Data Analysis and Visualization

IS465: Data Management and Governance

Outline

- Introduction to Data Analysis
- Data Analysis Techniques
- Data Analysis Tools
- Data Visualization Principles and Best Practices
- Business Intelligence and Data Dashboards
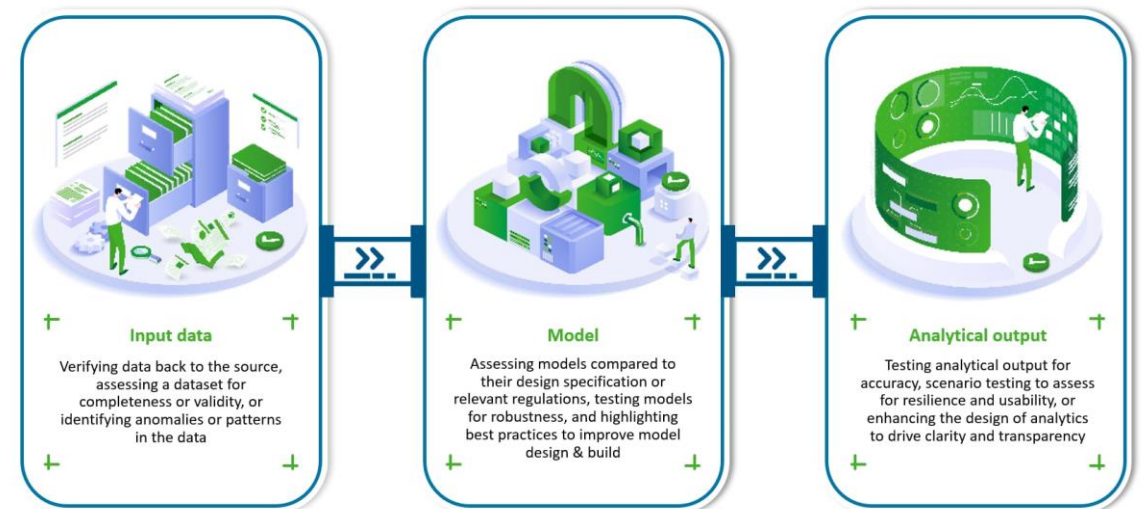
Introduction to Data Analysis

Unlocking Insights from Data

- Data analysis is a crucial step in extracting insights from data.
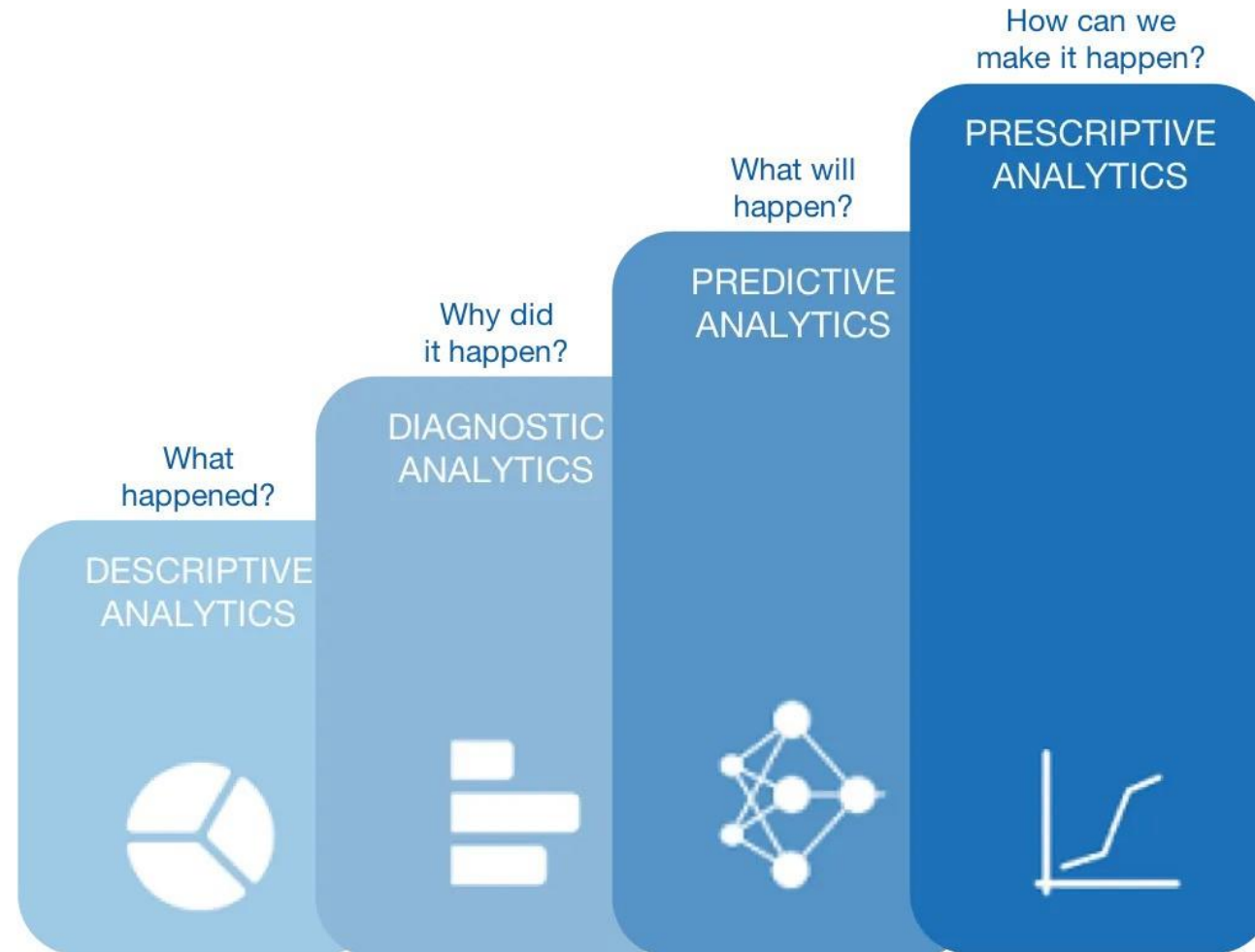
What is Data Analysis?

- Data analysis is the process of extracting insights and patterns from data to inform business decisions, solve problems, or answer questions.

- Examines data to identify trends, relationships, and correlations

- Uses various techniques and tools to extract insights

- Involves data cleaning, transformation, and modeling

Input data
Verifying data back to the source, assessing a dataset for completeness or validity, or identifying anomalies or patterns in the data

Model
Assessing models compared to their design specification or relevant regulations, testing models for robustness, and highlighting best practices to improve model design & build

Analytical output
Testing analytical output for accuracy, scenario testing to assess for resilience and usability, or enhancing the design of analytics to drive clarity and transparency

Why is Data Analysis Important?

- Informs business decisions with data-driven insights
- Identifies areas for improvement and optimization
- Supports data quality and governance initiatives
- Enhances data security and privacy by detecting anomalies

What are the different types of analysis?

How can we make it happen?

PRESCRIPTIVE ANALYTICS

What will happen?

PREDICTIVE ANALYTICS

Why did it happen?

DIAGNOSTIC ANALYTICS

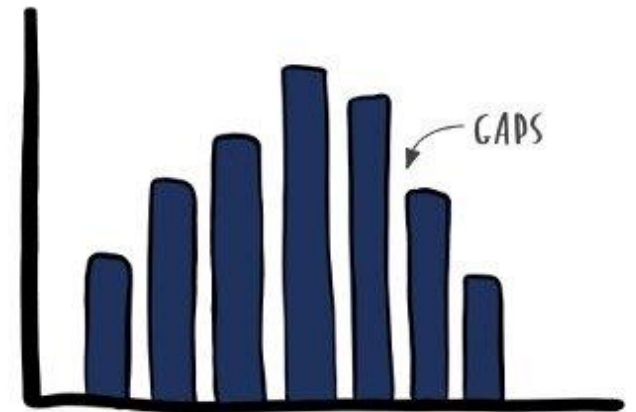What happened?

DESCRIPTIVE ANALYTICS

Descriptive Data Analysis

- Descriptive data analysis involves summarizing and describing the basic features of data to understand what happened.

- Examples:
  - Calculating means, medians, and modes
  - Creating reports and dashboards
  - Identifying trends and patterns

This is a **histogram**...

NO GAPS

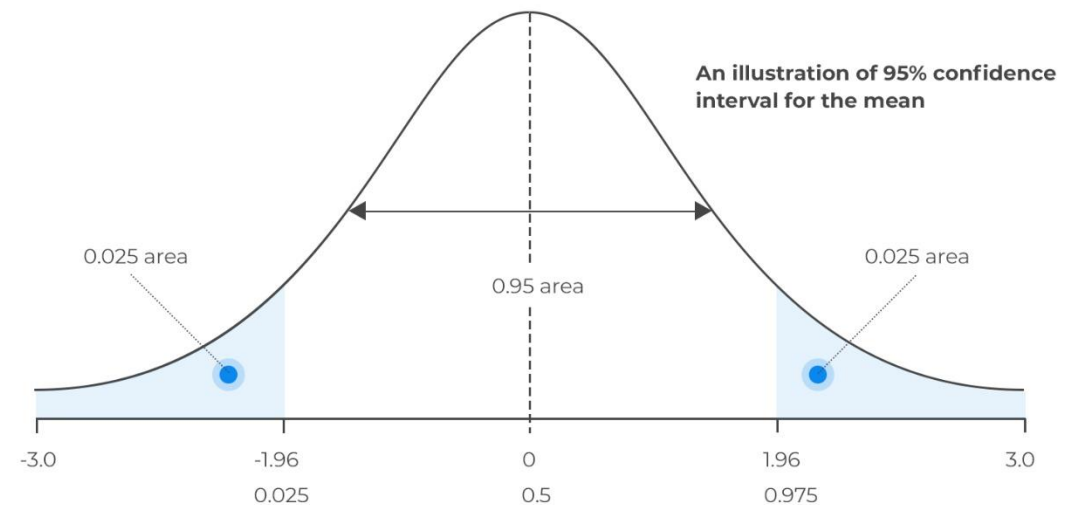This is a **bar chart**...

GAPS
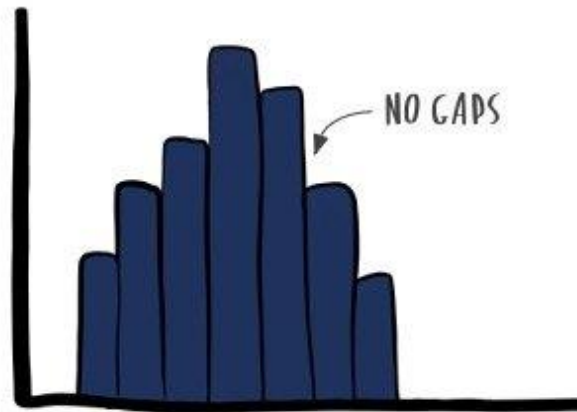
Inferential Data Analysis

- Inferential data analysis involves making conclusions or inferences about a larger population based on a sample of data.

- Examples:
  - Hypothesis testing
  - Confidence intervals
  - Regression analysis

95% Interval

An illustration of 95% confidence interval for the mean

0.025 area   0.95 area   0.025 area

-3.0   -1.96   0   1.96   3.0
        0.025   0.5   0.975

Predictive and Prescriptive Data Analysis

- Predictive data analysis involves using data to forecast what may happen, while prescriptive data analysis involves recommending actions based on data insights.

- Examples:
  - Predictive: forecasting sales, customer churn, or credit risk
  - Prescriptive: recommending product offers, pricing, or inventory levels
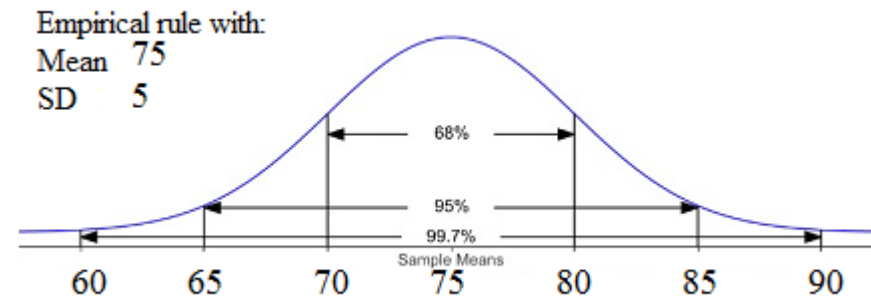
Data Analysis Techniques

Descriptive Analytics

- Understanding Your Data
- Descriptive analytics involves summarizing and describing the basic features of data to understand what happened.

Measures of Central Tendency

- Mean: The mean is the average value of a dataset.
- Formula: $\mu = (\Sigma x) / n$
- Example: calculating the mean of exam scores

Empirical rule with:
Mean 75
SD 5

68%
95%
99.7%
Sample Means

60   65   70   75   80   85   90

Measures of Central Tendency

- Median and Mode: The median is the middle value of a dataset, while the mode is the most frequent value.

- Formula:
  - median = middle value
  - mode = most frequent value

- Example: calculating the median and mode of a dataset
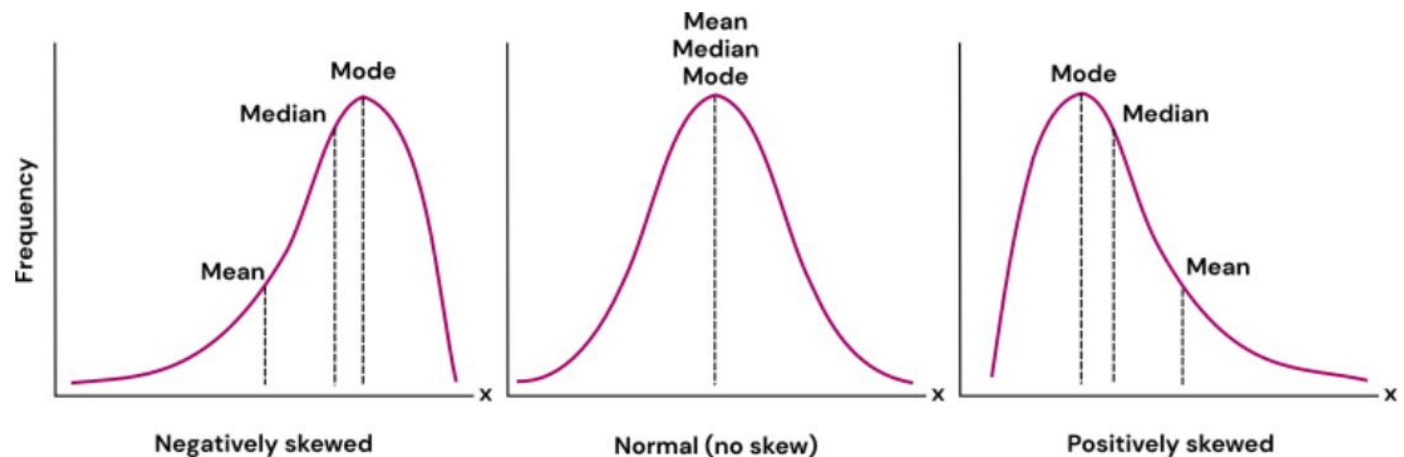
Measures of Variability

- Range and Variance: The range is the difference between the largest and smallest values, while variance measures the spread of a dataset.

- Formula: range = max - min, variance = $\Sigma(x - \mu)^2 / n$

- Example: calculating the range and variance of a dataset

Average phone use per day in minutes

Sample A   Sample A   Sample A
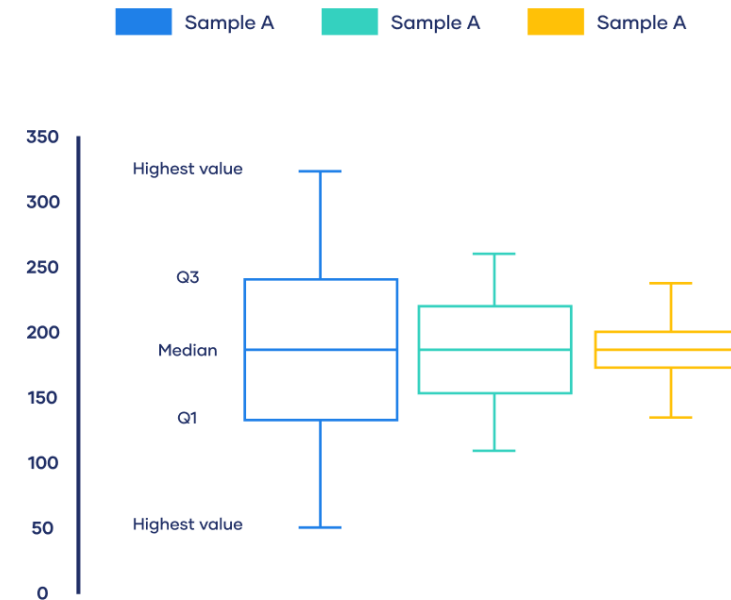
Measures of Variability

- Standard Deviation: The standard deviation is the square root of the variance, measuring the spread of a dataset.

- Formula: $\sigma = \sqrt{\sum(x - \mu)^2 / n}$

- Example: calculating the standard deviation of a dataset
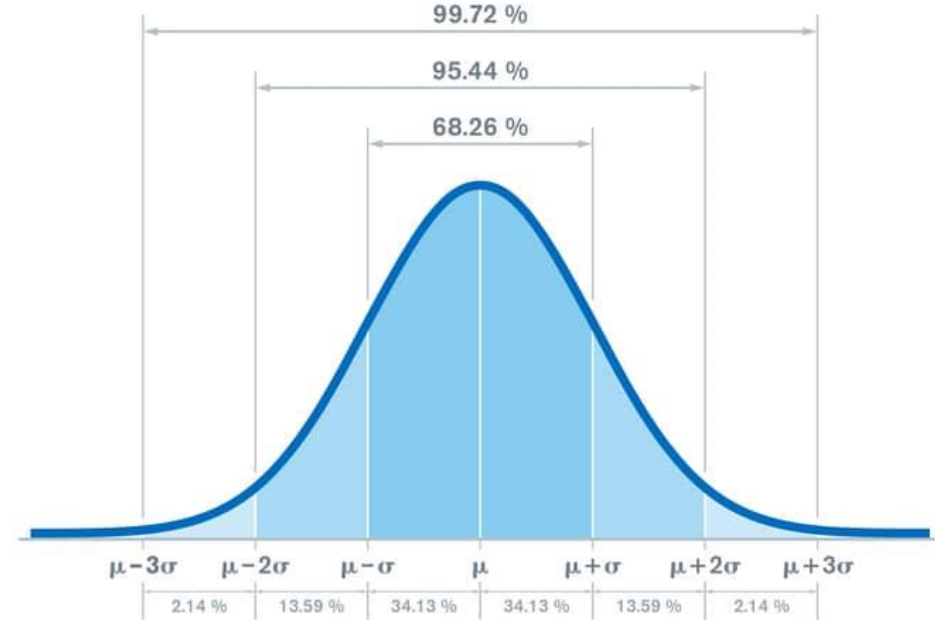
Data Summarization

- Data summarization involves reducing data to its most important features.
- Examples:
  - Summarizing data with mean, median, and mode

Positively skewed distribution: Number of books read in the last year
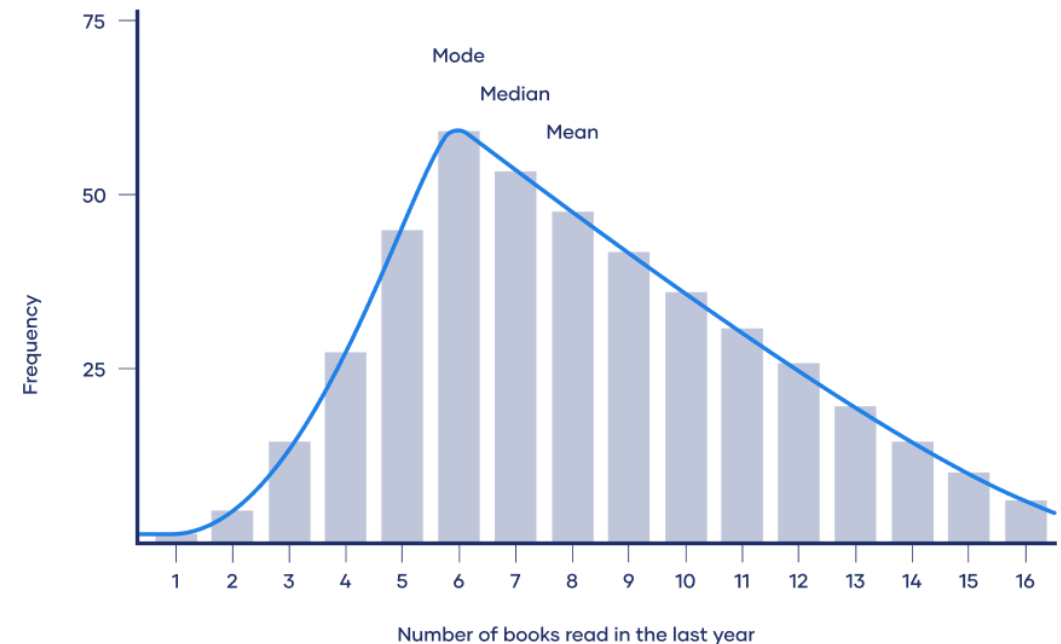
Data Aggregation

- Data aggregation involves combining data to create new insights.
- Examples:
  - Aggregating data by group or category

Type	Property
Pivot Table ▾	

Row

📅 timestamp	DAY	🗑
Choose...	-	

Value

# (Number of Rows)	-	☰
Choose...	-	

Column

Choose...	-

Color By (What is this?) ⚙

+ Filter

Showing 237 rows, 1 columns

| timestamp	(Number of Rows)
2016-11-01	480
2016-11-02	481
2016-11-03	579
2016-11-04	590
2016-11-05	392
2016-11-06	531
2016-11-07	589
2016-11-08	740

Inferential Analytics

- Making Conclusions from Data
- Inferential analytics involves making conclusions or inferences about a larger population based on a sample of data.
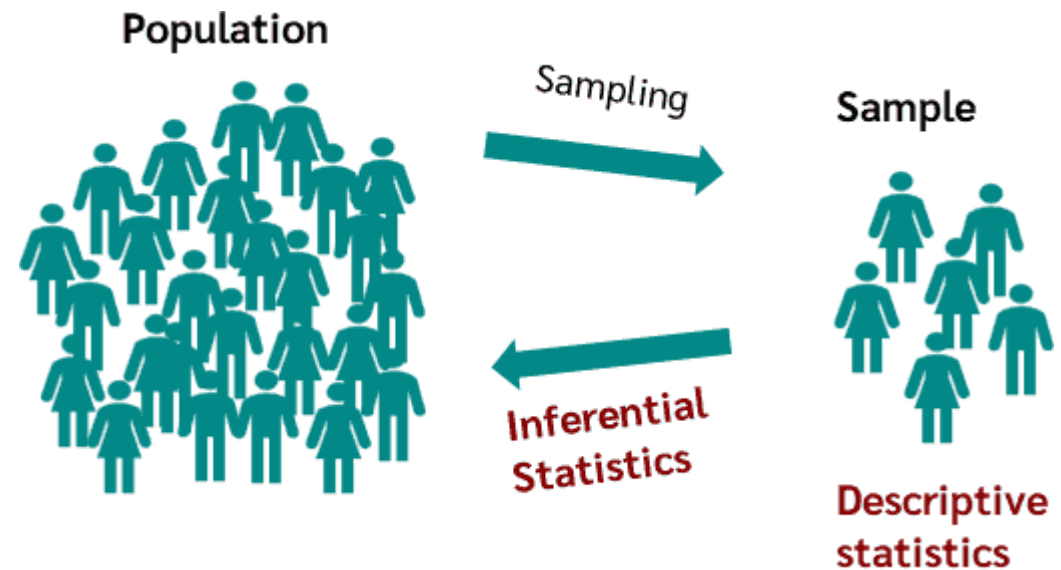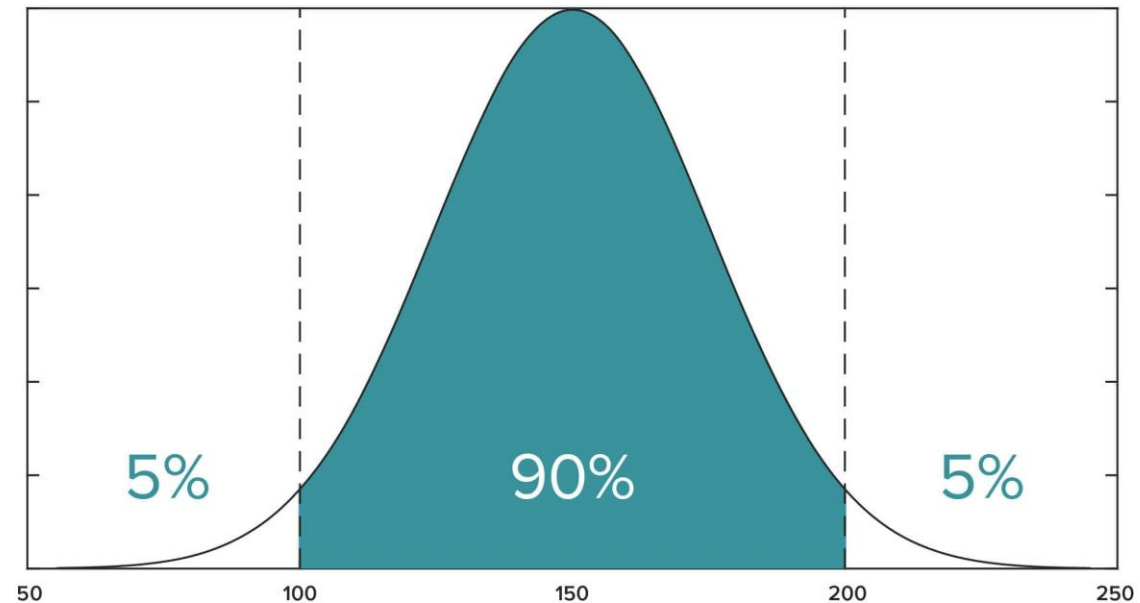
Hypothesis Testing

- Hypothesis testing involves testing a hypothesis about a population based on a sample of data.

- Example: testing the mean of a population
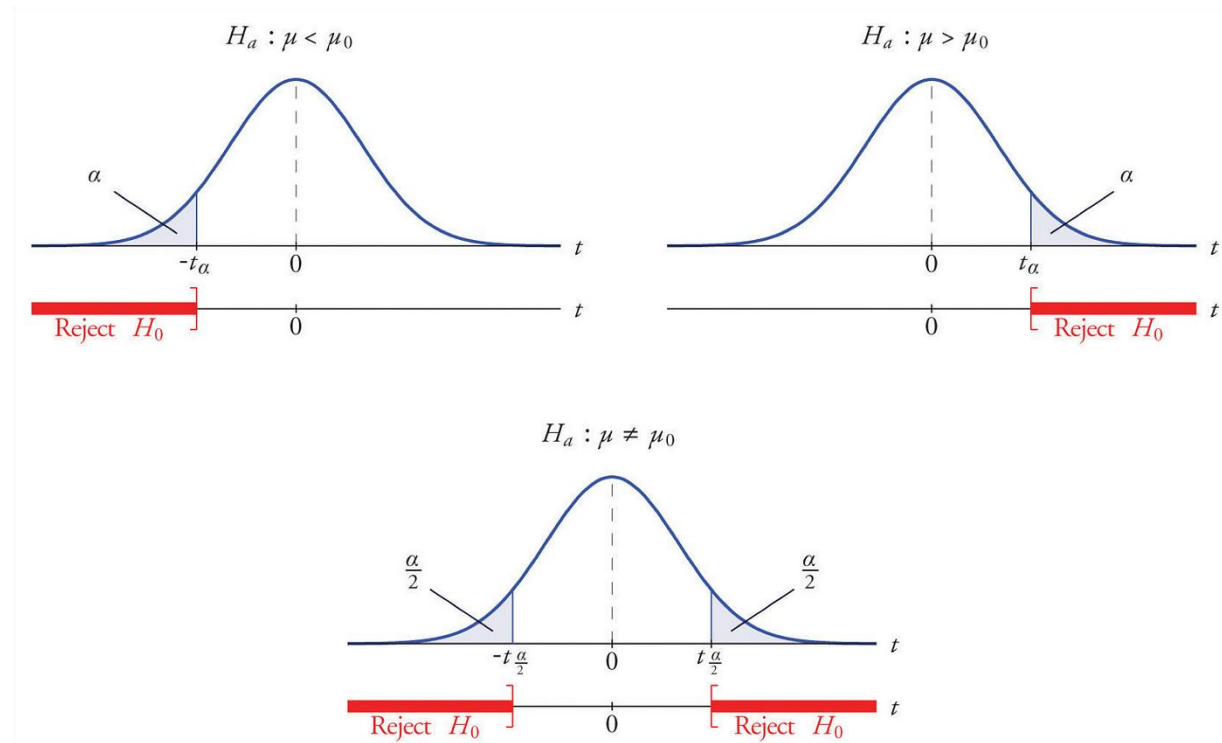
Confidence Intervals

- A confidence interval is a range of values within which a population parameter is likely to lie.

- Formula: CI = x̄ ± (Z * σ / √n)

- Example: calculating a confidence interval for a population mean

Regression Analysis

- Regression analysis involves modeling the relationship between a dependent variable and one or more independent variables.

- Example: modeling the relationship between exam scores and hours studied

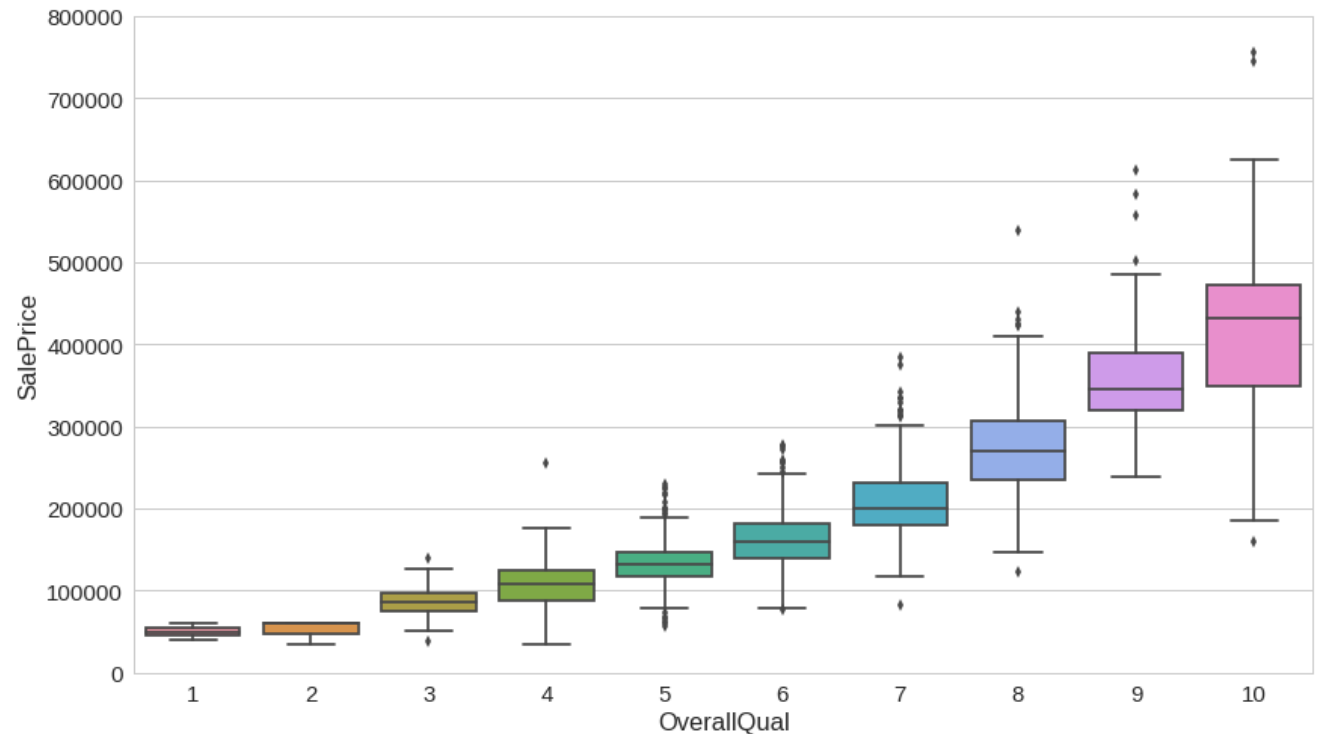Scatter Plot Showing Exam Scores Versus Period of Study
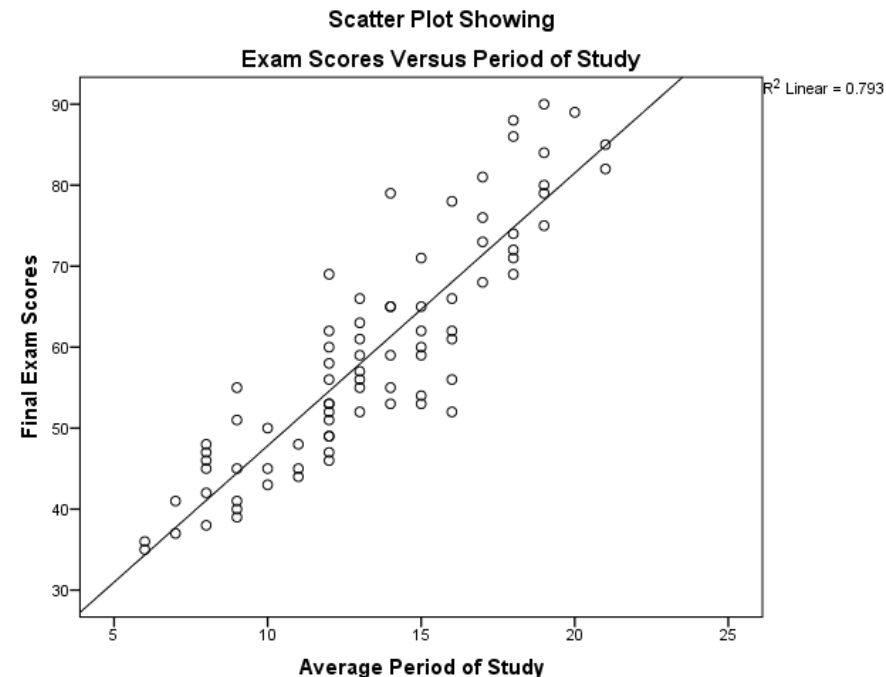
Predictive Analytics

- Forecasting the Future
- Predictive analytics involves using data to forecast what may happen in the future.
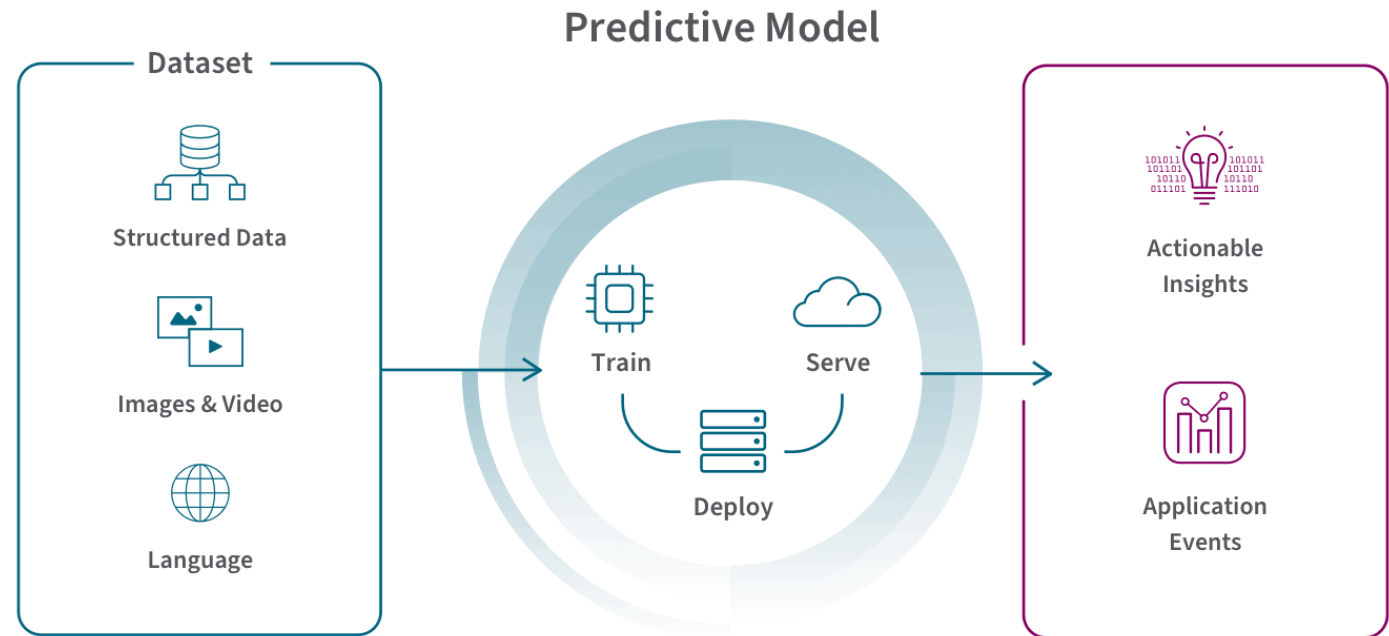
Regression Analysis

- Regression analysis can be used for predictive modeling, forecasting continuous outcomes.

- Example: predicting house prices based on features

Decision Trees

- Decision trees are a type of predictive model that use a tree-like structure to classify or predict outcomes.

- Example: predicting customer churn using a decision tree
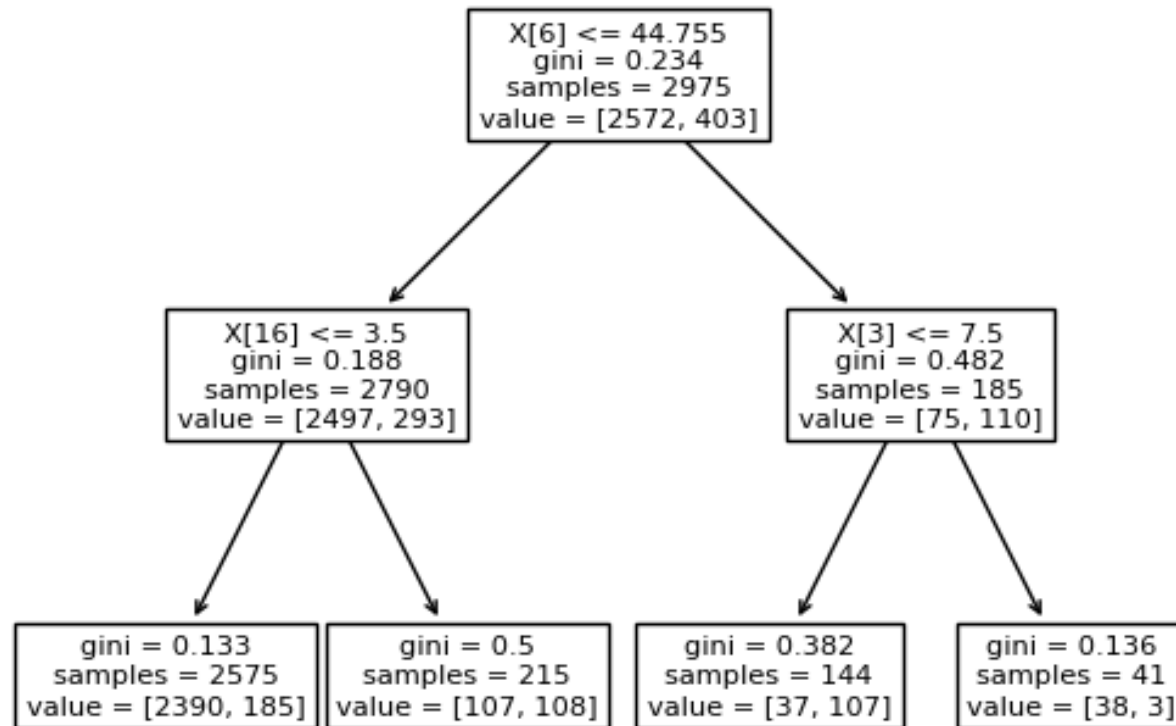
Clustering

- Clustering involves grouping similar data points into clusters to identify patterns or relationships.

- Example: clustering customers based on demographics and behavior
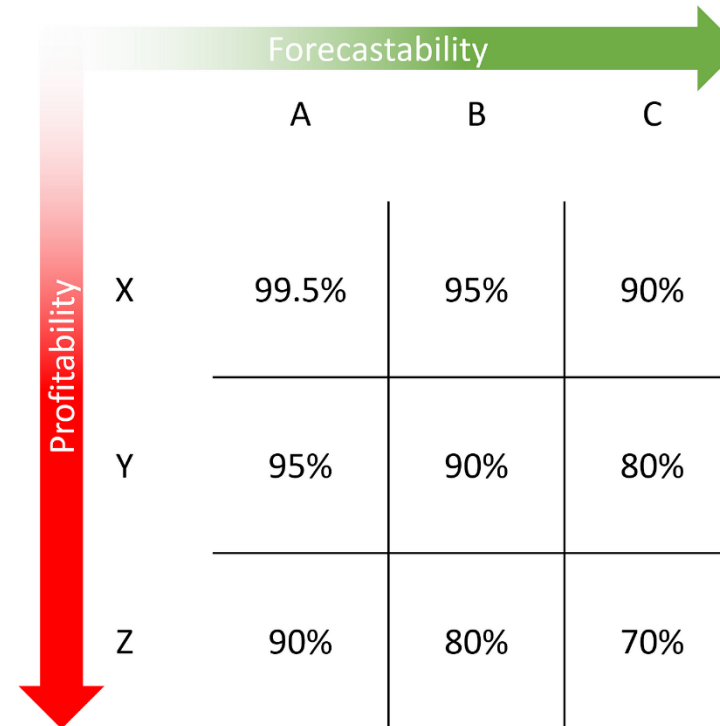
Prescriptive Analytics

- Recommending Actions
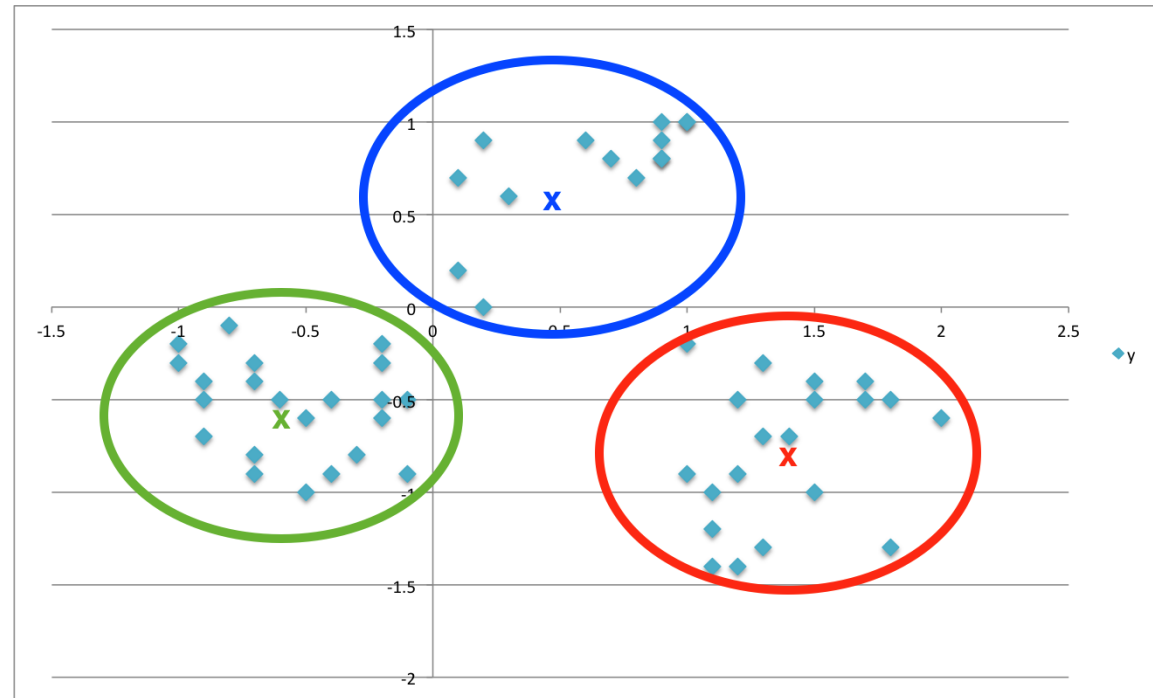- Prescriptive analytics involves using data to recommend actions or decisions.

Time Period

Future

Objectives

Optimization
Simulation
Actionable Information
for Decision Making

Prescriptive Analytics

What actions do we take?
How do we optimize?

Deliverables

Reports
Algorithms
Models & Code
Recommended Actions

Techniques

Streamlining
Simulation
Game Theory
Decision Analysis
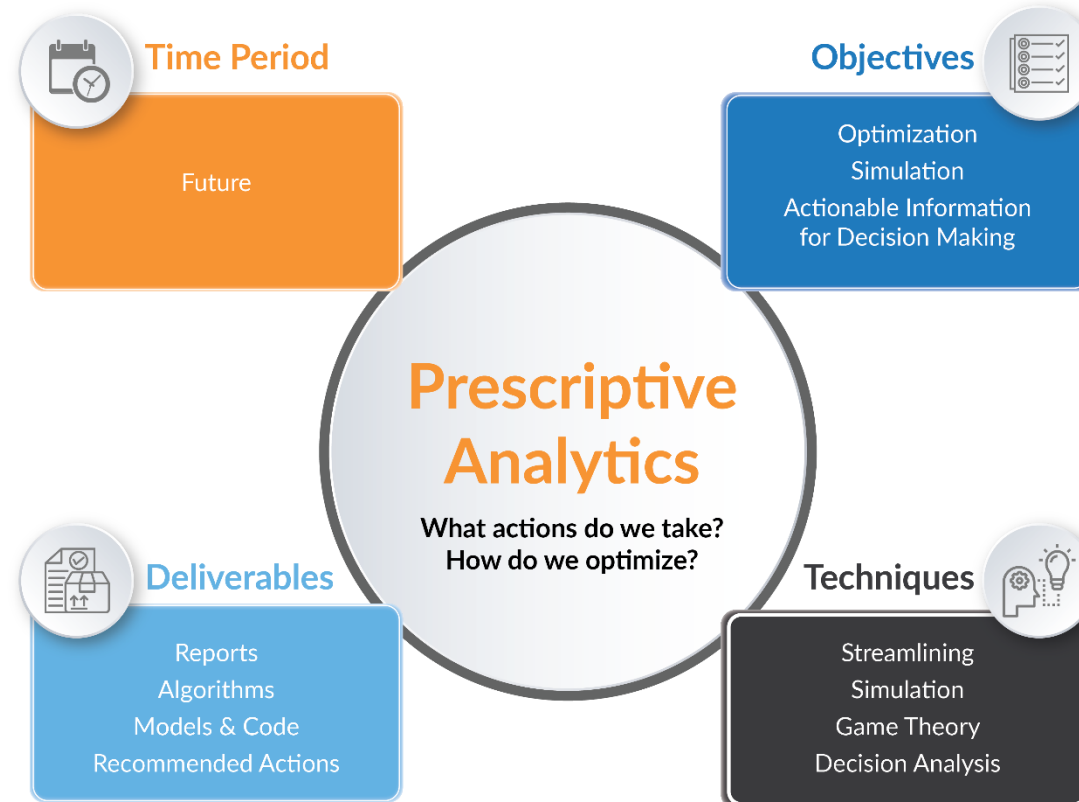
Optimization Techniques

- Optimization techniques involve finding the best solution among a set of possible solutions.

- Example: optimizing inventory levels using linear programming

	Forecastability →		
Profitability ↓	A	B	C
X	99.5%	95%	90%
Y	95%	90%	80%
Z	90%	80%	70%

Simulation Modeling

- Simulation modeling involves using data to simulate different scenarios and predict outcomes.

- Example: simulating the impact of different marketing strategies

Data Analysis Tools

Selecting the Right Tool for the Job

- Data analysis tools are essential for extracting insights from data.

Spreadsheets - Microsoft Excel

- Microsoft Excel is a popular spreadsheet software for data analysis and visualization.

- Features:
  - Data manipulation and calculation
  - Charting and graphing
  - Pivot tables and data analysis

Spreadsheets - Google Sheets

- Google Sheets is a cloud-based spreadsheet software for data analysis and collaboration.

- Features:
  - Real-time collaboration
  - Automatic saving and versioning
  - Integration with Google Drive and other Google tools

Statistical Software - R

- R is a popular programming language and environment for statistical computing and graphics.

- Features:
  - Statistical modeling and analysis
  - Data visualization and graphing
  - Large community of users and developers

Statistical Software - Python

- Python is a popular programming language with extensive libraries for statistical analysis and machine learning.

- Features:

- NumPy and Pandas for data manipulation

- Scikit-learn for machine learning

- Matplotlib and Seaborn for data visualization

Statistical Software - SAS

- SAS is a software suite for data management, predictive analytics, and business intelligence.

- Features:
  - Data manipulation and analysis
  - Statistical modeling and machine learning
  - Reporting and visualization

Statistical Software - SPSS

- SPSS is a software package for statistical analysis and data management.
- Features:
  - Data manipulation and analysis
  - Statistical modeling and machine learning
  - Reporting and visualization

Data Mining Tools - Apache Mahout

- Apache Mahout is a project of the Apache Software Foundation for building scalable machine learning algorithms.

- Features:
  - Classification and clustering
  - Recommendation systems
  - Integration with Hadoop and other big data tools

Data Mining Tools - Weka

- Weka is a collection of machine learning algorithms for data mining tasks.
- Features:
  - Classification and clustering
  - Regression and association rule mining
  - Integration with Java and other programming languages

Business Intelligence Tools - Tableau

- Tableau is a data visualization and business intelligence software for data analysis and reporting.

- Features:
  - Data connection and blending
  - Drag-and-drop interface for visualization
  - Integration with various data sources

Business Intelligence Tools - Power BI

- Power BI is a business analytics service by Microsoft for data visualization and business intelligence."

- Features:
  - Data connection and modeling
  - Interactive dashboards and reports
  - Integration with Microsoft Office and Azure

Power BI

Business Intelligence Tools - QlikView

- QlikView is a business intelligence software for data analysis and reporting.
- Features:
  - Data association and modeling
  - Interactive dashboards and reports
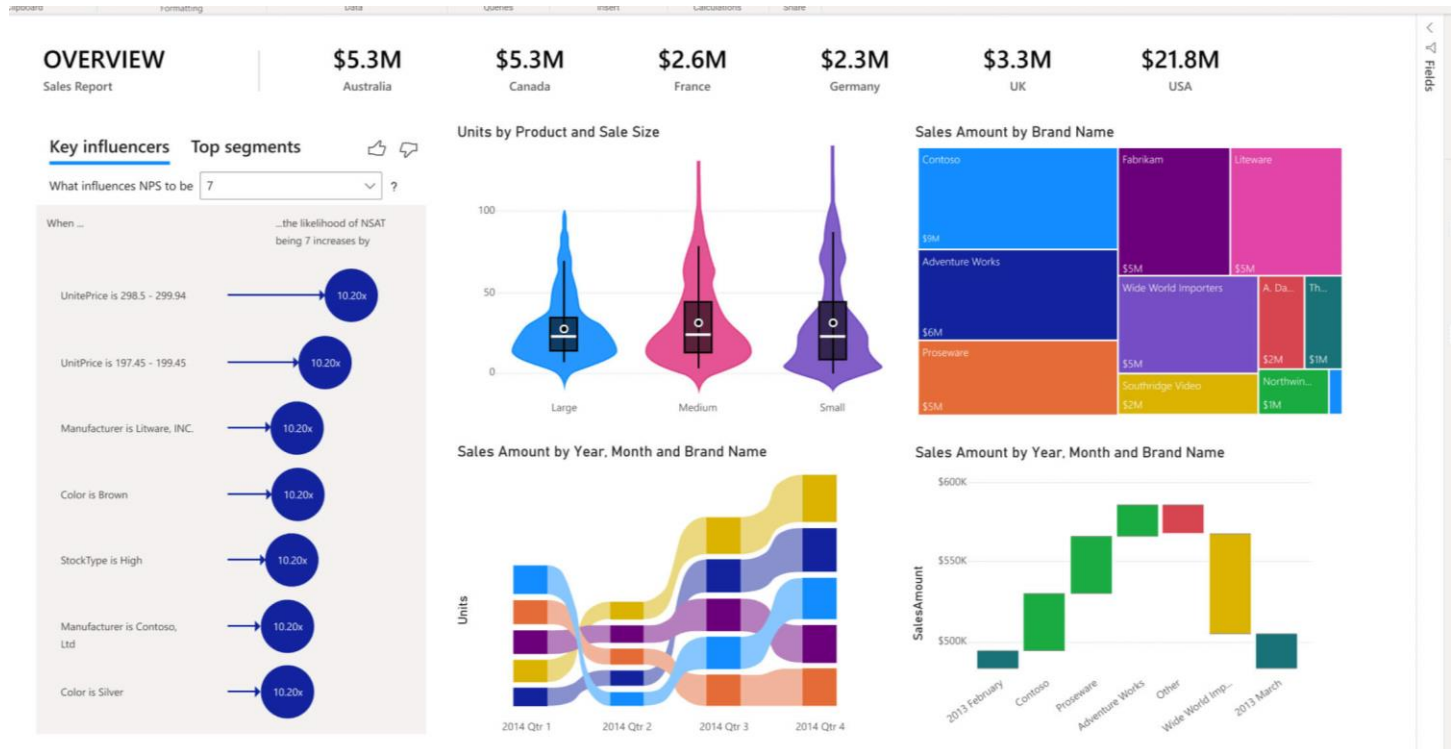 - Integration with various data sources

Communicating Data Insights Effectively

- Data visualization is the process of using visual representations to communicate data insights and trends.

Data Visualization Definition

- Using visual representations to communicate data insights and trends.
- Example: a simple bar chart or line graph

Importance of Data Visualization

- Effective communication of data insights
- Faster insights and decision-making
- Better understanding of complex data

Data Visualization Best Practices

- Choose the right chart type
- Use color effectively
- Avoid 3D charts and unnecessary embellishments
- Use interactive visualizations

Data Visualization Best Practices

- Choose the right chart type
  - Bar chart for categorical data
  - Line graph for time series data
  - Scatter plot for correlation analysis

Data Visualization Best Practices

- Use color effectively
  - Use a limited color palette
  - Avoid 3D effects and gradients
  - Use color to highlight important insights

Data Visualization Best Practices

- Avoid 3D charts and unnecessary embellishments
  - Avoid 3D pie charts and bar charts
  - Avoid unnecessary animations and transitions
  - Focus on simplicity and clarity

Data Visualization Best Practices

- Use interactive visualizations
    - Interactive dashboards and reports
    - Drill-down capabilities
    - Filtering and sorting options

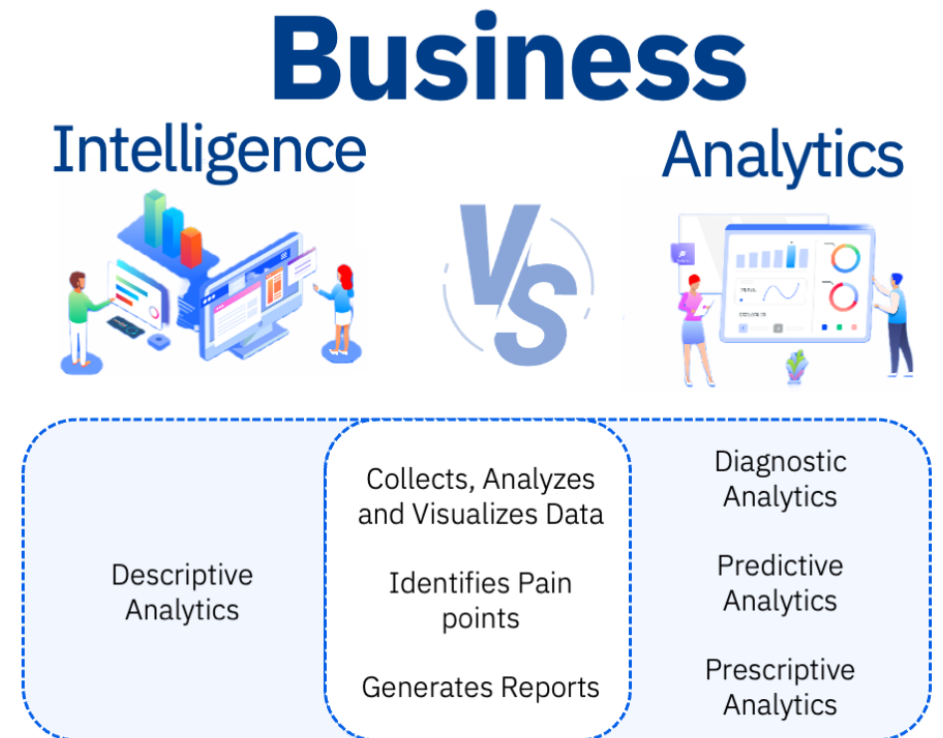# Business Intelligence and Data Dashboards

Business Intelligence

- Using Data to Support Business Decisions
- Business intelligence refers to the process of using data to support business decisions and drive organizational success.

Business Intelligence

- Using data to support business decisions and drive organizational success.

- Examples:
  - Analyzing customer behavior to inform marketing strategies
  - Identifying trends to optimize operations
  - Developing predictive models to forecast sales

Business Intelligence vs Analytics

Descriptive Analytics

Collects, Analyzes and Visualizes Data

Identifies Pain points

Generates Reports

Diagnostic Analytics

Predictive Analytics

Prescriptive Analytics

Data Dashboard

- A visual representation of key performance indicators (KPIs) that provide insights into business performance.

- Examples:
  - Sales dashboard to track revenue and customer acquisition
  - Operations dashboard to monitor production and supply chain performance
  - Marketing dashboard to track campaign performance and customer engagement

Types of Data Dashboards

- Operational dashboards: monitor and control daily operations
- Tactical dashboards: support tactical decision-making and planning
- Strategic dashboards: inform strategic decision-making and goal-setting

Operational Dashboards

- Monitor and control daily operations to ensure efficiency and effectiveness.

- Examples:
  - Production monitoring dashboard
  - Inventory management dashboard
  - Customer service dashboard

Tactical Dashboards

- Support tactical decision-making and planning to achieve short-term goals.
- Examples:
    - Sales performance dashboard
    - Marketing campaign dashboard
    - Financial performance dashboard
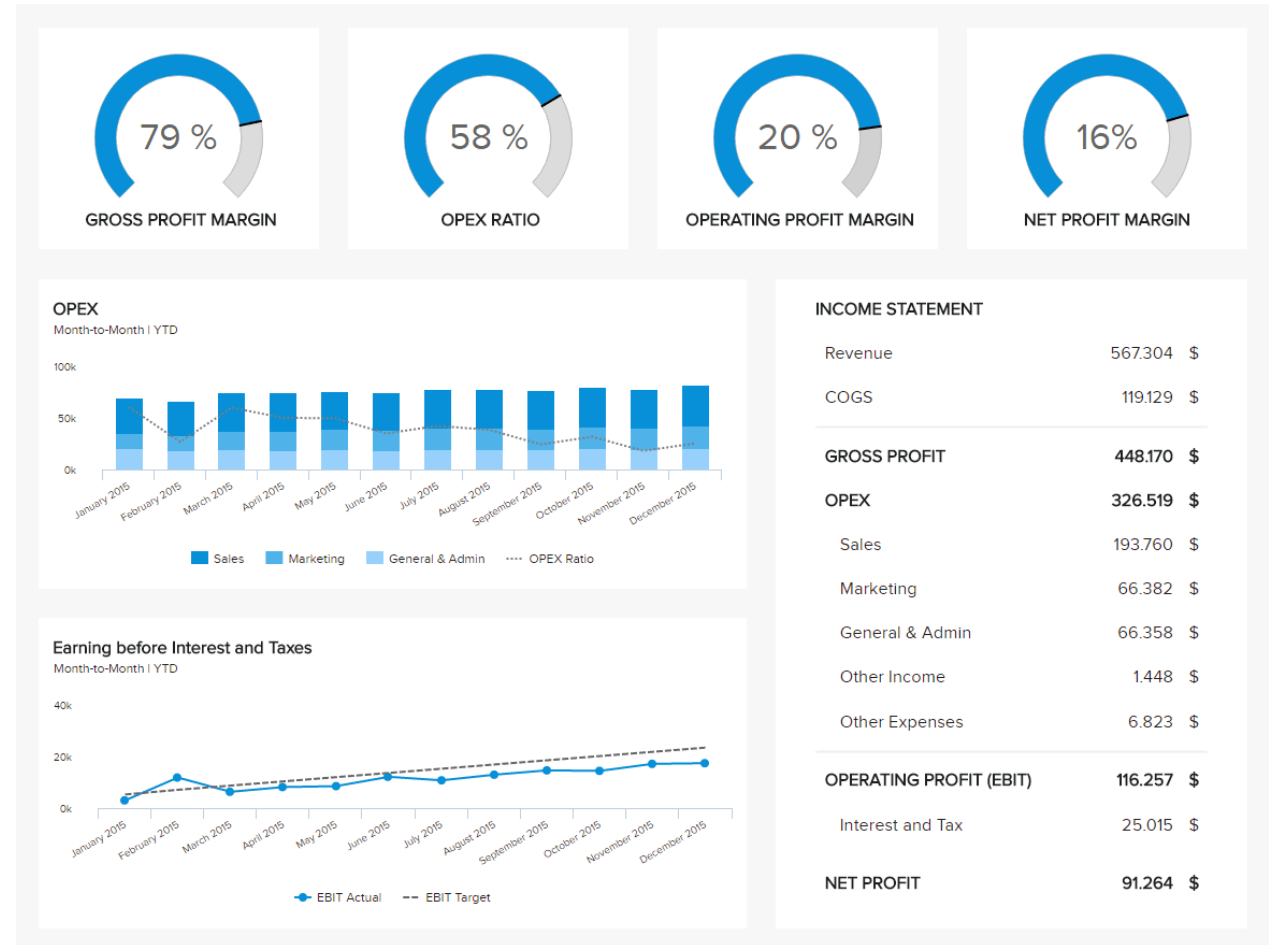
Strategic Dashboards

- Inform strategic decision-making and goal-setting to achieve long-term objectives.

- Examples:
  - Executive dashboard to track overall business performance
  - Product development dashboard to track innovation and R&D
  - Market analysis dashboard to track industry trends and competitors

Designing Effective Data Dashboards

- Identify key performance indicators (KPIs)
- Choose the right visualization tools
- Ensure data quality and integrity
- Keep it simple and intuitive

Identify Key Performance Indicators (KPIs)

- Metrics that measure business performance and progress towards goals.

- Examples:
  - Revenue growth rate
  - Customer satisfaction rate
  - Production efficiency rate
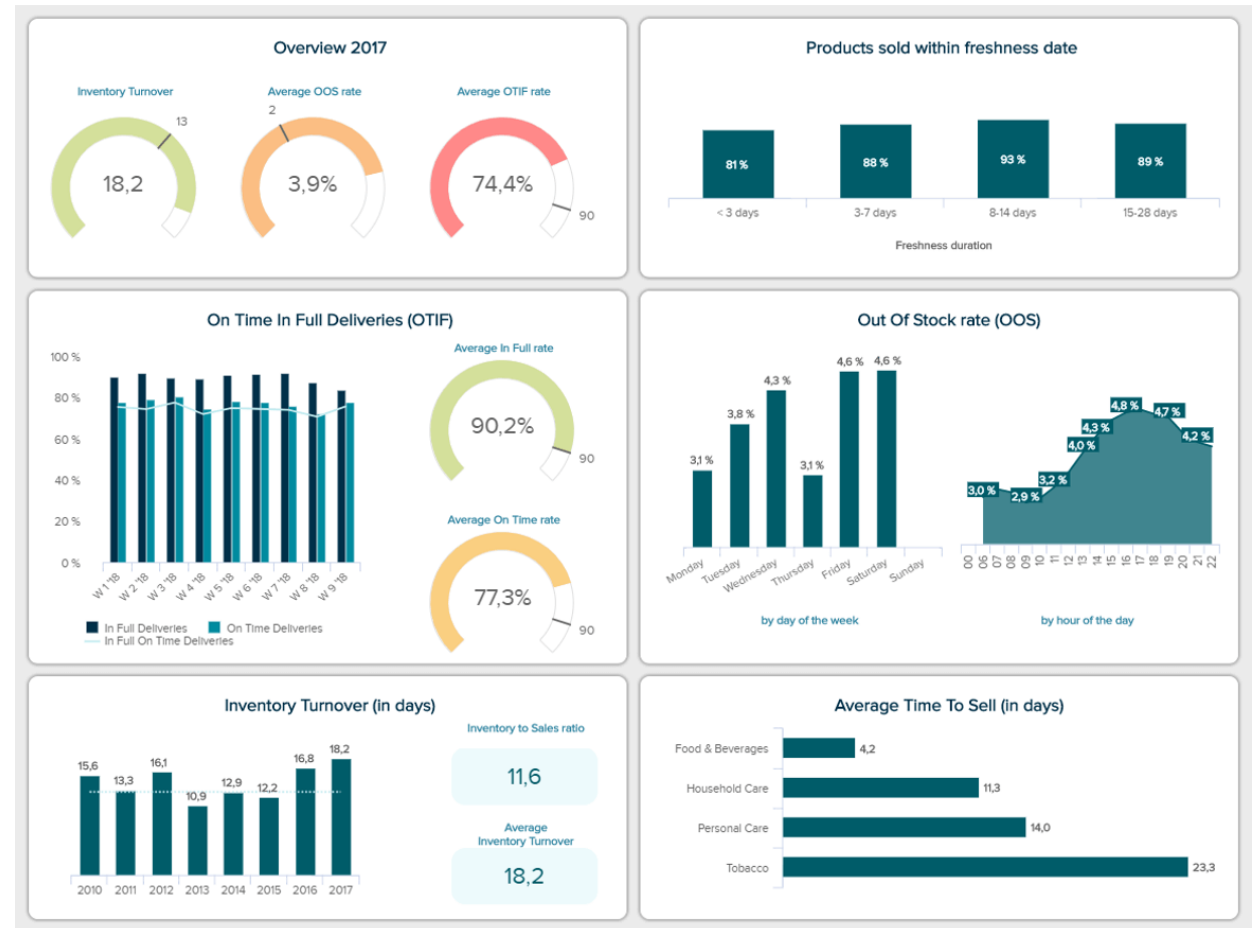
Ensure Data Quality and Integrity

- Data that is accurate, complete, and consistent."

- Examples:
  - Data validation and cleaning
  - Data normalization and transformation
  - Data backup and recovery