COURSE CERTIFICATE

Jan 14, 2025

PRIYANKA P

has successfully completed

Introduction to Data Analytics

an online non-credit course authorized by IBM and offered through Coursera

Rav Ahuja
Global Program Director,
Skills Network

Verify at:
https://coursera.org/verify/EV1TNU49CWZ8

Coursera has confirmed the identity of this individual and their participation in the course.

Introduction to Data Analytics

Module 1: What is Data Analytics

Data is at the core of decision-making in today's world. Businesses leverage data analytics to understand trends, optimize strategies, and drive success. In this module, I explored the fundamentals of data analytics, the role of a Data Analyst, different types of data analysis, and how AI is transforming the field.

1.1 The Modern Data Ecosystem

The **data ecosystem** is a system where different technologies and people work together to make sense of data. It includes **data sources** like databases and social media, an **enterprise data environment** where data is stored and cleaned, **end users** like analysts and business teams who use data for decisions, and **emerging technologies** like AI and cloud computing that help process data better.

Data professionals play a crucial role in extracting and interpreting insights. The key roles include:

- **Data Engineers** – Build and maintain data pipelines and storage infrastructure.
- **Data Analysts** – Collect, clean, and analyze data to generate insights.
- **Data Scientists** – Use advanced models and AI to predict trends.
- **Business Analysts & BI Analysts** – Use data-driven insights to support business strategies.

1.2 Types of Data Analysis

There are four primary types of data analysis, each serving a unique purpose in decision-making:

1. **Descriptive Analytics – "*What happened*?"**
    - Summarizes past data to identify trends and patterns.
    - Example: Sales reports showing last year's revenue growth.
2. **Diagnostic Analytics – "*Why did it happen*?"**
    - Identifies reasons behind past trends using deeper analysis.
    - Example: Investigating a sudden drop in customer engagement.
3. **Predictive Analytics – "*What will happen next*?"**
    - Uses historical data and statistical models to forecast outcomes.
    - Example: Predicting customer churn rates for an online business.

4. **Prescriptive Analytics – "*What should be done*?"**
    - Suggests optimal solutions based on data-driven insights.
    - Example: AI-powered recommendations for inventory management.

These analytical methods are essential in helping businesses make strategic and informed decisions.

1.3 The Data Analysis Process

A structured approach ensures accurate and meaningful insights. The data analysis process involves:

1. **Defining the Problem** – Understanding the business challenge and objectives.
2. **Setting Metrics** – Identifying key indicators for measurement and evaluation.
3. **Collecting & Cleaning Data** – Ensuring data accuracy, completeness, and consistency.
4. **Analyzing & Extracting Insights** – Identifying correlations, trends, and anomalies.
5. **Interpreting & Presenting Findings** – Communicating results through dashboards, reports, and visualizations.

Effective data analysis requires a combination of technical expertise, critical thinking, and storytelling skills to drive business value.

1.4 The Role of Generative AI in Data Analytics

AI is revolutionizing data analytics, making it faster and more efficient. Generative AI helps in:

- **Data Augmentation** – Creating synthetic data to improve machine learning models.
- **Anomaly Detection** – Identifying unusual patterns in fraud detection and cybersecurity.
- **Automated Report Generation** – Summarizing insights in plain language for better decision-making.
- **Simulation & Forecasting** – Predicting future trends based on historical data.

However, AI also poses challenges, such as **bias, misinformation, and ethical concerns**. As AI tools evolve, responsible usage and transparency remain crucial.

1.5 A Day in the Life of a Data Analyst

The daily responsibilities of a Data Analyst can vary, but a typical workflow includes:

1. **Acquiring & Cleaning Data** – Ensuring data accuracy and consistency.
2. **Analyzing Trends** – Identifying insights through statistical tools.
3. **Building Reports & Dashboards** – Presenting findings to stakeholders.
4. **Collaborating with Teams** – Discussing business needs and refining analysis.

Module 2: The Data Ecosystem

In today's world, data is everywhere, and understanding how to store, process, and analyze data is essential for anyone working in data analytics. Module 2 introduced me to the data ecosystem, covering data types, databases, file formats, processing tools, and Big Data technologies.

2.1 Understanding Different Types of Data

Before working with data, it's important to understand its structure. Data is categorized into three main types:

1. **Structured Data** – Organized in a tabular format with rows and columns.
   Examples: Databases, spreadsheets, transaction records.
2. **Semi-Structured Data** – Partially organized but lacks a rigid schema.
   Examples: Emails, XML & JSON files, log files.
3. **Unstructured Data** – No fixed format, complex to analyze.
   Examples: Images, videos, PDFs, social media posts.

Each type requires different storage and processing techniques, which leads to the use of databases, data warehouses, and data lakes.

2.2 Data Repositories – Where Data is Stored

A data repository is where information is collected, organized, and stored for easy access and analysis. The main types include:

- **Databases** – Store structured data, allowing easy querying (e.g., SQL databases).
- **Data Warehouses** – Centralized repositories storing clean, structured data for reporting.

- **Data Marts** – A smaller subset of a data warehouse, specific to a business function (e.g., finance, sales).
- **Data Lakes** – Store raw, structured, and unstructured data for flexible analysis.
- **Big Data Stores** – Handle massive data volumes for real-time processing and analytics.

2.3 File Formats Used in Data Analytics

Data is stored in different file formats based on the type of data and use case:

- **Delimited Text Files** – CSV, TSV (Common for data exchange).
- **Excel Files (.XLSX)** – Used for data analysis in spreadsheets.
- **JSON & XML** – Common in web data & APIs.
- **PDFs** – Used in documents but harder to analyze.

Knowing how to work with different file types is essential for a Data Analyst since data can come from multiple sources.

2.4 How Data is Collected – Common Data Sources

Data comes from various sources, and understanding where data originates is key:

- **Databases** – Transaction records, CRM data, HR records.
- **APIs & Web Services** – Used to pull data from social media, stock markets, weather reports, etc.
- **Web Scraping** – Extracting data from websites for analysis (e.g., pricing comparisons).
- **Data Streams** – Real-time data from IoT devices, GPS, social media feeds, and financial markets.
- **RSS Feeds** – Used for news aggregation and continuous updates.

Example:

A retail company might use:

Sales transactions from a database,

Customer reviews from web scraping,

Stock prices from an API,

Social media mentions from data streams

2.5 Programming Languages for Data Analytics

To work with data efficiently, analysts use different languages:

- **SQL (Structured Query Language)** – Used for querying and managing databases.
- **Python & R** – Popular for data analysis, visualization, and machine learning.
- **Java** – Used in Big Data frameworks like Hadoop and Spark.
- **Shell Scripting (Unix/Linux, PowerShell)** – Automates repetitive tasks.

Example:

A Data Analyst might use SQL to retrieve customer data, Python to clean it, and PowerShell to automate repetitive processes.

2.6 The ETL Process – Turning Raw Data into Insights

ETL (Extract, Transform, Load) is a core process in data analytics:

- Extract – Collecting data from different sources (databases, APIs, flat files).
- Transform – Cleaning, formatting, and standardizing data.
- Load – Storing processed data into a database, data warehouse, or data lake.

There are two types of data movement:

1. Batch Processing – Moves large amounts of data at scheduled times (e.g., daily sales reports).
2. Stream Processing – Processes data in real-time (e.g., stock market prices).

2.7 Big Data & NoSQL Databases

With the explosion of data, traditional databases can struggle to keep up. Big Data technologies help solve this problem.

1. **NoSQL Databases (for handling large, flexible datasets):**
    - Key-Value Stores (Redis, DynamoDB) – Used for caching and real-time recommendations.
    - Document Databases (MongoDB, CouchDB) – Store flexible records for eCommerce, CRM.
    - Column-Based Stores (Cassandra, HBase) – Best for time-series and sensor data.
    - Graph Databases (Neo4J, CosmosDB) – Used for social networks, fraud detection.
2. **The 5 V's of Big Data (Big Data is defined by these characteristics):**
   Volume – Huge amounts of data are collected daily.

Velocity – Data is generated at high speed (e.g., live stock market data).
Variety – Data comes in different formats (text, images, videos, etc.).
Veracity – Ensuring data accuracy and reliability.
Value – Turning raw data into useful insights.

3. **Big Data Processing Tools**

   To analyze Big Data, companies use tools like:

   - Hadoop : Stores and processes data across multiple servers.

   - Hive : A data warehouse for querying large datasets.

   - Spark : Processes data faster with in-memory computing.

   Example:

   - A streaming platform like Netflix uses Spark for real-time recommendations.

   - An eCommerce site uses Hive to analyze customer behavior.

   - A finance company stores historical transactions in Hadoop for fraud detection.

Module 3: Gathering and Wrangling Data

Data is the foundation of analysis, but before we can derive insights, we need to collect, clean, and prepare it properly. This module focused on the essential steps of gathering, transforming, and ensuring data reliability.

3.1 Understanding Data Sources

Data comes from many different places and can be categorized as:

- **Primary Data** – Directly collected through surveys, interviews, or internal company systems (e.g., CRM, HR databases).
- **Secondary Data** – Collected by others but available for use (e.g., research papers, public records, industry reports).
- **Third-Party Data** – Purchased from data aggregators who collect and sell datasets (e.g., financial data providers).

Common Data Sources:

- Databases – Internal company systems, external databases.
- Cloud Storage – Real-time access to global data.
- Web Data – Public records, government data, social media, and websites.

- Sensors & IoT Devices – Data from smart devices, wearables, GPS, medical tools, etc.
- Surveys & Interviews – Insights from direct user interactions.
- Data Exchanges – Platforms where organizations share data securely.

The variety of sources means that combining different types of data can help businesses make better decisions and find new opportunities.

3.2 Gathering & Importing Data

Once we identify the right data sources, we need to collect and import the data into a system for analysis.

Methods to Gather Data:

- **SQL Queries** – Used for extracting data from relational databases.
- **APIs (Application Programming Interfaces)** – Help fetch data from web services, apps, and other sources.
- **Web Scraping** – Automates the collection of website data.
- **Data Streams & Feeds** – Live data from social media, GPS, financial markets, etc.
- **ETL Tools** – Extract, Transform, and Load (ETL) tools help clean and move data into databases.

Where Do We Store Data?

- **Relational Databases (SQL)** – Best for structured data like sales records.
- **NoSQL Databases** – Ideal for flexible and semi-structured data like JSON files.
- **Data Lakes** – Store large amounts of structured and unstructured data for deep analysis.

Choosing the right storage method depends on the type, size, and purpose of the data.

3.3 What is Data Wrangling?

Before data can be analyzed, it needs to be cleaned, formatted, and structured. This process is called Data Wrangling (or Data Munging) and consists of four key steps:

1. **Discovery** – Understanding the data: What's missing? What's inconsistent?
2. **Transformation** – Converting data into a usable format (e.g., merging datasets, changing formats).
3. **Validation** – Checking data for errors and inconsistencies.
4. **Publishing** – Making the clean data available for analysis.

Common Transformations:

- **Structuring Data** – Organizing it into tables or logical formats.
- **Cleaning Data** – Removing duplicates, fixing typos, handling missing values.
- **Enriching Data** – Adding external data for deeper insights.

Without proper wrangling, data analysis can lead to incorrect conclusions.

There are several tools available to clean and transform data efficiently:

- Excel & Google Sheets – Basic data cleaning and analysis.
- OpenRefine – A free tool for cleaning messy datasets.
- Google DataPrep – An AI-powered tool for preparing structured and unstructured data.
- IBM Watson Data Refinery – Cleans and organizes large datasets.
- Trifacta Wrangler – Cloud-based tool for transforming raw data into usable formats.
- Python & R – Programming languages with powerful libraries for data manipulation:
    - Pandas (Python) – For handling large datasets.
    - NumPy (Python) – For numerical computations.
    - Dplyr (R) – For manipulating large data tables.

The choice of tools depends on the data type, team expertise, and infrastructure.

3.4 Data Cleaning - The Most Important Step

Steps to Clean Data:

- **Inspect** – Identify missing values, inconsistencies, and duplicates.
- **Clean** – Correct errors, remove unnecessary data, fix formatting issues.
- **Verify** – Check whether the cleaned data is accurate and useful.

Common Data Issues & Fixes:

- **Missing Values** – Remove or fill them using statistical methods.
- **Duplicate Entries** – Identify and eliminate redundant data points.
- **Incorrect Data Types** – Convert text to numbers, fix date formats, etc.
- **Outliers** – Detect extreme values that may distort analysis.

Good data cleaning ensures reliable analysis and better decision-making.

3.5 Ensuring Data Reliability

Even after cleaning, we must verify data reliability to avoid misleading insights.

Ways to Ensure Reliable Data:

- Run summary statistics – Check if numbers make sense (e.g., no negative website visits).
- Perform logic checks – Verify if trends align with expectations.
- Cross-check sources – Ensure the data is unbiased and accurate.
- Use tracking – Ensure no missing or outdated records.

If data is unreliable, insights and business decisions will be flawed.

Module 4: Mining & Visualizing Data and Communicating Results

4.1 What is Statistical Analysis?

Data analysis is all about making sense of numbers. Statistics is a branch of mathematics that helps us collect, analyze, and interpret data to find patterns and trends. Every day, companies and researchers use statistics to make better decisions.

Statistical analysis means applying statistical methods to data samples to understand what they represent. A sample is a smaller group taken from a larger population. For example, if a company wants to understand customer preferences, they might survey 1,000 people instead of the entire country.

There are two main types of statistical analysis:

1. **Descriptive Statistics** – This helps summarize and present data in a simple way using charts, graphs, and numbers.
    - Central Tendency (Mean, Median, Mode) shows where most data points are.
    - Dispersion (Variance, Standard Deviation, Range) tells us how spread out the data is.
    - Skewness shows if data is symmetrical or leaning towards one side.
2. **Inferential Statistics** – This helps make predictions based on sample data.
    - Hypothesis Testing determines if findings are significant or just by chance.
    - Confidence Intervals show the range in which the true population value likely falls.
    - Regression Analysis helps predict relationships between different variables.

Statistical analysis is important to ensures that our insights are reliable and not random.

4.2 What is Data Mining?

Data mining is the process of uncovering patterns, trends, and useful information from large sets of data. It is the heart of data analysis, used in everything from customer behavior predictions to fraud detection.

Patterns and Trends

- Patterns are repeated behaviors or occurrences in data. For example, customers who buy laptops often buy laptop bags too.
- Trends show long-term changes over time. For instance, global temperatures rising over the last century is a trend.

Common Data Mining Techniques

1. **Classification** – Categorizing data into different groups, like classifying emails as spam or not spam.
2. **Clustering** – Grouping similar data together, like customer segmentation in marketing.
3. **Anomaly Detection** – Finding unusual patterns, like detecting fraud in banking transactions.
4. **Association Rule Mining** – Finding relationships between items, like "people who buy chips often buy soda."
5. **Sequential Patterns** – Identifying event sequences, like tracking how users navigate a website.
6. **Decision Trees** – A flowchart-like model that helps in decision-making.
7. **Regression Analysis** – Predicting numerical values based on existing data, like predicting house prices based on size and location.

Data mining is important because it helps businesses make better decisions by revealing hidden patterns in data.

Tools for Data Mining

There are many tools that help with data mining. Some of the most commonly used ones are:

1. Spreadsheets (Excel, Google Sheets) – Great for small datasets, sorting, filtering, and creating charts.

2. R – A powerful programming language with built-in functions for statistical modeling and data mining.
3. Python – Popular for data science, with libraries like Pandas (data manipulation) and NumPy (mathematical computing).
4. IBM SPSS Statistics – A tool used for statistical analysis and data mining.
5. IBM Watson Studio – A cloud-based platform for advanced analytics and machine learning.
6. SAS (Statistical Analysis System) – Used for big data analytics and predictive modeling.

Choosing the right tool depends on your needs—for example, Excel is great for beginners, while Python and R are better for handling large datasets.

4.3 Storytelling and Data Visualization

Storytelling helps make data meaningful by turning numbers into insights people can relate to and act on. Instead of just showing statistics, a good story connects facts with emotions, making the message more persuasive. For example, a smartwatch company found that fitness trackers encouraged users to upgrade their devices. Instead of listing numbers, they shared a customer's journey to highlight this trend. Data visualization also makes data easier to understand by using charts, graphs, and dashboards to show patterns and trends. Common tools like Excel, Tableau, and Power BI help businesses create reports that support better decisions. When combined, storytelling and visualization make data clear, engaging, and useful.