# Week 4

Wednesday, April 28, 2021 2:32 PM

# Overview of Statistical Analysis

## Statistical Analysis

- **Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data.**
- **Statistical Analysis is the application of statistical methods to a sample of data in order to develop an understanding of what that data represents**

    - Calculations such as average income, average age, highest-paid professions
    - Analyzing vaccine data to ensure safety and efficacy
    - Gaining greater insight into customer requirements to reduce customer churn

**Sample -** A representative selection drawn from a total population.

**Population -** A discrete group of people or things that can be identified by at least one common characteristic for purposes of data collection and analysis

**There are two types of Statistical analysis:**

- **Descriptive Statistics -** Summarizing information about the sample
    - Enables you to present data in a meaningful way
    - Allows for simpler interpretation of data
    - Does not attempt to draw conclusions about the population from which the sample is taken

    - Common measures of **Descriptive Statistical Analysis:**
        - **Central Tendency** - Locating the center of a data sample. (Mean, Median and Mode)
        - **Dispersion** - Measure of variability in a data set (Variance, Standard Deviation, Range)
        - **Skewness** - Is the measure of whether the distribution of values is symmetrical around a central value or skewed left or right.

  **Variance** - defines how far away the data points fall from the center.
  **Standard Deviation** - tells you how tightly your data is clustered around the mean.
  **Range** - gives you the distance between the smallest and largest values in your datasets

- **Inferential Statistics** - Making inferences or generalizations about the broader population
    - Takes data from a sample to make inferences about the larger population from which the sample was drawn
    - Helps draw generalizations that apply the results of the sample to the population as a whole

    - Common methodologies of **Inferential Statistics** include:
        - **Hypothesis Testing** - Can tell you whether the efficacy of a vaccine observed in a control group is likely to exist in the population as well.
        - **Confidence Intervals -** Incorporate the uncertainty and sample error to create a range of values the actual population value is likely to fall within.
        - **Regression Analysis -** Incorporates hypothesis test that help determine whether the relationships observed in the sample data actually exist in the population rather than just the sample.

# What is Data Mining

## Data Mining

- The process of extracting knowledge from data.
- An interdisciplinary field that involves the use of pattern recognition technologies, statistical analysis, and mathematical

techniques
- Aims to identify correlations in data, find patterns and variations, understand trends, and predict probabilities.

**Patterns and Trends**

- A **Pattern** recognition is the discovery of regularities, or commonalities, in data.
- A **Trend** is the general tendency of a set of data to change over time.

**Applications of Data Mining**

- Profiling customers behaviors, needs, and disposable income in order to offer targeted campaigns.
- Financial institutions tracking customer transactions for unusual behaviors and flagging fraudulent transactions using data mining models.
- The use of statistical models to predict a patient's likelihood for specific health conditions and prioritizing treatment.
- Accessing performance data of students to predict achievement levels and make a focused effort to provide support where required.
- Helping investigation agencies deploy police force where the likelihood of crime is higher.
- Aligning supply and logistics with demand forecasts.

**Data Mining Techniques**

- **Descriptive**
- **Diagnostic**
- **Predictive**
- **Prescriptive Modeling**

Commonly used data mining techniques:

- **Classification -** Classifying attributes into target categories.
- **Clustering -** Involves grouping data into clusters so they can be treated as groups.
- **Anomaly or Outlier Detection -** Finding patterns in data that are not normal or unexpected.
- **Association Rule Mining -** Establishing a relationship between two data events.
- **Sequential Patterns -** Tracing a series of events that take place in a sequence.
- **Affinity Grouping -** Discovering co-occurrence in relationships.
- **Decision trees -** Building classification models in the form of a tree structure with multiple branches, where each branch represents a probable occurrence.
- **Regression -** Identifying the nature of the relationship between two variables, which could be casual or correlational.

**Data mining helps separate the noise from the real information and helps businesses focus their energies on only what is relevant.**

# Tools for Data Mining

## Tools for Data Mining

Some of the commonly used software and tools for data mining:

- **Spreadsheets**
  - Hosting data that has been exported from other systems in an easily accessible and easy-to-read format.
  - Creating pivot tables to showcase specific aspects of your data.
  - Drawing comparisons between different sets of data.
  - Data Mining Client, XLMiner, KnowledgeMiner - Allow you to perform common mining tasks such as classification, regression, association rules, clustering, and model building.

- **R-Language** - one of the most widely used languages for performing statistical modeling and computations by statisticians and data miners. Using R-libraries you can perform data mining operations such as:
  - Regression

- Classification
- Data Clustering
- Association Rule Mining
- Text Mining
- Outlier Detection
- Social Network Analysis

- **Python** - is an interpreted high-level general-purpose programming language.
  - **Pandas**
    - Perform basic numerical computations such as mean, median, mode and range.
    - Calculate statistics and answer questions regarding correlation between data and distribution of data.
    - Explore data visually and quantitatively
    - Visualize data with help from other Python libraries
  - **NumPy**
    - A tool for mathematical computing and data preparation in Python
    - Offers a host of built-in functions and capabilities for data mining
  - **Jupyter**
    - Jupyter Notebooks have become the tool of choice for Data Scientists and Data Analysts when working with Python to perform data mining and statistical analysis.

- **IBM SPSS Statistics** - Statistical Process For Social Sciences
  - Popularly used for advanced analytics, text analytics, trend analysis, validation of assumptions, and translation of business problems into data science solutions
  - Is closed-source
  - Requires a license for use
  - Has an easy to use interface
  - Requires minimal coding for complex tasks
  - Has efficient data management tools
  - Is popular for its in-depth analysis capabilities and accurate data results

- **IBM Watson Studio -** leverages a collection of open source tools such as Jupyter notebooks, and extends them with closed source IBM tools that make it a powerful environment for data analysis and data science.
  - Is available through a web browser on the public cloud, private cloud, and as a desktop app
  - Enables team members to collaborate on projects
  - Includes SPSS Modeller flows that enable you to quickly develop predictive models for your business data.

- **SAS Enterprise Miner -** is a comprehensive, graphical workbench for data mining.
  - Provides powerful capabilities for interactive data exploration
  - Can manage information from various sources, mine and transform data, and analyze statistics
  - Offers a graphical user interface for non-technical users.

  - Identify patterns in the data using a range of available modeling techniques
  - Explores relationships and anomalies in data
  - Analyze big data
  - Validate the reliability of finding from the data analysis process

**Conclusion and key considerations for selection the right data mining tool:**

- Data size and structures supported by the tool
- Key features
- Data visualization capabilities
- Infrastructure needs
- Ease of use
- Learnability

# Summary and Highlights

**In this lesson, you have learned the following information:**

Some of the commonly used software and tools for data mining:

Statistics is a branch of mathematics dealing with the **collection**, **analysis**, **interpretation**, and **presentation** of numerical or quantitative data.

**Statistical Analysis** involves the use of statistical methods in order to develop an understanding of what the data represents.

**Statistical Analysis** can be:

- **Descriptive**; that which provides a summary of what the data represents. Common measures include Central Tendency, Dispersion, and Skewness.
- **Inferential**; that which involves making inferences, or generalizations, about data. Common measures include Hypothesis Testing, Confidence Intervals, and Regression Analysis.

**Data Mining**, simply put, is the process of extracting knowledge from data. It involves the use of pattern recognition technologies, statistical analysis, and mathematical techniques, in order to identify correlations, patterns, variations, and trends in data.

There are several techniques that can help mine data, such as, classifying attributes of data, clustering data into groups, establishing relationships between events, variables, and input and output.

A variety of software and tools are available for analyzing and mining data. Some of the popularly used ones include Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio, and SAS, each with their own set of characteristics, strengths, limitations, and applications.

# Overview of Communicating and Sharing Data Analysis Findings

**Data Analysis Process**

- The success of your communication depends on how well others can understand and trust your insights to take further action. Combination of Story - Visualization - Data

**Who is my audience?**

- A diverse group of people representing different business functions and roles

**What is important to them?**

- Understanding the information needs of your audience will help you decide what, and how much, information is essential to enable a better understanding of your findings.

**What will help them trust me?**

- Begin your presentation by demonstrating your understanding of the business problem to your audience. Speak in the language of the organization's business domain.

**Structure your presentation**

- The next step in designing your communication is to structure and organize your presentation for maximum impact.
  - Reference your data
  - State your assumptions
  - Organize your presentation
  - Identify the best formats for presenting your data

**The Role of Visuals**

- **Trust, Understanding, Relatability**.
  - Establish credibility of your findings
  - Present data within a narrative
  - Support the narrative with visual

# Introduction to Data Visualization

**Overview**

- Data visualization is the discipline of communicating information through the use of visual elements such as graphs, charts, and maps. Its goal is to make information easy to comprehend, interpret, and retain.

- What is the relationship that I am trying to establish?
- Do I want to compare multiple values, such as the number of products sold, and revenues generated over the last three years?
- Do I want to detect anomalies in the data?

**What is the question I'm trying to answer?**

- What should be the key takeaway for my audience?
- What does my audience need to know?
- What are the questions they might have?

**Common types of graphs:**

- **Bar Charts** - great for comparing related data sets or parts of a whole.
- **Column Charts -** compare values side-by-side.
- **Pie Charts -** show the breakdown of an entity into its sub-parts and the proportion of the sub-parts in relation to one another.
- **Line Charts -** display trends.

- **Dashboards -** organized and display reports and visualizations coming from multiple data sources into a single graphical interface. Dashboards can present both operational and analytical data. Dashboards can present both operational and analytical data.
  - Are easy to comprehend by an average user
  - Make collaboration easy between teams
  - Allow you to generate reports on the go

# Introduction to Visualization and Dashboarding Software

**Overview**

**Commonly used data visualization software and tools include:**

- **Spreadsheets**
  - Most commonly used software for graphical representation of data set
  - Easy to learn
  - Documentation and video tutorials for ready reference
  - Provides several chart types - bar/line/pie/pivot/scatter/Gantt/Waterfall/combination
  - Provides recommendations on visual representation
  - Can add chart title, change colors of elements, and add labels to data

- **Jupyter Notebook and Python libraries**

        - Is an open-source web application that provides a great way to explore data and create visualizations
        - **Matplotlib**
            - Widely used Python data visualization library
            - Provides different kinds of 2D and 3D plots and the flexibility to create plots in several different ways
            - Helps create high-quality interactive graphs and plots with just a few lines of code
        - **Bokeh**
            - Provides interactive charts and plots
            - Delivers high-performance interactivity over large or streaming datasets
            - Offers flexibility for applying interaction, layouts, and different styling options to visualization
            - Can transform visualizations written in other Python libraries, such as **Matplotlib, Seaborn and Ggplot**
        - **Dash**
            - A python based framework for creating interactive web-based visualizations
            - Helps build highly interactive web applications using Python code
            - Does not require knowledge of HTML and JavaScript
            - Is easily maintainable, cross-platform, and mobile-ready

- **R-studio and R-Shiny**
    - Basic visualizations such as histograms, bar charts, line charts, box plots, and scatter plots
    - Advanced visualizations such as heat maps, mosaic maps, 3D graphs, and correlograms
    - **Shiny**
        - is an R package that helps build interactive web apps that can be hosted as standalone apps on a webpage
        - you can also build dashboards using Shiny
        - The ease of working with Shiny is what popularized it among data professionals

- **IBM Cognos Analytics**
    - Is an end-to-end analytics solution
    - Importing custom visualizations
    - A forecasting feature that provides time-series data modeling and forecasts
    - Recommendation for visualizations based on your data
    - Conditional formatting which allows you to see the distribution of your data and highlight exceptional data points
    - Cognos is known for its superior visualizations and overlaying data on the physical world using its geospatial capabilities.

- **Tableau**
    - Is a software company that produces interactive data visualization products.
    - Create interactive graphs and charts in the form of dashboards and worksheets, with drag and drop gestures.
    - Publish results in the form of stories.
    - Import R and Python scripts.
    - Compatible with Excel files, Text files, Relational databases, Google Analytics, Amazon Redshift, Cloud databases.

- **Microsoft Power BI**
    - Is a cloud-based business analytics service from Microsoft that enables you to create reports and dashboards.
    - A powerful and flexible tool known for its speed.
    - Has a drag and drop interface.
    - Is compatible with multiple sources, Excel, SQL Server, and cloud-based data repositories.
    - Provides the ability to collaborate and share dashboards and reports securely.

# Summary and Highlights

**In this lesson, you have learned the following information:**

Data has value through the stories that it tells. In order to communicate your findings impactfully, you need to:

- Ensure that your audience is able to trust you, understand you, and relate to your findings and insights.
- Establish the credibility of your findings.
- Present the data within a structured narrative.
- Support your communication with strong visualizations so that the message is clear and concise, and drives your audience to take action.

Data visualization is the discipline of communicating information through the use of visual elements such as graphs, charts, and maps. The goal of visualizing data is to make information easy to comprehend, interpret, and retain.

For data visualization to be of value, you need to:

- Think about the key takeaway for your audience.
- Anticipate their information needs and questions, and then plan the visualization that delivers your message clearly and impactfully.

There are several types of graphs and charts available for you to be able to plot any kind of data, such as bar charts, column charts, pie charts, and line charts.

You can also use data visualization to build dashboards. Dashboards organize and display reports and visualizations coming from multiple data sources into a single graphical interface. They are easy to comprehend and allow you to generate reports on the go.

When deciding which tools to use for data visualization, you need to consider the ease-of-use and purpose of the visualization. Some of the popularly used tools include Spreadsheets, Jupyter Notebook, Python libraries, R-Studio and R-Shiny, IBM Cognos Analytics, Tableau, and Power BI.