Exploratory Data Analysis

Dr.G.Malathi
Associate Professor, SCSE
Image Processing and Vision – Head
Vellore Institute of Technology, Chennai

Syllabus

- CSE5007 Exploratory Data Analysis L,T,P,J,C 2,0,0,4,3
- Objective:
- This course introduces the methods for data preparation and data understanding. It covers essential exploratory techniques for understanding multivariate data by summarizing it through statistical methods and graphical methods

Syllabus

- Expected Outcomes
- After successfully completing the course the student should be able to
- 1. Handle missing data in the real world data sets by choosing appropriate methods
- Summarize the data using basic statistics
- Visualize the data using basic graphs and plots.
- Identify the outliers if any in the data set.
- Choose appropriate feature selection and dimensionality reduction techniques for handling multi-dimensional data.

Syllabus

- **Module1 INTRODUCTION TO EXPLORATORY DATA ANALYSIS**
- Data Analytics life cycle,
- Exploratory Data Analysis (EDA) –Definition,
- Motivation,
- Steps in data exploration,
- The basic data types,
- Data Type Portability

Syllabus

- **Module2 PREPROCESSING - TRADITIONAL METHODS AND MAXIMUM LIKELIHOODESTIMATION**
- Introduction to Missing data,
- Traditional methods for dealing with missing data,
- Maximum Likelihood Estimation – Basics,
- Missing data handling,
- Improving the accuracy of analysis

Syllabus

- **Module3 PREPROCESSING - BAYESIAN ESTIMATION**
- Introduction to Bayesian Estimation,
- Multiple Imputation - Imputation Phase, Analysis and Pooling Phase,
- Practical Issues in Multiple Imputation,
- Models for Missing Not at Random Data

Syllabus

- **Module4 DATA SUMMARIZATION & VISUALIZATION**
- Statistical data elaboration,
- 1-D Statistical data analysis,
- 2-D Statistical data Analysis,
- N-D Statistical data analysis

Syllabus

- **Module5** OUTLIERANALYSIS
- Introduction,
- Extreme Value Analysis,
- Clustering based, Distance Based and Density Based outlier analysis,
- Outlier Detection in Categorical Data

Syllabus

- **Module6 FEATURE SUBSET SELECTION**
- Feature selection algorithms:
- filter methods,
- wrapper methods and
- embedded methods,
- Forward selection,
- backward elimination,
- Relief,
- greedy selection,
- genetic algorithms for feature selection

Syllabus

- **Module7** DIMENSIONALITY REDUCTION
- Introduction,
- Principal Component Analysis (PCA),
- Kernel PCA,
- Canonical Correlation Analysis,
- Factor Analysis,
- Multidimensional scaling,
- Correspondence Analysis

Reference Books

Reference Books

- 1. Charu C. Aggarwal , “Data Mining The Text book”, Springer, 2015.
- 2. Craig K. Enders, “Applied Missing Data Analysis”, The Guilford Press, 2010.
- 3. Inge Koch, “Analysis of Multivariate and High dimensional data”, Cambridge University Press, 2014.
- 4. Michael Jambu, “Exploratory and multivariate data analysis”, Academic Press Inc. , 1990.
- 5. Charu C. Aggarwal, “Data Classification Algorithms and Applications”, CRC press, 2015

Projects

- **Team: 5 members per team**
- 1. Exploring the data sets for Data Science problems from Kaggle website
- 2. Applying exploratory data analysis in the field of biometrics for reliable and robust identification of humans from their personal traits, mainly for security and authentication purposes
- 3. Analyze the dataset for Fraud Detection, Customer segmentation etc.
- Note: Students can down load real-time data sets for different Machine Learning Tasks fromhttps://archive.ics.uci.edu/ml/datasets.htmland http://sci2s.ugr.es/keel/datasets.php#sub1 and do the projects

Data Analytics

- It’s a process of examining data sets to draw conclusion with the aid of specialized systems and softwares
- Data Analytics initiatives can help businesses
  - increase revenues,
  - improve operational efficiency,
  - optimize marketing campaigns,
  - customer service efforts,
  - quick response to emerging market trends,
  - gain information over rivals

Data Analytics

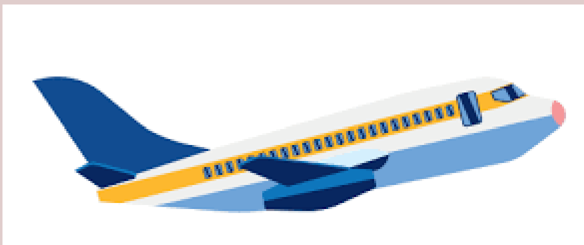
Applications of Data Analytics

- It supports variety of business uses.
- Example1: banks and credit card companies analyze withdrawal and spending patterns to prevent fraud and identity theft.
- Example 2: E-commerce companies and marketing services providers do clickstream analysis to identify website visitors who are more likely to buy a particular product or service based on navigation and page-viewing patterns

Applications of Data Analytics

- Example 3:Healthcare organizations mine patient data to evaluate the effectiveness of treatments for cancer and other diseases

Data Science Domain

Domain	Usage
	Predicting flight delay
	Predicting life time value of a customer Cross selling Up selling
	Disease Prediction Medicine effectiveness

Data Science Domain

Domain	Usage
	Sentiment Analysis Digital Marketing
	Discount Offering Demand Forecasting
	Self driving cars Pilotless aircrafts Drones

Data Analytics Process

- It starts with data collection, in which data scientist identify the information they need for a particular analytics application and then assemble it for use
- Data from different source systems need to be combined via data integration routines, transformed into a common format and loaded into the analytics system such as Hadoop or NoSQL Database

Data Analytics Process

- Alternatively, the collection process may consist of pulling a relevant subset out of a stream of raw data that flows into Hadoop and moving it to a separate partition so that it can be analyzed without affecting the overall data
- The quality issues in the data may affect the accuracy of analytics application. So it is subjected to data profiling and data cleansing.

Data Analytics Process

- This is required to make sure that the information in a data set is consistent, errors and duplicate entries are eliminated
- Data Governance policies are applied to ensure that the data matches the corporate standards
- The data scientist builds an analytical model using predictive modelling tools and or other programming languages such as python, Scala, R etc..

Data Analytics Process

- The model is initially run against a partial data set to test its accuracy and tested as a process known as training.
- Finally the model is run in production mode against the full data set
- At last the results generated by analytical models are communicated to business executives and other end users to aid in their decision making using visualization techniques

Machine Learning Methodology

Identify Problem: Define the problem statement and the end outcome expected

Gather data: Identify, Collect and prepare data available for the use case

Perform EDA , Build Features: Explore , Analyze and Study the length and depth of data

Build Machine Learning Models: Train and develop machine learning models for the use case

Productionize solution: Develop data products, deploy automated solutions

Exploratory Data Analysis

- It is the process of studying the data by leveraging various statistical and visualization techniques
- It aims to find patterns and relationships in data.

Data Analytics

- Analyzing big data is the process of
  - examining large data sets
  - to uncover hidden patterns,
  - show changes over time, and
  - confirm or challenge theories

What is Data Analytics? Contd..

Examples:

- BigData in Child Welfare
- Dr. John Snow was skeptical of the theory that foul air was the cause of a significant cholera outbreak.
- He collected interview data on those infected and identified patterns from the data he collected, concluding that the water pump on the street was the source of the outbreak.
- The pump was turned off and the outbreak stopped.

What is Data Analytics? Contd..

Examples:

- BigData in Healthcare
- Dr. John Halamka, one of the foremost health care professional in the world shares a personal situation demonstrating how data and analytics can benefit patients and catalyze positive changes in health care delivery.
- His wife, of Asian descent, was diagnosed with stage IIIA breast cancer in December 2011.

What is Data Analytics? Contd..

- He and his team queried data from all the Harvard hospitals on treatment and outcomes for the last 10,000 Asian females with a tumor similar to his wife's.
- Data revealed a medication that would be most effective for her specific case.
- Halamka credits his wife's full recovery to the findings that he was able to draw from this analysis.

Data Science - Definition

- It is an interdisciplinary field about
  - scientific methods
  - processes, and
  - systems to extract knowledge or
  - insights from data in various forms, either
  - structured or unstructured

Data Science

- Data Science is a field that encompasses anything related to
- data cleansing
- preparation and
- Analysis
- To get knowledge from the data

Applications of Data Science

Why we need Data Science

- Tradition data was structure and small in size.
- Business Intelligence Tools was enough for analyzing it
- Current trend, data is either unstructured or semi-structured.

Sources of data

- It is generated from financial logs, multimedia forms, sensors and instruments
- BI tools are not capable of processing huge data
- So advanced analytical tools and algorithms for processing, analyzing and drawing insights

What makes Data Scientist Special?

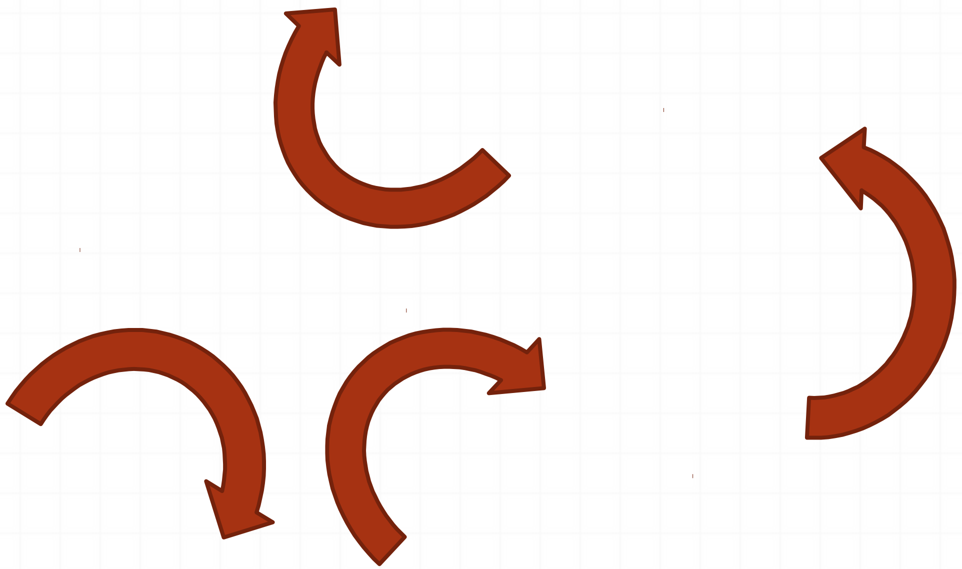
- Traditional Data Analyst explains what is going on by processing the history of the data.
- Data Scientist
  - does an Exploratory Data Analysis
  - Makes use of machine learning algorithms to identify the occurrence of a particular event in the future

Life cycle of Data Science

Myth:

- Directly go for data collection and analysis

Life cycle of Data Science



Life Cycle of Data Science

Phase1 – Discovery: Frame the business problem and formulate initial hypothesis

- Understand:
  - Specifications
  - Requirements
  - Priorities
  - Required budget

Life Cycle of Data Science

Phase1 – Discovery: Frame the business problem and formulate initial hypothesis

- Ability to ask the right question
- Assess required resources present in terms of
  - People, Technology, Time , Data to support the project

Life Cycle of Data Science

Phase1 – Discovery: Frame the business problem and formulate initial hypothesis

- Identify people and key stakeholders
- Identify the data sources

Life Cycle of Data Science

Phase2 – Data Preparation:

- Prepare Analytic Sandbox/workspace
- Perform ELT(Extract, Load, Transform)
- Learning about the data
- Data Conditioning
- Survey & Visualize

Life Cycle of Data Science

Phase2 – Data Preparation:

It includes steps:

- to explore,
- preprocess and
- condition data

Most labour intensive step in the analytical life cycle

Its generally the most iterative phase

Life Cycle of Data Science

Phase2 – Data Preparation:

Require analytics sandbox to perform analytics for the entire duration of the project.

Life Cycle of Data Science

- Phase2 – Data Preparation: Require analytics sandbox to perform analytics for the entire duration of the project.

- A secure private cloud is set up in an organization behind the firewall
- It is interconnected to the IT production environment for data sharing
- The users create simple functions using a language built for specific purpose
- Sandbox gives the analyst a safe place to build models and run experiments to analyse how the company works under such conditions

Life Cycle of Data Science

Phase2 – Data Preparation:

Life Cycle of Data Science

Phase2 – Data Preparation:

- Perform ELT to get the data into sandbox:
  - Extract
  - Load
  - Transform
- Data cleaning, transformation and visualization can be done using R which will help to
  - Spot outliers
  - Establish a relationship between the variables

Life Cycle of Data Science

Phase2 – Data Preparation: Learning about the data

Its important to get familiar with the data

- List your data sources
- List what is needed vs. what is available
- Highlight gaps – Identifies data not currently available
- Identifies the data outside the organization that might be useful

Life Cycle of Data Science

Phase2 – Data Preparation: Data Conditioning

Clean and normalize the data

Discern what you keep and what you discard

Life Cycle of Data Science

Phase2 – Data Preparation: Survey & Visualize

- Overview, filter then maintain data of interest
- Descriptive statistics
- Use data visualization tools to gain an overview of data
  - Does the data contains unexpected values?

Life Cycle of Data Science

- Learning about the Data: Sample Dataset Inventory

Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Center Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

Life Cycle of Data Science

Phase3 – Model Planning:

- Data Exploration
- Variable Selection
- Model Selection

Life Cycle of Data Science

Phase3 – Model Planning: Data Exploration

- Assess the data the understand the relationship between the variables
- Assess the structure of the data – this dictates the tools and analytic techniques for the next phase

Life Cycle of Data Science

Phase3 – Model Planning: Variable Selection

- Explore the data to select variables and methods using visualization tools
- Inputs from stakeholders and domain experts
- Iterative testing to confirm the most significant variables

Life Cycle of Data Science

Phase3 – Model Planning: Model Selection

- Choose an analytical technique based on the end goal of the project

Life Cycle of Data Science

Phase3 – Model Planning:

- Determine the methods and techniques to draw the relationships between the variables
- relationships will set the base for choosing the algorithms which you will implement in the next phase.
- insights into the nature of your data and have decided the algorithms to be used

Life Cycle of Data Science

Phase3 – Model Planning:

The problem to Solve	The category of techniques
I want to group items by similarity. I want to find structures(commonalities) in the data	Clustering
I want to discover relationships between actions or items	Association Rules
I want to determine relationship between the outcome and the input variables	Regression
I want to assign (known) labels to objects	Classification

Life Cycle of Data Science

Phase3 – Model Planning:

The problem to Solve	The category of techniques
I want to fine the structure in a temporal process. I want to forecast the behavior of a temporal process	Time series analysis
I want to analyze my text data	Text analysis

Life Cycle of Data Science

Phase4 – Model Building:

- develop datasets for training and testing purposes.
- consider whether your existing tools will suffice for running the models
- analyze various learning techniques like classification, association and clustering to build the model.

Life Cycle of Data Science

Phase5 – Operationalize:

- deliver final reports, briefings, code and technical documents
- a pilot project is also implemented in a real-time production environment
- To get a clear picture of the performance and other related constraints on a small scale before full deployment

Life Cycle of Data Science

Phase6– Communicate Results:

- Evaluate the results achieved with the goals mentioned in phase1.
- By communicating the key findings with the stake holder

Case study: Diabetes Prevention

Aim: Predict the occurrence of diabetes.

Step1: Data collection based on medical history of the patient

Attributes:

1. ngravid - No. of times pregnant
2. glu - Plasma glucose concentration
3. bp - Blood pressure
4. skin - Triceps skinfold thickness
5. bmi - Body mass index
6. ped - Diabetes pedigree function
7. age - Age
8. income - Income

Case study: Diabetes Prevention

Patient id	ngravid	glu	bp	skin	bmi	ped	age	income
1	6	148	72	35	33.6	0.627	50	
2	**one**	85	66	29	26.6	0.351	31	
3	1	97	**6600**	15	23.2	0.487	22	

Case study: Diabetes Prevention

Step2: Clean and prepare the data for data analysis

1. Data consists of inconsistencies like missing values, blank columns and abrupt values
2. Data cleaning required

Case study: Diabetes Prevention

This data has a lot of inconsistencies.

1. In the column ngravid, 1 is written as ‘one’
2. In the column bp, one of the values is 6600 which is a huge value
3. Income column is blank. Does not give sense in contributing to diabetes. It has to be removed from the table.
4. Clean and preprocess the data by removing outliers so that it can be used for analysis.

Case study: Diabetes Prevention

Step3: Analysis

1. load the data into the analytical sandbox
2. apply various statistical functions to know about the number of missing values,
3. Visualize the data to get an idea of distribution of the data

Case study: Diabetes Prevention

Step4:Analysis the problem to fix the algorithm

1. we already have the major attributes for analysis like *npreg, bmi,* etc., so we will use supervised learning technique to build a model here.
2. decision tree because it takes all attributes into consideration in one go, like the ones which have a linear relationship as well as those which have a non-linear relationship.
3. linear relationship between *npravid* and

Case study: Diabetes Prevention

Step5:Run a small pilot project

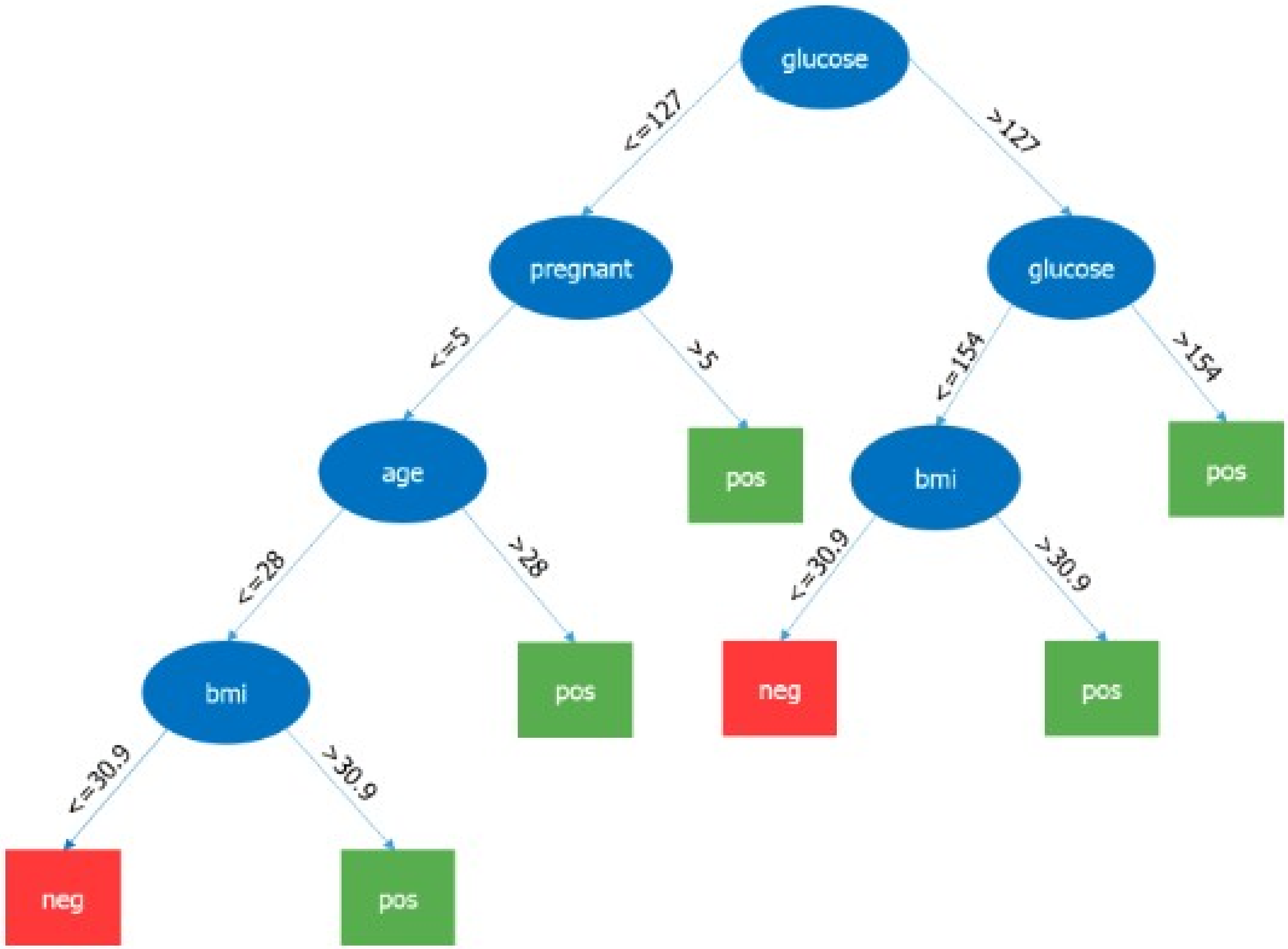
1. to check if our results are appropriate
2. look for performance constraints
3. Re-plan and rebuild the model if the results are not accurate

Case study: Diabetes Prevention

Step6: Full deployment

1. Once we have executed the project successfully, we will share the output for full deployment.

Case study: Diabetes



Applications: Healthcare

- Example: Reducing hospital readmissions
- In a pilot study, Manhattan's Mount Sinai Medical Center cut readmissions with the help of a computer model that predicts which patients have the highest chances of returning to the hospital.
- The model draws information on factors like disease and past hospital visits from hospital claims data.
- Caregivers give follow-up calls and other assistance to those likely to come back to the hospital.

Applications: Healthcare contd…

Medical Exams by Bathroom Mirrors:

- Bathroom mirrors, for example, could read one's skin temperature, pulse and blood pressure, alerting doctors to early indicators of health problems.

Applications: Urban Living

- Urban Informatics:
- System that allows police officers to remotely monitor traffic flows at key junctions via a network of cameras and sensors.
- Passing real-time information about traffic conditions to each other, re-routing themselves to keep traffic moving.

Applications: Crime Prevention

Predictive Policing:

- uses a constantly-calibrated feed of data on criminal incidents to tell officers where and when future crimes may occur.

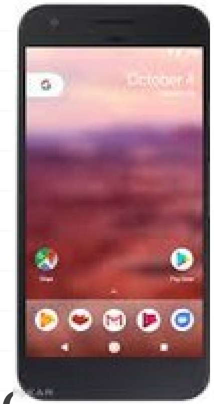
Case Study: Marketing Analytics

- Cross sell: It involves the sale of multiple products offered by a single product/service provider to a new or existing customer.

- Up sell: selling higher value products/services to an existing customer.

Case Study: Marketing Analytics

- Cross Selling:



- Up Selling:

Next Best Product to Recommend Model Framework

- This model provide answers to the following process:
  - What – choice of product
  - Whom – selection of customers
  - When – timing
  - How – contact strategy

Next Best Product to Recommend Model Framework

Target variable is response = 1 if the customer has expressed intent to buy, 0 otherwise. Get the response rate.

Response rate is no. of responders divided by customers contacted for the offer.

Next Best Product to Recommend Model Framework

- Response Model: The following table shows the variables description and relationship to cross-sell response in descending order of importance

Next Best Product to Recommend Model Framework

Variable Description	Relationship with new product response
% of times customer has reacted positively when contacted for an offer	positive
Total credit card limit	positive
Ratio of international spend on the card	positive
Is a premium card holder	positive
Average cash usage on credit card	positive

Next Best Product to Recommend Model Framework

- **Implementation**:
- Based on the response model, a cut-off of the score can be decided
- Customers exceeding the cut-off should be considered for marketing.

Machine Learning

- Machine learning is a set of algorithms that train
  - on a data set to make predictions or
  - take actions in order to optimize some systems

Machine Learning Algorithms

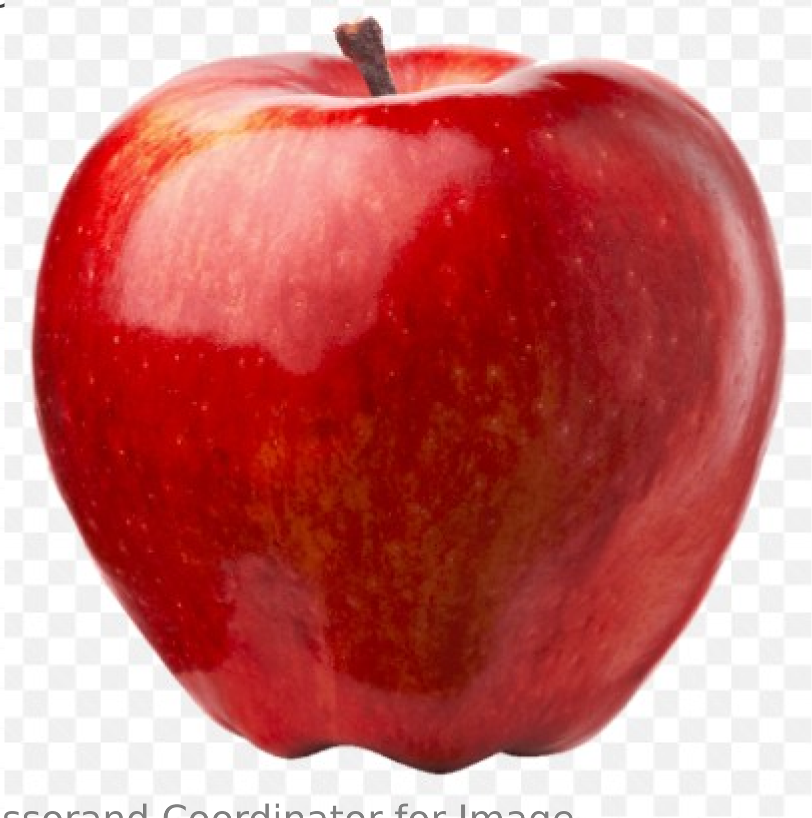
Supervised Learning

Supervised Learning

Supervised learning

Supervised Learning

- This algorithm consist of a target / outcome variable (or dependent variable)



Supervised Learning

- It is to be predicted from a given set of predictors (independent variables)

Supervised Learning

- Using these set of variables, we generate a function that map inputs to desired outputs

Supervised learning

The training process continues until the model achieves a desired level of accuracy on the training data

Unsupervised Learning

Unsupervised Learning

we do not have any target or outcome variable to predict / estimate

Unsupervised Learning

It is used for clustering population in different groups

Unsupervised Learning

- widely used for segmenting customers in different groups for specific intervention

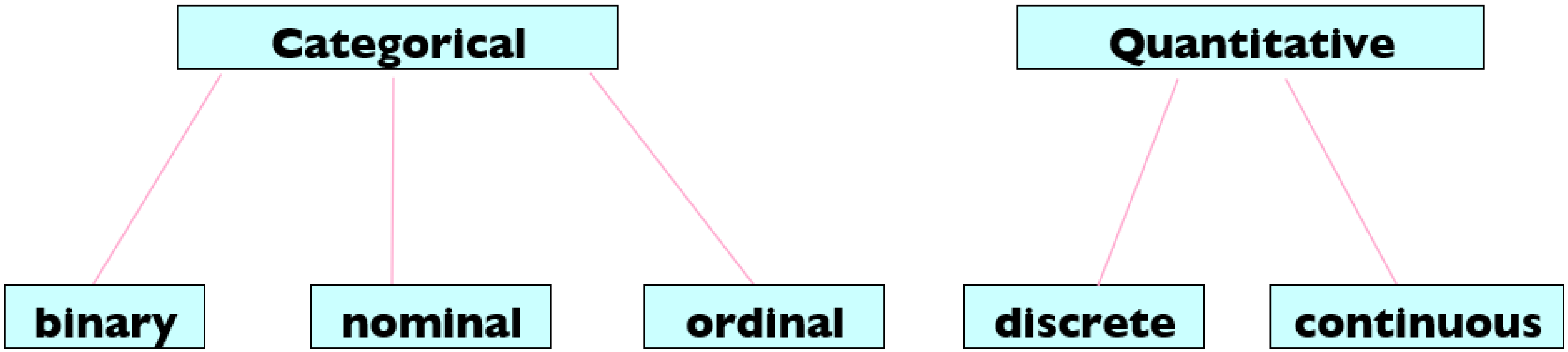
Reinforcement Learning

- the machine is trained to make specific decisions
- the machine is exposed to an environment where it trains itself continually using trial and error
- This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions

EDA

- Before making inferences from data it is essential to examine all your variables.
- Why?
  - to catch mistakes
  - to see patterns in the data
  - to find violations of statistical assumptions
  - to generate hypotheses

Types of Data



2 categories

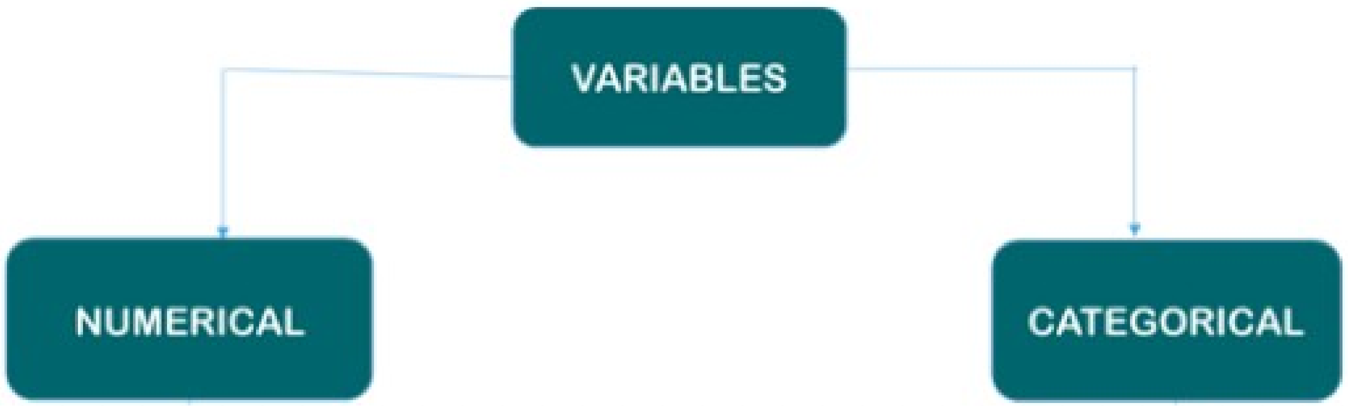
more categories

order matters

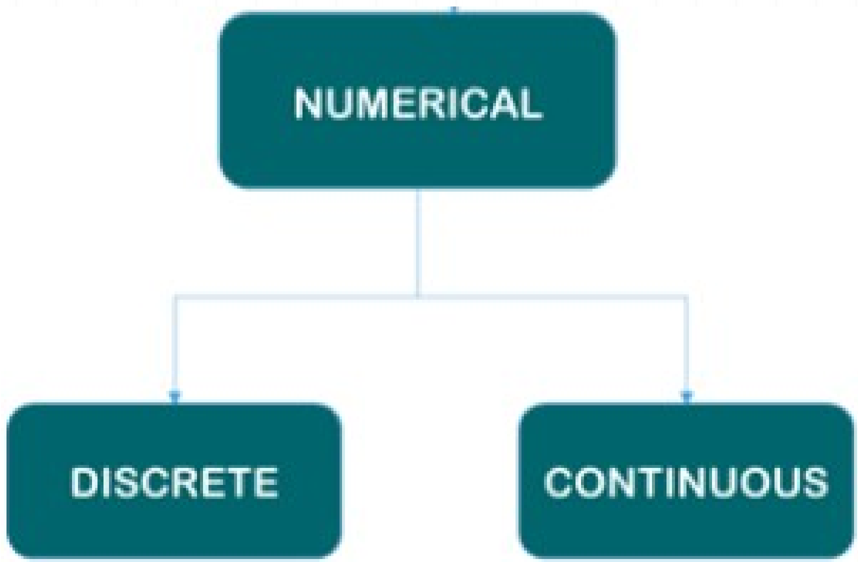
numerical

uninterrupted

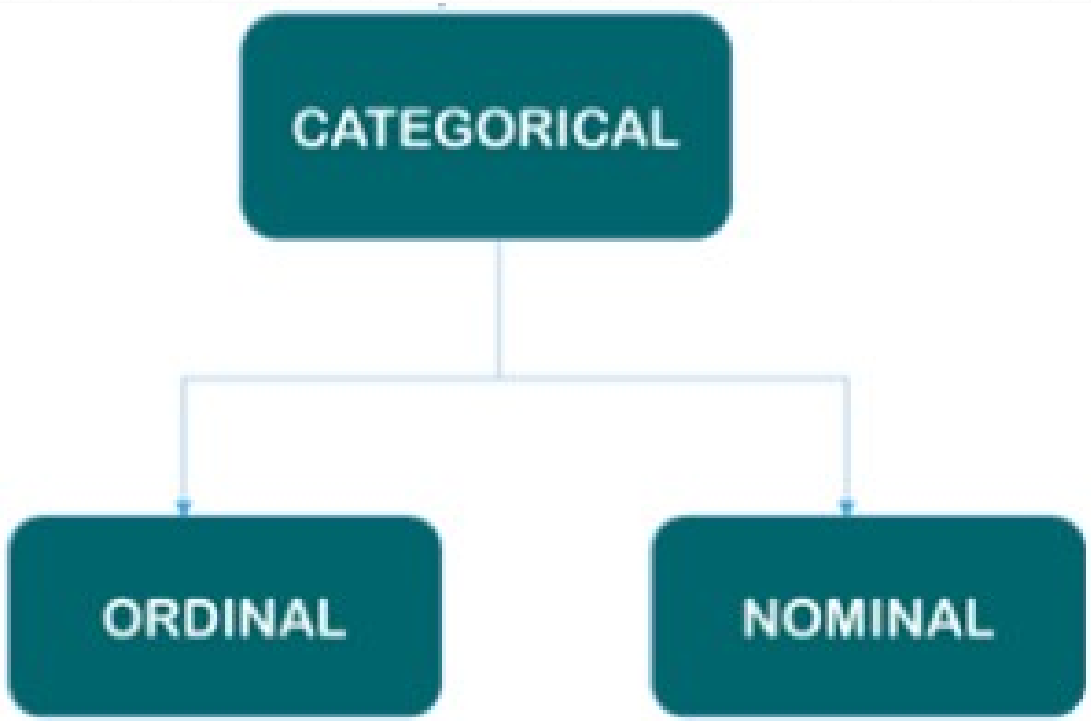
Exploring Data Types



Exploring Data Types



Exploring Data Types



Exploring Data Types

Exploring Data Types

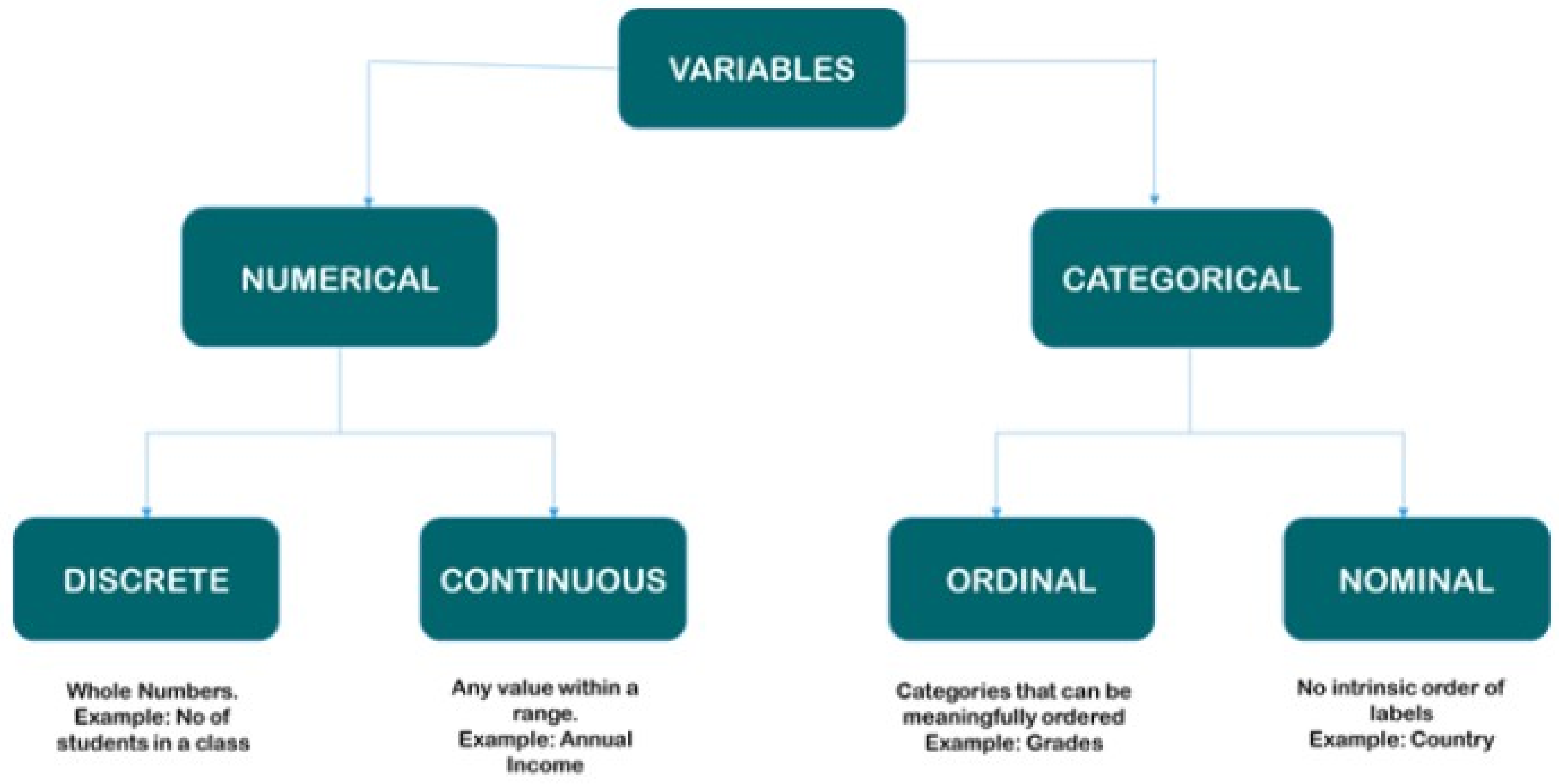
ORDINAL

Categories that can be meaningfully ordered
Example: Grades

NOMINAL

No intrinsic order of labels
Example: Country

Classification of Variables



Dimensionality of Data Sets

- Univariate: Measurement made on one variable per subject
- Bivariate: Measurement made on two variables per subject
- Multivariate: Measurement made on many variables per subject

Univariate, Bivariate, Multivariate Data and its analysis

- Univariate Data:
- This type of data consists of **only one variable**
- It does not deal with causes or relationships
- the main purpose of the analysis is to describe the data and find patterns that exist within it

Univariate, Bivariate, Multivariate Data and its analysis

- The example of a univariate data can be height.

Heights (in cm)	**164**	**167.3**	**170**	**174.2**	**178**	**180**	**186**

- Conclusions can be drawn using central tendancy measures (mean, median,mode)
- Dispersion or spread of data (range, min, max, quantile, variance and SD)
- Graphical Representation using histogram, pie-chart, bar chart

Univariate, Bivariate, Multivariate Data and its analysis

- Bivariate Data:
- This type of data involves **two different variables**
- the analysis is done to find out the relationship among the two variables
- Example of bivariate data can be temperature and ice cream sales in summer season.
- one of these variables is independent while the other is dependent.

Univariate, Bivariate, Multivariate Data and its analysis

- the relationship is visible from the table that temperature and sales are directly proportional to each other
- as the temperature increases, the sales also incr

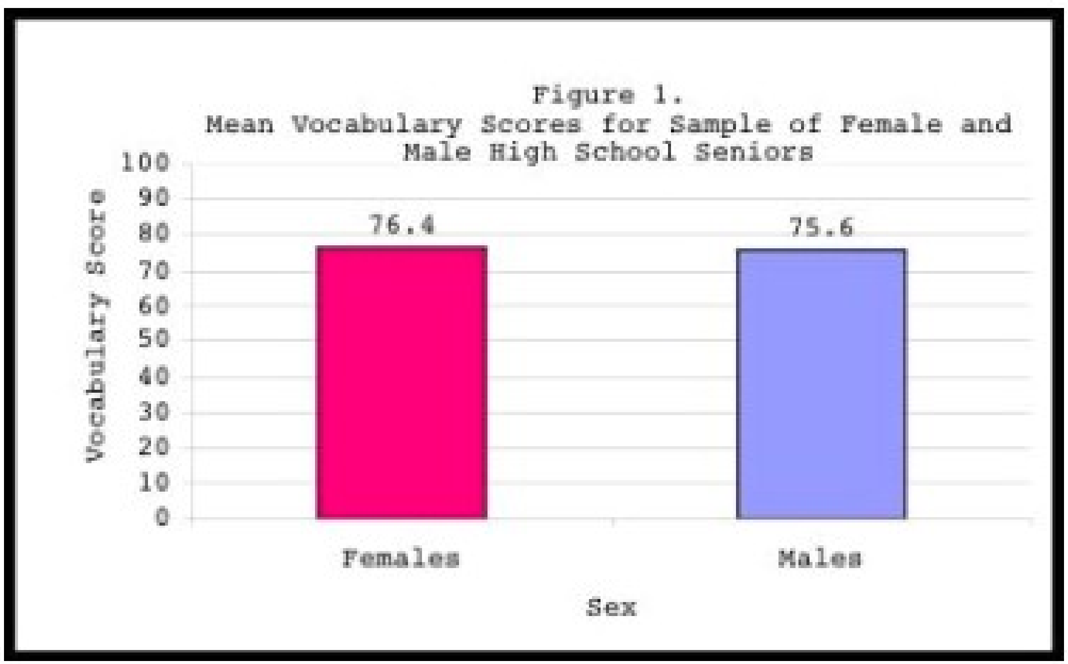
TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

Bivariate Analysis

- Relationship between two variables
- Independent variable and dependent variable
- Independent variable: are not affected by anything
- Dependent variable: That can be changed by the outside factors
- Control variable:It may alter either a dependant variable or independent variable

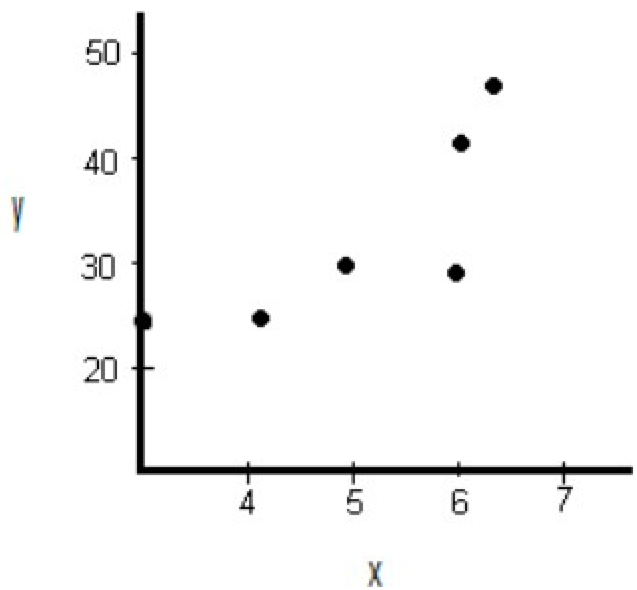
Bivariate Analysis - Example

- Consider the given graph where gender is independent variable and mean vocabulary score is a dependant variable
- the mean vocabulary scores depend on the independ ant is male or f



Types of Bivariate Analysis

- Scatter Plots: These give you a visual idea of the pattern that your variables follow.



Types of Bivariate Analysis

- Correlation coefficient
- This coefficient tells you if the variables are related
- zero means they aren’t correlated
- while a 1 (either positive or negative) means that the variables are perfectly correlated
- Example:
- P

$$r = \frac{\text{mean}(XY) - \text{mean}(X) \times \text{mean}(Y)}{\text{SD}(X) \times \text{SD}(Y)}$$

Attribute Selection-Example

X	Y
10	30
15	45
20	60
25	65
30	80

x	y	xy	x2	y2
10	30			
15	45			
20	60			
25	65			
30	80			

= [(Xy/n) - (x/n)*(y/n)]/sqrt(((x2/n) – sqr(x/n)) * ((y2/n)-sqr(y/n)))

Attribute Selection - Example

- If correlation is ≤ 0.5 it is interpreted as very low.
- If correlation is between 0.51 to 0.79 it is interpreted as low.
- If correlation is between 0.80 to 0.89 it is interpreted as moderate.
- If correlation is ≥ 0.90 it is interpreted as high.

- Inference: Variables x,y are highly correlated then any one can be removed

Univariate, Bivariate, Multivariate Data and its analysis

- Multivariate Data:
- involves **three or more variables**
- Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.
- It contains more than one dependent variable
- Technique: PCA

Numerical Summaries of Data

- Central Tendency measures: They are computed to give a “center” around which the measurements in the data are distributed.
- Variation or Variability measures. They describe “data spread” or how far away the measurements are from the center.
- Relative Standing measures. They describe the relative position of specific measurements in the data.

Estimates of Location

- Variables with measured or count data might have thousands of distinct values.
- A basic step in exploring your data is getting a
  - "typical value" for each feature (variable):
  - an estimate of where most of the data is located (i.e., its central tendency).

Key terms for estimates of location

- Mean:
- The most basic estimate of location is the mean, or *average* value
- The mean is the sum of all the values divided by the number of values
- Consider the following set of numbers: {3 5 1 2}.
- The mean is (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75.
- Mean or x-bar represents the mean of a sample from a population.

Estimates of location

- Mean:

To calculate the average $\bar{x}$ of a set of observations, add their value and divide by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Estimates of location

- Trimmed mean:
- Trimmed means are widely used, and in many cases, are preferable to use instead of the ordinary mean
- A trimmed mean eliminates the influence of extreme values
- *trimmed mean* is calculated by dropping a fixed number of sorted values at each end and then taking an average of the remaining values.
- the sorted values
  by x (1) , x (2) , ... , x (n)where x (1) is the smallest value and x (n) the largest, trimmed mean = sum(x(2)+x(3)+....x(n-1))/(n-2)

Estimates of location

- Example: Consider an international diving competition in which 5 judges are employed. The top and bottom scores from five judges are dropped, and the final score is the average of the three remaining judges.
- This makes it difficult for a single judge to manipulate the score, perhaps to favor his country’s contestant.

Estimates of location

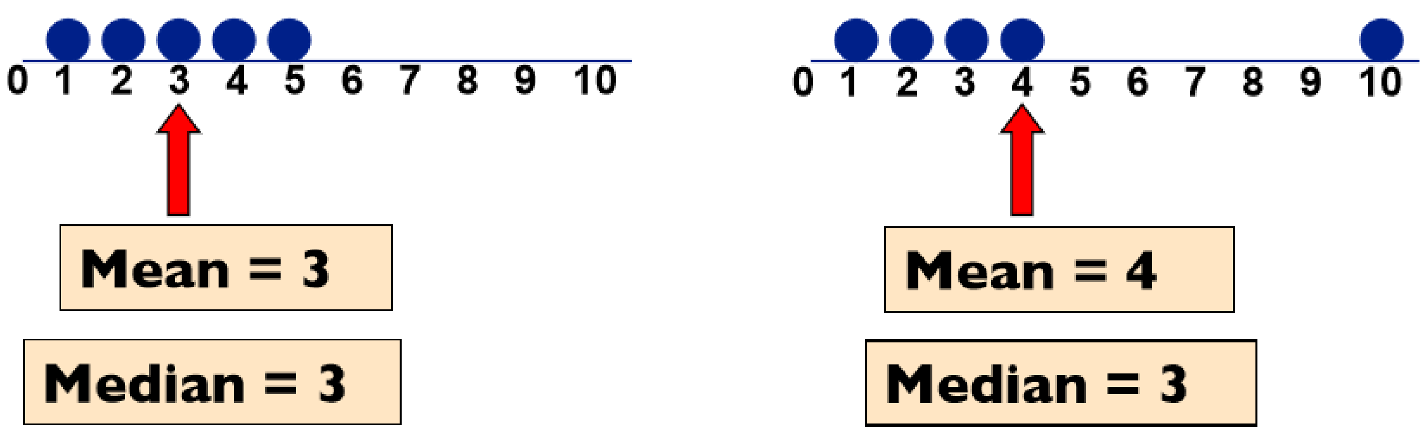
- Weighted mean is calculated by multiplying each data value (xi) by a weight(wi) and dividing their sum by the sum of weights

Estimates of location

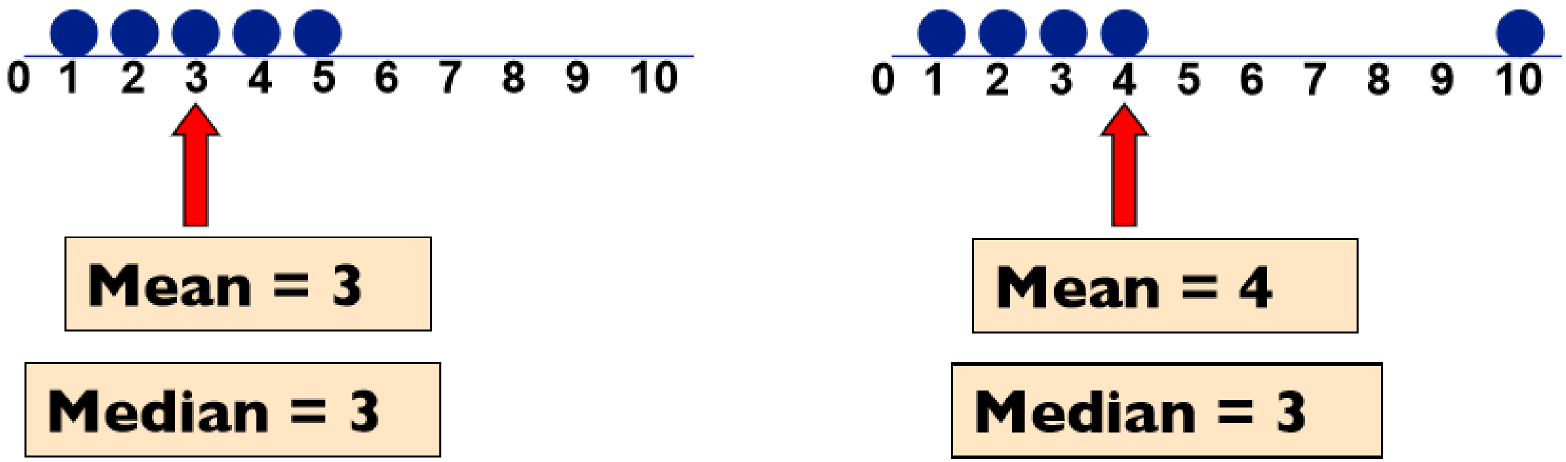
- Median is the middle number on a sorted list of the data
- If there are an odd number of observations, find the middle value
- \- If there are an even number of observations, find the middle two values and average them
- Age of participants: 17 19 21 22 23 23 23 38
- Median = (22+23)/2 = 22.5

Which Location is best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers
- Assume that 1 to 10 refers to the salary of a person in x place where a celebrity is also in



Which Location is best?



Outliers

- An outlier is any value that is very distant from the other values in a data set
- When outliers are the result of bad data, the mean will result in a poor estimate of location, while the median will be still be valid
- In anomaly detection, the points of interest are the outliers

Standard Deviation

- *deviations*, between the estimate of location and the observed data.
- For a set of data {1, 4, 4}, the mean is 3 and the median is 4
- The deviations from the mean are the differences: 1 – 3 = –2, 4 – 3 = 1 , 4 – 3 = 1
- The variance is an average of the squared deviations
- standard deviation is the square root of the variance.

Variance

- Average of squared deviations of values from the mean

$$\hat{\sigma}^2 = \frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard Deviation

- Standard Deviation is the square root of variance

$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}}$$

Estimates based on percentile

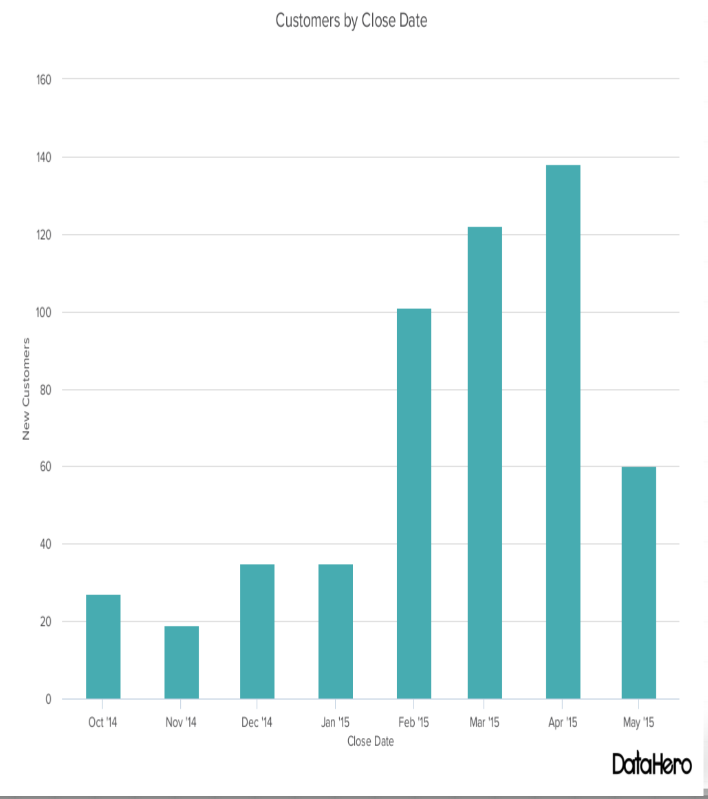
- Statistics based on sorted (ranked) data are referred to as *order statistics*.
- In a data set, the *P*th percentile is a value such that at least *P* percent of the values take on this value or less
- For example, to find the 80th percentile, sort the data. Then, starting with the smallest value, proceed 80 percent of the way to the largest value.

IQR

- Difference between the 25th percentile and the 75th percentile, called the *interquartile range* (or IQR)
- Example:
- Consider 3,1,5,3,6,7,2,9
- Sort 1,2,3,3,5,6,7,9
- 25th percentile is at 2.5 (1,2,3,3,5,6,7,9) median (50%)
- So 25th percentile is (1,2,3,3,5,6,7,9) = 2.5
- So 75th percentile is (1,2,3,3,5,6,7,9) = 6.5
- Interquantile range is 6.5 – 2.5 = 4

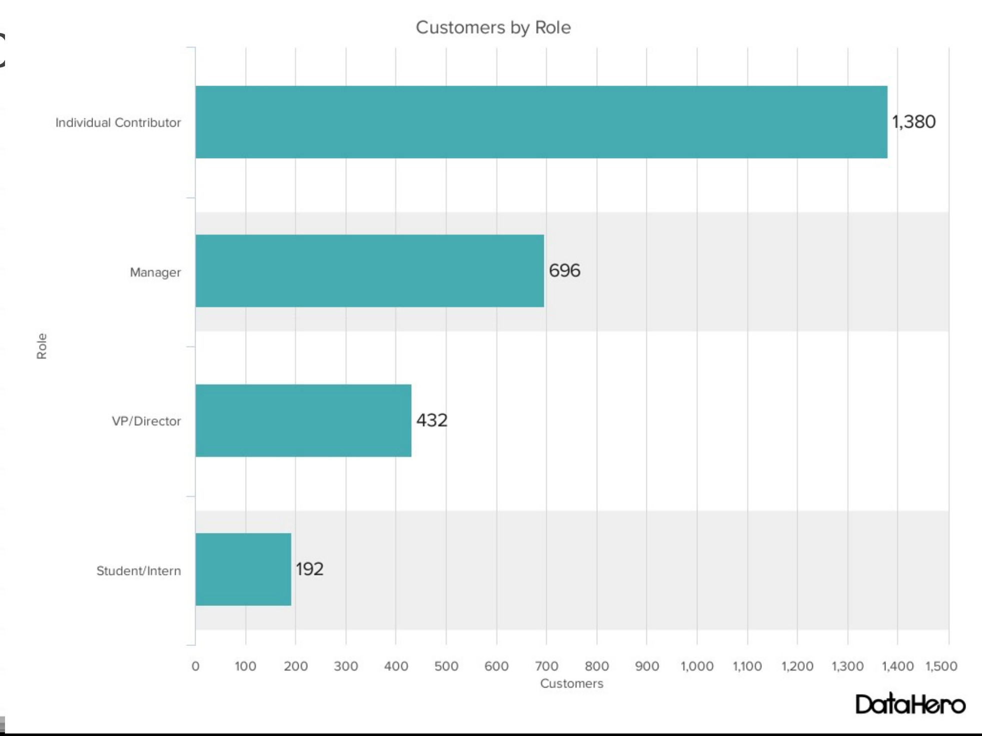
Exploring Data Distribution

- Column Chart
- A column chart is used to show a comparison among different items
- Example: Customers by close



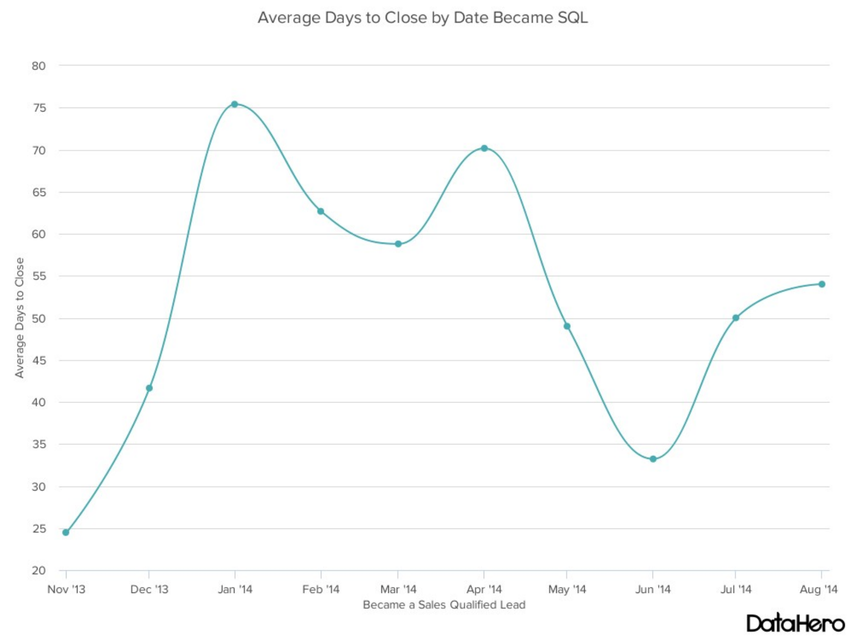
Exploring Data Distribution

- Bar Graph
- basically a horizontal column chart
- To avoid clutter when one data label is long and another is shc



Exploring Data Distribution

- Line Graph
- continuous data set.
- A line graph reveals trends or progress over time

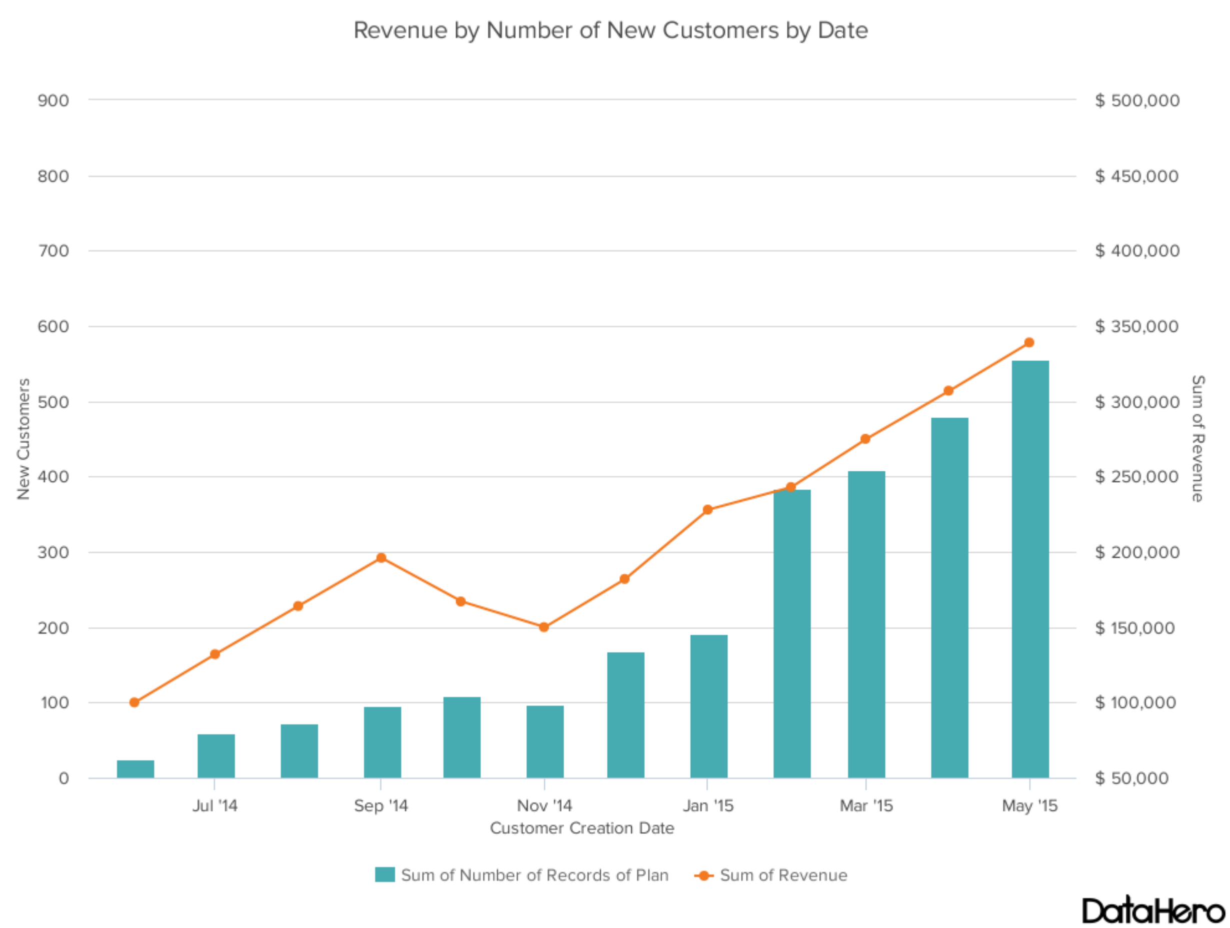


Exploring Data Distribution

- Dual Axis Chart
- allows you to plot data using two y-axes and a shared x-axis
- It's used with three data sets, one of which is based on a continuous set of data and another which is better suited to being grouped by category
- his should be used to visualize a correlation

Exploring Data Distribution

- Dual Axis Chart

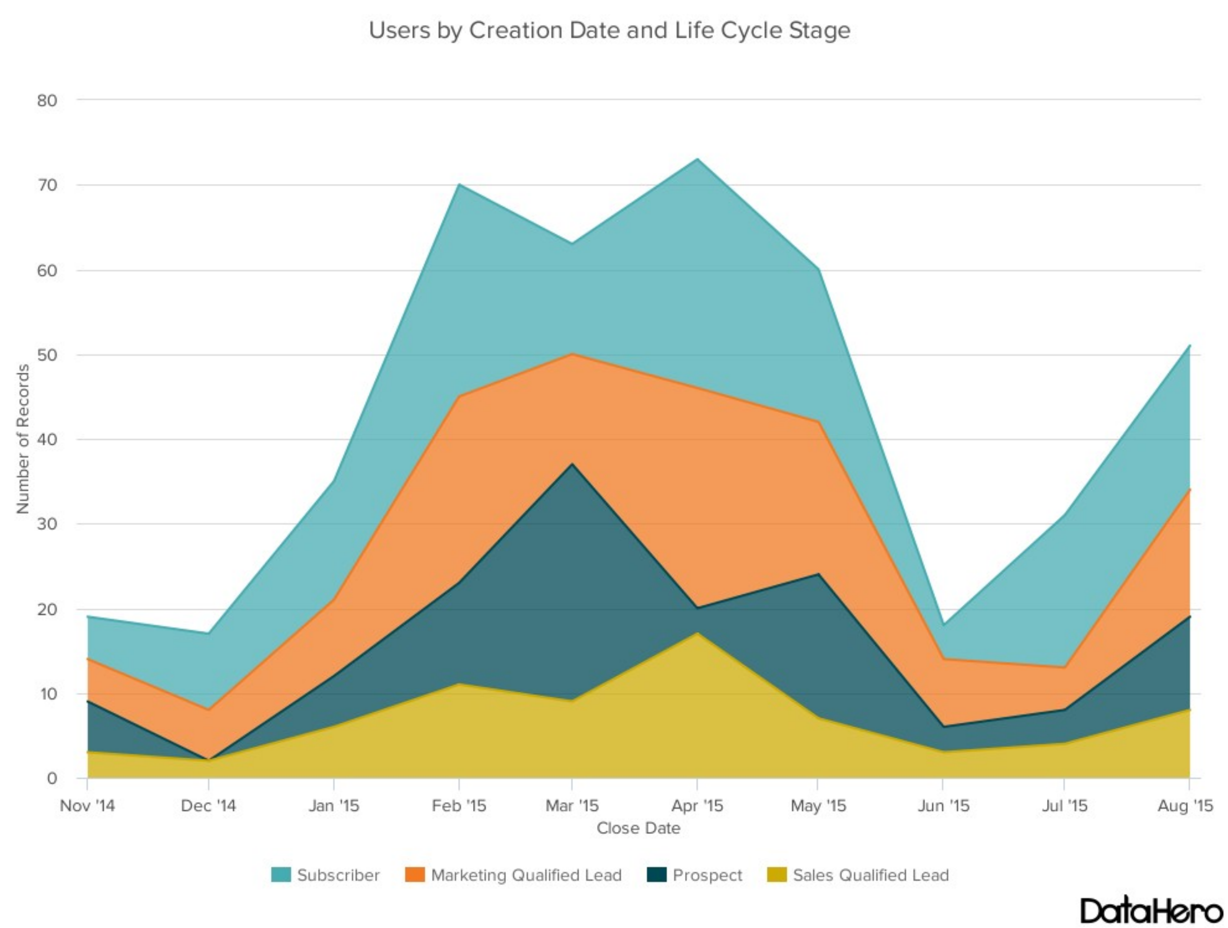


Exploring Data Distribution

- Area Chart
- An area chart is basically a line chart, but the space between the x-axis and the line is filled with a color or pattern
- It is useful for showing part-to-whole relations
- Example: individual sales reps' contribution to total sales for a year

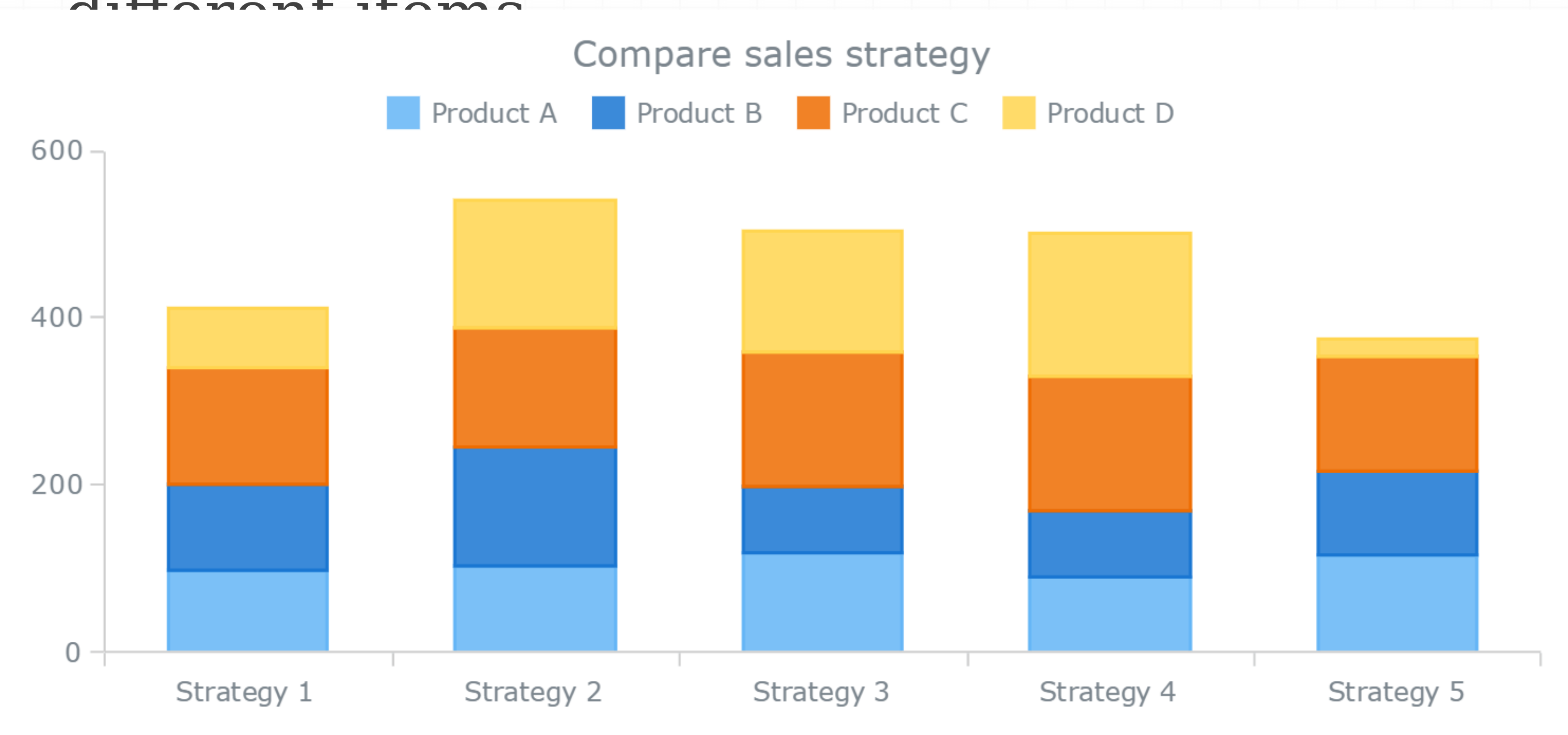
Exploring Data Distribution

- Area Char



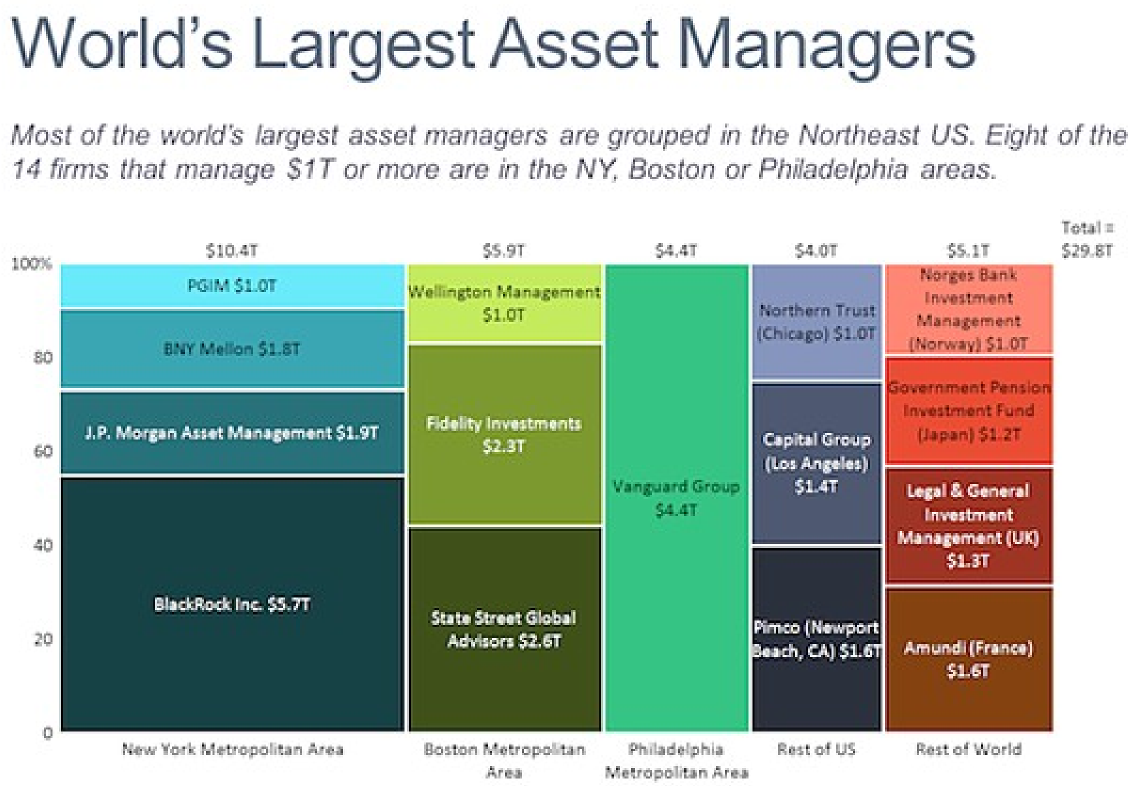
Exploring Data Distribution

- Stacked Bar Chart
- This should be used to compare many different items



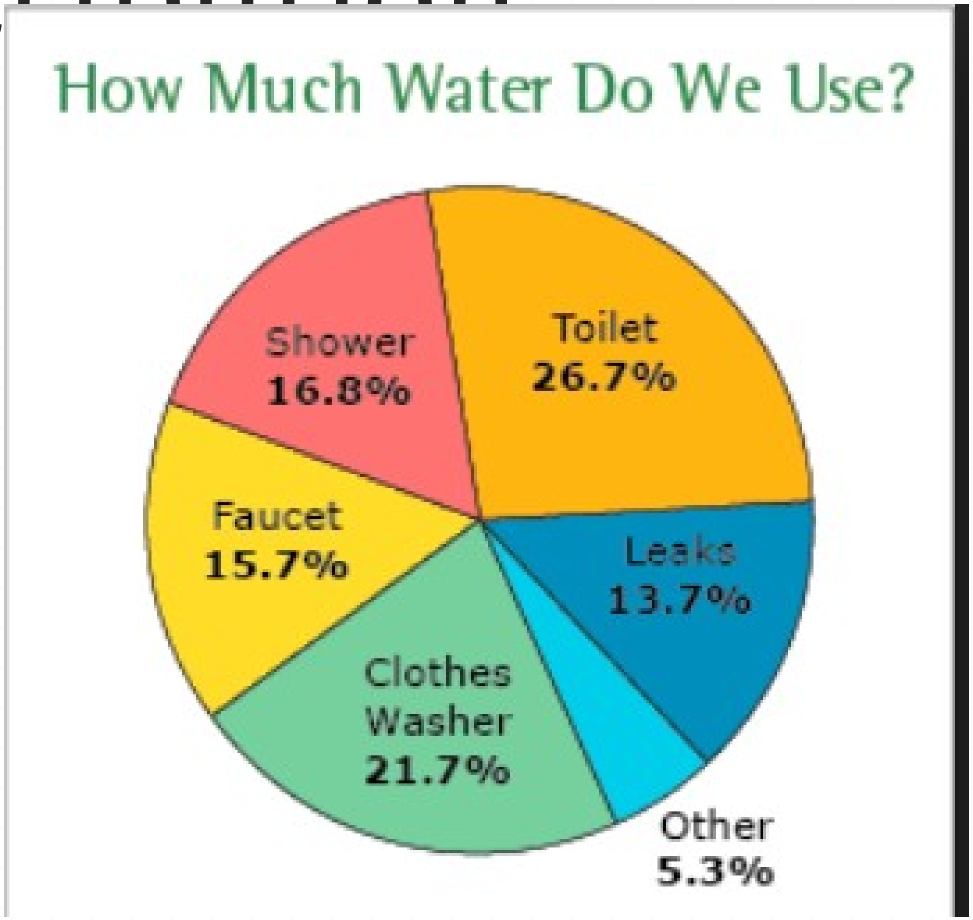
Exploring Data Distribution

- Mekko Chart
- This type of graph can compare values, measure each one's composition
- To show how your data is distributed across each one.



Exploring Data Distribution

- Pie Chart
- shows a static number
- how categories represent part of a whole
- the total sum of all segments needs to equal 100%



Exploring Data Distribution

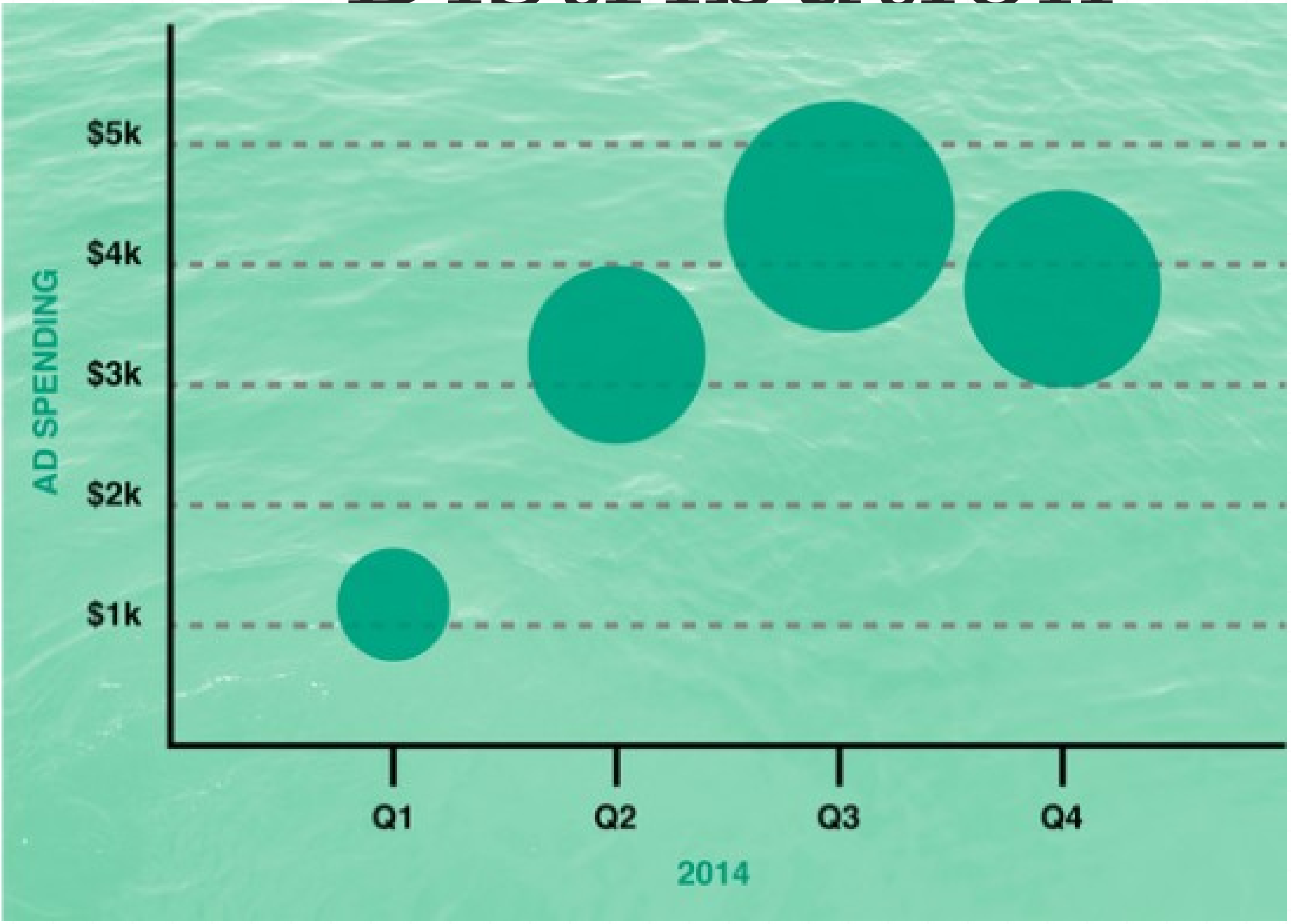
- Scatter Plot Chart
- Show the relationship between two different variables
- It should be used when there are many different data points, and you want to highlight similarities in the data set.
- This is useful when looking for outliers

Exploring Data Distribution

Exploring Data Distribution

- Bubble Chart
- is similar to a scatter plot in that it can show distribution or relationship.
- There is a third data set, which is indicated by the size of the bubble or circle.

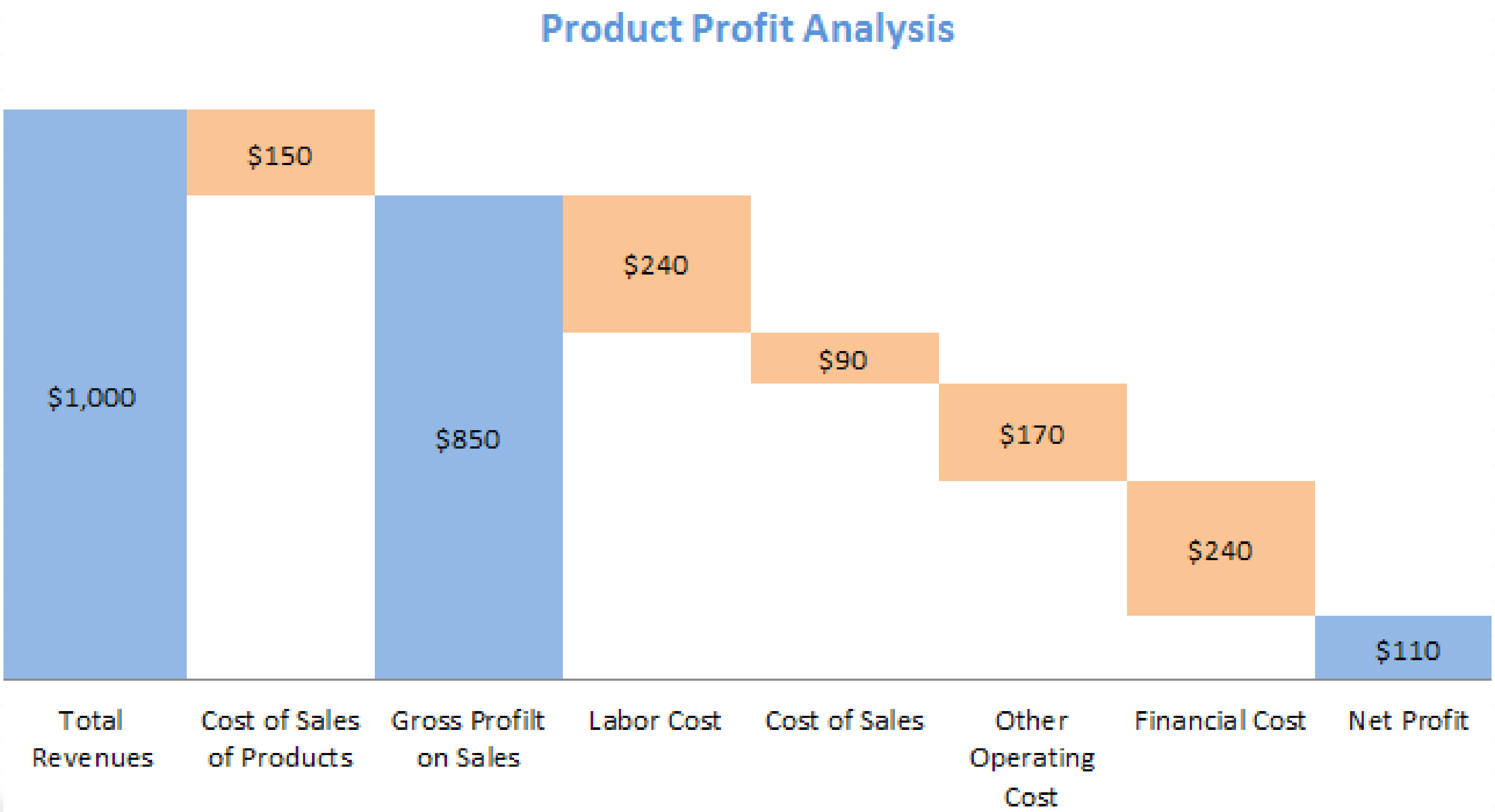
Exploring Data Distribution



Exploring Data Distribution

- Waterfall Chart
- A waterfall chart should be used to show how an initial value is affected by intermediate values -- either positive or negative -- and resulted in a final value.

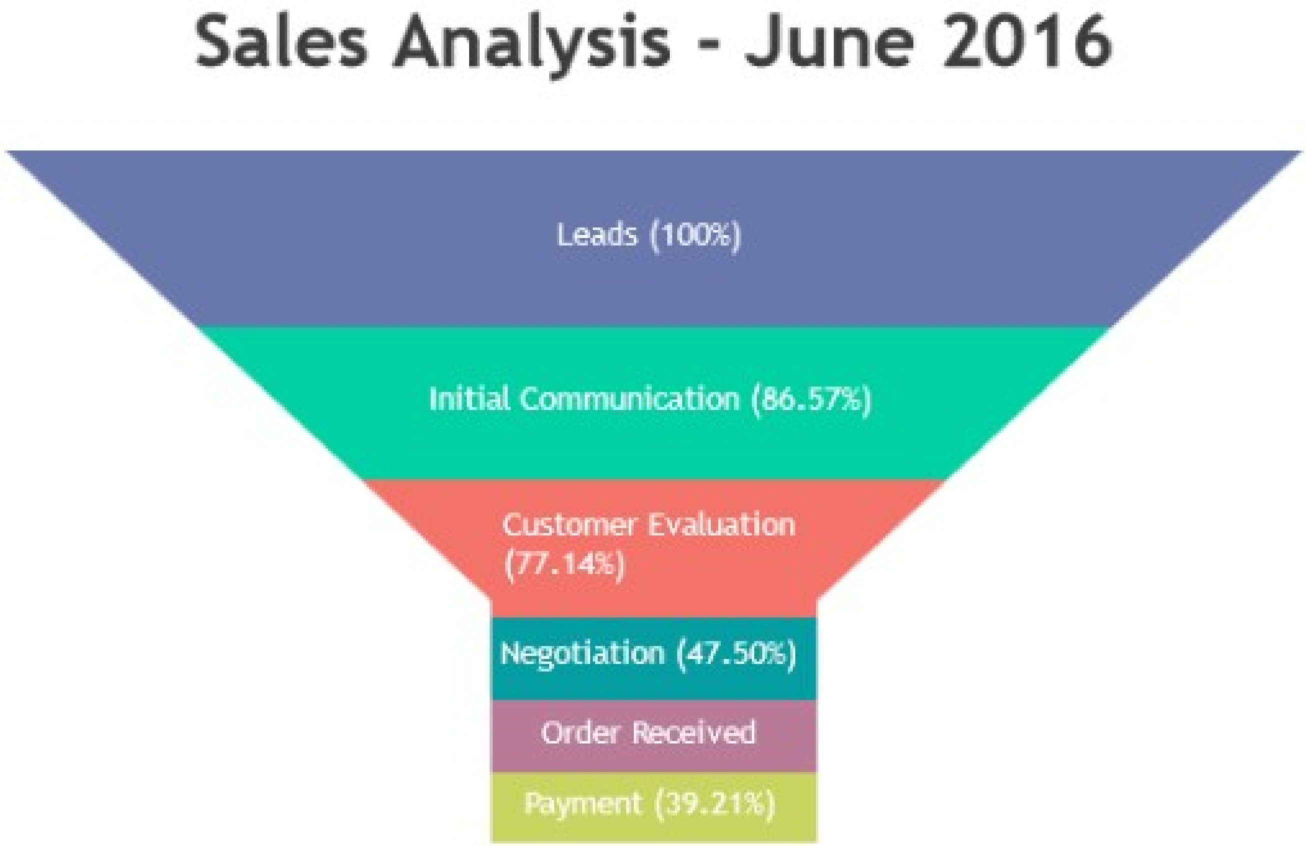
Exploring Data Distribution



Exploring Data Distribution

- Funnel Chart
- shows a series of steps and the completion rate for each step
- This can be used to track the sales process or the conversion rate across a series of pages or steps.

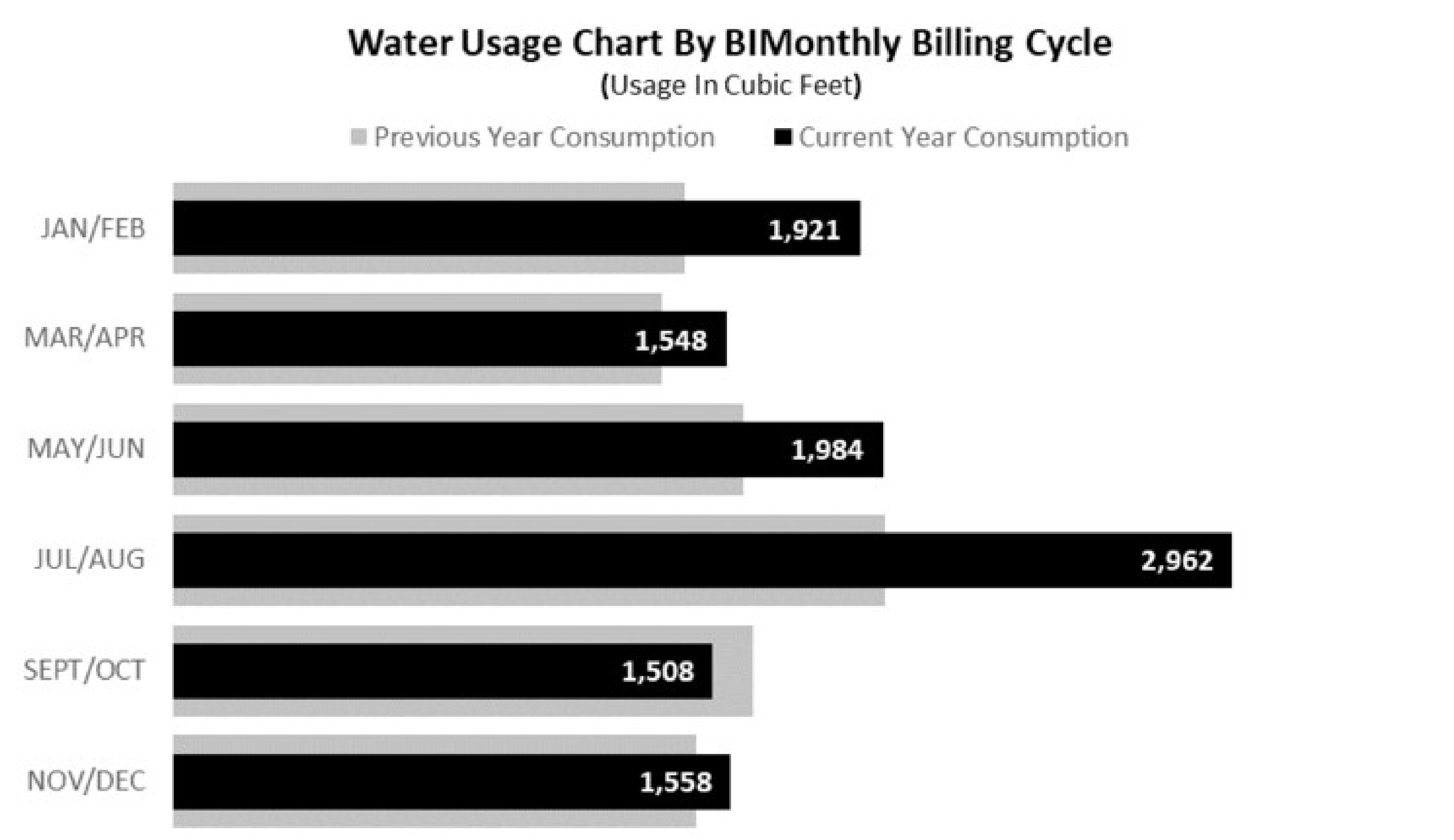
Exploring Data Distribution



Exploring Data Distribution

- Bullet Chart
- A bullet graph reveals progress toward a goal, compares this to another measure,
- provides context in the form of a rating or performance.

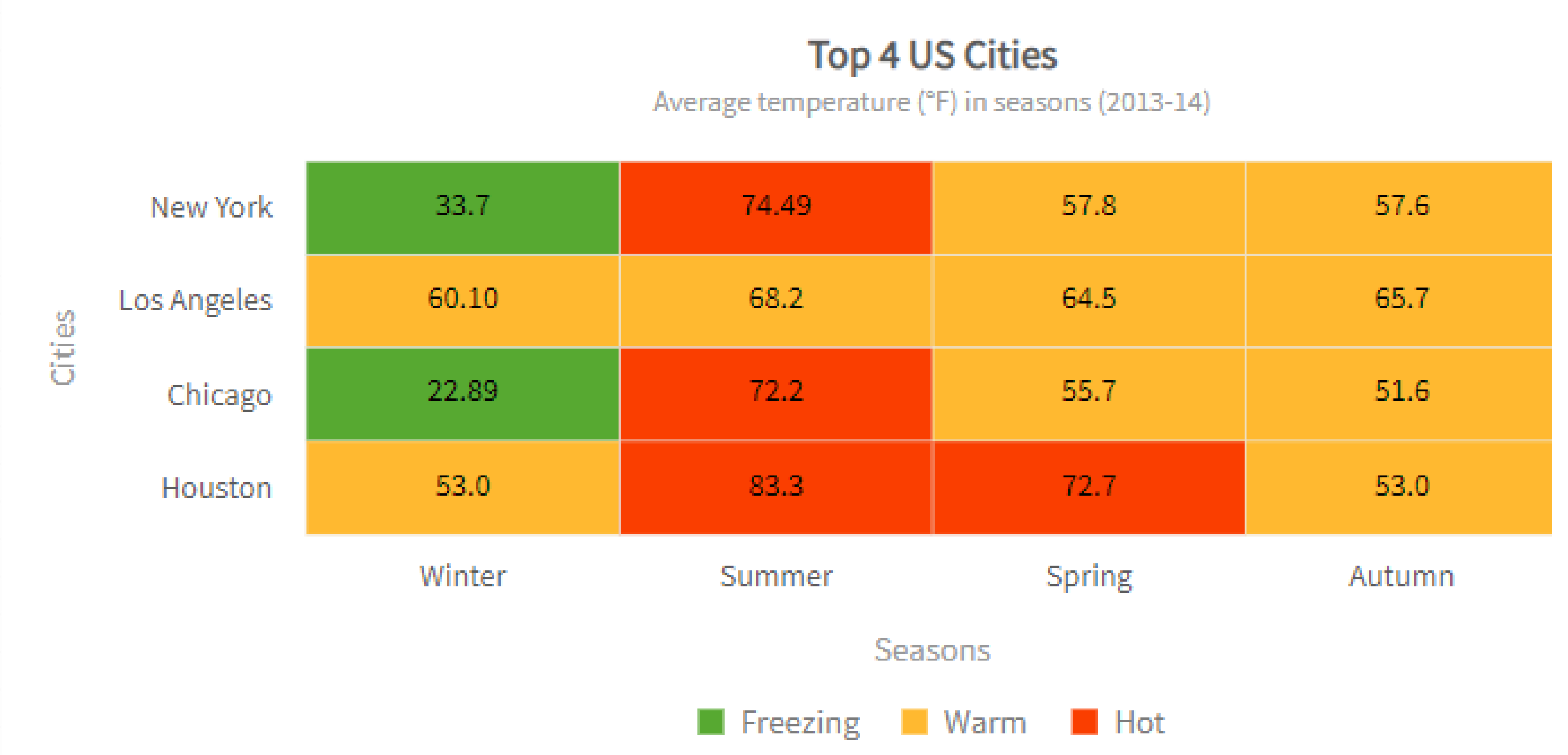
Exploring Data Distribution



Exploring Data Distribution

- Heat Map
- A heat map shows the relationship between two items and provides rating information, such as high to low or poor to excellent.
- to plot data like employee attendance records
- you can use colors like red, yellow, blue and green to indicate a bad, average, good, and excellent grade

Exploring Data Distribution



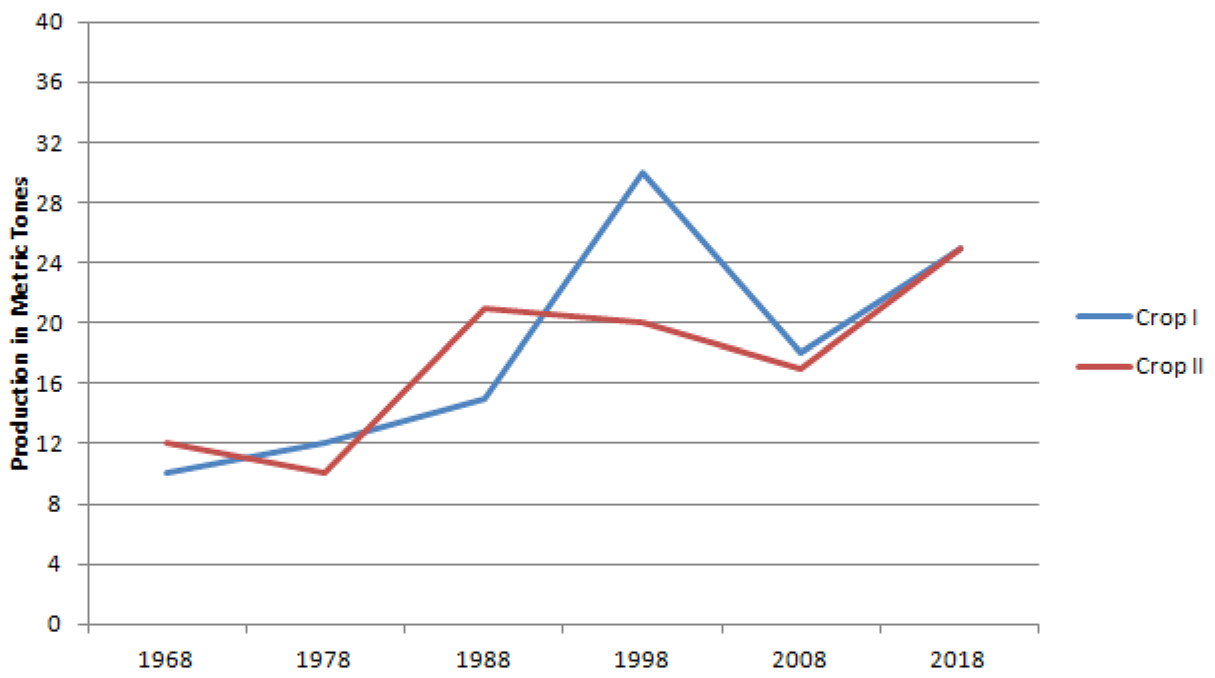
Example1

- Problem: Draw a line graph for the production of two types of crops for the given years.

	Production in metric tones	
Year	Crop I	Crop II
1968	10	12
1978	12	10
1988	15	21
1998	30	20
2008	18	17
2018	25	25

Example1

- Solution: The required graph is



Example 2

- Draw the histogram for the given data.

Marks	No. of Students
15 – 18	7
19 – 22	12
23 – 26	56
27 – 30	40
31 – 34	11
35 – 38	54
39 – 42	26
43 – 46	37
47 – 50	7
Total	250

Example2

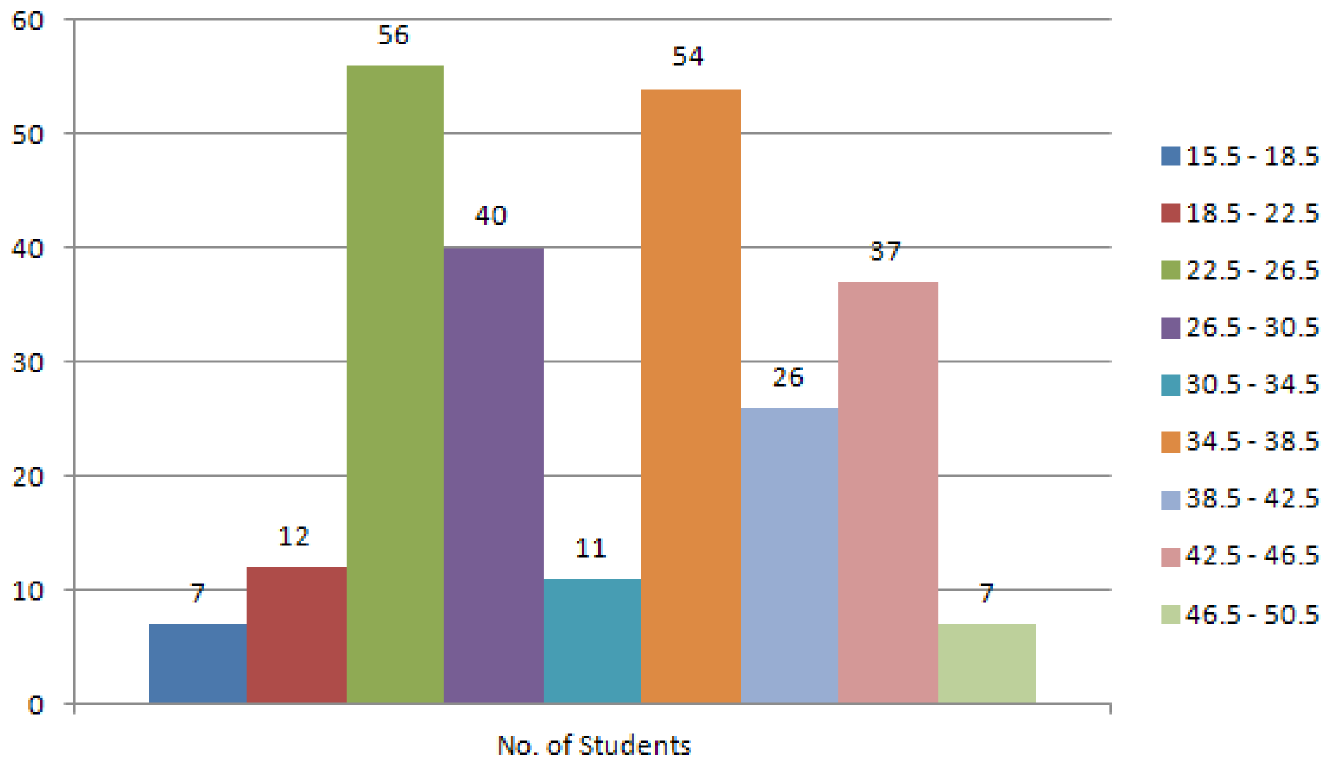
- **Solution:** This grouped frequency distribution is not continuous. We need to convert it into a continuous distribution with exclusive type classes. This is done by averaging the difference of the lower limit of one class and the upper limit of the preceding class. Here, $d = \frac{1}{2}(19 - 18) = \frac{1}{2} = 0.5$. We add 0.5 to all the upper limits and we subtract 0.5 from all the lower limits.

Example2

Marks	No. of Students
14.5 – 18.5	7
18.5 – 22.5	12
22.5 – 26.5	56
26.5 – 30.5	40
30.5 – 34.5	11
34.5 – 38.5	54
38.5 – 42.5	26
42.5 – 46.5	37
46.5 – 50.6	7
Total	250

Example2

- The corresponding histogram is:



Take Home:
Case Study: XYZ MNC Company

- Case: Possibilities of an enquiry being converted into a buyer for sure.
- Software Design Requirements:
- How many people in family
- Per month salary
- Loans if any
- Two wheeler/four wheeler possession if any
- Residing in rental or own
- Has life insurance policies
- Has dogs at home?
- Dog's breed?

CONTACT DETAILS

T
H
A
N
K

Y
O
U

Dr. G. Malathi
Associate Professor and Coordinator for
Image Processing Research Group
School of Computing Sciences and Engineering
VIT University
Chennai Campus
malathi.g@vit.ac.in
9840833337

Take Home:
Case Study: XYZ MNC Company

- Case: Possibilities of an enquiry being converted into a buyer for sure.
- Knows to drive
- Is married or lone
- Does he/she has children
- Frequent place of visits
- Which type of hotel preferred for stay?
- School in which the kids are studying?
- Restaurants they visit
- Frequency of visit to restaurant

Take Home:
Case Study: XYZ MNC Company

- Case: Possibilities of an enquiry being converted into a buyer for sure.
- Surprising Inferences from the above data obtained from the enquiry:
- If a person has insurance policies, he cares his family a lot and does not spend on higher end cars
- If a family has dogs at home, they are trustable. They will surely pay the lone amount
- The types of hotels or resorts they stay has 5 stars, then they can be pushed to choose higher end if they opt for sedan

Take Home:
Case Study: XYZ MNC Company

- Case: Possibilities of an enquiry being converted into a buyer for sure.
- Surprising Inferences from the above data obtained from the enquiry:
- If the person comes for enquiry more than once, then the chance of conversion to be a buyer is more if negotiation is initiated or offer value added services
- People with kids have 90% chances of buying

Why EDA?

- Detection of mistakes
- Checking of assumptions
- Preliminary selection of appropriate models
- Determining relationship among the variables

Exploratory Data Analysis (EDA) – An introduction

- Through EDA
  - get to know about the data
    - distributions,
    - Data quality problems
    - Outliers
    - Correlations and inter-relationships

164

Steps to Understand, clean and prepare

- Variables identification
- Univariate Analysis
- Bi-variate Analysis
- Missing values treatment
- Outliers treatment

165

Variable Identification

- Identify
  - Predictor(Input) variables
  - Target(Output) variables
  - Data type of the variables
  - Category of the variables

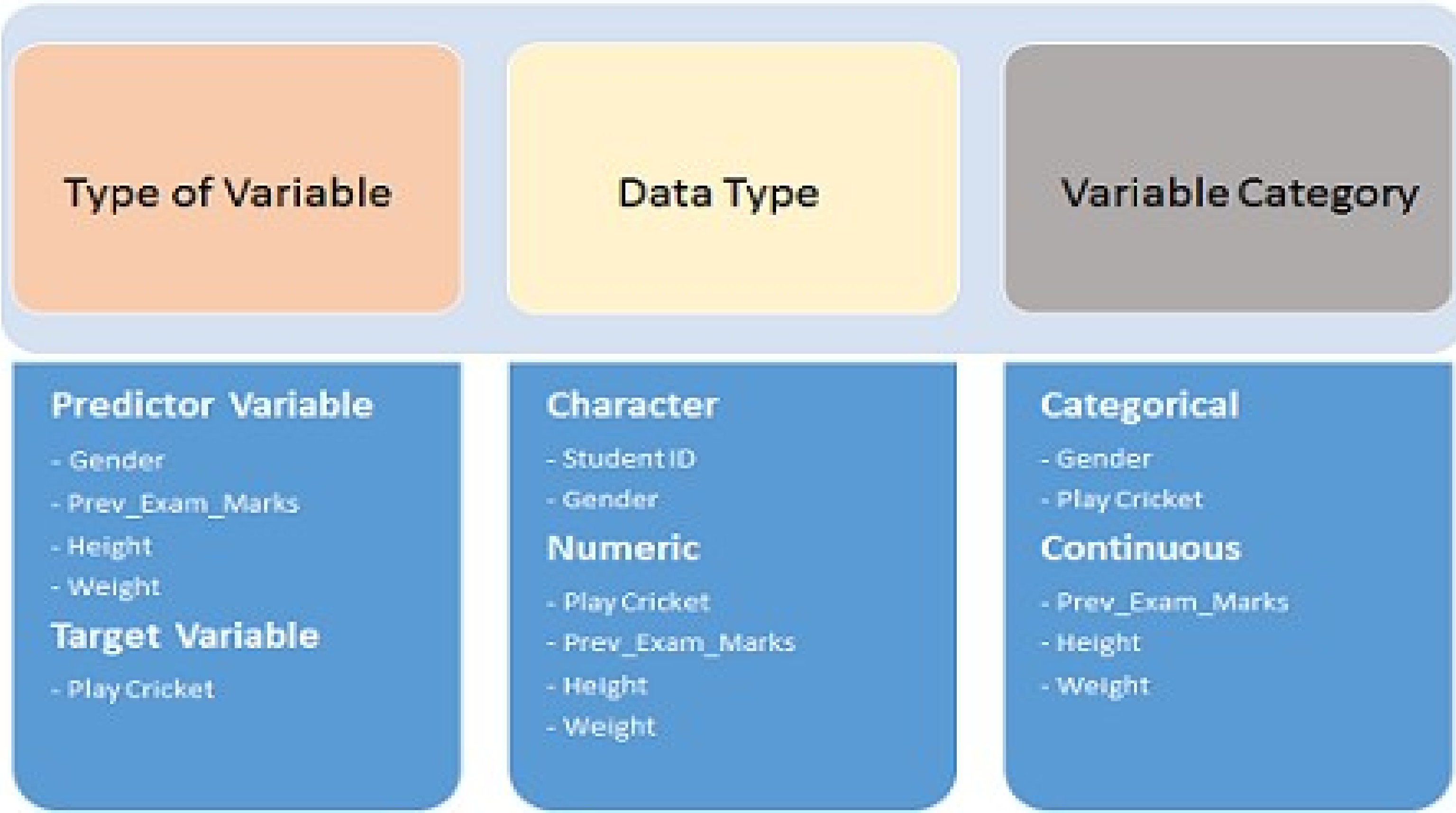
Variable Identification cont…

- To predict, whether the students will play cricket or not. Need to identify predictor variables, target variable, data type of variables and category

Target Variable

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

Variable Identification cont…



Types of Data

- Categorical data
  - Nominal data
  - Ordinal data
  - Binary data
- Measurement data
  - Discrete
  - Continuous

169

Categorical Data

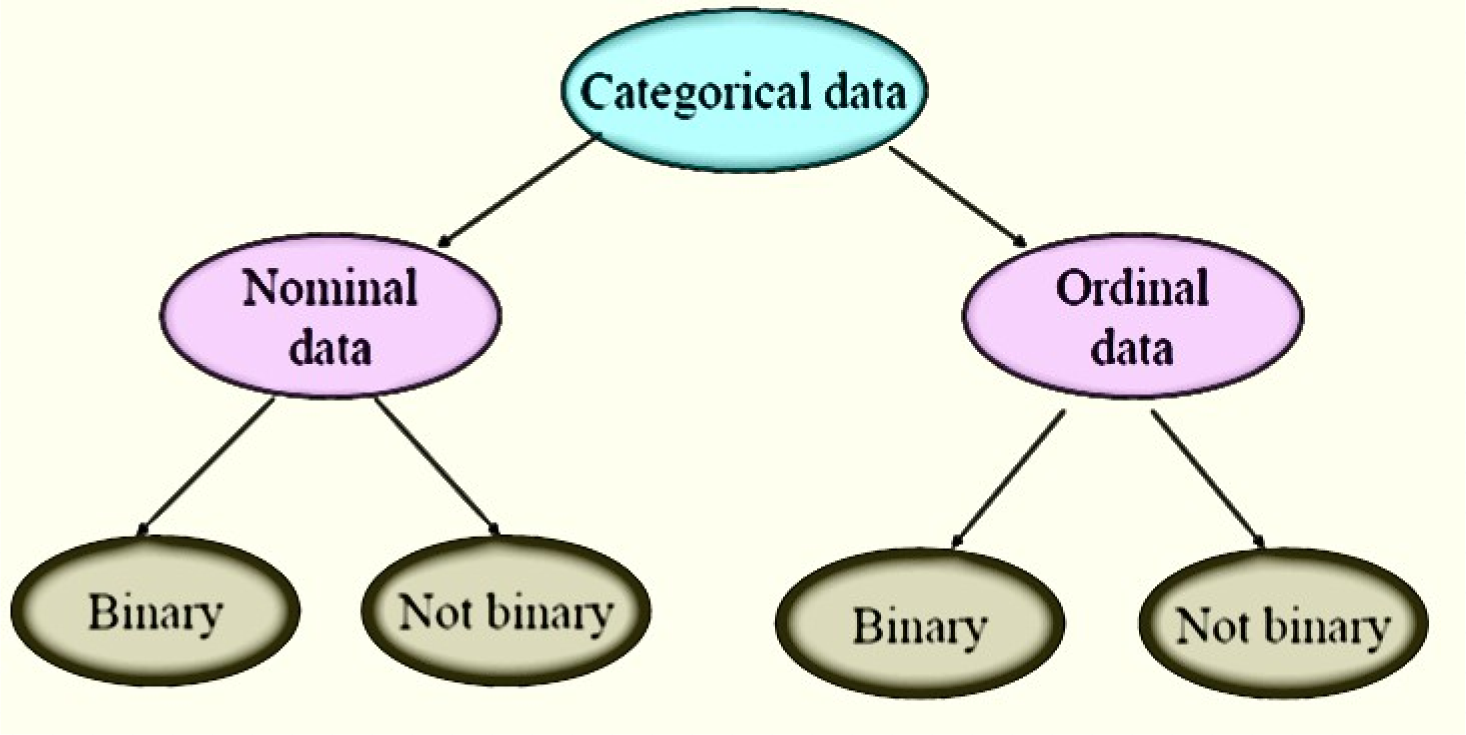
- The objects being studied are grouped into categories based on some qualitative trait
- The resulting data are merely labels or categories

170

Examples: Categorical Data

- Hair color
  - brown, red, black, white, etc.
- Nationality
  - Indian, non-Indian

Categorical data classified as Nominal, Ordinal, and/or Binary



172

Nominal Data

- A type of categorical data in which objects fall into unordered categories.

0Hair color

  – brown, red, black, etc.

- Race
  – Caucasian, African-American, Asian, etc.
- Nationality
  – Indian, non-Indian

Ordinal Data

- A type of categorical data in which order is important.
- Class
  - fresh, junior, senior, super senior
- Degree of illness
  - none, mild, moderate, severe, …, going, going, gone

174

Binary Data

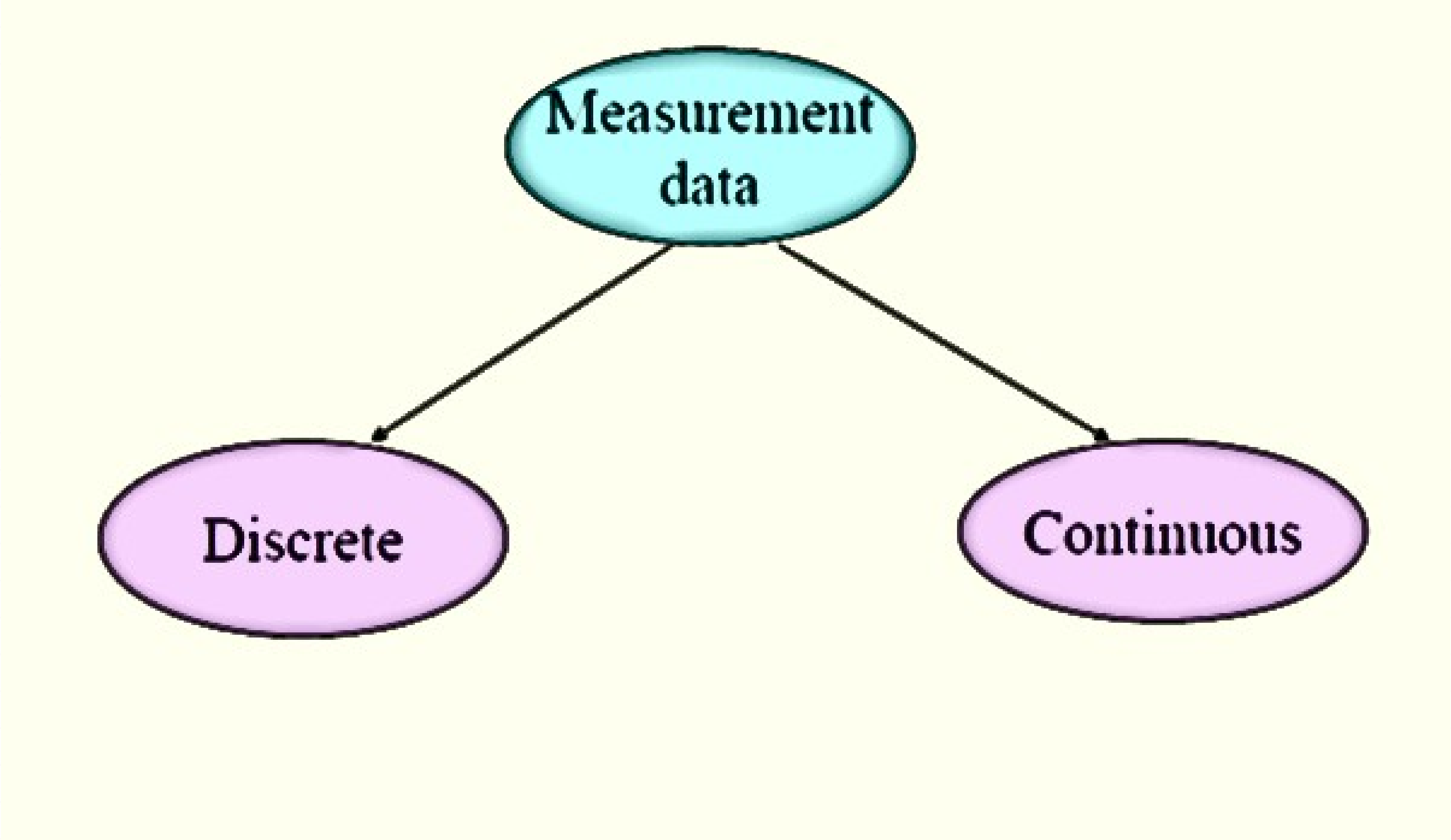
- A type of categorical data in which there are only two categories.
- Binary data can either be nominal or ordinal.
- Nationality
  - Indian, non-Non Indian
- Attendance
  - present, absent

175

Measurement Data

- The objects being studied are measured based on some quantitative trait.
- The resulting data are set of numbers.
  - Bio markers like cholestrol level, insulin level etc
  - Height
  - Age
  - No. of students absent for the exam
  - No. of parents not turned up for parents meeting

Classification of Measurement Data

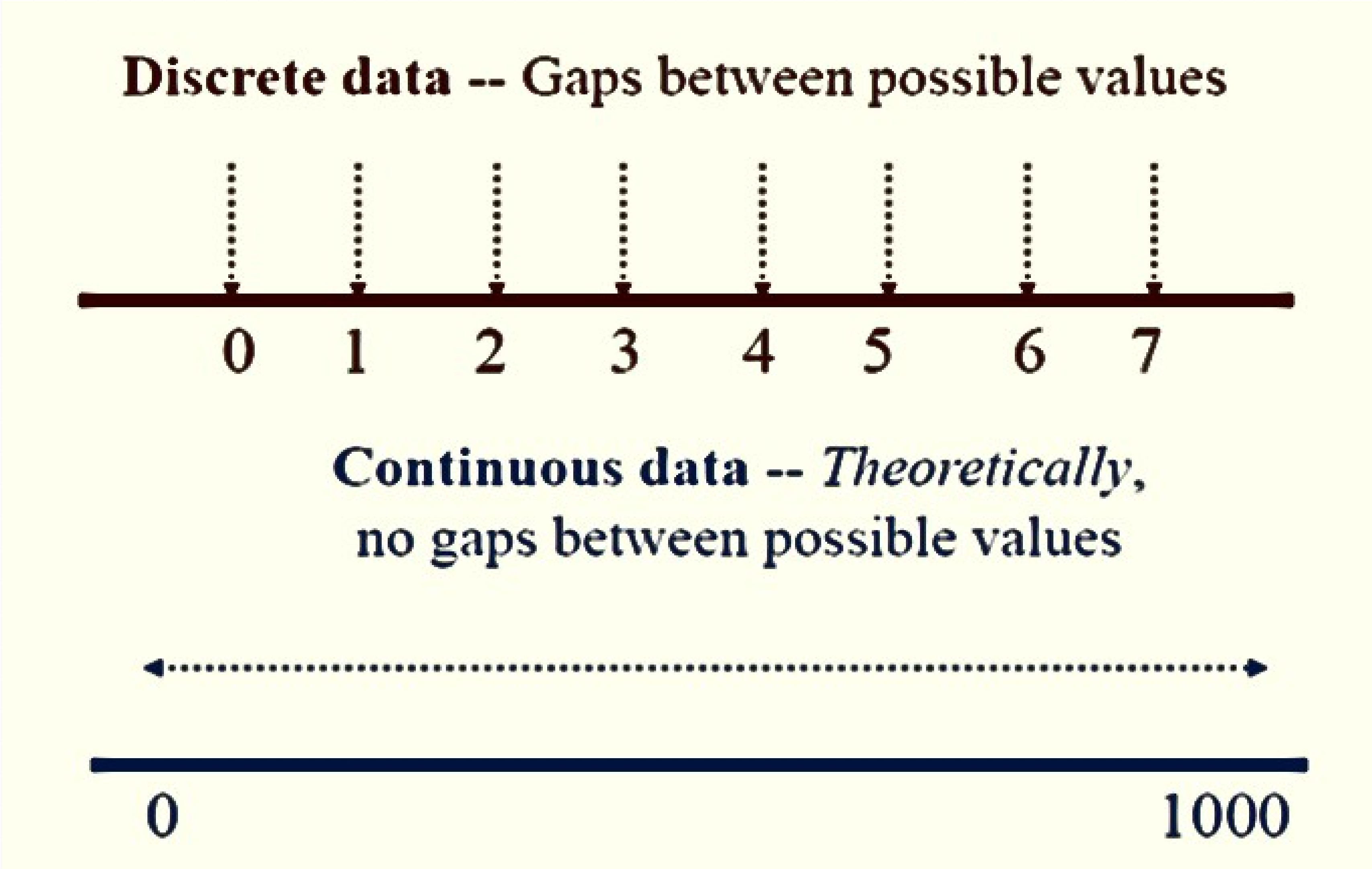


177

Classification of Measurement Data

- Discrete Measurement Data
  - Only certain values are possible (there are gaps between the possible values).
- Continuous Measurement Data
  - Theoretically, any value within an interval is possible with a fine enough measuring device.

178


Discrete data -- Gaps between possible values
0 1 2 3 4 5 6 7
Continuous data -- *Theoretically*,
no gaps between possible values
0
1000

179

Examples:Discrete Measurement Data

- No. of students absent for the exam
- No. of parents not turned up for the parents meeting
- Generally, discrete data are counts.

180

Examples: Continuous Measurement Data

- Bio markers like cholestrol level, insulin level etc
- Height
- Age
- Generally, continuous data come from measurements.

181

Statistical Analysis

- Categorical data are commonly summarized

using "**percentages**" (or "**proportions**").

  – 10% of students cleared exam was boys

182

Statistical Analysis

- Measurement data are typically summarized using averages (or means).
  - – Average weight of male students of II Year BE CSE is 70 KG.
  - – Average weight of female students of II Year BE CSE is 50 KG.

Missing Value Treatment

- Missing data in the training data set can reduce the power/fit of a model
- Lead to a biased model because the relationship between the other variables are not analyzed properly
- May lead to wrong prediction or classification

184

Missing Value Treatment cont...

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%

chances of playing cricket by males is higher than females – in first table

chances of playing cricket by females is higher than males – in second table

Why data set has missing values?

- May occur at two stages:
  - Data Extraction
  - Data Collection
- **Data Extraction**:
  - Double-check for correct data with data guardians.
  - Some hashing procedures can be used to make sure data extraction is correct.
  - Typically easy to find and can be corrected easily as well.

186

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
  - Missing at random
  - Missing that depends on unobserved predictors
  - Missing that depends on the

Data collection cont…

- **Missing completely at random:**
  - This is a case when the probability of missing variable is same for all observations.
  - Example:
    - Respondents of data collection process decide that they will declare their earning after tossing a fair coin.
- **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.
  - Example:
    - Collecting data for age
      - Female has higher missing value compare to male.

188

Data collection cont…

- **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.
  - 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.
  - Example:

Which are the methods to treat missing values ?

- Deletion
- Imputation

Which are the methods to treat missing values ?

- **Deletion:** Two types:
  - List Wise Deletion
    - Delete observations where any of the variable is missing.
    - This method reduces the power of model because it reduces the sample size.
  - Pair Wise Deletion.
    - 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

- listwise deletion will remove a case completely if it is missing a value for one of the variables included in the analysis.
- you are conducting analyses using cumulative high school GPA, hours of study for first semester, SAT score, and first semester grade in college algebra.
- Participant X is missing data for cumulative high school GPA, therefore, Participant X will be completely removed from the analyses because the participant does not have complete data for all the variables.

List-wise deletion

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

- pairwise deletion will not omit a case completely from the analyses.
- Pairwise deletion omits cases based on the variables included in the analysis
- If you are conducting analyses using cumulative high school GPA, hours of study for first semester, SAT score, and first semester grade in college algebra
- Participant X is missing data for cumulative high school GPA

Pair-wise deletion

- Participant X will be omitted from any analyses using cumulative high school GPA
- but they will not be omitted from analyses for which the participant has complete data.

pair-wise deletion

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

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.

Mean/ Mode/ Median Imputation

- 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.
- Two types:-
  - **Generalized Imputation:**
    - Calculate the mean or median for all non missing values of that variable then replace missing value with mean or median.
    - Average of all non missing values of "**Manpower"** (**28.33**) and then replace missing value with it.
  - **Similar case Imputation:**
    - 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.

Maximum Likelihood Estimation

- Maximum likelihood estimation is a method that will find the values of μ and σ that result in the curve that best fits the data.
- The values that we find are called the maximum likelihood estimates (MLE).

- *How do we calculate the maximum likelihood estimates of the parameter values using the log likelihood*
- *Solution:*
- *Consider the data points:* 9, 9.5 and 11

$$\ln(P(x;\mu,\sigma)) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\left[(9-\mu)^2 + (9.5-\mu)^2 + (11-\mu)^2\right]$$

- This expression can be differentiated to find the maximum.
- find the MLE of the mean, μ.
- take the partial derivative of the function with respect to μ, giving

$$\frac{\partial \ln(P(x;\mu,\sigma))}{\partial \mu} = \frac{1}{\sigma^2}\left[9 + 9.5 + 11 - 3\mu\right].$$

- setting the left hand side of the equation to zero and then rearranging for μ gives:

$$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

- Similarly fin… σ …

Multiple Imputation

- Multiple Imputation follows:
  - Imputation Phase
  - Analysis Phase
  - Pooling Phase

Multiple Imputation

- All multiple imputation methods follow three steps.
- **Imputation** – missing values are imputed. However, the imputed values are drawn *m* times from a distribution rather than just once. At the end of this step, there should be *m* completed datasets.
- **Analysis** – Each of the *m* datasets is analyzed. At the end of this step there should be *m* analyses.
- **Pooling** – The *m* results are consolidated into one result by calculating the mean, variance, and confidence interval of the variable of concern.

Multiple Imputation

- In multiple imputation, the imputatin process is repeated multiple times resulting in multiple imputed datasets
- In this method the imputation uncertainty is accounted for by creating these multiple datasets
- In the imputation model, the <u>variables that are related to missingness</u>, can be included.

Imputation Model

- Imputation Phase
- several copies of the data set are created each containing different imputed values
- The imputed values are estimated using the means and covariance of the observed data
- Regression equations are used to predict the incomplete values from the complete values and a normally distributed residual term is added to each value to restore variability

- This process is iterated several times, updating the regression parameters after every iteration, to obtain different imputed values each time.
- auxiliary variables can be included to improve the estimation of the imputed values.
- imputed dataset is stored until the required number of imputed datasets is reached.
- it is important to include the correct variables in the imputation process
- imputation model should fits the distribution assumptions of the data.

- when incomplete data are continuous and normally distributed, a multivariate normal distribution or linear regression can be used for the imputation
- when data are not normal, or not continuous other imputation algorithms should be applied

Analysis Phase

- the statistical analysis is carried out
- On each imputed dataset, the analysis is carried out that would have been applied had the data been complete
- That way as many sets of results are created as the number of imputed datasets created in the imputation phase

Pooling Phase

- the multiple sets of results or parameter estimates are combined into a single set of results
- When the estimates are pooled by Rubin's Rules, the parameter estimates are summarized by taking the average over the parameter estimates from all imputed datasets
- The standard errors are pooled by combining the within imputation variance and the between imputation variance

Pooling Phase

$$Var_{within} = \frac{\sum_{i=1}^{M} SE^2{}_i}{M}$$

$$Var_{between} = \frac{\sum_{i=1}^{M} (\beta_i - \bar{\beta})^2}{M - 1}$$

$$Var_{total} = Var_{within} + Var_{between} + \frac{Var_{between}}{M}$$

Pooling formula's: Var is variance; SE is standard error; M is the number of imputed datasets; Beta is the parameter estimate.

Rubin’s Rules

- Rubin´s Rules (RR) are designed to pool parameter estimates, such as mean differences, regression coefficients, standard errors and to derive confidence intervals and p-values.

Multiple Imputation

- Multiple imputation can be used in cases where the data is
- missing completely at random,
- missing at random,
- and even when the data is missing not at random.

Observed variables

- Observed variables are actually measured by the researcher
- Its classified into observed exogenous variable ie. It is not controlled by other variables similar to independent variable
- Endogenous variable ie it is controlled by other variables similar to dependant variable
- Example: job satisfaction scale, happiness measurement scale etc

Why Bayesian Estimation is used?

- **Bayesian** inference is therefore just the process of deducing properties about a population or probability distribution from data using **Bayes**' theorem

Missing Completely at Random (MCAR)

- Missing data are MCAR when the probability of missing data on a variable is unrelated to any other measured variable and is unrelated to the variable with missing values itself
- Missingness on the variables are completely unsystematic

Missing Completely at Random (MCAR)

- when data are missing for respondents for which their questionnaire was lost in the mail
- separating the missing and the complete cases and examine the group characteristics. If characteristics are not equal for both groups, the MCAR assumption does not hold.

Examples of MCAR Data

Complete data	
Age	**IQ score**
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	**IQ score**
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Missing at Random (MAR)

- When the probability of missing data on a variable is related to some other measured variable in the model, but not to the value of the variable with missing values itself.
- For example, only younger people have missing values for IQ. In that case the probability of missing data on IQ is related to age.

- It is recommended to incorporate correlates of missingness into the missing data handling procedure to diminish bias and improve the chances of satisfying the MAR assumption.

Example of Missing at Random Data

Complete data	
Age	**IQ score**
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	**IQ score**
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

Missing not at Random

- Data are missing not at random (MNAR) when the missing values on a variable are related to the values of that variable itself, even after controlling for other variables.
- For example, when data are missing on IQ and only the people with low IQ values have missing observations for this variable.
- A problem with the MNAR mechanism is that it is impossible to verify that scores are MNAR without knowing the missing values.

Missing not at Random (MNAR) Data

Complete data		Incomplete data	
Age	IQ score	Age	IQ score
25	133	25	133
26	121	26	121
29	91	29	
30	105	30	
30	110	30	110
31	98	31	
44	118	44	118
46	93	46	
48	141	48	141
51	104	51	
51	116	51	116
54	97	54	

Introduction to Bayesian Estimation

- It is based on probability of occurrence of an event.
- Example:
- Tossing a coin,
- rolling a die, and
- drawing a card out of a well-shuffled pack of cards

Bayes Theorem

- Sample Space:
- The result of an experiment is called an **outcome**.
- **The set of all possible outcomes of an event is called the sample space.**
- For example, if our experiment is throwing dice and recording its outcome,
- the sample space will be: S1 = {1, 2, 3, 4, 5, 6}
- What will be the sample when we're tossing a coin?
- S2 = {H, T}

Bayes Theorem

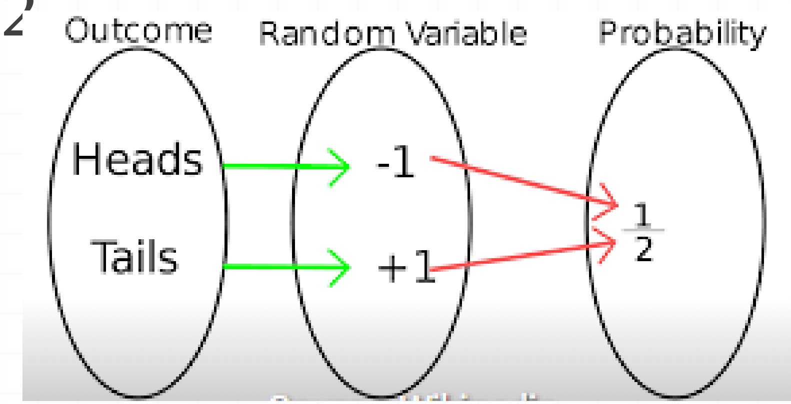
- Event:
- **An event is a set of outcomes (i.e. a subset of the sample space) of an experiment.**
- Example: Consider rolling of a dice
- E = An even number is obtained = {2, 4, 6}
- F = A number greater than 3 is obtained = {4, 5, 6}
- The probability of these events:
- P(E) = Number of favourable outcomes / Total number of possible outcomes = 3 / 6 = 0.5
- P(F) = 3 / 6 = 0.5
- Then, $E \cup F = \{2, 4, 5, 6\}$ and $E \cap F = \{4, 6\}$

Bayes Theorem

- Now consider an event G = An odd number is obtained: Then $E \cap G$ = empty set = Φ
- Such events are called **disjoint** events
- These are also called **mutually exclusive** events because only one out of the two events can occur at a time

Bayes Theorem

- Random Variable:
- Define a random variable X on the sample space of the experiment of tossing a coin.
- It takes a value +1 if “Heads” is obtained and -1 if “Tails” is obtained.
- Then, X takes on values +1 and -1 with equal probability of 1/2



Bayes Theorem

- **Exhaustive Events**
- **A set of events is said to be exhaustive if at least one of the events must occur at any time.**
- Thus, two events A and B are said to be exhaustive if $A \cup B = S$, the sample space.
- Let's say that A is the event that a card drawn out of a pack is red and B is the event that the card drawn is black.
- Here, A and B are exhaustive because the sample space S = {red, black}.

Bayes Theorem

- **Independent Events**
- **If the occurrence of one event does not have any effect on the occurrence of another, then the two events are said to be independent**
- two events A and B are said to be independent if:
- $P(A \cap B) = P(AB) = P(A)*P(B)$
- if A is obtaining a 5 on throwing a die and B is drawing a king of hearts from a well-shuffled pack of cards, then A and B are independent just by their definition.

Bayes Theorem

- **Conditional Probability**
- **Conditional probability is defined as the probability of an event A, given that another event B has already occurred (i.e. A conditional B)**.
- This is represented by P(A|B) and we can define it as:
- $P(A|B) = P(A \cap B) / P(B)$
- Let event A represent picking a king, and event B, picking a black card. Then, we find P(A|B) using the above formula:

Bayes Theorem

- $P(A|B) = P(A \cap B) / P(B)$
- $P(A \cap B)$ = P(Obtaining a black card which is a King) = 2/52 $P(B)$ = P(Picking a black card) = ½
- Thus, $P(A|B) = 4/52$.

Bayes Theorem

- **Marginal Probability**
- **It is the probability of an event A occurring, independent of any other event B, i.e. marginalizing the event B.**
- Marginal probability $P(A) = P(A|B)*P(B) + P(A|\sim B)*P(\sim B)$
- $P(A) = P(A \cap B) + P(A \cap \sim B)$ //conditional probability
- where ~B represents the event that B does not occur.

Bayes Theorem

- the probability that a random card drawn out of a pack is red (event A) = ½
- calculate the same through marginal probability with event B as drawing a king.
- $P(A \cap B) = 2/52$ (because there are 2 kings in red suits)
- $P(A \cap \sim B) = 24/52$ (remaining cards from the red suit)
- $P(A) = 2/52 + 24/52 = 26/52 = 1/2$

Bayes Theorem

- Consider that A and B are any two events from a sample space S where $P(B) \neq 0$.
- Using conditional probability,
- $P(A|B) = P(A \cap B) / P(B)$
- $P(B|A) = P(A \cap B) / P(A)$
- $P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$
- $P(A|B) = P(B|A)*P(A) / P(B)$

Bayes Theorem

- P(A) and P(B) are probabilities of observing A and B independently of each other (marginal probabilities)
- P(B|A) and P(A|B) are conditional probabilities
- P(A) is called **Prior probability** and
- P(B) is called **Evidence**
- P(B) = P(B|A)*P(A) + P(B|~A)*P(~A)
- P(B|A) is called **Likelihood** and P(A|B) is called **Posterior probability**
- Bayes Theorem can be written as:
- posterior = likelihood * prior / evidence

Bayes Theorem

- There are 3 boxes labeled A, B, and C:
- Box A contains 2 red and 3 black balls
- Box B contains 3 red and 1 black ball
- And box C contains 1 red ball and 4 black balls
- The three boxes are identical and have an equal probability of getting picked.
- Consider that a red ball is chosen. Then what is the probability that this red ball was picked out of box A?

Bayes Theorem

- Let E denote the event that a red ball is chosen and A, B, and C denote that the respective box is picked
- We are required to calculate the conditional probability $P(A|E)$
- We have prior probabilities $P(A) = P(B) = P(C) = 1 / 3$, since all boxes have equal probability of getting picked.
- $P(E|A)$ = Number of red balls in box A / Total number of balls in box A = $2 / 5$
- Similarly, $P(E|B) = 3 / 4$ and
- $P(E|C) = 1 / 5$

Bayes Theorems

- evidence $P(E) = P(E|A)*P(A) + P(E|B)*P(B) + P(E|C)*P(C)$
- $= (2/5) * (1/3) + (3/4) * (1/3) + (1/5) * (1/3) = 0.45$
- what is the probability that this red ball was picked out of box A?
- $P(A|E) = P(E|A) * P(A) / P(E) = (2/5) * (1/3) / 0.45 = 0.296$

T
H
A
N
K

Y
O
U

CONTACT DETAILS

Dr. G. Malathi

Associate Professor

Coordinator Image Processing Research Group

School of Computing Sciences and Engineering

VIT University

Chennai Campus