

# Data Analysis Techniques

## Course Description:

This 40-hour intensive training program equips early-career data professionals with the practical skills to collect, clean, analyze, and present data-driven insights. Emphasizing tool fluency, problem-solving, and real-world application, participants will walk away with project-ready competencies across key data platforms. The course blends guided instruction with hands-on labs to develop confidence in data storytelling, trend analysis, and predictive modeling.

**Duration:** 40 hours

## Learning Objectives:

By the end of this program, learners will be able to:

- Profile, cleanse, and transform raw datasets for analysis
- Design and execute exploratory data analysis (EDA) in Python and SQL
- Build effective data visualizations using Power BI and matplotlib
- Use correlation and trend analysis to identify key business drivers
- Apply statistical techniques like regression, clustering, and hypothesis testing
- Automate repetitive data prep using Power Query [DEMO]
- Write efficient queries to join, aggregate, and pivot data across sources [DEMO]
- Develop and present a full-cycle analytics project with actionable outcomes [DEMO]

## Content Coverage

### Module 1: Foundations of Data Analysis

- What is data analysis and its business value
- Types of data: qualitative vs quantitative
- The data analysis lifecycle
- Framing analytical questions using the CRISP-DM model
- Overview of tools and environments
- Defining success criteria and stakeholder goals

### Module 2: Working with Data

- Sourcing structured vs unstructured data [DEMO]
- Data profiling and quality checks
- Missing data treatment: deletion vs imputation
- Outliers and data validation techniques
- Data exploration using Power Query and pandas [DEMO]
- Qualitative data treatment (tagging, transcription basics)

### Module 3: Regression & Predictive Modeling

- Introduction to regression: linear, multiple, and logistic

- Feature selection and multicollinearity
- Model building with scikit-learn [DEMO]
- Evaluating model accuracy (MAE, RMSE,  $R^2$ )
- Predictive use cases: demand forecasting, lead scoring
- Interpreting coefficients for business impact

#### **Module 4: Monte Carlo Simulation & Risk Analysis**

- Probability distributions and sampling
- Why and when to use Monte Carlo techniques
- Building a simulation using Python's random and numpy [DEMO]
- Visualizing simulation results with histograms
- Real-world scenarios: budget forecasting, inventory risk
- Sensitivity analysis and scenario modeling

#### **Module 5: Segmentation & Clustering**

- Unsupervised learning: K-means, DBSCAN, Hierarchical Clustering
- Determining optimal clusters using the elbow method
- Practical application: customer segmentation [DEMO]
- Visualizing clusters in 2D/3D space
- Interpreting clustering results for campaign design
- Comparing with traditional personas

#### **Module 6: Cohort and Time Series Analysis**

- Grouping users by acquisition/behavioral events
- Retention and churn visualization [DEMO]
- Time series components: trend, seasonality, noise
- Forecasting with moving averages and ARIMA models [DEMO]
- Real-world case: analyzing product usage trends
- Pitfalls: overfitting and false seasonality detection

#### **Module 7: Sentiment & Text Analysis**

- Overview of Natural Language Processing (NLP)
- Data cleaning: tokenization, stop words, stemming
- Sentiment scoring using TextBlob or VADER [DEMO]
- N-gram analysis and topic extraction
- Real-world use: brand monitoring via social media
- Visualization of sentiment shifts over time

#### **Module 8: Synthesis & Capstone**

- Selecting the right technique for the problem
- End-to-end project execution [DEMO]