COURSE 1: INTRODUCTION TO DATA ANALYTICS

Module 1: What is Data Analytics

1.1 What is Data Analytics & Why It Matters

Data Analytics is the practice of **examining data to extract insights** that support better decisionmaking.

According to Forrester, businesses now see data as a **core competitive advantage**, leading to:

- High demand for skilled data analysts
- Strong salary growth
- Continuous upskilling initiatives

1.1.1 **Why Data Analytics is a Great Career**

- Supply–demand mismatch → high job opportunities
- Can be a **career** or a **stepping stone** to:
    - Data Science ○ Data Engineering ○ Business Analytics
    - Business Intelligence (BI)

1.2 Who This Course is For

This course is suitable for:

- Fresh graduates (any stream)
- Working professionals planning a career switch
- Managers & decision-makers using data
- Anyone in an analytics-enabled role

1.3 The Modern Data Ecosystem

A **data ecosystem** is a network of interconnected components that collect, process, analyze, and deliver data.

Key Components

1. **Data Sources** ○

   Structured & unstructured data

   ○ Examples:

   - Text, images, videos
   - Social media, clickstreams
   - IoT devices, sensors
   - Legacy databases
   - Real-time streaming data

2. **Data Repository (Enterprise**

   Data Environment) ○

   Raw data is:

   - Collected
   - Cleaned
   - Organized
   - Standardized ○ Must follow:
   - Security
   - Privacy laws (health, biometric, household data)
   - Master data standards

3. **Data Consumers** ○ Data

   Analysts ○ Data Scientists ○

   Business Users

   ○ Applications & APIs

4. **Enabling Technologies** ○

   Cloud Computing ○ Machine

   Learning ○ Big Data

According to Forbes, data growth is continuous and self-reinforcing.

1.4 Roles in the Data Ecosystem

Data Engineer

Role: Makes raw data usable **Responsibilities:**

- Extract, transform, load (ETL) data
- Design & maintain data pipelines
- Manage data storage systems

Skills Required:

- Programming
- Databases (SQL & NoSQL)
- System architecture

Data Analyst

Role: Converts data into insights **Responsibilities:**

- Clean and inspect data
- Identify patterns & correlations
- Perform statistical analysis
- Create dashboards & reports **Key Questions Answered:**
- Why did sales drop?
- Is user experience improving?
- What patterns exist in customer behavior?

Skills Required:

- Excel / Sheets
- SQL
- Statistics
- Visualization tools
- Programming (Python / R)
- Storytelling with data

Data Scientist

Role: Predicts future outcomes

Responsibilities:

- Build ML & AI models
- Create predictions & classifications **Key Questions Answered:**
- What will happen next quarter?
- Which customers may churn?
- Is this transaction fraudulent?

Skills Required:

- Mathematics & Statistics
- Machine Learning
- Programming
- Domain knowledge

Business Analyst & BI Analyst • Translate insights into **business actions**

- Focus on:
  - Business performance
  - Market trends
  - External influences

Simple Summary

Role	**Function**
Data Engineer	Raw → Usable Data
Data Analyst	Usable Data → Insights
Data Scientist	Insights → Predictions
BA / BI Analyst	Decisions & Strategy

1.5 What is Data Analysis?

Data Analysis is the process of:

1. Gathering data
2. Cleaning data
3. Analyzing & mining data
4. Interpreting results
5. Reporting findings

Goal: Discover patterns & correlations to support decisions

1.6 Types of Data Analytics

Descriptive Analytics — *What happened?*

- Summarizes historical data
- Example: Monthly sales report

Diagnostic Analytics — *Why did it happen?*

- Finds causes behind outcomes
- Example: Traffic drop analysis

Predictive Analytics — *What might happen?*

- Uses trends & probability
- Example: Sales forecasting

Predictions are **probabilistic**, not guaranteed.

Prescriptive Analytics — *What should we do?* •

Recommends actions

- Examples:
    - Airline price optimization ○

      Self-driving cars

1.7 Data Analysis Process (Very Important)

1. **Understand the problem**
2. **Define success metrics**
3. **Gather data**
4. **Clean data** ○ Handle missing values ○ Remove outliers
5. **Analyze & mine data**
6. **Interpret results**
7. **Present insights** ○ Dashboards ○ Reports ○ Visuals

Communication is as important as analysis.

1.8 Data Analysis vs Data Analytics Although often used interchangeably:

Term	Meaning
Analysis	Detailed examination
Analytics	Computational data analysis

In this course, **both terms mean the same thing**.

1.9 Data Analyst Responsibilities

- Data acquisition
- Writing SQL queries
- Cleaning & standardizing data
- Statistical analysis
- Pattern recognition
- Reporting & documentation

1.10 **Data Analyst Skill Set**

Technical Skills

- Excel / Google Sheets
- Visualization tools (Power BI, Tableau)
- Programming (Python, R)
- SQL
- Data Warehouses & Lakes
- Big Data tools (Hadoop, Spark)

Functional Skills

- Statistics
- Analytical thinking
- Problem-solving
- Data visualization
- Project management

Soft Skills (Key Differentiator)

- Communication
- Storytelling
- Collaboration
- Curiosity
- Intuition
- Stakeholder management

1.11 Generative AI for Data Analysts

What is Generative AI?

AI systems that **create new content** (text, images, code, music).

Examples:

- Chatbots
- Content generation
- Synthetic data creation

Key Techniques

- GANs
- VAEs
- Transformers (e.g., GPT-3, BERT)

Applications in Data Analytics

- Data augmentation
- Anomaly detection
- Forecasting & simulation
- Automated reporting
- Insight summarization

Limitations & Concerns

- Bias in training data
- Hallucinated outputs
- Lack of source transparency
- Ethical & misuse risks

Responsible AI usage is essential.

Module 2: The Data Ecosystem

2.1 What is a Data Analyst Ecosystem?

A **Data Analyst Ecosystem** includes:

- Infrastructure
- Software & tools
- Frameworks
- Languages
- Processes

used to **gather, clean, analyze, mine, and visualize data**.

2.2 Understanding Data

Definition of Data

Data is **raw, unorganized information** such as:

- Facts, observations
- Numbers, symbols
- Text, images, videos

Data becomes meaningful only **after processing and analysis**.

2.3 Types of Data (Based on Structure)

Structured Data

- Well-defined schema
- Organized into **rows & columns**
- Easy to store, query, and analyze

Examples:

- SQL databases
- OLTP systems
- Spreadsheets (Excel, Google Sheets)
- Online forms

- Sensors (GPS, RFID)
- Web & server logs

Stored mainly in **Relational Databases (RDBMS)**
Queried using **SQL**

Semi-Structured Data

- Partially organized
- Uses **tags, metadata, or hierarchy**
- No rigid schema

Examples:

- Emails
- XML
- JSON
- Zipped files
- Network packets
- Integrated multi-source data

Common formats: **XML & JSON**

Unstructured Data

- No predefined structure
- Mostly qualitative
- Cannot fit rows & columns

Examples:

- Social media posts
- Images, videos, audio
- PDFs & documents
- Web pages
- Surveys

Stored in:

- File systems

- NoSQL databases

Summary Table

Type	Structure	Examples
Structured	Fixed schema	Databases, Excel
Semi-Structured	Partial schema	XML, JSON, Emails
Unstructured	No schema	Images, Videos, social media

2.4 Data File Formats

Delimited Text Files

- Plain text
- Values separated by delimiters

Common Types:

- CSV (comma)
- TSV (tab) **Advantages:**
- Simple
- Platform-independent
- Widely supported

XLSX (Excel Open XML)

- Spreadsheet format
- Multiple worksheets
- Rows, columns & cells

Advantages:

- Open format
- Secure
- Rich functionality

XML (Extensible Markup Language)

- Tag-based hierarchical format
- Human & machine readable
- Platform-independent

Use cases:

- Data exchange
- Surveys
- Bank statements

PDF (Portable Document Format)

- Device & OS independent
- Common in legal & financial data
- Harder to extract data from

JSON (JavaScript Object Notation)

- Lightweight text format
- Widely used in APIs
- Easy to parse

Most **APIs & Web Services** return data as JSON.

2.5 Data Sources

Internal Data Sources

- Relational databases: ○ SQL Server ○ Oracle
  - MySQL ○

    IB

    M DB2

- Enterprise systems:
    - CRM ○ E RP
    - Transaction systems

External Data Sources

- Government datasets
- Financial & weather data providers
- POS & market data vendors

APIs & Web Services

- Data accessed via requests
- Return data in:
    - JSON ○ X ML ○ H TML
    - Text

Use cases:

- Sentiment analysis
- Stock price analysis
- Data validation

Web Scraping

Extracts data from web pages

Popular Tools:

- BeautifulSoup
- Scrapy
- Selenium

- Pandas

Applications:

- Price comparison
- Lead generation
- ML dataset creation

Data Streams & Feeds

- Continuous real-time data
- Time-stamped & geo-tagged

Examples:

- IoT sensors
- Stock tickers
- Social media feeds
- GPS tracking

Streaming Tools:

- Apache Kafka
- Apache Spark Streaming
- Apache Storm

2.6 Languages Used by Data Professionals

Query Languages

- Designed for databases
- Example: **SQL**

Uses:

- Retrieve data
- Insert, update, delete records
- Create tables & views

Programming Languages

Python

- Open-source
- Easy to learn
- Huge analytics ecosystem **Libraries:**
- Pandas → data analysis
- NumPy & SciPy → statistics
- Matplotlib & Seaborn → visualization
- BeautifulSoup → web scraping
- OpenCV → image processing

Ideal for beginners & big data analytics

R

- Statistical & visualization-focused
- Strong plotting libraries (ggplot2, Plotly)
- Used heavily in research & analytics

Java

- Object-oriented & platform-independent
- Used in big data frameworks:
  - Hadoo
    p ○
    Hive ○
    Spark
- Suitable for performance-critical systems

Shell & Scripting Languages

Unix/Linux Shell

- Automates repetitive tasks
- File handling
- System administration

PowerShell

- Microsoft automation tool
- Object-based
- Works well with:
  - JSON
  - CSV
  - XML
  - REST APIs

2.7 Data Repositories

What is a Data Repository?

A **central location** where data is:

- Collected
- Organized
- Stored
- Isolated for analytics & reporting

Databases

- Managed by **DBMS**
- Used for storing & querying data

Types:

1. **Relational Databases (RDBMS)** ○ Structured schema ○ SQL-based
2. **Non-Relational Databases (NoSQL)** ○ Schema-less ○ Flexible & scalable ○ Used for big data

Data Warehouses

- Centralized analytics repository
- Uses **ETL process**:
    - Extract ○ Transform ○ Load

Optimized for reporting & BI

Big Data Stores

- Distributed storage & computation
- Handle:
    - Large volume
    - High velocity
    - Variety of data

2.8 Relational Database (RDBMS)

Definition

A **Relational Database** stores data in **tables (relations)** consisting of **rows (records)** and **columns (attributes)**.
Tables are **related using common fields** (keys).

Example

- **Customer Table**: Customer ID, Name, Address, Phone
- **Transaction Table**: Transaction Date, Customer ID, Amount
- Relationship created using **Customer ID**

Key Characteristics

- Data stored in **structured format**
- Uses **SQL (Structured Query Language)**
- Relationships between tables via **primary & foreign keys**
- Fixed **schema**

Advantages

- **Reduced redundancy** – data stored once and referenced
- **Data integrity & consistency** using constraints
- **ACID compliance** ○ Atomicity ○ Consistency ○ Isolation
  - Durability
- **High security & controlled access**
- **Fast querying** on millions of records
- Easy **backup & disaster recovery**

Popular Relational Databases

- IBM DB2
- Microsoft SQL Server
- MySQL
- Oracle Database
- PostgreSQL

Cloud Relational Databases (DBaaS)

- Amazon RDS
- Google Cloud SQL
- SQL Azure

Use Cases

- **OLTP (Online Transaction Processing)**

- **Data Warehouses (OLAP)**
- **IoT systems (lightweight structured data)**

Limitations

- Poor support for **unstructured & semi-structured data**
- Schema must match during migration
- Field length limitations

2.9 NoSQL Databases

Definition

NoSQL (Not Only SQL) databases are **non-relational**, **schema-flexible**, and designed for **scalability, performance, and big data**.

Key Features

- Schema-less / flexible schema
- Supports **structured, semi-structured & unstructured data**
- Distributed & cloud-friendly
- Not primarily SQL-based

Types of NoSQL Databases

Key-Value Stores

- Data stored as **key : value**
- Extremely fast

Use Cases

- Session management
- Caching
- User preferences

Examples

- Redis
- Memcached
- Amazon DynamoDB

Document-Based Databases

- Data stored as **JSON-like documents**
- Flexible indexing & queries

Use Cases

- eCommerce
- Medical records
- CRM systems

Examples

- MongoDB
- CouchDB
- Cloudant

Column-Based Databases

- Stores data by **columns instead of rows**
- Fast read/write for large datasets

Use Cases

- Time-series data
- IoT & weather data

Examples

- Apache Cassandra
- Apache HBase

Graph Databases

- Uses **nodes & relationships**
- Best for connected data

Use Cases

- Social networks
- Fraud detection
- Recommendation systems

Examples

- Neo4j
- Cosmos DB

Advantages of NoSQL

- Handles **huge data volumes**
- Horizontally scalable
- Cost-effective
- Cloud & distributed ready
- High availability

2.10 RDBMS vs NoSQL (Exam Favorite)

Feature	**RDBMS**	**NoSQL**
Schema	Fixed	Flexible
Data Type	Structured	All types
Scalability	Vertical	Horizontal
ACID	Supported	Limited
Cost	High	Lower
Maturity	Very mature	Relatively newer

2.11 Data Warehouses, Marts & Lakes

Data Warehouse

- Central repository
- **Cleaned, structured, historical data**
- Used for **reporting & BI**
- Single source of truth

Data Mart

- Subset of data warehouse
- Business-specific (Sales, Finance)
- Faster performance
- Isolated security

Data Lake

- Stores **raw data** in native format
- Supports **structured, semi-structured, unstructured**
- Used for **advanced & predictive analytics**
- Often staging area for data warehouse

2.12 ETL Process (Extract – Transform – Load)

Extract

- Collects data from sources • Batch tools: Stitch, Blendo
- Streaming tools: ○ Apache Kafka ○ Apache Storm **Transform**
- Data cleaning
- Standardization
- Validation
- Enrichment
- Business rules

Load

- Initial load
- Incremental load
- Full refresh
- Load verification & recovery

2.13 Data Pipeline

- Broader concept than ETL
- Moves data from **source → destination**
- Supports **batch + streaming**
- Used for real-time systems

Tools

- Apache Beam
- Google Dataflow

2.14 Big Data

Definition

Big Data refers to **large, fast, and diverse data** that cannot be processed using traditional tools.

5 V's of Big Data

- **Velocity** – speed of data generation
- **Volume** – amount of data
- **Variety** – data types
- **Veracity** – data quality
- **Value** – usefulness

2.15 Big Data Technologies

Apache Hadoop

- Distributed storage & processing
- Runs on clusters
- Cost-effective

HDFS (Hadoop Distributed File System)

- Splits files into blocks
- Replicates data
- Fault-tolerant
- Data locality

Apache Hive

- SQL-like querying on Hadoop
- High latency
- Best for ETL & analytics
- Not for transactions

Apache Spark

- In-memory processing
- Real-time analytics
- Supports Java, Python, Scala, SQL
- Streaming + ML + ETL

Module 3: Gathering and Wrangling Data

3.1 Identifying Data Requirements

Starting Point

At this stage, you already know:

- **Where you are** (current state)
- **Where you want to be** (desired outcome)
- **What will be measured** and **how it will be measured**

The next step is **identifying the right data** for your use case.

How to Identify Data

Begin by determining:

1. **What information is needed**
2. **Where the data can come from**

These decisions are driven entirely by **your goal**.

Example: Targeted Marketing Campaign

Goal:
Identify the age group that buys products most and design targeted campaigns.

Data Identified:

- Customer profile
- Purchase history
- Location
- Age, education, profession
- Income, marital status
- Customer complaints
- Customer service survey ratings
- Social media data (likes, shares, comments)

Insight: Supplementing core data with **complaints and social media sentiment** helps improve recommendations and advocacy.

3.2 Data Collection Planning

Why a Plan Is Needed

A data collection plan ensures **clarity for execution**.

Key Planning Decisions

- **Timeframe**
    - Real-time (e.g., website visitors)
    - Fixed duration (e.g., event-based data)
- **Volume**
    - Entire population (e.g., age group 21–30)
    - Sample size (e.g., 100,000 users)
- **Dependencies & Risks**
- **Mitigation strategies**

3.3 Data Collection Methods

Purpose

Define **how** data will be collected from identified sources.

Factors Affecting Method Choice

- Type of data
- Volume of data
- Time sensitivity
- Source systems

Once finalized:

- Implement the strategy
- Continuously update the plan as conditions evolve

3.4 Data Quality, Governance & Privacy

Data Quality

Reliable data must be:

- Accurate
- Error-free
- Complete
- Relevant • Accessible ✓ Define:
- Quality traits
- Metrics
- Validation checkpoints

Data Governance

Concerns:

- Security
- Compliance
- Regulations

Poor governance can result in:

- Legal penalties
- Loss of credibility
- Invalid insights

Data Privacy Ensure:

- Confidentiality
- License to use
- Regulatory compliance
- Audit trails

⚠ Loss of trust in data compromises the entire analysis process.

3.5 Types of Data Sources

Based on Ownership

- **Internal**
- **External**

Based on Origin

Type	Description
Primary Data	Collected directly by you
Secondary Data	Collected by others, reused
Third-Party Data	Purchased from aggregators

Examples

Primary Data

- Internal databases (CRM, HR, workflows)
- Surveys
- Interviews
- Observations
- Focus groups

Secondary Data

- Research papers
- Public records
- External databases
- Published surveys

Third-Party Data

- Purchased datasets from aggregators

3.6 Common Data Sources

Databases

- Internal business systems
- External subscription databases
- Cloud databases (real-time insights)

Web & social media

- Public websites
- Government portals
- Platforms like Facebook, Twitter, YouTube, Instagram

Used for:

- Customer sentiment
- Engagement analysis
- Market trends

Sensor & IoT Data

- Wearables
- Smart cities
- Medical devices
- Smartphones
- Home appliances

Data Exchange Platforms Enable

secure, governed data sharing.

Examples:

- AWS Data Exchange
- Crunchbase
- Lotame
- Snowflake

Surveys, Interviews & Observation

- Surveys → Quantitative insights
- Interviews → Qualitative insights
- Observation → Behavioral insights

These can be **primary, secondary, or third-party** data.

3.7 Tools & Methods for Gathering Data

SQL

- Used to query relational databases
- Supports filtering, grouping, ordering, and limiting data

Non-Relational Query Tools

- CQL for Apache Cassandra
- GraphQL for Neo4j

APIs (Application Programming Interfaces) •

Access data from databases, services, marketplaces

- Used for:
    - Data extraction
    - Data validation (e.g., address verification)

Web Scraping

- Extracts data from web pages
- Used for text, images, videos, product data

RSS Feeds

- Continuous data updates from forums and news sites

Data Streams

- Real-time data from:
    - IoT devices ○ Applications
    - GPS
    - Social media platforms

3.8 Importing Data into Repositories

Why Import?

To:

- Combine data from multiple sources
- Enable querying, wrangling, and analysis
- Provide a unified interface

Repository vs Data Type

Data Type	Storage Option
Structured	Relational DB, NoSQL
Semi-structured	NoSQL
Unstructured	NoSQL, Data Lake

Data Formats

- XML
- JSON (preferred for web services)

3.9 Tools for Importing Data

- ETL tools:
    - Talend
    - Informatica
- Programming languages:
    - Python
    - R

3.10 What is Data Wrangling?

Definition

Data Wrangling (Data Munging) is an **iterative process** that involves:

- Data exploration
- Data transformation
- Data validation
- Making data available for analysis

Its purpose is to convert **raw data** (collated from multiple sources and stored in repositories) into **analysis-ready data** that produces **credible and meaningful insights**.

3.11 Phases of Data Wrangling (4-Step Process)

1. Discovery (Exploration)

- Understand the data in relation to the **use case**
- Identify:
    - Structure
    - Quality issues
    - Missing fields
    - Mapping requirements
- Decide **how to clean, organize, and transform** the data

2. Transformation (Core Phase) This phase forms the **bulk of data wrangling**.

a) Structuring

- Changes the **form or schema** of data
- Required when data comes from **different sources** (DBs, APIs, files)

Common Techniques

- **Joins** → Combine **columns**
- **Unions** → Combine **rows**

Exam Tip:

Join = columns | Union = rows

b) Normalization

- Removes redundancy
- Improves consistency
- Common in **transactional systems (OLTP)**

c) Denormalization

- Combines multiple tables
- Improves query performance
- Used in **reporting & analytics (OLAP)**

d) Data Cleaning

Fixes data issues that affect analysis accuracy.

Common Issues

- Missing values
- Duplicate records
- Inaccurate or biased data
- Null values
- Outliers

Handling Missing Data

- Remove records
- Source missing data
- Imputation (statistical estimation)

e) Data Enrichment

- Adds **additional data points** to increase insight value **Examples**
- Adding business performance data from public sources
- Sentiment score from customer feedback

- Geo-weather data for occupancy analysis
- Metadata (timestamps, tags, scores)

3. Validation

- Verifies **quality, consistency, and security**
- Uses validation rules and checks
- Ensures data meets defined standards after transformation

4. Publishing

- Delivers:
    - Transformed & validated dataset
    - Metadata and documentation
- Used for **downstream analytics, reporting, ML models**

Important:
All steps must be **documented** for repeatability and auditability.

3.12 Iterative Nature of Data Wrangling

- All phases are **iterative**
- Changes and decisions must be:
    - Reproducible
    - Documented
    - Justified

3.13 Data Wrangling Tools & Software

Spreadsheets

- Microsoft Excel
- Google Sheets
- Power Query (Excel)
- Query functions (Sheets)

✓ Manual wrangling
✓ Easy to use
Not scalable for big data

OpenRefine

- Open-source
- Menu-based operations
- Supports CSV, TSV, XLS, XML, JSON
- Easy to learn

Google DataPrep

- Cloud-based
- Auto-detects schema & anomalies
- Suggests next steps
- No infrastructure management

IBM Watson Studio Refinery

- Cleans & transforms large datasets
- Automatic governance enforcement
- Supports multiple data sources

Trifacta Wrangler

- Interactive & cloud-based
- Collaboration support
- Exports to Excel, Tableau, R

Python-Based Tools

- Jupyter Notebook
- NumPy
- Pandas

✓ Fast
✓ Scalable
✓ Automation friendly

R-Based Tools

- Dplyr
- Data.table
- Jsonlite

3.14 Data Quality & Importance

According to a report by **Gartner**:

Poor data quality weakens business competitiveness and leads to faulty decisions.

Common Data Quality Issues

- Missing values
- Inconsistent records
- Duplicates
- Syntax errors
- Outliers
- Incorrect delimiters

3.15 Data Cleaning vs Data Wrangling

Aspect	Data Cleaning	Data Wrangling
Scope	Subset	End-to-end process
Includes	Fixing errors	Cleaning + Transforming + Validating
Phase	Transformation	Entire workflow

3.16 Data Cleaning Workflow

1. Inspection

- Data profiling
- Rule-based checks
- Visualization for outliers

2. Cleaning

Techniques

- Handle missing values
- Remove duplicates
- Remove irrelevant data
- Convert data types
- Standardize formats
- Fix syntax errors
- Handle outliers

3. Verification

- Re-check constraints
- Validate accuracy
- Ensure data health

All changes and reasons **must be documented**.

Module 4: Mining and Visualizing Data and Communicating Results

4.1 What is Statistics?

Statistics is a branch of mathematics that deals with:

- **Collection**
- **Analysis**
- **Interpretation**
- **Presentation** of **numerical (quantitative) data**.

Real-life Examples

- Average income
- Average age
- Highest-paid professions
- Business decisions, healthcare research, customer behavior analysis

Importance

Statistics helps organizations make **data-driven decisions** instead of assumptions.

4.2 What is Statistical Analysis?

Statistical Analysis is the application of statistical methods to a **sample of data** to understand what the data represents.

Key Terms

- **Population**: Entire group with a common characteristic *Example*: All licensed drivers in a state
- **Sample**: Subset of the population *Example*: Male drivers over age 50

Purpose

- Ensure relationships in data are **meaningful**
- Avoid conclusions based on **random chance**

4.3 Types of Statistical Analysis

A. Descriptive Statistics

Used to **summarize and describe data**

No predictions or generalizations

Common Measures:

1. Central Tendency

Shows where most values lie

- **Mean**: Average
- **Median**: Middle value (not affected by outliers)
- **Mode**: Most frequent value

Median is preferred when outliers exist

2. Dispersion

Shows **variability** in data

- **Range**: Max – Min
- **Variance**: Spread of data points
- **Standard Deviation**: How close values are to the mean

3. Skewness

- Measures **asymmetry** of data distribution
- Can be **left-skewed** or **right-skewed**
- Affects which statistical tests are valid

Other descriptive tools:

- Tables
- Charts
- Graphs
- Correlation
- Scatter plots

B. Inferential Statistics

Used to **draw conclusions about a population** based on a sample

Common Techniques:

1. **Hypothesis Testing**
    - Tests whether observed results are statistically significant

- Example: Vaccine effectiveness

2. **Confidence Intervals**
   - Provides a **range** where the true population value lies
3. **Regression Analysis** ○ Determines relationships between variables ○ Helps predict outcomes

---

4.4 Statistics & Data Mining Relationship

Statistics forms the **foundation of Data Mining** by:

- Providing analytical methods
- Separating **noise from meaningful patterns**
- Supporting better decision-making

4.5 What is Data Mining?

Data Mining is the process of **extracting useful knowledge from data**.

Goals

- Identify **patterns**
- Discover **relationships**
- Detect **trends**
- Predict **future outcomes**

Pattern vs Trend

- **Pattern**: Repeating regularities in data *Example*: Login behavior of users
- **Trend**: Direction of change over time
  Example: Global warming

4.6 Applications of Data Mining

- Customer behavior analysis
- Fraud detection
- Healthcare predictions
- Student performance prediction

- Crime prevention
- Demand forecasting

4.7 Data Mining Techniques

Common Techniques

- **Classification**: Assigning categories *Example*: Low / Medium / High spenders
- **Clustering**: Grouping similar data
- **Anomaly Detection**: Identifying unusual behavior
- **Association Rule Mining**: Finding relationships *Example*: Laptop → Cooling pad
- **Sequential Pattern Mining**: Order of events
- **Affinity Grouping**: Cross-selling & up-selling
- **Decision Trees**: Tree-based classification
- **Regression**: Predicting continuous values

Key objective: Focus on relevant information only

4.8 Tools for Statistical Analysis & Data Mining

Spreadsheets

- Microsoft Excel, Google Sheets
- Pivot tables, comparisons
- Add-ons for mining tasks

Programming & Analytics Tools

- **R**
    - Statistical modeling, mining packages
- **Python**
    - Libraries: Pandas, NumPy
    - Jupyter Notebook support

Enterprise Tools

- **IBM SPSS Statistics**
    - Advanced analytics, minimal coding

- **IBM Watson Studio**
    - Collaboration, ML & AI models
- **SAS**
    - Enterprise-grade data mining & statistics

Tool selection depends on:

- Data size
- Features
- Visualization
- Infrastructure
- Ease of use

4.9 **Communicating & Sharing Data Analysis Findings**

Importance of Communication in Data Analysis

- The **data analysis lifecycle** starts with understanding the problem and **ends with communicating insights** that drive decisions.
- Data projects are **collaborative**, involving multiple business functions.
- Insights create value **only when stakeholders understand and trust them**.

Role of a Data Analyst

- Tell a **story with data**
- Convert analysis into **actionable insights**
- Use **clear visuals + structured narrative**

Understanding Your Audience

Before creating any presentation, ask:

- **Who is my audience?**
- **What matters to them?**
- **What will make them trust my findings?**

Audience Characteristics

- Different business functions
- Strategic vs operational roles
- Varying levels of domain knowledge

- Different impact levels from the problem

Key Rule: A presentation is **not a data dump**

Structuring an Effective Data Story

Best Practices

- Start by **clearly stating the business problem**
- Explain the **desired outcome**
- Speak the **business language**, not technical jargon
- Share only **relevant data**

Build Credibility

- Share:
    - Data sources ○ Assumptions ○ Hypotheses
    - Validation steps
- Avoid hiding assumptions
- Treat data as a **"black box"** for the audience—open it carefully **Organizing the Narrative**
- Group information logically: ○ Qualitative vs Quantitative
- Choose one approach:
    - **Top-down** (insights → data)
    - **Bottom-up** (data → insights)
- Be **consistent**

Choose the Right Format

- Executive summary
- Fact sheet
- Detailed report
  (Depends on how the audience will use the information)

Role of Data Visualization

Why Visualization Matters

- Visuals create **clarity and impact**

- A single chart can replace **thousands of words**
- Helps reveal:
    - Patterns ○ Trends ○ Relationships
    - Anomalies

Data has value through the stories it tells

4.10 **Introduction to Data Visualization**

What is Data Visualization?

Data Visualization is the discipline of communicating information using:

- Graphs
- Charts
- Maps
- Diagrams

Goals

- Easy comprehension
- Better interpretation
- Improved retention

Choosing the Right Visualization Ask yourself:

- Am I comparing parts of a whole?
- Am I comparing multiple values?
- Am I showing change over time?
- Am I showing correlation?
- Am I detecting anomalies?

Every visualization must answer a **specific question**

Common Types of Charts

Bar Chart

- Compare related datasets

- Compare parts of a whole

Column Chart

- Compare values side-by-side
- Ideal for showing change over time
- Handles negative & positive values better

Pie Chart

- Shows proportion of sub-parts
- Total = 100%

Line Chart

- Shows trends over time
- Best for continuous variables

4.11 **Dashboards**

What is a Dashboard?

A **dashboard** is a single interface displaying:

- Multiple visualizations
- Data from multiple sources

Benefits

- Real-time monitoring
- Bird’s-eye view + drill-down
- Easy collaboration
- Quick decision-making

Use Cases

- Marketing performance
- Sales conversions
- Operational health monitoring

4.12 **Visualization & Dashboarding Tools**

Spreadsheets

- Microsoft Excel
- Google Sheets
- Easy to learn
- Auto-updating charts
- Best for quick analysis & collaboration

Python & Jupyter

- **Jupyter Notebook**
- Python libraries:
    - Matplotlib – basic & advanced plots
    - Bokeh – interactive visualizations ○ Dash – interactive web apps

R Tools

- **RStudio**
- **Shiny**
- Used for:
    - Statistical visuals ○ Interactive dashboards ○ Web-based applications

Enterprise & BI Tools

- **IBM Cognos Analytics** ○ Forecasting ○ Conditional formatting ○ Geospatial analytics
- **Tableau** ○ Drag-and-drop dashboards ○ Storytelling with data ○ Integrates Python & R
- **Microsoft Power BI** ○ Cloud-based

- Interactive dashboards

Mobile & collaboration

support

Module 5: Career Opportunities and Data Analysis in Action

5.1 Career Opportunities in Data Analysis

1. Demand for Data Analysts

- Data analyst roles exist across **all industries**:
    - Banking & Finance ○ Insurance ○ Healthcare ○ Retail ○ Information Technology
    - Government & Academia
- Demand for skilled data analysts **far exceeds supply**
- Companies are willing to **pay a premium** for skilled professionals
- Big data analytics market is growing rapidly → long-term career stability

2. Classification of Data Analyst Roles A. Data Analyst Specialist

Roles Focused on **technical & analytical growth**

Career progression path:

- Junior / Associate Data Analyst
- Data Analyst
- Senior Data Analyst
- Lead Analyst
- Principal Analyst

Key characteristics:

- Continuous improvement in:
    - Technical skills ○ Statistical skills ○ Analytical thinking

Exposure depends on: ○

Organization

size ○ Team

size ○ Industry

Small teams:

- End-to-end exposure: ○ Data collection ○ Cleaning ○ Analysis ○ Visualization ○ Presentation

Large organizations:

- Roles divided by function
- Deep expertise in one phase at a time

Senior responsibilities (Lead/Principal):

- Define team processes
- Recommend tools & platforms
- Upskill team members
- Team expansion & mentoring

B. Domain Specialist (Functional Analyst) Roles

Focused on **industry/domain expertise**

Examples:

- Healthcare Analyst
- Sales Analyst
- Marketing Analyst
- Finance Analyst
- Social Media Analyst **Key points:**
- Strong domain knowledge

May not be highly technical

- Seen as **subject matter experts**
- Titles reflect domain specialization

C. Analytics-Enabled Job Roles Roles where analytics **enhances performance**

Examples:

- Project Managers
- Marketing Managers
- HR Managers

Why important?

- Data-driven decision making
- Higher efficiency & effectiveness
- Many openings fall into this category

D. Transition to Other Data Professions

With additional upskilling, data analysts can move into:

- Data Engineer
- Big Data Engineer
- Data Scientist
- Business Analyst
- Business Intelligence Analyst

Example paths:

- Interest in data lakes & big data → Big Data Engineer
- Interest in business decisions → Business Analytics / BI

3. Key Skills Required for Growth

- Technical: SQL, Python, querying tools, data repositories
- Visualization tools
- Communication & presentation skills

  Stakeholder & project management skills
- Ability to work with multiple tools & platforms over time

5.2 The Many Paths to Data Analysis

1. Academic Path

Degrees that give a strong foundation:

- Data Analytics
- Statistics
- Computer Science
- Management Information Systems
- IT Management

2. Online Learning Path

Multi-course specializations offered by:

- Coursera
- edX
- Udacity

Benefits:

- Designed by industry experts
- Hands-on projects & assignments
- Portfolio-ready projects
- Suitable even without a formal degree

3. Mid-Career Transition Path

From Non-Technical Roles

- Explore **Domain Specialist** roles
- Example:
    - Sales → Sales Analyst
- Use existing industry knowledge

  Upskill in:
    - Statistics
    - Programming

- Analytics tools

From Technical Roles

- Faster transition
- Easier tool adoption
- Existing domain understanding is an advantage

4. Transferable Skills

Many skills are already used in other jobs:

- Problem-solving
- Communication
- Project management
- Storytelling

These can be enhanced through:

- Online courses
- Trainings
- Communities & forums

Key takeaway:
Formal qualifications are helpful—but **curiosity, learning mindset, and consistency matter more**.

5.3 **Generative AI for Data Analytics**

1. What is Generative AI?

- A type of AI that **creates new synthetic data**
- Unlike traditional AI (predict/classify), it **generates new content**

2. Applications of Generative AI in Data Analytics

A. Synthetic Data Generation

- Creates artificial datasets • Useful when real data is limited
- Helps in: ○ Testing ○ Model training
    - Data augmentation

B. Handling Missing Data

- Fills missing values
- Provides a more complete dataset
- Improves analysis quality

C. Data Transformation

- Converts data formats:
    - Text ↔ Images
- Enables creative data representation

D. Data Preparation & Cleaning

- Automates: ○ Cleaning ○ Normalization ○ Transformation
- Speeds up journey from raw data to insights

E. Querying & Q&A Systems

- Helps generate complex queries
- Enables **natural language interaction Examples:**
- OpenAI GPT – powerful Q&A and language tasks
- BERT – strong contextual understanding

F. Data Visualization & Dashboards

- Creates adaptive & interactive visualizations
- Enhances aesthetics and clarity

- Personalized dashboards

Examples:

- Tableau AI
- IBM Cognos
- Looker AI

G. Data Storytelling

- Generates narratives from data
- Highlights key insights
- Converts raw data into compelling stories

3. Key Takeaway

Generative AI is:

- Not just a tool
- A **catalyst for innovation**
- Reshaping how data is: ○ Created ○ Analyzed ○ Visualized
    - ○ Communicated

5.4 **Using Data Analysis for Detecting Credit Card Fraud**

1. Problem Context

- Most credit card fraud occurs via **credential theft**, not physical card theft
- Goal: **Early detection & mitigation**

2. Type of Analysis Used

Descriptive Analytics
(Analyzing historical data to understand patterns & anomalies)

3. Common Fraud Indicators (Anomalies)

- Sudden increase in transaction frequency

- Transactions significantly higher than user average
- Bulk purchases with minor variations (size/color)
- Change in delivery address (home → PO box/warehouse)
- IP address mismatch with billing location

4. Required Data Points (Examples)

- Cardholder details
- Transaction amount
- Transaction time & frequency
- Delivery address
- IP address & location
- Merchant details

5. Data Preparation Steps 1. Identify relevant data points

2. Clean data:
    - Missing values
    - Incorrect entries
    - Format inconsistencies (e.g., dates)
3. Analyze patterns & anomalies
4. Visualize findings for stakeholders

6. Visualization Purpose

- Highlight hidden trends
- Communicate insights clearly
- Support fraud detection decisions

FINAL ASSIGNMENT

Using Data Analysis for Detecting Credit Card Fraud

Companies today are employing analytical techniques for the early detection of credit card frauds, a key factor in mitigating fraud damage. The most common type of credit card fraud does not involve the physical stealing of the card, but that of credit card credentials, which are then used for online purchases.

Imagine that you have been hired as a Data Analyst to work in the Credit Card Division of a bank. And your first assignment is to join your team in using data analysis for the early detection and mitigation of credit card fraud.

In order to prescribe a way forward, that is, suggest what should be done in order for fraud to get detected early on, you need to understand what a fraudulent transaction looks like. And for that you need to start by looking at historical data.

Here is a sample data set that captures the credit card transaction details for a few users.

IP Address	User ID	Account Number	Age	Shipping Address	Transaction Date	Transaction Time	Transaction Value	Product Category	Units Purchased
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	15-5-20	15:00:05	$121.58	Clothing	1
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	10-6-20	10:23:10	$79.23	Electronics	2
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	1-6-20	07:12:45		Home Décor	1
1.186.52.7	johnp	25671147	32	In-store	3-6-20	01:11:10	$2,009.99	Electronics	10
	johnp	25671147	32	In-store	2020-06-03	01:15:12	$4,131.00	Electronics	15
1.186.52.7	johnp	25671147	32	P.O. Box 1049	03-06-2020	01:22:24	$3,010.50	Tools	20
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	15 May 2020	17:02:08	$234.20	Furniture	1
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	18 May 2020	19:12:45	$141.00	Kithcen Supplies	3
	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	01 June 2020	17:34:15	$157.25	Car Spares	2
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	13 June 2020	18:02:10	$59.99	Kithcen Supplies	1
172.165.10.1	ellend	11568528		P.O. Box 1322	07 June 2020	15:53:12	$99.99	Clothing	1
172.165.10.1	ellend	11568528		P.O. Box 1322	08 June 2020	17:15:30	$53.15	Beauty	1
1.167.255.10	ellend	11568528		P.O. Box 5401	02 July 2020	00:05:10	$4,895.00	Laptop	1

Descriptive techniques of analysis, that is, techniques that help you gain an understanding of what happened, include the identification of patterns and anomalies in data. Anomalies signify a variation in a pattern that seems uncharacteristic, or, out of the ordinary. Anomalies may occur for perfectly valid and genuine reasons, but they do warrant an evaluation because they can be a sign of fraudulent activity.

Past studies have suggested that some of the common events that you may need to watch out for include:

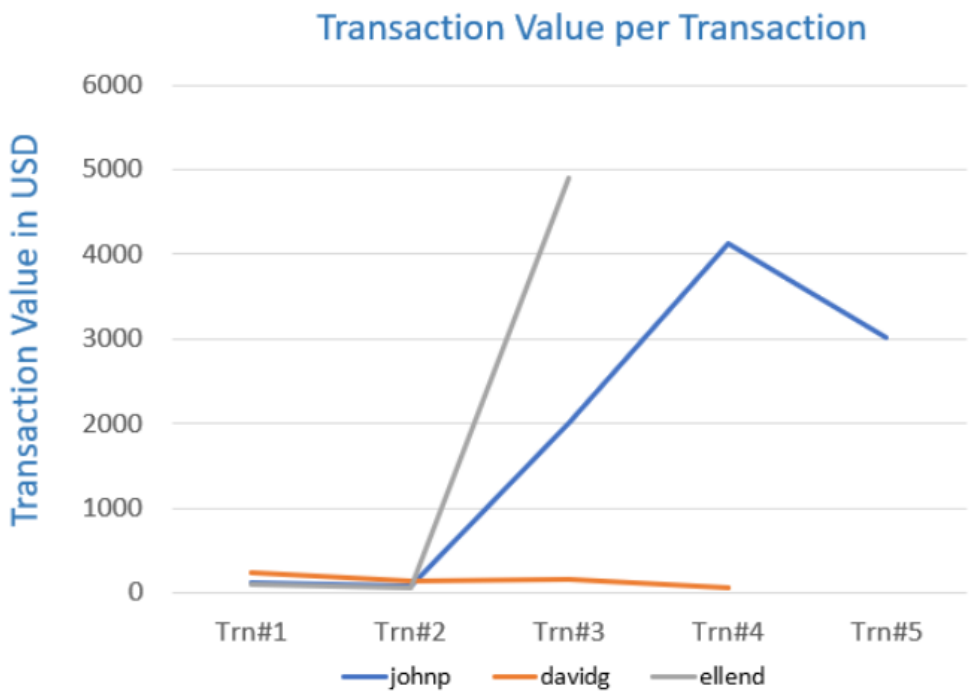
- A change in frequency of orders placed, for example, a customer who typically places a couple of orders a month, suddenly makes numerous transactions within a short span of time, sometimes within minutes of the previous order.
- Orders that are significantly higher than a user's average transaction.
- Bulk orders of the same item with slight variations such as color or size—especially if this is atypical of the user's transaction history.

- A sudden change in delivery preference, for example, a change from home or office delivery address to in-store, warehouse, or PO Box delivery.
- A mismatched IP Address, or an IP Address that is not from the general location or area of the billing address.

Before you can analyze the data for patterns and anomalies, you need to:

- **Identify and gather all data points that can be of relevance to your use case.** For example, the card holder's details, transaction details, delivery details, location, and network are some of the data points that could be explored.
- **Clean the data.** You need to identify and fix issues in the data that can lead to false or incomplete findings, such as missing data values and incorrect data. You may also need to standardize data formats in some cases, for example, the date fields.

Finally, when you arrive at the findings, you will create appropriate visualizations that communicate your findings to your audience. The graph below samples one such visualization that you would use to capture a trend hidden in the sample data set shared earlier on in the case study.



In the next section you will be asked to answer the following 5 (five) questions based on this case study:

1. **List at least 5 (five) data points that are required for the analysis and detection of a credit card fraud. (3 marks)**
   - Transaction amount
   - Transaction date and time
   - Merchant category
   - Cardholder location
   - Transaction frequency

2. **Identify 3 (three) errors/issues that could impact the accuracy of your findings, based on a data table provided. (3 marks)**
   - Inconsistent date formats (e.g., 15-5-20, 03-06-2020, 2020-06-03)
   - Inconsistent shipping addresses for the same user (home address, in-store, P.O. Box)
   - Missing transaction values (blank amount for one Home Décor transaction)

3. **Identify 2 (two) anomalies, or unexpected behaviors, that would lead you to believe the transaction may be suspect, based on a data table provided. (2 marks)**
   - Sudden high-value transactions for johnp (e.g., $4,131 and $3,010.50) compared to earlier low amounts
   - Multiple large transactions in a short time window, especially electronics/tools with unusually high unit counts

4. **Briefly explain your key take-away from the provided data visualization chart. (1 mark)**
   - johnp and ellend show spikes in transaction value, indicating potential fraudulent activity, while davidg's spending remains consistent and low-risk.

5. **Identify the type of analysis that you are performing when you are analyzing historical credit card data to understand what a fraudulent transaction looks like. [Hint: The four types of Analytics include: Descriptive, Diagnostic, Predictive, Prescriptive] (1 mark)**
   - Descriptive Analytics