## *What is Data Analytics?*

In this lesson, you will learn about the different types of data analysis and the key steps in a data analysis process. You will gain an understanding of the different components of a modern data ecosystem, and the role Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts play in this ecosystem. You will also learn about the role, responsibilities, and skillsets required to be a Data Analyst, and what a typical day in the life of a Data Analyst looks like.

By the end of this lesson, you will:

* Explain the different components of a modern data ecosystem
* Describe and differentiate between the role different data professionals play in this ecosystem
* Explain what data analysis is, the different types of data analysis, and the key steps in the data analysis process
* Describe the responsibilities and skillsets of a Data Analyst
* Understand what a day in life of a Data Analyst can look like
* Understand the different emerging technologies shaping data ecosystems
* Identify technology that allows for limitless storage and high-performance computing
* Differentiate between data roles and outline each role’s duties
* Differentiate between the four types of data analytics
* Understand which of the four types of data analytics are being used in any given
* Explain what a modern data ecosystem is and what it involves scenario
* Discuss how a data analyst works with a data ecosystem
* Outline the different steps involved in the data analysis process 1.1.3.c Discuss the tech and soft skills required to be a successful data analyst
* Understand and apply the strategies available to ensure credibility in data findings

Representative Image

**What is Data Analysis?**

Data analysis is a process of inspecting, cleansing, transforming, and modeling data to discover useful information, informing conclusions, and support decision-making. This helps businesses understand past performance and inform their future decisions.

It helps businesses to:

* Understand their past performance and informs their decision-making for future actions
* looking for patterns in financial transactions to detect fraud, using recommendation
* Engines to drive conversion, mining, social media posts for customer voice or brands personalizing their offers based on customer behavior analysis, business leaders realized that data holds the key to competitive advantage.

**Data Analysis Process**

There are several approaches, techniques, and processes that can be used for analyzing data.  Although, the following steps are considered a standard in the process of data analysis.

* ***Step 1***

**Developing an understanding of the problem and the desired outcome.**

Understanding the problem that needs to be solved, ensure that you are working on the right direction and know desired/expected outcome  outcome

* ***Step 2***

**Setting a clear metric for evaluating outcomes.**

Setting a clear metric. This stage of the process includes deciding what will be measured.

* ***Step 3***

**Gathering, cleaning, analyzing, and mining data to interpret results.**

Identifying all sources of data, how this data would be stored, what tools would be used to analyze the same

* ***Step 4***

**Communicating the findings in ways that impact decision-making.**

Add a description or ways in which this communication can be established eg: reports, presentations, documentation etc, in a nutshell - Storytelling with the help of data.

Can you think of different business case use scenarios of Data Analysis? A few are listed below:

**Fraud Detection**

+

**Recommendation System**

+

**Data Mining**

+

**Can you sort the following steps in the order of the Data Analysis Process that you learnt above?**

**0/4 Cards Correct**

REPLAY

**Data Analysis and Mining**

**Gather, Extracting, Transforming and Loading Data**

**Data Visualization**

**Data Wrangling and Cleaning**

**Step 1**

**Step 2**

**Step 3**

**Step 4**

**What is the Modern Data Ecosystem?**

The modern data ecosystem includes a network of interconnected and continually evolving entities that include:

* **Data:** Data is available in various formats, structures, and sources. Documents, spreadsheets, structured, unstructured, primary, secondary  - these are all the words you will commonly hear associated with data.
* **Enterprise Data Environment:**Enterprise Data Environment is the one in which raw data is staged so it can be organized, cleaned, and optimized for use by end-users.
* **End-Users:**End-users are people such as business stakeholders, analysts, and programmers who consume data for various purposes.

Emerging technologies such as Cloud Computing, Machine Learning, and Big Data, are continually reshaping the data ecosystem and the possibilities it offers.

Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts; all play a vital role in the ecosystem for deriving insights and business results from data.

**Primary Types of Data Analysis**

**Prescriptive Analytics**helps answer the question, What should be done about it?

ex. Airlines automatically adjust ticket prices based on customer demand.

**Diagnostic analytics**helps to answer the question. Why did it happen?

It takes the insights from descriptive analytics to dig deeper to find the cause of the outcome. For example, a sudden change in traffic to a website without an obvious cause or an increase in sales in a region where there has been no change in marketing.

**Predictive analytics**helps answer the question, What will happen next?

Like risk assessment and sales forecasts.

**Descriptive analytics** is the process of using current and historical data to identify trends and relationships. It’s sometimes called the simplest form of data analysis because it describes trends and relationships but doesn’t dig deeper. It tries to answer the questions "What", "How much?" Ex: How much has the number of flights surged to a specific location since the past year? What have been the sales volumes?

**Data Professionals**

**Data Engineers**: A technical person who designs and build systems for collecting, storing and analysing data at scale.

**Data Analysts:** A professional who works with data to provide insights. They collect and analyse various sources and translate it into simple English or meaningful information that can be used for decision-making.

**Data Scientists:** Determine the question their team should be asking and figure out how to answer those questions using data. They often develop predictive models for theorizing and forecasting.

**Business Analysts:** A person who helps organizations achieve its goals by analysing data, documenting and improving business processes, services and software acting as a bridge between various stakeholders.

## *The Data Ecosystem*

In this lesson, you will learn about the different types of data structures, file formats, sources of data, and the languages data professionals use in their day-to-day tasks. You will gain an understanding of various types of data repositories such as Databases, Data Warehouses, Data Marts, Data Lakes, and Data Pipelines. In addition, you will learn about the Extract, Transform, and Load (ETL) Process, which is used to extract, transform, and load data into data repositories.

By the end of this lesson, you will:

* Describe and differentiate between relational and non-relational database management systems
* Explain the different types of data structures, file formats, and sources of data
* Explain the features and use of the different languages used by data professionals
* Describe how Data Warehouses, Data Marts, Data Lakes, and Data Pipelines work
* Explain how the Extract, Transform, and Load process works to make raw data ready for analysis
* Explain what Big Data is and summarize the features and use of some of the Big Data processing tools
* Identify the shell and scripting languages most commonly used for different data functions
* Differentiate between structured and unstructured data
* Differentiate between file formats and explain how each can be viewed on a device
* Identify the return data from each data source
* Identify which programming language supports multiple programming paradigms
* Identify the technology that has allowed Data Marts and Data Warehouses to be used for non-relational data
* Discuss the advantages of RDBMS
* Identify the NoSQL database type that is used for visualizing, analyzing, and finding connections between different pieces of data
* Differentiate between the different data repositories
* Define “veracity” in the context of Big Data
* Explain the key use cases of Apache Spark

**Databases**

By now, you are aware of the fact that data gets stored in databases. There are two types of databases where the data can be stored: Relational databases (RDBMS) and non-relational databases. Let's see the basic difference both of these:

A screenshot of a computer program

Description automatically generated

Relational and Non-relational Database

|  |  |
| --- | --- |
| Relational Database | Non-relational Database |
| Stores data in tables | Does not use tables |
| Relationship between tables and field types is called a schema | Optimized for the type of data its storing |
| Also called SQL Databases | Also called NoSQL Databases. There are 4 different types: Document oriented, Key-value stores, Wide-column stores, Graph stores |
| Popular relational/SQL Databases: MS SQL Server, MySQL, PostgreSQL | Popular non-relational/NoSQL databases: MongoDB, Redis, |

**Types of Data**

Data is broadly categorized into three entities:

1. Structured Data: Data that follows a rigid format and can be organized neatly into rows and columns is structured data. This is the data that you see typically in databases and spreadsheets, for example: Dates, Phone numbers, ZIP codes, Customer names
2. Unstructured Data: Complex, and mostly qualitative information that is impossible to reduce to rows and columns.
3. Semi-Structured Data: Semi-structured data is a mix of data that has consistent characteristics and data that doesn’t conform to a rigid structure. For example: **emails.**An **email** has a mix of structured data, such as the name of the sender and recipient, but also has the contents of the email, which is unstructured data.

**Data Sources and Structure:**

Data sources have never been as dynamic and diverse as they are today. Banks, government organizations, political data - there are numerous sources from which data can be extracted. With every data that is sourced, as a Data Analyst, you need to ensure its reliability and security. As varied are the sources of data, so are the data extraction methods from these sources.

**What is Big Data?**

In simple words, Big Data is larger, more complex datasets, especially from new data sources. These datasets are so voluminous that traditional data processing software cannot manage them. Although, there is no fixed definition for Big Data, it is featured by the 5 Vs:

**Volume**:  With big data, you’ll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

**Velocity**: Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.

**Variety**: Variety refers to the many types of data that are available. Unstructured and semi structured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

**Value**: Data has intrinsic value. But it’s of no use until that value is discovered.

**Veracity**:  It refers to the quality and accuracy of data. Gathered data could have missing pieces, may be inaccurate or may not be able to provide real, valuable insight. Veracity, overall, refers to the level of trust there is in the collected data.

**Big Data Processing Tools**

Big Data processing technologies provide ways to work with large sets of structured, semi-structured, and unstructured data so that value can be derived from big data.

* Apache Hadoop
* Apache Hive
* Apache Spark

**Big Data use cases**

Big data can help you address a range of business activities, from customer experience to analytics. Here are just a few. .

Product Development: Companies like Netflix and Procter & Gamble use big data to anticipate customer demand. They build predictive models for new products and services by classifying key attributes of past and current products or services and modelling the relationship between those attributes and the commercial success of the offerings. In addition, P&G uses data and analytics from focus groups, social media, test markets, and early store rollouts to plan, produce, and launch new products.

Predictive Maintenance: Factors that can predict mechanical failures may be deeply buried in structured data, such as the year, make, and model of equipment, as well as in unstructured data that covers millions of log entries, sensor data, error messages, and engine temperature. By analyzing these indications of potential issues before the problems happen, organizations can deploy maintenance more cost effectively and maximize parts and equipment uptime.

Customer Experience: The race for customers is on. A clearer view of customer experience is more possible now than ever before. Big data enables you to gather data from social media, web visits, call logs, and other sources to improve the interaction experience and maximize the value delivered. Start delivering personalized offers, reduce customer churn, and handle issues proactively.

Machine Learning: Machine learning is a hot topic right now. And data—specifically big data—is one of the reasons why. We are now able to teach machines instead of program them. The availability of big data to train machine learning models makes that possible.

A screenshot of a web page

Description automatically generated

**What is Big Data?**

In simple words, Big Data is larger, more complex datasets, especially from new data sources. These datasets are so voluminous that traditional data processing software cannot manage them. Although, there is no fixed definition for Big Data, it is featured by the 5 Vs:

**Volume**:  With big data, you’ll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

**Velocity**: Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.

**Variety**: Variety refers to the many types of data that are available. Unstructured and semi structured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

**Value**: Data has intrinsic value. But it’s of no use until that value is discovered.

**Veracity**:  It refers to the quality and accuracy of data. Gathered data could have missing pieces, may be inaccurate or may not be able to provide real, valuable insight. Veracity, overall, refers to the level of trust there is in the collected data.

**Big Data Processing Tools**

Big Data processing technologies provide ways to work with large sets of structured, semi-structured, and unstructured data so that value can be derived from big data.

* Apache Hadoop
* Apache Hive
* Apache Spark

**Big Data use cases**

Big data can help you address a range of business activities, from customer experience to analytics. Here are just a few. .

**Product Development**

Companies like Netflix and Procter & Gamble use big data to anticipate customer demand. They build predictive models for new products and services by classifying key attributes of past and current products or services and modeling the relationship between those attributes and the commercial success of the offerings. In addition, P&G uses data and analytics from focus groups, social media, test markets, and early store rollouts to plan, produce, and launch new products.

**Predictive Maintenance**

Factors that can predict mechanical failures may be deeply buried in structured data, such as the year, make, and model of equipment, as well as in unstructured data that covers millions of log entries, sensor data, error messages, and engine temperature. By analyzing these indications of potential issues before the problems happen, organizations can deploy maintenance more cost effectively and maximize parts and equipment uptime.

**Customer Service**

The race for customers is on. A clearer view of customer experience is more possible now than ever before. Big data enables you to gather data from social media, web visits, call logs, and other sources to improve the interaction experience and maximize the value delivered. Start delivering personalized offers, reduce customer churn, and handle issues proactively.

**Machine Learning**

Machine learning is a hot topic right now. And data—specifically big data—is one of the reasons why. We are now able to teach machines instead of program them. The availability of big data to train machine learning models makes that possible.

Big data sure is the future of data since there are millions of bit flying everywhere in a single second. In the next lesson, we will discuss how to gather this data and how to process it.

## *Gathering & Wrangling Data*

**Data Gathering**

Data gathering is the process of collecting, pooling, and analyzing data or information for a specific purpose. It can relate to various areas of interest, such as business, education, or research. Data gathering requires a clear objective, a plan, and a data collection tool or technique.

## **Process for Identifying Data**

## **Determine the information you want to collect**

## The first thing you need to do is choose what details you want to collect. As an example, you may decide to collect data about which type of articles are most popular on your website among visitors who are between the ages of 18 and 34. You might also choose to gather information about the average age of all of the customers who bought a product from your company within the last month.

## Define a plan for collecting data

1. Establish a timeframe - how long will the entire process take?

2. How much data is sufficient for credible analysis?

3. Define dependencies, risk, and mitigation plan

## Define your data collection method

Data can be qualitative or quantitative. The methods of data collection depend upon a number of factors:

1. Source of data
2. Type of data
3. The timeframe over which you need data
4. Volume of data
5. Degree of accuracy required
6. Funds

***SOURCES OF DATA***

Primary Data: Primary data refers to information obtained directly from the source. Data from the organization’s CRM, HR, or workflow applications. This can be obtained directly from surveys, focus groups, interviews, or questionnaires

Secondary Data: Secondary data refers to the information obtained from existing sources such as external databases, research articles, and training material. This can also be obtained directly from surveys, focus groups, interviews, or questionnaires.

Third Party: Third-party refers to data that is purchased from aggregators who collect data from various sources and combine it into comprehensive datasets.

Some of the data sources from which you could be gathering data include databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys and observation studies.

A survey results with a survey

Description automatically generated with medium confidence

**Activity Time!**

Suggest a correct source of gathering data for the below-specified scenarios. There are no incorrect answers, just remember to explain your rationale behind your choice.

* Scenario 1:

      A marketing company is interested in the proportion of people who will buy a particular product.

* Scenario 2:

 A Community College instructor is interested in the mean number of days math students are absent from class during a quarter.

* Scenario 3:

 John requires an information about mechanical, orderly tasks, like checking the number of manual interventions required in a day to keep an assembly line functioning smoothly.

* Scenario 4:

Sneha wants to launch her online store. She wants to run a quick analysis  through which she can decide on the type of online purchases made by users frequently.

**How do we gather and import data?**

There are many ways in which the data can be gathered and imported from various sources. Some of these methods are listed below. Please click on the flashcards to know more about these:

API is the acronym for application programming interface — a software intermediary that allows two applications to talk to each other. APIs are an accessible way to extract and share data within and across organizations.

Whenever you use any app, it connects to the Internet and sends data to a server. The server then retrieves that data, interprets it, performs the necessary actions, and sends it back to your phone. The application then interprets that data and presents you with the information you wanted in a readable way.

Data exchange is the process of taking data structured under a source schema and transforming it into a target schema, so that the target data is an accurate representation of the source data. Data exchange allows data to be shared between different computer programs.

Also known as event stream processing, streaming data is the continuous flow of data generated by various sources. By using stream processing technology, data streams can be processed, stored, analyzed, and acted upon as it's generated in real-time.  The term "streaming" is used to describe continuous, never-ending data streams with no beginning or end, that provide a constant feed of data that can be utilized/acted upon without needing to be downloaded first.

Web scraping refers to the extraction of data from a website. This information is collected and then exported into a format that is more useful for the user. It can be a spreadsheet or an API.

**What is Data Wrangling?**

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw” data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data.

Lets take a look at the Data Wrangling Process below.

A diagram of data wrangling

Description automatically generated

Discovery Phase: The Discovery phase, also known as the Exploration phase, is about understanding your data better with respect to your use case.

To make your data easier to use and analyze, you must look at it and consider how you would like the data to be arranged.

The data may show trends or patterns during the discovery process. This is a crucial step because it will influence all subsequent actions. It also identifies obvious problems, like values that are missing or incomplete.

Transformation: The transformation phase forms the bulk of the data-wrangling process. It involves the task to transform the data, such as:

1. Structuring - unprocessed data is formatted to make sense, can make use of spreadsheets or other tools
2. Normalizing, De-Normalizing - Getting rid of any unstructured or redundant data
3. Cleaning - Adding missing values, getting rid of outliers
4. Enriching data - Adding context to the data

Validation:

1. Checking the quality of data after cleansing, normalizing, denormal zing, and enriching the data
2. Verifying consistency, quality, and security of data
3. Once the transformations are complete, you can view the Results Summary, which displays detailed statistics of the transformations applied over the entire dataset. You can then export the results of this transformation into the appropriate output format best fit for your visualization or analytics tool of choice

Publishing phase: When your data has been successfully structured, cleaned, enriched and validated, it’s time to publish your wrangled output for use in downstream analytics processes. Through the wrangling process, a wider variety of data sources can be used in different statistics, analytics and data visualization applications. This broadens the usage of data throughout the organization and enhances the potential value of data to the business.

Now you’re ready to deliver the output of your data wrangling efforts into the appropriate format for downstream analytic uses. You can publish your data for business analysis, insights using your tools of choice.

**Let's dive deeper into Data Transformation**

Data transformation is nothing but processing the raw data you have collected from several different sources that gives some semblance. This semblance is essential before to undertake the next steps towards Data Analysis. There are many ways in which data can be transformed, some of them are listed below:

Structuring: Structuring refers to actions that change the form or schema of your data. Splitting columns, pivoting rows, and deleting fields are all forms of structuring.

Join and Union are the most common structural transformation to combine data from one or more tables when you are using SQL to structure data.

Normalizing: Normalization, from a statistical view, often has to do with calculating new values from a dataset to standardize the data on a particular scale.

It can also imply how well the transaction data is handled for reducing redundancy and inconsistency.

De-normalizing: Denormalization is used to combine data from multiple tables into a single table so that it can be queried faster.

Cleaning: During the cleaning stage, users identify data quality issues, such as missing or mismatched values, and apply the appropriate transformation methods to correct or delete these values from the dataset.

Data cleaning is the most important steps of Data Transformation. It involves 3 further steps: Inspection, Data profiling and Verification.

Fun Fact: Data Scientists spend 60% of their time cleaning data - Gartner Research.

Enchancing/Enrichment: Data enrichment is the process of improving the accuracy and reliability of your raw customer data. Teams enrich data by adding new and supplemental information and verifying the information against third-party sources.

Data enriching (also called data appending) ensures your data accurately and thoroughly represents your audience.

**Tools Used For Data Wrangling**

Data wrangling is an important step in Data Analysis as it provides financial insights, improves reporting, provides a unified format to the otherwise crude data, helps understand customer base and elevates the data quality. There are several tools used for Data wrangling and some salient ones are listed below:

1. Microsoft Excel is a spreadsheet editor developed by Microsoft for Windows, macOS, Android, iOS and iPadOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It is the most widely used tool for Data Wrangling.
2. R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical :linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering,  and graphical techniques, and is highly extensible.
3. Trifacta is a Cloud-based Interactive platform for profiling data and applying Machine Learning and Analytics models to it. Regardless of how chaotic or complex the datasets are, this data engineering tool tries to create intelligible data. Deduplication and Linear Transformation techniques allow users to delete duplicate entries and fill blank cells in datasets.
4. Watson Studio provides the environment and tools for you to collaborately work on data to solve your business problems. You can choose the tools you need to analyze and visualize data, to cleanse and shape data, to ingest streaming data, or to create and train machine learning models.
5. Tableau is a data visualization tool that has a variety of eye-catching visualizations, including Treemaps, Gantt Charts, Histograms, and Motion Charts. It’s important to note that it’s not primarily a Data Wrangling Tools, but it does have some Data Preparing and Cleaning Tools that aid in the creation of the flashy visuals.

**Activity Time: Case Analysis**

Maria is a 25-year-old US Army veteran, newly returned to the civilian workforce. She has recently completed a six-year commitment with the Army. During her time in the Army, she worked in supply management and logistics. She has decided to pursue a degree in Management Systems and Information Technology.

Maria has asked you to use your data skills to help her search for the best school for her. She is willing to relocate anywhere in the continental United States, but she has a few criteria that her ideal schools must satisfy:

1. Safety of the city
2. Schools should be offering a degree in IT
3. Ranking of the school

Please refer to the following datasets.

**A screenshot of a login

Description automatically generated**

**Your Next Steps**

1. **Clean the Data for**
   1. Any duplicates
   2. Missing Values
   3. Inconsistent values
2. **Data Enrichment**
   1. Calculate School Ranking
   2. Calculate the overall crime rate
3. **Structure the Data:**Merge the tables and produce the dataset which must have:
   1. Top 5 schools on rankings for It college
   2. Be in a city that is below 50th percentile in overall crime
   3. Remove unnecessary columns

# **Mining & Visualizing Data and Communicating Results**

**Statistical Analysis**

Statistical knowledge helps you use the proper methods to collect the data, employ the correct analyses, and effectively present the results. Statistics is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions.

Statistical Analysis can be:

* Descriptive: Descriptive analysis allows for a simple and meaningful representation of data. Common measures of descriptive analysis include:
  1. Central Tendency
  2. Skewness
  3. Dispersion
* Inferential: Takes data from a sample to make inferences about the larger population from which the sample was drawn. Common Methodologies for Inferential Statistics include:
  1. Hypothesis Testing
  2. Confidence Intervals
  3. Regression Analysis

**How is Statistics related to Data Analysis?**

**Data analysis and statistics are extremely important to many fields. In fact, statistics and data analysis are crucial for conducting internal audits and performance reviews.**

**In many industries like marketing, research, financial services and medical or clinical research, statistics and data analysis are crucial. This is true because of the inherent power of statistics in the analyzing of data.**

**Statistics can also be essential to achieve extremely important growth and efficiency objectives. Because it is so important for businesses to use statistics to lead to meaningful results, interpretations and business decisions, the importance of statistics can not be overstated.**

There are various software packages to perform statistical data analysis, such as Statistical Analysis System (or SAS), Statistical Package for the Social Sciences (or SPSS), and Stat Soft.

All these packages perform the function of processing complex raw data and generate meaningful insights. This helps the organization to make better decisions. It helps us to mine data from various sources, compile it and analyze it. It does the digging, cleaning, processing, and packaging for data.

A red square with white text

Description automatically generated

**What is Data Mining?**

Data mining is also called knowledge discovery and data mining (KDD). It is an extraction of useful patterns from data sources, e.g., databases, texts, web, image. Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

The patterns that are generated must be valid, novel, potentially useful, understandable.

**Let's see how Data Mining works in the field of Market Analysis and Management**

Where does the data come from? It can come from varied sources that include but are not limited to Credit card transactions, loyalty cards, discount coupons, customer complaint calls, brand surveys, customer satisfaction studies etc.

When conducting Market Analysis, a lot of other analyses also need to be factored in.

Target Marketing:

* Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
* Determine customer purchasing patterns over time

Cross Market Analysis:

* Find associations/co-relations between product sales, & predict based on such association

Customer Requirement Analysis:

* Identify the best products for different customers
* Predict what factors will attract new customers

Provision of Summary Information:

* Multidimensional summary reports:
  + Eg: Summarize all transactions of the first quarter from three different branches
* Statistical Summary information
  + E.g. What is the average age for customers who buy product A?

Fraud Dection:

* Find outliers of unusual transactions

Financial Planning:

* Summarize and compare the resources and spending

**Data Mining Techniques**

A diagram of a data algorithm

Description automatically generated

**Classification**

* Classification is the process finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown
* The derived model may be represented in various forms, such as classification (if – then) rules, decision trees, and neural networks

**Clustering**

* Involves grouping Data into Clusters
* Similar data is grouped in the same cluster
* Dissimilar data is grouped in the same cluster

**Regression Analysis**

* Regression deals with the prediction of a value, rather than a class.
* Regression is a data mining function that predicts a number.
* For example, a regression model could be used to predict children's height, given their age, weight, and other factors.
* There are two types of regression models: Linear regression and multiple linear regression models.

**Association Rules / Affinity Grouping**

* An association algorithm creates rules that describe how often events have occurred together.
* Association analysis is widely used for a market basket or transaction data analysis.
* Example: When a customer buys a Computer, then 90% of the time they will buy the software.

**Outlier Detection**

* This technique is used for identifying unusual or suspicious cases that deviate from the projected pattern or expected norm
* The applications of outlier or anomaly detection lie in the identification of credit fraud, taxation fraud, etc.

**Decision Trees**

* A decision tree is a flow-chart-like tree structure, where each node represents a test on an attribute value, each branch denotes an outcome of a test, and tree leaves represent classes or class distributions.
* Decision work best with a simple data set.

**Sequential Pattern Mining**

* Sequential pattern mining, also known as GSP (Generalized Sequential Pattern) mining, is a technique used to identify patterns in sequential data.
* The goal of GSP mining is to discover patterns in data that occur over time, such as customer buying habits, website navigation patterns, or sensor data.

**Communicating Data Analysis Findings**

The process of Data Analysis starts with *Understanding the problem* and ends with *communicating the findings*in a way that impacts decision making.

To reconnect with your audience. Begin by asking yourself these questions:

1. Who is my audience?
2. What is important to them?
3. What will help them trust me?

Next, **Structure your Presentation**

* Reference your data
* State your assumption
* Organize your presentation
* Identify the best formats for presenting your data