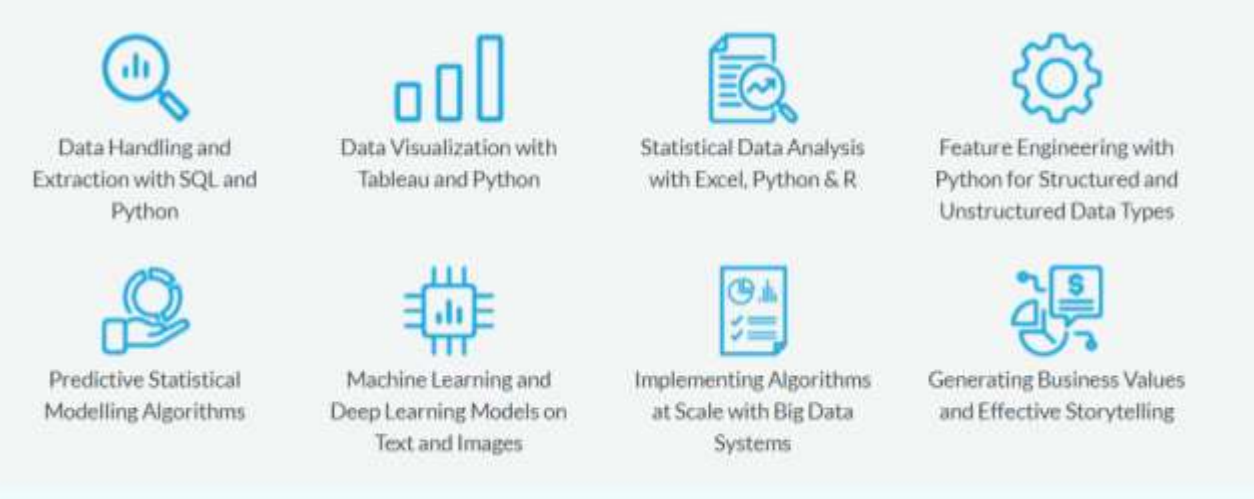
Data Science Syllabus

PROGRAM CURRICULUM

This curriculum is designed to be a comprehensive program covering technical and business aspects of the application of analytics and data science. It starts by laying a strong foundation of essential tools and techniques including descriptive and inferential statistics, data extraction and manipulation with SQL, data manipulation and processing with Python & R, and data visualization with Tableau.

Students will also have access to Tableau's cutting-edge visual analytics software for free while enrolled in the program for study and practice purposes, through Tableau's global Academic Program.

The program builds on this foundation by extending the analysis capabilities to predictive models using statistical modelling and machine learning, and covering data types that are both traditional and structured, to more unstructured types including text and images.



STATISTICS FOR DATA SCIENCE

- Introduction To Statistics
- Terminologies In Statistics
- Categories In Statistics
- Understanding Descriptive Analysis
- Descriptive Statistics In R
- Understanding Inferential Analysis
- Inferential Statistics In R

FEATURE ENGINEERING

- Data Exploration – Sanity Checks
- Preparing Data Quality Reports
- Data Preparation -Outliers and Missing Value Treatments
- Variable Profiling Using Information Value

BIG DATA AND MACHINE LEARNING WITH SPARK

- Introduction to Big Data Ecosystem
- Hadoop and HDFS
- Querying with Hive
- Data Engineering Case Study
- Introduction to Spark
- Spark Streaming
- PySPark
- ML with Spark Case Studies

SQL

- SQL Servers as Data Sources
- Data Normalization and Consequences
- Basic SQL DML Queries
- SQL Joins

R PROGRAMMING

- Basics of R Programming
- Data Manipulation
- Data Preprocessing
- Scatter Plot
- Histogram
- Bar & Stack Bar Chart
- Box Plot
- Area Chart
- Heat Map
- Correlogram

PREDICTIVE STATISTICAL MODELLING IN R

- Linear Regression – BLUE Estimators and Interpreting Model Results
- Linear Regression – Checking Model Assumptions and Improving Models

- Logistic Regression – Logistic Cost Function and Interpreting Model Results
- Logistic Regression – Measuring Classification Performance – AUC, ROC, Confusion Matrix
- Poisson Regressions – Cost Function, Overdispersion and Zero Inflation
- Poisson regression – Interpreting Model Results

PYTHON

- Python Fundamentals
- Basic Data Structures
- Data Manipulation
- Pandas
- Seaborn and MatPlotLib

DATA VISUALIZATION WITH PYTHON

- Introduction to Matplotlib
- Basic Plotting with Matplotlib
- Dataset on Immigration
- Line Plots
- Area Plots
- Histograms
- Bar Charts
- Pie Charts
- Box Plots
- Scatter Plots
- Bubble Plots

MACHINE LEARNING WITH PYTHON

- Tree Models – Regression Trees and Classification Trees
- Feature Importance
- Purity Measures – GINI
- Purity Measures – Entropy MSE
- Building and Pruning Trees
- Ensemble Methods – Bagged Ensembles
- Ensemble Methods – Random Forests
- Ensemble Methods – Gradient Boosting – Xg Boost
- Clustering – K Means and Hierarchical Models

INTRODUCTION TO NLP

- Handling text data
- Handling image data
- NLP

DATA VISUALIZATION WITH TABLEAU

- Basics of Data Visualization
- Models to Value
- Pitfalls of Predictive Models in Business
- Storytelling with Data

Job Profile after this course

- **Data Scientist**
- **Data Analyst**
- **ML Developer**
- **NLP Expert**