Analyzing and Mining Data

* Statistical Analysis
  + Dealing with collection, analysis, interpretation, and presentation of numerical or qualitative data.
  + Examples
    - Calculation of average income
    - Vaccine data to ensure safety
    - Getting insight into customer requirements
  + Definition
    - Statistical analysis is the application of stats methods to a sample of data to develop an understanding of what that data represents.
      * Sample - Selection of people from a total population
      * Population – Discrete group that can be identified by at least one common characteristics.
  + Ensure
    - Data is interpreted correctly
    - Identify apparent relationships
  + Types of Statistics
    - Descriptive Statistics – Summarizing information about the sample
      * Enables simple presentation of the data
      * Easy to understand raw data rather than making conclusions
        + Central Tendency – Mean, Median, Mode
        + Dispersion – Variance, Standard Deviation, Range
        + Skewness – Measure of how the distribution of values is symmetrical around.
    - Inferential Statistics – Making Inference or generalizations about the broader population with a smaller representative sample.
      * Hypothesis testing – Studying the effectiveness of a vaccine by using a control group.
      * Regression Analysis – Whether the relationships observed in the sample data actually exist in the population rather than just the sample.

Packages – SAS, SPSS, StatSoft

Data Mining

* Process of extracting knowledge from data
* Pattern recognition is discovery of regularities, or commonalities, in data.
  + User log data
  + Analysis of user behavior
* A Trend
  + General trend of a set of data to change over time
* Techniques
  + Classification – Classifying attributes into target categories
  + Clustering – Grouping data into cluster to be treated as groups
  + Anomaly or outlier – Finding patterns in data that are not normal or expected.
  + 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 – Classification models in the form of a tree with multiple branches where each branch represents a probable occurrence.
  + Regression – Nature of the relationship between two variables, which could be causal or correlational.

Tools for Data Mining

* Software and Tools
  + Spreadsheets
    - Basic data mining tasts
    - Host data that has been exported
    - Creating Pivot tables
    - Drawing Comparison between different sets of data
    - Add In (Classification, regression, association rules, clustering and model building)
      * Data mining client
      * XLMiner
      * KnowledgeMiner
  + R-language
    - Language for performing statistical modeling and computation by statistician and data miners.
    - Used for
      * Regression
      * Classification
      * Data Clustering
      * Association Rule Mining
      * Text Mining
      * Outlier Detection
      * Social Network Analysis
    - Packages
      * Tm – Framework for text mining application within R
      * twitter – Mining tweets
    - IDE
      * RStudio is a popularly used open-source IDE for working with the R programming language
  + Google Sheets
    - Add-ons
      * Text Analysis
      * Text Mining
      * Google Analytics
  + Python
    - Packages
      * Pandas
        + Open-Source model for working with data structures and analysis.
        + Allows the user to upload data in any format and provides a simple platform to organize, sort and manipulate data.
        + Perform basic numerical computation such a mean, media, mode and range
        + Calculate statistics and answer question regarding correlation between data and distribution of data.
        + Explore data visually and quantitatively.
      * NumPy
        + Tool for mathematical computing and data preparation in Python
        + Built-in function and capabilities for data mining.
      * Jupyter Notebooks
        + Tool of choice for data Scientist and Data analyst when working with python to perform data mining and statistical analysis.
  + IBM SPSS Statistics
    - Popular for advance analytics, text analytics, trend analysis, validation of assumptions, and translation of business problem into data science solution
      * Closed source
      * License for use
      * Easy to use interface
      * Minimal coding for complex task
      * Efficient data management tools
      * Popular for indept analysis capabilities and accurate data results
  + IBM Watson Studio
    - Included in IBM cloud pak
      * Available through a web browser on the public cloud, private cloud and as a desktop app
      * Collaborate with team members on projects
      * SPSS Modeler flows that enable you to quickly develop predictive models for your business data.
  + SAS
    - SAS Enterprise Miner graphical workbench
      * Capabilities for interactive data exploration
      * Manage information from various sources, mine and transform data and analyze statistics.
      * Offer graphical interface for non tech users.
        + Allows for

Identify patterns

Explore relationships anomalies in data

Analyze big data

Validate reliability of findings from the data analysis process.

* + - * + Easy to use and easy to debug.

Summary

In this lesson, you have learned the following information:

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.

Communicating Data Analysis Findings

* Understanding the problem
* Communicating findings
  + Story
    - Trust, understanding, relatability
    - Establish credibility of findings
    - Present data within a narrative
    - Support the narrative with visual
  + Visualization
    - Graphs
    - Charts
    - Diagrams
  + Data
  + Intended Audience
    - What is important to them?
    - Include data that is relevant to the audience.
    - Structure Presentation
      * Reference your data
      * State your assumptions
      * Organize your presentation
      * Identify the best formats for presenting your data
* Data projects
  + Collaborative effort across business functions
  + Multi-Disciplinary skill
  + Finding incorporated in a larger business initiative.

Data Visualization

* Data visualization is the discipline of communicating information using visual elements such as graphs, charts, and maps. Its goal is to make information easy to comprehend, interpret and retain.
  + Choose visualization that is best
    - Relationship that I’m trying to establish
    - Compare multiple values
    - Do I need audience to see the correlation between two variables?
    - Do I want to detect anomalies?
  + What is the question I’m trying to answer?
    - Key takeaway for my audience
    - Audience needs to know what?
    - What questions they have?
  + Basic charts
    - Bars
      * Great for comparing related data sets or parts of a whole.
    - Columns
      * Show change over time (Comparison side by side)
    - Pie Charts
      * Shows breakdown of entity into its sub parts and the proportion of the sub-parts in relation to one another.
    - Line Charts
      * Display Trends continuously with respect to time.
  + Dashboards
    - Organize and display reports and visualization coming from multiple data sources into a single physical interface.
    - Can present operation and analytical data.
    - Presents a bids eye view of the complete picture
      * Easy to comprehend
      * Make collaboration easy
      * Generate Reports on the go.
* Tools and software
  + Spreadsheets
    - Most used software for graphical representations of data sets
    - Easy to learn
    - Provide several charts, bar,line,pie,pivot,scatter,trendline,gantt, waterfall, combination.
    - Charts title, change of color of elements and add labels to data.
    - Similar to google sheets but sheets is more useful for collaborative workloads.
  + Jupyter Notebook and Python Libraries
    - Open source appl for exploring data.
    - Matplotlib: Widly used python data visualization library
      * Provides different kinds of 2d and 3d plots and the flexibility to create plots in several different ways.
      * Helps to create high quality graphs.
      * Large Community support
    - Bokeh
      * Interactive charts and plots
      * Deliver high-performance interactive over large for streaming data sets.
      * Flexibility for applying interaction, layouts, and different styling options to visualization
      * Can transform visualization written in other python libraires such as Matplotlib, Seaborn and Ggplot.
    - Dash
      * Creating interactive web-based visualization
      * Doesn’t require knowledge of HTML and JS
      * Easily maintainable, cross-platform, and mobile-ready.
  + R-Studio and R-Shiny
    - Create basic visualization such as histogram, bar, charts, line, box and scatter
    - Creating heatmaps, mosaic maps, 3d graphs and correlograms.
    - Shiny is an R package that helps build interactive web apps
      * Build dashboards using shiny
      * Ease of working.
  + IBM Cognos Analytics
    - End to end analytics solutions
    - Importing custom visualization
    - Forecasting feature that provides time-series data modeling and forecast
    - Recommendation for visualization based on your data.
    - Conditional formatting
    - Known for superior visualization and overlaying data on the physical world.
  + Tableau
    - Create interactive graphs as dashboards and worksheets with drag and drop features.
    - Publish results in the form of stories
    - Import R and Python Scripts.
    - Compatible with
      * Excel files
      * Text files
      * Relational Database
      * Cloud database (Google Analytics, Amazon Redshift)
  + Microsoft Power BI
    - Cloud based BI from Microsoft that enables you to create reports and dashboards.
    - Powerful tool known for its speed and efficiency.
    - Drag and drop features.
    - Compatible with
      * Excel
      * SQL Server
      * Cloud base data repositories.
    - Provides ability to collaborate and share dashboards and reports securely.

Summary

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.