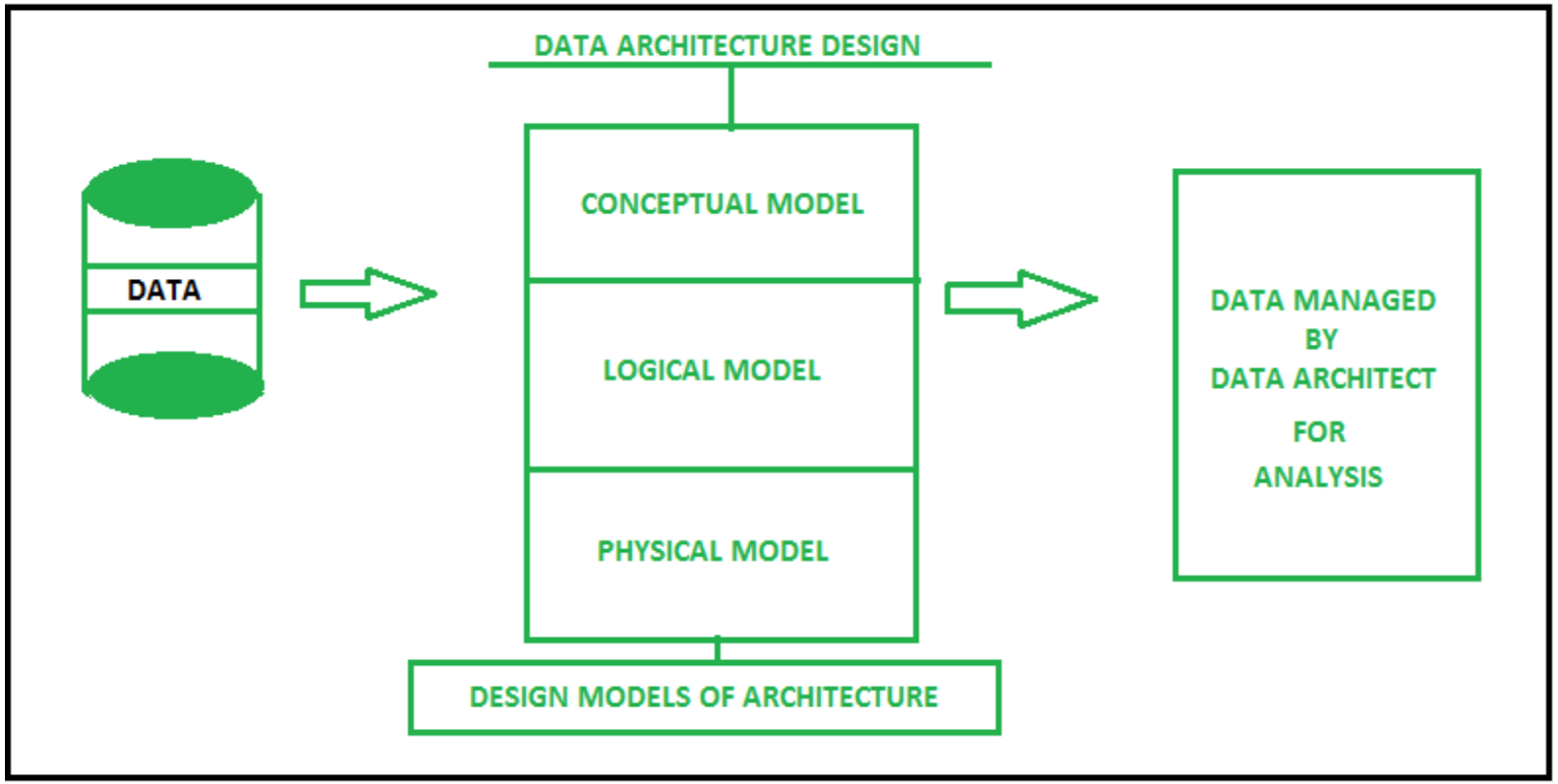
UNIT_I

Data Management

Data architecture



Common Types of Data Sources

- **Databases:** Relational (SQL, MySQL, PostgreSQL) and NoSQL databases.

- **Files:** Spreadsheets (Excel, Google Sheets), CSVs, text documents, PDFs.

- **Web & APIs:** Websites, web scraping, and Application Programming Interfaces (APIs) for real-time data.
- **Sensors & Devices:** IoT devices, physical measurements, live readings.

- **Social Media:** Posts, likes, shares from platforms like Twitter, Facebook.

- **Surveys & Interviews:** Direct data collection from people (structured or unstructured).

- **Third-Party Providers:** External data vendors, data warehouses (Snowflake, BigQuery)

Data quality (noise, outliers, missing values, duplicates)

Common Data Quality Problems

Noise: Meaningless, random errors or variations in data, often from measurement issues or system imperfections (e.g., a temperature reading of 1000°C).

Outliers: Extreme data points far from the norm, which can be genuine but unusual values or errors, skewing statistical results (e.g., a CEO's salary in an employee dataset).

Missing Values: Gaps where data should be, due to sensor failures, skipped questions, or system crashes, requiring imputation or exclusion.

Duplicates: Identical or near-identical records, causing over-representation, skewed trends, and increased storage.

Data Processing

- Data processing is the **sequence of operations** that convert **raw, unorganized data** into **useful, structured information**. It is a core step in **data analytics, machine learning, business intelligence, and database systems**.
- Because raw data is messy and unreadable.
  Data processing helps:
- Improve accuracy
- Reduce redundancy
- Enable machine learning models
- Extract patterns and insights
- Support strategic decisions

UNIT-II
Data Analytics

I) Qualitative Data:

As the name suggests Qualitative Data tells the features of the data in the statistics. Qualitative Data is also called Categorical Data and it categorizes the data into various categories. Qualitative data includes data such as the gender of people, their family names, and others in a sample of population data.

- Qualitative data is further categorized into two categories that include,

a) **Nominal Data** (Example:*ender (Male or female), Race (White, Black, Asian), Religion (Hinuduism, Christianity, Islam, Judaism), Blood type (A, B, AB, O), etc.)*

b) **Ordinal Data**(Example:*Education level (Elementary, Middle, High School, College),Job position (Manager, Supervisor, Employee), etc.)*

II) Quantitative Data (Numerical Data):

- Quantitative Data is the type of data that represents the numerical value of the data. They are also called Numerical Data. This data type is used to represent the height, weight, length, and other things of the data. Quantitative data is further classified into two categories that are,

a) Discrete Data(Example:*Height of Students in a class,Marks of the students in a class test&Weight of different members of a family, etc.)*

b) Continuous Data(Example:*Temperature Range & Salary range of Workers in a Factory, etc.)*

Data modeling techniques

Common data modeling techniques:

a) Dimensional Modeling:

Organizes data for data warehouses and business intelligence, using facts (measures) and dimensions.

b) Entity-Relationship (ER) Modeling:

Represents entities (like "customer" or "product") and their relationships within a database, often visualized using an ER diagram.

c) Graph Data Modeling:

Uses nodes to represent entities and edges to represent relationships, ideal for analyzing complex connections, notes Coursera.

d) Relational Modeling:

Organizes data into tables with rows and columns, based on relational algebra, notes Atlan.

e) Hierarchical Modeling:

Structures data in a tree-like format with a single parent for each child element.

f) Object-Oriented Modeling:

Represents data as objects, which have both data (attributes) and actions (methods).

Missing data imputation

- Data Imputation is the process of handling missing values or null values in a dataset for the purpose of enhancing efficiency and accuracy in the model training process.
- This can be done by either replacing estimated or predicted values. Missing values occur due to various reasons, including, missing information, data entry errors, data deletion, or other inconsistencies.
- It is essential to handle these values to ensure unbiased predictions and data compatibility with all models. In this article, we will see What is Data Imputation and Techniques to perform it in Machine Learning.

These can arise due to various reasons. Some common causes are:

- Human error at time of Data entry
- Data collection issues
- Mismatched features while Dataset merging
- Data corruption or Data loss
- Inconsistent Data types
- Unanswered questionnaires

Business modeling

- Business modeling is a strategic process in modern analytics that helps companies understand the relationships between their organization's data and business objectives.
- This process produces business models that reflect an organization's business processes and logic.
- Business modelling is used to design current and future state of an enterprise.
- This model is used by the Business Analyst and the stakeholders to ensure that they have an accurate understanding of the current As-Is model of the enterprise.
- It is used to verify if, stakeholders have a shared understanding of the proposed To-be of the solution.
- Analyzing requirements is a part of business modelling process and it forms the core focus area. Functional Requirements are gathered during the Current state.
- These requirements are provided by the stakeholders regarding the business processes, data, and business rules that describe the desired functionality which will be designed in the Future State.

UNIT_III
Regression

Regression Analysis is a statistical method used to understand the relationship between input features and a target value that varies across a continuous numeric range. It helps measure how changes in different factors affect the outcome, allowing better predictions, planning and decision-making across various fields.

- **Types of Regression**
- Some commonly used regression techniques are:
- **Linear Regression:** Models straight-line relationships between predictors and outputs.($Y=\beta 0+\beta 1X+\epsilon$, Where: Y is the predicted value,$\beta 0$ is the intercept,$\beta 1$is the coefficientaffecting X,$\epsilon$is the error term.)
- **Multiple Regression:** Uses multiple input features to predict one continuous outcome.($Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$)
- **Polynomial Regression:** Captures non-linear patterns by transforming input variables.($y=\beta 0+\beta 1x+\beta 2x2+\beta 3x3+\cdots+\beta nxn+\epsilon$)

logistic regression

- Logistic regression is a statistical method in data analytics used for predicting the probability of a binary outcome, such as "yes" or "no," "pass" or "fail," or "spam" or "not spam". It works by modeling the relationship between independent variables and a dependent variable, using a sigmoid function to produce a probability between 000 and 111. A threshold, often 0.50.50.5, is then used to classify the outcome.

- **How it works**

- **Predicts probability:**

Logistic regression calculates the probability of a specific event occurring based on a set of input variables.

- **Uses the sigmoid function:**

Instead of a straight line like linear regression, it uses an S-shaped curve called the sigmoid function, which maps the linear combination of inputs to a probability value between 000 and 111. The formula is: P(x)=11+e−xcap P open paren x close paren equals the fraction with numerator 1 and denominator 1 plus e raised to the negative x power end-fraction $P(x)=11+e-x$

- **Classifies outcomes:**

A threshold is applied to the predicted probability to classify the outcome into one of two categories. For example, if the predicted probability is greater than 0.50.50.5, it's classified as one outcome (e.g., "pass"); otherwise, it's the other (e.g., "fail").

least squares Regression

- Least Squares Regression in data analytics finds the "line of best fit" for data by minimizing the sum of the squared differences (residuals) between actual data points and the predicted line, enabling modeling of relationships between variables for predictions, widely used in simple/multiple regression for forecasting trends in finance, economics, etc., though sensitive to outliers and assumes linearity.
- **Key Concepts**
- **Best-Fit Line:** A straight line drawn through a scatter plot that best represents the trend of the data points, minimizing overall error.
- **Residuals:** The vertical distances (errors) between each actual data point and the regression line.
- **Minimizing Squared Errors:** Instead of summing absolute errors, it squares them to penalize larger errors more heavily and avoid negative/positive errors canceling out, hence the name "least squares".

model building/fitting, analytics applications across domains.

- Model building is an essential part of data analytics and is used to extract insights and knowledge from the data to make business decisions and strategies.
- In this phase of the project data science team needs to develop data sets for training, testing, and production purposes. These data sets enable data scientists to develop an analytical method and train it while holding aside some of the data for testing the model.

Key Data Analytics Applications Across Domains:

- **Finance & Banking:** Fraud detection & risk management Analyzing transaction data to spot anomalies and assess risks associated with loans and investments.
- **Healthcare:** Diagnosis & treatment optimization Utilizing massive patient datasets to identify patterns, predict disease trends, and improve patient outcomes.
- **Retail & E-Commerce:** Customer insights & inventory managementAnalyzing browsing and purchase history to provide personalized recommendations, optimize pricing, and predict product demand.
- **Manufacturing:** Quality control & process optimizationUsing predictive analytics to anticipate equipment failures, reduce waste, and refine production workflows.
- **Marketing:** Customer segmentation & campaign analysisSegmenting customers based on behavior and demographics to launch more effective and targeted advertising campaigns.

The Role of "Fitting" in Data Analytics:

"Fitting a model" is a core step in the data analytics process, typically falling under **predictive** and **prescriptive analytics**.

- **Process**: It involves selecting an appropriate modeling technique (like regression or clustering) and training the model using historical data to identify relationships and patterns.
- **Purpose**: A well-fitted model can then be used to forecast future trends, events, and behaviors, enabling proactive and data-driven decisions.
- **Tools**: Analysts use various tools like Python, R, SAS, and platforms such as Tableau and Power BI to perform data preparation, model fitting, and visualization.
- **Importance of Domain Knowledge**: Successful model fitting requires significant domain expertise to ensure the right variables are used and the results are interpreted accurately within the specific industry context.

Unit IV

Data segmentation

- Object segmentation in data analytics, often from computer vision, divides images into distinct parts (pixels or regions) for detailed analysis, unlike general data segmentation which groups users/records by traits (age, behavior); it's used in medical imaging (tumor detection), autonomous vehicles (road/sign ID), and manufacturing (defect finding), enabling precise object understanding beyond bounding boxes for richer insights and model building.
- Tree-based models are a class of supervised machine learning algorithms used for both classification and regression tasks in data analytics. They operate by recursively partitioning the data into smaller subsets based on the values of input features, forming a decision tree structure.

How they work:

- **Root Node:**

The process begins with a single root node representing the entire dataset.

- **Splitting:**

At each internal node, the algorithm evaluates different features and their values to find the optimal split that best separates the data based on the target variable. This split creates branches leading to child nodes.

- **Recursive Partitioning:**

This splitting process continues recursively down the tree until a stopping criterion is met, such as a maximum depth, a minimum number of samples in a leaf node, or a lack of significant improvement in purity.

- **Leaf Nodes:**

The final nodes, called leaf nodes, represent the predicted class or value for the data points that fall into that specific partition.

supervised & unsupervised learning

Supervised learning as the name suggests, works like a **teacher** or **supervisor** guiding the machine. In this approach we teach or train the machine using the labelled data(correct answers or classifications) which means each input has the correct output in the form of answer or category attached to it. After that machine is provided with a new set of examples (data) so that it can analyses the training data and produces a correct outcome from labeled data.

For example: a labeled dataset of images of Elephant, Camel and Cow would have each image tagged with either "**Elephant**", "**Camel**" or "**Cow**."

Types of Supervised Learning: 1.Regression 2.Classification

- Unsupervised learning: is a part of machine learning which works differently from supervised because there is no teacher(supervisor) involved to guide the machine. In this approach the machine is given with data that has **no labels or categories**. It analyzes the data on its own to find patterns, groups or relationships without any prior knowledge. The machine learns by discovering hidden structures within the data without being told what the correct output should be.

For example: unsupervised learning can analyze animal data and group the animals by their traits and behavior. These groups might represent different species which allows the machine to organize animals without any prior labels or categories.

- **Types of Unsupervised Learning: 1.Clustering and 2.Association rule learning**

Decision trees (regression & classification)

- Decision trees in data analytics are supervised machine learning algorithms used for both classification and regression, creating a flowchart-like structure to model decisions.
- They work by recursively splitting data based on feature values, with classification trees predicting a categorical outcome and regression trees predicting a continuous, numerical value.
- These algorithms are valued for their interpretability, ability to handle various data types, and use as a foundation for more complex ensemble methods like random forests.

Classification vs. Regression trees:

- **Output:** Predicts a categorical (e.g., "yes" or "no") or discrete outcome Predicts a continuous, numerical outcome (e.g., price, temperature)
- **Splitting Goal :**To maximize the purity of nodes, often using criteria like Gini impurity or Information Gain to make splits where the majority of data points belong to a single class To minimize the variance or mean squared error within the subsets at each split
- **Prediction:** Assigns the class that is most frequent in the final leaf node for a given data point Assigns the average value of the target variable of all the data points in the final leaf node

Over fitting & pruning

- In data analytics, **overfitting** happens when a model learns training data too well, including noise, failing to generalize to new data (high training accuracy, low test accuracy).
- **Pruning** is the key solution, simplifying complex models (like decision trees) by removing unnecessary branches or nodes to reduce complexity, control variance, and improve performance on unseen data.

Over fitting Explained

- **Definition:** A model becomes too complex, essentially "memorizing" the training data's specific instances and noise rather than learning underlying patterns.
- **Symptoms:** Excellent performance on training data but poor performance (high error) on new, unseen data (test/validation data).
- **Analogy:** Like a student who memorizes answers for a test but doesn't understand the concepts, failing when questions are rephrased.

Pruning Explained (Focus on Decision Trees)

- **Definition:** The process of reducing the size of a decision tree by removing sections (branches/nodes) that provide little predictive power, often fitting noise in the data.
- **Goal:** Simplify the tree to improve its ability to generalize (generalization error).

Types of Pruning

- **Pre-Pruning (Early Stopping):** Stops tree growth *before* it fully fits the training data, using hyperparameters like max_depth or min_samples_leaf to set limits.
- **Post-Pruning (Backward Pruning):** Grows a full tree first, then removes branches from the bottom up, using a validation set to decide which nodes to cut, balancing error and complexity.

Time Series methods

What is a Time Series?

- **A time series is a sequence of data points collected, recorded, or measured at successive, evenly-spaced time intervals.**
- Each data point represents observations or measurements taken over time, such as stock prices, temperature readings, or sales figures. Time series data is commonly represented graphically with time on the horizontal axis and the variable of interest on the vertical axis, allowing analysts to identify trends, patterns, and changes over time.

Importance of Time Series Analysis:

- **Predict Future Trends:** Time series analysis enables the prediction of future trends, allowing businesses to anticipate market demand, stock prices, and other key variables, facilitating proactive decision-making.
- **Detect Patterns and Anomalies:** By examining sequential data points, time series analysis helps detect recurring patterns and anomalies, providing insights into underlying behaviors and potential outliers.
- **Risk Mitigation:** By spotting potential risks, businesses can develop strategies to mitigate them, enhancing overall risk management.
- **Strategic Planning:** Time series insights inform long-term strategic planning, guiding decision-making across finance, healthcare, and other sectors.
- **Competitive Edge:** Time series analysis enables businesses to optimize resource allocation effectively, whether it's inventory, workforce, or financial assets. By staying ahead of market trends, responding to changes, and making data-driven decisions, businesses gain a competitive edge.

ARIMA

- ARIMA (Autoregressive Integrated Moving Average) model is used for forecasting time series data. It combines three components: autoregression (AR), differencing (I) and moving averages (MA). These components allow the model to capture patterns such as trends and seasonality, helping to predict future values based on historical data. It combines three key components to model data:

1. Autoregression (AR):

- The autoregressive part (AR) of an ARIMA model is represented by the parameter p. It signifies the dependence of the current observation on its previous values. Mathematically, an AR(p) model can be represented as: $Yt=c+\phi 1Yt-1+\phi 2Yt-2+\ldots+\phi pYt-p+\epsilon t$
  Here:
- YtYt is the current observation
- cc is a constant
- ϕ1$\phi 1$ to ϕ2$\phi 2$ are the autoregressive parameters
- ϵtϵt represents the error term at time tt

ARIMA

2. Differencing (I):

- The differencing part of ARIMA is represented by the parameter d. It involves transforming a non stationary time series into a stationary one by differencing consecutive observations. We can apply the differencing operation multiple times until stationarity is achieved. The formula for differencing is: straightforward

$Yt' = Yt - Yt-1$ $Yt' = Yt - Yt-1$

Here:

- Yt' is the differenced series at time t
- Yt is the original series at time t
- $Yt-1$ is the value of the series at the previous time step
- The differencing process is typically applied multiple times until stationarity is achieved. The notation I(d) indicates the order of differencing required for stationarity.

ARIMA

3. Moving Average (MA):

The moving average part (MA) of an ARIMA model is represented by the parameter q. It indicates the dependence of the current observation on the previous forecast errors. Mathematically, an MA(q) model can be represented as:

$$Yt=c+\epsilon t+\theta 1\epsilon t-1+\theta 2\epsilon t-2+\ldots+\theta q\epsilon t-q Yt=c+\epsilon t+\theta 1\epsilon t-1+\theta 2\epsilon t-2+\ldots+\theta q\epsilon t-q$$

Here:

- $Yt Yt$ is the current observation
- $c c$ is a constant
- $\epsilon t \epsilon t$ is the error at time $t t$
- $\theta 1 \theta 1$ to $\theta q \theta q$ are the moving average parameters

Forecasting& error measures

Forecast error measures in data analytics quantify prediction accuracy by comparing forecasts to actuals, with key metrics including **MAD (Mean Absolute Deviation)**, **MSE (Mean Squared Error)**, **RMSE (Root Mean Squared Error)**, and **MAPE (Mean Absolute Percentage Error)**

Common Forecast Error Measures:

- **Forecast Error (e):** The basic difference: Actual Value - Forecasted Value.
- **Mean Absolute Deviation (MAD):** Average of absolute errors, shows error in original units.
- **Mean Squared Error (MSE):** Average of squared errors, heavily penalizes large errors.
- **Root Mean Squared Error (RMSE):** Square root of MSE, also in original units but sensitive to large errors.
- **Mean Absolute Percentage Error (MAPE):** Average of absolute errors as a percentage of actuals, easy to interpret.
- **Mean Percentage Error (MPE):** Average of errors (not absolute), indicates forecast bias (over/under-forecasting).
- **Tracking Signal:** Monitors bias over time, helps detect systematic errors.

Feature extraction& prediction.

Feature extraction in data analytics transforms raw, complex data (like images, text, audio) into simpler, meaningful numerical features, reducing noise and dimensionality for better model efficiency, while **prediction** uses these extracted features with machine learning models (like regression, classification) to forecast future outcomes, trends, or patterns, making data actionable for decisions

Feature Extraction (The "What")

- **Goal**: Convert raw data into a concise set of informative features (variables/attributes).
- **Why**: Raw data is often too complex, noisy, or in the wrong format for direct ML use; extraction simplifies it without losing critical info.

Examples:

  - **Image**: Edges, textures, shapes.
  - **Text**: Word counts, sentiment scores.
  - **Audio**: Pitch, tone, frequency components (MFCCs).

- **Methods**: Manual (domain expertise) or automated (Deep Learning, PCA).

Prediction (The "So What")

- **Goal**: Use the extracted features to build models that forecast future events or classify new data.
- **Process**: Applying ML algorithms (like Logistic Regression, Decision Trees, Neural Networks) to the engineered features.
- **Types**: Classification (predicting categories), Regression (predicting values), Time Series (predicting future points).
- **Outcome**: Actionable insights, risk assessment, demand forecasting, etc..

Unit V
Data Visualization techniques

Data Visualization:

- Pixel-Oriented Visualization Techniques,
- Geometric Projection Visualization Techniques,
- Icon-Based Visualization Techniques,
- Hierarchical Visualization Techniques,

Visualizing Complex Data and Relations:

Data visualization: Data visualization is the graphical representation of

- information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- **Uses of data visualization**
- Powerful way to explore data with presentable results. Primary use is the preprocessing portion of the data mining process.Supports in data cleaning process by finding incorrect and missing values.General Data visualization Techniques:
- Box plots
- Histograms
- Heat maps
- Charts
- Tree maps

Data visualization Techniques:

1. Pixel Oriented Visualization technique
2. Geometric Projection Visualization technique-

a. Scatter plot matrices

b. Hyper slice

c. Parallel coordinates

3. Icon based visualization techniques
4. Hierarchical Visualization techniques

Pixel-Oriented Visualization Techniques:

- A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value.
- For a data set of m dimensions, pixel-oriented techniques create m windows on the screen, one for each dimension.The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows.
- The colors of the pixels reflect the corresponding values. Inside a window, the data values are arranged in some global order shared by all windows.
- The global order may be obtained by sorting all data records in a way that's meaningful for the task at hand.
- Pixel-based visualizations use that approach and are capable of displaying large amounts of data on a single screen.

Geometric Projection visualization techniques:

- A drawback of pixel-oriented visualization techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space. Geometric projection techniques help users find interesting projections of multidimensional data sets.
- Geometric projection techniques are a good choice for finding outliers and correlation between attributes in multivariate data. A geometric projection technique does this by using transformations and projections of the data.
- When using large data sets a clustering algorithm is usually necessary to apply before the visualization technique to avoid cluttered and unclear data caused by the too much information. Some widely used geometric projection techniques are:

Scatter plots:

- A scatter plot is one of the most common visualization techniques and can be visualized both in 3D and 2D.
- The scatter plot visualizes different attributes of the data on the x,y axis for 2D visualizations and also along the z-axis in 3D.
- Scatter plots are usable to find correlations between attributes in arbitrary small data sets.
- If the data set gets too big or contains too many attributes the scatter plot gets cluttered and hard to interpret.

Hyper Slice:

- Hype rSlice is a new method for the visualization of scalar functions of many variables. With this method the multi-dimensional function is presented in a simple and easy to understand way in which all dimensions are treated identically. The central concept is the representation of a multi-dimensional function as a matrix of orthogonal two-dimensional slices.
- These two- dimensional slices lend themselves very well to interaction via direct manipulation, due to a one to one relation between screen space and variable space.
- Parallel coordinates:

To visualize n-dimensional data points, the parallel coordinates technique draws n equally spaced axes, one for each dimension, parallel to one of the display axes.A data record is represented by a polygonal line that intersects each axis at the point corresponding to the associated dimension value.A major limitation of the parallel coordinate's technique is that it cannot effectively show a data set of many records.

Icon-based visualization techniques:

- Icon-based techniques visualize data by changing the properties of an icon or glyph according to the data. An early version was Chernoff faces where data is mapped to different face parts as nose, mouth, eyes and more. For example how rich people are can be mapped to the mouth of the Chernoff face. Rich people represented by a happy mouth and and poor people by a sad mouth. Other methods are:

Hierarchical Visualization Techniques:

- The visualization techniques discussed so far focus on visualizing multiple dimensions simultaneously.
- However, for a large data set of high dimensionality, it would be difficult to visualize all dimensions at the same time.
- Hierarchical visualization techniques partition all dimensions into subsets (i.e., subspaces). The subspaces are visualized in a hierarchical manner. "Worlds-within-Worlds," also known as n-Vision, is a representative hierarchical visualization method.
- Suppose we want to visualize a 6-D data set, where the dimensions

are F,X1, : : : ,X5.

- Given more dimensions, more levels of worlds can be used, which is why the method is called "worlds-within-worlds."
- As another example of hierarchical visualization methods, tree-maps display

hierarchical data as a set of nested rectangles. For example, a tree-map visualizing Google news stories.

Data visualization choices:

- Five factors that influence data visualization choices:
- Audience: It's important to adjust data representation to the specific target audience.
- Content: The type of data you are dealing with will determine the tactics.
- Context: You can use different data visualization approaches and read data depending on the context.
- Dynamics: There are various types of data, and each type has a different rate of change.
- Purpose: The goal of data visualization affects the way it is implemented. In order to make a complex analysis, visualizations are compiled into dynamic and controllable dashboards that work as visual data analysis techniques and tools.

Tools for Data visualization:

- Data visualization tools for different types of users and purposes.

Tableau

R

Python

Plotly

IBM watson analytics

Tools for complex data visualization

- The growing adoption of connected technology places a lot of opportunities
- before the companies and organizations. To deal with large volumes of multi-source often unstructured data, businesses search for more complex visualization and analytics solutions. This category includes Power BI, Kibana and Grafana.
- Power BI is exceptional for its highly intuitive drag-and-drop interface, short
- learning curve and large integration capabilities, including Salesforce and MailChimp.
- Kibana is the part of the Elastic Stack that turns data into visual insights.
- It's built on and designed to work on Elasticsearch data only. This exclusivity, however, does not prevent it from being one of the best data visualization tools for log data.
- Grafana a professional data visualization and analytic tool that supports up to 30 data sources, including AWS, Elastic search and Prometheus. Grafana is more flexible in terms of integrations compared to Kibana, each of the systems works best with its own type of data.

Data Visualization Process

- Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.
- The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.

visualizing complex data & relationships

- Visualizing complex data in analytics turns overwhelming numbers into understandable graphics (charts, maps, dashboards) to reveal trends, patterns, outliers, and relationships, using techniques like heat maps for correlations, network diagrams for connections, and tree maps for hierarchies, with interactivity (filters, drill-downs) making it accessible for faster, smarter decisions across teams by telling stories and simplifying complex insights.

Key Techniques for Complex Data & Relationships:

- **Heatmaps:** Use color intensity to show correlations/interactions between variables.
- **Network Diagrams (Graphs):** Nodes (dots) and edges (lines) map out connections, showing central/isolated elements.
- **Treemaps:** Nested rectangles show hierarchical data, size representing value.
- **Matrix Visualizations:** Grid format for viewing relationships between multiple datasets.
- **Scatter Plots (Enhanced):** Use color, size, and shape to show interactions among multiple factors.
- **Sankey Diagrams/Flow Charts:** Illustrate flow and connections between stages or entities.
- **Parallel Coordinates:** Plot multiple variables for each data