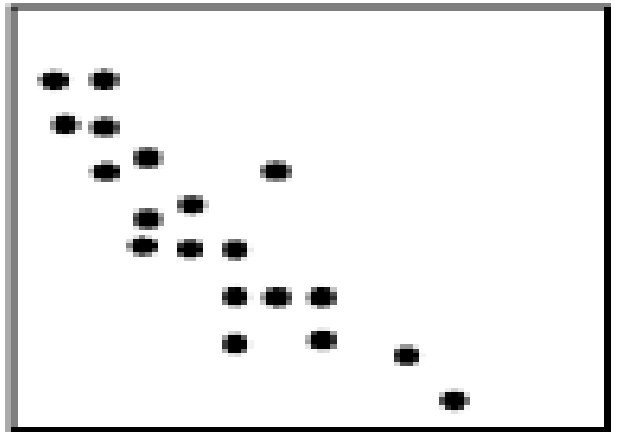
Strong positive correlation



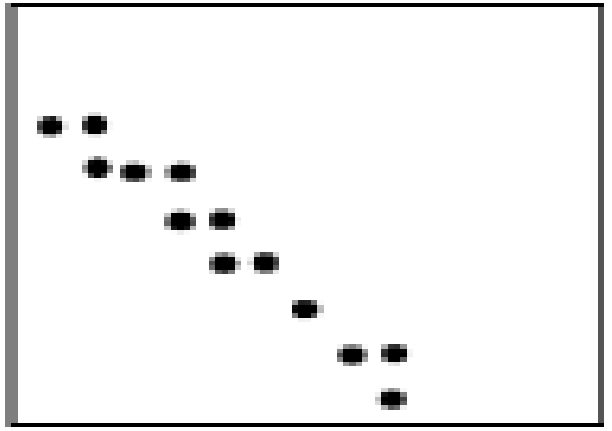
Moderate positive correlation



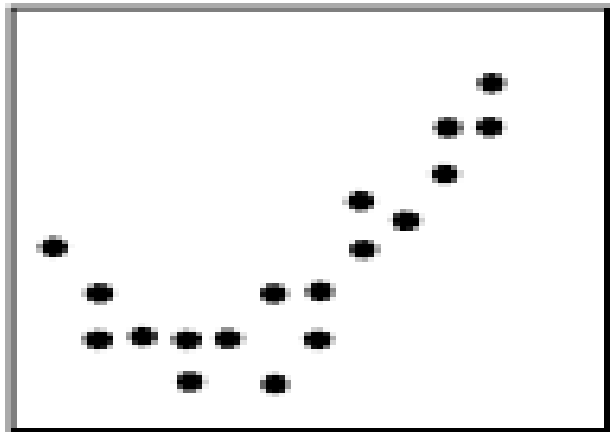
No correlation



Moderate negative correlation



Strong negative correlation



Curvilinear relationship

- Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them.
- To find the strength of the relationship, we use Correlation.

  Correlation varies between -1 and +1.

  -1: perfect negative linear correlation

  +1:perfect positive linear correlation and

  0: No correlation

Correlation can be derived using following formula:

- **Correlation = Covariance(X,Y) / SQRT(Var(X)* Var(Y))**

X	65	72	78	65	72	70	65	68
Y	72	69	79	69	84	75	60	73

Metrics	Formula	Value
Co-Variance (X,Y)	=COVAR(E6:L6,E7:L7)	18.77
Variance (X)	=VAR.P(E6:L6)	18.48
Variance (Y)	=VAR.P(E7:L7)	45.23
Correlation	=G10/SQRT(G11*G12)	0.65

Categorical & Categorical: To find the relationship between two categorical variables, we can use following methods:

Two-way table: We can start analyzing the relationship by creating a two-way table of count and count%.

The rows represents the category of one variable and the columns represent the categories of the other variable.

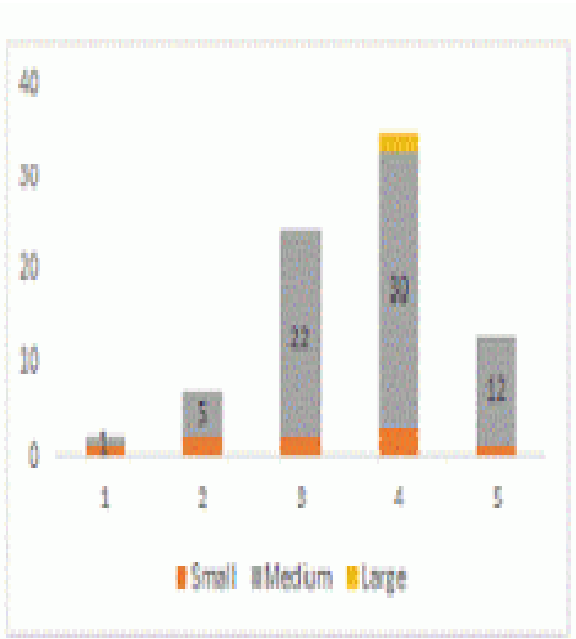
We show count or count% of observations available in each combination of row and column categories.

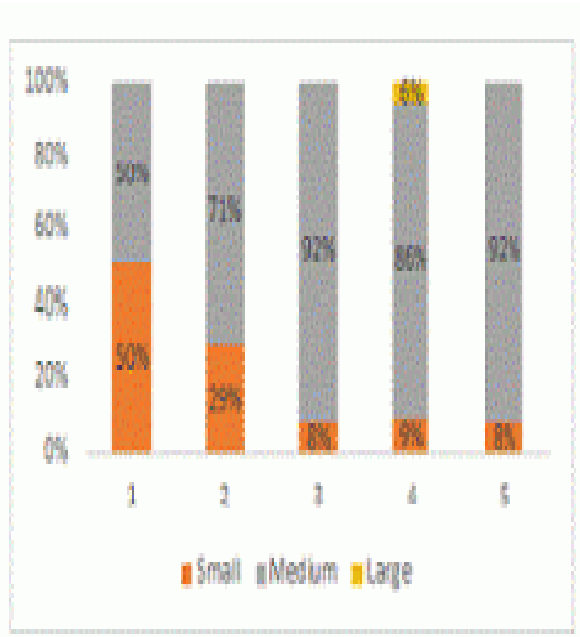
Stacked Column Chart: This method is more of a visual form of Two-way table.

Product Category

Frequency Row Pct	1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77





- **Chi-Square Test: This test is used to derive the statistical significance of relationship between the** variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well.
- Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table.
- It returns probability for the computed chi-square distribution with the degree of freedom.
- Probability of 0: It indicates that both categorical variable are dependent
- Probability of 1: It shows that both variables are independent.

- Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence.
- The chi-square test statistic for a test of independence of two categorical variables is found by

$$X^2 = \sum (O - E)^2 / E$$

where *O represents* the observed frequency.
E is the expected frequency under the null hypothesis and computed by:

$$E = \frac{row\ total \times column\ total}{sample\ size}$$

Categorical & Continuous: While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables.
If levels are small in number, it will not show the statistical significance.
To look at the statistical significance we can perform Z-test, T-test or ANOVA.
Z-Test/ T-Test:- Either test assess whether mean of two groups are statistically different from each other or not.

$$z = \frac{\left| \bar{x}_1 - \bar{x}_2 \right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- If the probability of Z is small then the difference of two averages is more significant.
- The T-test is very similar to Z-test but it is used when number of observation for both categories is less than 30.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

where:

- $\overline{X}_1, \overline{X}_2$: *Averages*
- S_1^2, S_2^2 : *Variances*
- N_1, N_2 : *Counts*
- t : has t distribution with $N_1 + N_2 - 2$ degree of freedom

MISSING VALUE TREATMENT

- Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

- In the left scenario, we have not treated missing values.
- The inference from this data set is that the chances of playing cricket by males is higher than females. On the other hand, if you look at the second table, which shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.
- We looked at the importance of treatment of missing values in a dataset. Now, let's identify the reasons for occurrence of these missing values. They may occur at two stages:

- 1. Data Extraction: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.

- 2. Data collection: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:

❑ Missing completely at random: This is a case when the probability of missing variable is same for all observations.

For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin. If an head occurs, respondent declares his / her earnings & vice versa.

Here each observation has equal chance of missing value.

❑ Missing at random: This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.

❑ Missing that depends on unobserved predictors: This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included "discomfort" as an input variable for all patients.

❑ Missing that depends on the missing value itself: This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

Which are the methods to treat missing values ?

1. **Deletion: It is of two types:**

List Wise Deletion & Pair Wise Deletion.

- In list wise deletion, we delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size.
- In pair wise deletion, we perform analysis with all cases in which the variables of interest are present. Advantage of this method is, it keeps as many cases available for analysis. One of the disadvantage of this method, it uses different sample size for different variables

List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Deletion methods are used when the nature of missing data is “Missing completely at random” else non random missing values can bias the model output.

2. **Mean/ Mode/ Median Imputation:** Imputation is a method to fill in the missing values with estimated ones.

   The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values.

   It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

It can be of two types:-

- **Generalized Imputation:** In this case, we calculate the mean or median for all non missing values of that variable then replace missing value with mean or median.

  "Manpower" is missing so we take average of all non missing values of "Manpower" (28.33) and then replace missing value with it.

- **case Imputation:** In this case, we calculate average for gender "Male" (29.75) and "Female" (25) individually of non missing values then replace the missing value based on gender. For "Male", we will replace missing values of manpower with 29.75 and for "Female" with 25.

3. **Prediction Model:** Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data.

4. **KNN Imputation:** In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing.

   The similarity of two attributes is determined using a distance function. It is also known to have certain advantage & disadvantages.

Advantages:

- k-nearest neighbour can predict both qualitative & quantitative attributes
- Creation of predictive model for each attribute with missing data is not required
- Attributes with multiple missing values can be easily treated
- Correlation structure of the data is taken into consideration

Disadvantage:

- KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances.
- Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes.

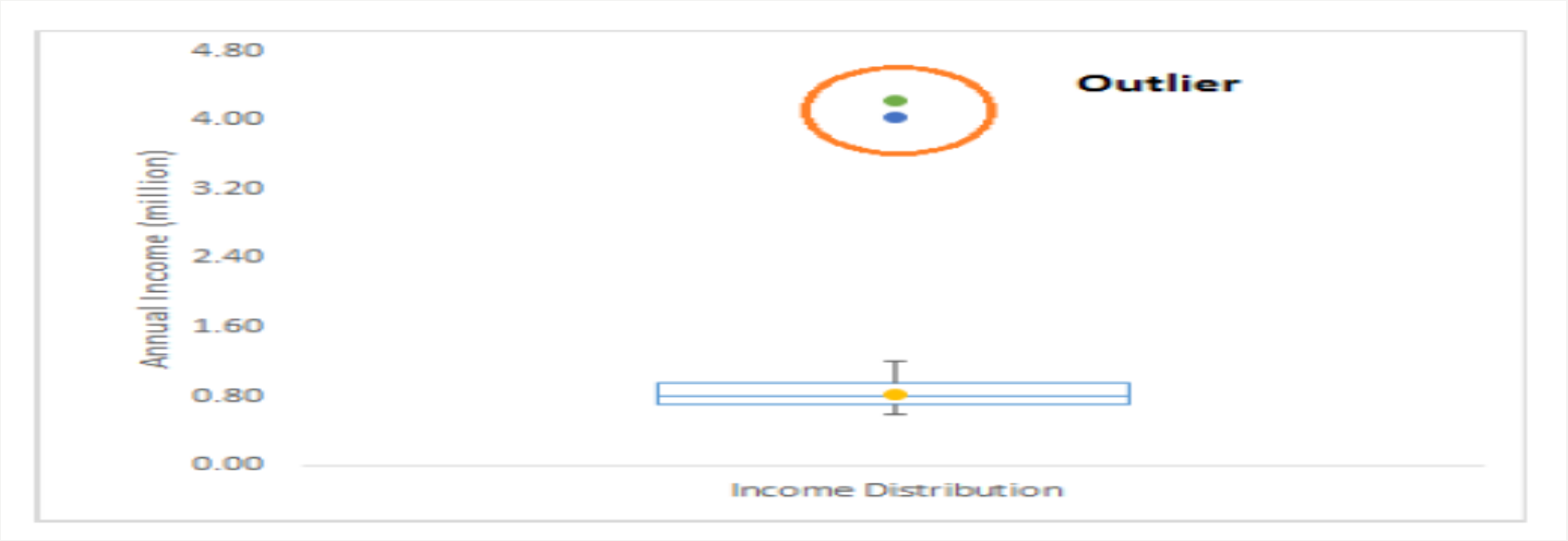
Techniques of Outlier Detection and Treatment

What is an Outlier?

Outlier is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations.

Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

- Let's take an example, we do customer profiling and find out that the average annual income of customers is $0.8 million. But, there are two customers having annual income of $4 and $4.2 million.
- These two customers annual income is much higher than rest of the population. These two observations will be seen as Outliers



classified in two broad categories:

1. **Artificial (Error) / Non-natural**
2. **Natural.**

Let’s understand various types of outliers in more detail:

Data Entry Errors:- Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.

For example: Annual income of a customer is $100,000. Accidentally, the data entry operator puts an additional zero in the figure. Now the income becomes $1,000,000 which is 10 times higher.

- **Measurement Error:** It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty.

  For example: There are 10 weighing machines. 9 of them are correct, 1 is faulty. Weight measured by people on the faulty machine will be higher / lower than the rest of people in the group.

- **Experimental Error:** Another cause of outliers is experimental error.

  For example: In a 100m sprint of 7 runners, one runner missed out on concentrating on the 'Go' call which caused him to start late. Hence, this caused the runner's run time to be more than other runners. His total run time can be an outlier.

- **Intentional Outlier:** *This is commonly found in self-reported measures that involves sensitive data.*

  For example: Teens would typically under report the amount of alcohol that they consume. Only a fraction of them would report actual value. Here actual values might look like outliers because rest of the teens are under reporting the consumption.

- **Data Processing Error:** Whenever we perform data mining, we extract data from multiple sources. It is possible that some manipulation or extraction errors may lead to outliers in the dataset.

- **Sampling error:** For instance, we have to measure the height of athletes. By mistake, we include a few basketball players in the sample. This inclusion is likely to cause outliers in the dataset.

- **Natural Outlier:** When an outlier is not artificial (due to error), it is a natural outlier.

  For instance: In my last assignment with one of the renowned insurance company, I noticed that the performance of top 50 financial advisors was far higher than rest of the population. Surprisingly, it was not due to any error.

DATA CLEANSING

INTRODUCTION

"A company's most important asset is information. A corporation's ability to compete, adapt, and grow in a business climate of rapid change is dependent in large measure on how well the company uses information to make decisions. Sharing information that isn't clean and consolidated to the fullest extent can substantially reduce the effectiveness of a system of significant investment and considerable pay-off potential."

WHAT IS DATA CLEANSING?

- **Data cleansing or data scrubbing** is the act of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant etc. parts of the data and then replacing, modifying or deleting this dirty data.

- Data cleansing can occur within a single set of records, or between multiple sets of data which need to be merged, or which will work together.

- Typos and spelling errors are corrected, mislabeled data is properly labeled and filed, and incomplete or missing entries are completed.

- In more complex operations, data cleansing can be performed by computer programs. These data cleansing programs can check the data with a variety of rules and procedures decided upon by the user

- •The goal of data cleansing is not just to clean up the data in a database but also to bring consistency to different sets of data that have been merged from separate databases.

Why is Data "Dirty" ?

- Dummy Values,
- Absence of Data,
- Multipurpose Fields,
- Cryptic Data,
- Contradicting Data,
- Inappropriate Use of Address Lines,
- Violation of Business Rules,
- Reused Primary Keys,
- Non-Unique Identifiers, and
- Data Integration Problems

Steps in Data Cleansing

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

Parsing

Parsing locates and identifies individual data elements in the source files and then isolates these data elements in the target files.

Parsing

Input Data from Source File
Beth Christine Parker, SLS MGR
Regional Port Authority
Federal Building
12800 Lake Calumet
Hedgewisch, IL

Parsed Data in Target File

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Correcting

Corrects parsed individual data components using sophisticated data algorithms and secondary data sources.

Correcting

Parsed Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Corrected Data

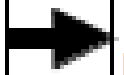
First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Standardizing

Standardizing applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom business rules.

Standardizing

Corrected Data	
First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398



Corrected Data	
Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Matching

Searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.

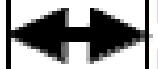
Match Patterns

Business Name	Street	Branch Type	Customer #/Tax ID	City	Vendor Code	Pattern	Pattern I.D.
Exact	Exact	Exact	Exact	Exact	Exact	AAAAAA	P110
Exact	VClose	Exact	VClose	Exact	Blanks	ABAAA-	P115
Exact	VClose	Exact	Blanks	Exact	Exact	ABA-AA	P120
Exact	VClose	Close	Close	Exact	Exact	ABCCAA	S300
VClose	VClose	Exact	Close	Exact	Exact	BBACAA	S310

Matching

Corrected Data (Data Source #1)

Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398



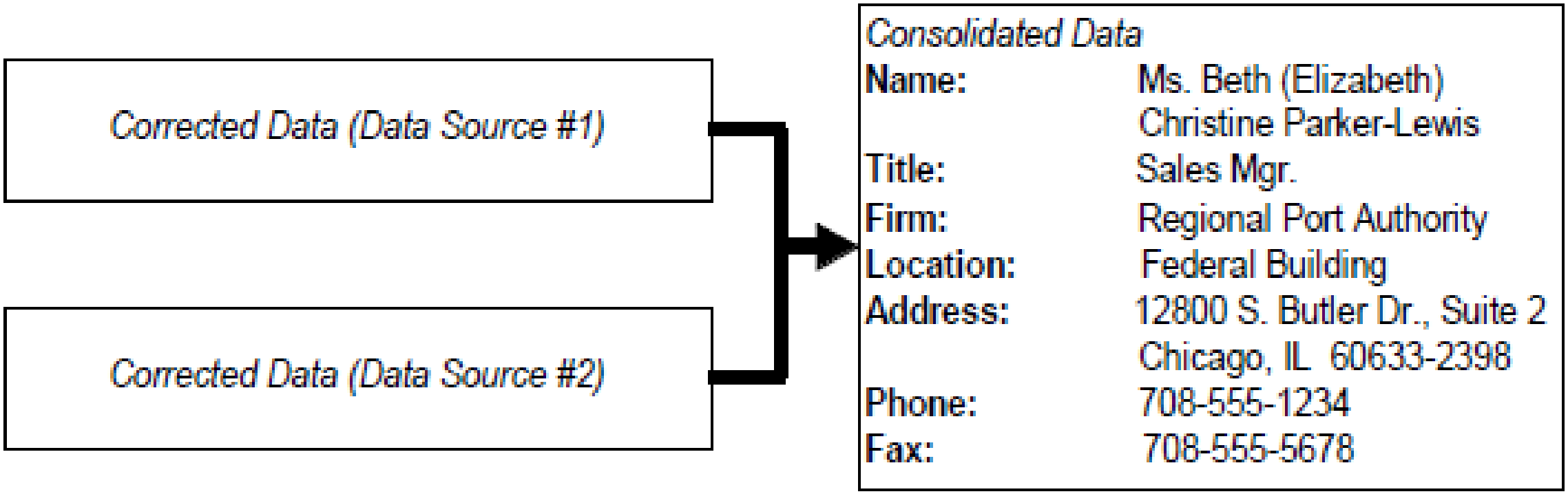
Corrected Data (Data Source #2)

Pre-name:	Ms.
First Name:	Elizabeth
1st Name Match Standards:	Beth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker-Lewis
Title:	
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr., Suite 2
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398
Phone:	708-555-1234
Fax:	708-555-5678

Consolidating

Analyzing and identifying relationships between matched records and consolidating/merging them into ONE representation.

Consolidating



Recommended Best Practices

1. Use metadata to document rules .

2. Determine data cleansing schedule .

3. Build quality into new and existing systems.

Arithematic - Mean :-

(1) Statistics - Arithematic Mean of individual data series :-

Eg :- rotation

Items :- 5 10 20 30 40 50 60 70

formula :-

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{}$$

alternatively written as. $\bar{x} = \frac{\Sigma x}{N}$

where $x_1, x_2, x_3 \ldots x_n$ = individual observations of variables

Σx = Sum of observations of the variable

N = No of observations.

Eg :-

Items :- 14 36 45 70 105

sol :- $\bar{x} = \frac{14+36+45+70+105}{5} = \frac{270}{5} = 54.$

(8 11) . Discrete data series :-

Eg: notation :-

Items	5	10	20	30	40	50	60	70
frequency	2	5	1	3	12	0	5	7

formula :-

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \dots f_nx_n}{N} = \frac{\Sigma fx}{N}$$

N = no of observations

$f_1, f_2, \dots f_n$ = Different values of frequency f.

$x_1, x_2 \dots x_n$ = Different values of variable x.

Items :-	14	36	45	70
frequency :-	2	5	1	3

Items	frequency f	fx
14	2	28
36	5	180
75	1	45
70	3	210
	$N = 11$	$\Sigma fx = 463$

$$\bar{x} = \frac{463}{11} = 42.09.$$

$$\bar{x} = \frac{463}{11} = 42.09.$$

(iii) Continuous Data Series:-

Notation:

Item.	0-5	5-10	10-20	20-30	30-40
Frequency.	2	5	1	3	12.

In Case Continous series, a mid point computed as $\frac{\text{lower limit} + \text{upper limit}}{2}$

Formula:

$$\bar{x} = \frac{f_1 m_1 + f_2 m_2 + f_3 m_3 \ldots\ldots f_n m_n}{N}$$

Where N = no. of observations

$f_1, f_2 \ldots f_n$ = Different values of frequency 'F'.

$m_1, m_2, m_3 \ldots m_n$ = Different values of mid points for ranges.

$m_1, m_2, m_3, \ldots m_n$ = Different values

Eg:-

Items	0-10	10-20	20-30	30-40
frequency	2	5	1	3

Year Given data

Items	midpoint	frequency	fm
0-10	5	2	10
10-20	15	5	75
20-30	25	1	25
30-40	35	3	105
		N = 11	$\Sigma fm = 215$

$$\bar{x} = \frac{215}{11} = 19.54.$$

(2) Arithematic - Median :-

(i) Individual series :-

Notation :-

Items	5	10	20	30	40	50	60	70

formula :-

$$\text{Mean} = \text{value of } \left(\frac{N+1}{2}\right)^{th} \text{item.}$$

N = No: of observations

Eg :-

Items	14	36	45	70	105

Given no: are 5, an odd no:- thus middle no:- is Arithematic median.

∴ Arithematic median of given no: is 45.

(ii) Disbete series :-

formula :- Median = value of $\left(\frac{N+1}{2}\right)^{th}$ item

∴ Arithmetic median of 3 ... =

(ii) Discrete series:-

Formula:- Median = value of $\left(\frac{N+1}{2}\right)^{th}$ item

N = No: of observations

Eg:-

Items	14	36	45	70	105	145
frequency	2	5	1	3	12	0

$$M = \text{value of } \left(\frac{N+1}{2}\right)^{th} \text{item}$$

$$= \text{value of } \left(\frac{6+1}{2}\right)^{th} \text{item}$$

$$= \text{value of } 3.5^{th} \text{ item}$$

$$= \left(\frac{45+70}{2}\right) = 57.5$$

$= \left(\frac{n}{2}\right)$

= value of 3.5^{th} term

$= \left(\frac{45+70}{2}\right) = 57.5$

(iii) Continous Median :-

Items	0-5	5-10	10-20	20-30	30-40
frequency	2	5	1	3	12

Formula:

$$\text{Median} = l + \frac{\left(\frac{n}{2} - c.f\right)}{f} \times i$$

l = Lower limit of median class, median class is that class where $\frac{n^{th}}{2}$ item is lying.

$c.f.$ = Cumulative frequency of class preceding the median class.

f = frequency of median class ; i = Class interval of median class

Eg:- In a study conducted on an organization, the distribution of income across the workers is observed. Find the median wage of the workers of the organization.

					Freq
06	men	get	less than	Rs 500	6
13	men	"	"	Rs 1000	7
22	"	"	"	Rs 1500	9
30	"	"	"	Rs 2000	8
34	"	"	"	Rs 2500	4
40	"	"	"	Rs 3000	6 .

Sol From Given data.

Income Rs/-	MP (midpoint) m	frequency f	$(m-250)/500$ d (distribution)	fd	c.f
0 - 500	250	6	-2	-12	6
500 - 1000	750	7	-1	-7	13
1000 - 1500	1250	9	0	0	22
1500 - 2000	1750	8	1	8	30
2000 - 2500	2250	4	2	8	34
2500 - 3000	2750	6	3	18	40
		$N = 40$		$\Sigma f_d = 15$	

In order to simplify calculation, a common factor $i = 500$ has been taken

$$\text{median} = L + \frac{\left(\frac{n}{2} - c.f\right)}{f} \times i$$

where $L = 1000$; $\frac{n}{2} = 20$ $\quad c.f = 13$

$f = 9$; $i = 500$

frequency of median class

$$\therefore \text{median} = 1000 + \frac{(20-13)}{9} \times 500$$

$N = 40$ $\quad \Sigma f_d = 13$

In order to simplify calculation, a common factor $i = 500$ has been taken

$$\text{median} = l + \frac{\left(\frac{n}{2} - c.f\right)}{f} \times i$$

where $L = 1000$; $\frac{n}{2} = 20$ $\quad c.f = 13$

$f = 9$; $i = 500$

Boundary of median class

$$\therefore \text{median} = 1000 + \frac{(20-13)}{9} \times 500$$

$$= 1000 + 388.9$$

$$= 1388.9 \text{ //}$$

∴ The median wage is Rs 1389/-

Arithematic Mode :-

(1) Individual series :-

Notation :-							
Item :-	5	10	20	30	40	50	60	70

Eg :-

Items :-	14	36	45	36	105	36

Arithematic mode gives = 36. (max no of time = 3) ✓

2) Discrete series :-

Eg :-

Items	14	36	45	70	105	165
Frequency	2	5	1	3	12	0

The arithematic mode of given no :- is 105 as the highest frequency 12 is associated with 105. ✓

M_o = mode

(iii) Continuous series :-

formula :- $M_o = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

M_o = mode
f_1 = frequency of modal class
f_0 = freq of Pre-modal class
f_2 = freq of class succeeding modal class
i = class interval.

In case there are 2 values of variable which have equal highest frequency then the series is bi-modal and mode is said to be ill defined. In such situations mode is calculated by the

mode = 3 median − 2 mean

Eg: Calculates Arithematic mode of following

Ages	No of workers
0-5	3
5-10	7
10-15	15
15-20	30
20-25	20
25-30	10
30-35	5

Sol Using formula:

$$M_o = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$L = 15$; $f_1 = 30$; $f_0 = 15$; $f_2 = 20$; $i = 5$

$$\therefore M_o = 15 + \frac{30 - 15}{2 \times 30 - 15 - 20} \times 5 = 15 + 3 = 18$$

Statistics - Relative Standard Deviation:-

In Probability theory and statistics, the Co-efficient of variation (CV) also known as relative standard deviation is a standard measure of dispersion of a probability distribution (or) frequency distribution

formula:- $100 \times \frac{s}{\bar{x}}$

s = sample standard deviation

$\bar{x}$ = sample mean

Eg:- find the RSD for following set of no: 49, 51.3, 52.7, 55.8 and standard deviation are 2.843765.

Sol step 1:- Standard deviations are rounded by 2-decimal places (2.84)

step 2:- multiply step 1 by 100.

$$2.84 \times 100 = 284.$$

step 3:- find the sample mean $\bar{x}$.

$$\frac{49 + 51.3 + 52.7 + 55.8}{4} = \frac{208.8}{4} = 52.2$$

step 4:- Divide step 2 by absolute value of step 3

$$\frac{284}{|52.2|} = 5.44$$

∴ the RSD is $52.2 \pm 5.4\%$

Statistics – Variance

- A variance is defined as the average of squared difference from mean value

$$\delta = \frac{\Sigma (M - n_i)^2}{n}$$

M = mean of items
n = no of items considered
n_i = items.

Eg:- Find variance of following data :
{600, 470, 170, 430, 300}

Sol

step 1:- $M = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$

step 2:-

$$\delta = \frac{\Sigma (M - n_i)^2}{n}$$

$$= \frac{(600 - 394)^2 + (470 - 394)^2 + (170 - 394)^2 + (430 - 394)^2 + (300 - 394)^2}{5}$$

$$= \frac{(206)^2 + (76)^2 + (-224)^2 + (36)^2 + (-94)^2}{5}$$

$$= \frac{42,436 + 5,776 + 50,176 + 1296 + 8836}{5} = \frac{108520}{5}$$

$$= \frac{(14)(13)(3)(11)}{2(1)}$$

$$= 21,704$$

$$= =.$$

$$\therefore \boxed{\bar{x} = \frac{\Sigma_{i=1}^{n} f_i x_i}{\Sigma_{i=1}^{n} f_i}}$$

eg: N

$\Sigma_{i=1}^{n} f_i x_i$ = sum of products of freq and corresponding observations

Statistical Methods:-

Arithematic Mean of grouped Data:-

Mean of data is also termed as "average"

2-types of basics of Presentation:-

1. ungrouped Data
2. Grouped Data

ungrouped data:- is raw data which is written in term of list of numbers

Eg:- 89, 55, 75, 90, 98, 46, 76, 64, 71, 97

Grouped data:-

is able to managed (or) grouped by construction of table

Eg:-

marks	frequency
45	4
30	1
85	3
90	5
75	3

class intervals	frequency
5-10	3
10-15	1
15-20	5
20-25	3
25-30	8

Methods and formulae :-

(i) Direct method :- used when grouped data is given. and given observations are denoted as x_i

$$\therefore \quad \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

$\bar{x}$ = mean of given data

$\sum_{i=1}^{n} f_i$ = sum of frequencies also denoted by 'N'

$\sum_{i=1}^{n} f_i x_i$ = sum of Products of freq and corresponding observation

(ii) Assumed mean method :-

is a method that uses an assumed mean in order to calculate actual mean

→ This method does cut short the calculations that are more in direct method

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

A = Middle value that is assumed as mean for calculations

$\sum_{i=1}^{n} f_i d_i$ = sum of Product of frequencies and corresponding deviations

$d_i = x_i - A$ this deviation makes the calculations quite easier Even if data is too big

(*) find the mean of following data using direct method

Daily wedges (in $)	90	100	110	120	130	140	150
No:- of employees	1	2	1	3	6	4	5

sol

Daily wedges x_i	No: of employees (f_i)	$f_i x_i$
90	1	90
100	2	200
110	1	110
120	3	360
130	6	780
140	4	560
150	5	750
	$\sum_{i=1}^{7} f_i = 22$	$\sum_{i=1}^{7} f_i x_i = 2850$

	$\sum_{i=1}^{7} f_i = 22$	$\sum_{i=1}^{7} f_i x_i = 2850$

$$\bar{x} = \frac{2850}{22} = 129.55$$

(*) Estimate the mean of following data scored

marks obtained :-	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No: of students :-	4	5	4	4	3	4	16

80-90	90-100	100-110
13	11	16

x-intervals	x_i	f_i	$d_i = x_i - A$	$f_i d_i$
10 - 20	15	4	15 - 55 = -40	-160
20 - 30	25	5	25 - 55 = -30	-150
30 - 40	35	4	35 - 55 = -20	-80
40 - 50	45	4	45 - 55 = -10	-40
50 - 60	55 (= A)	3	55 - 55 = 0	0
60 - 70	65	4	65 - 55 = 10	40
70 - 80	75	16	75 - 55 = 20	320
80 - 90	85	13	85 - 55 = 30	390
90 - 100	95	11	95 - 55 = 40	440
100 - 110	105	16	105 - 55 = 50	800
		$\sum f_i = 80$		1560

$$\therefore \bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i} = 55 + \frac{1560}{80} = 55 + 19.5 = 74.5 .$$