Data Science Beginner Track

12-Week Comprehensive Curriculum

DataVerse Africa Internship Division | Cohort 4.0

Program Overview

Item	Details
Duration	12 Weeks (3 Months)
Track Level	Beginner
Lecture Schedule	2 sessions per week (Tuesdays & Thursdays, 7:00 PM - 9:00 PM)
Tools Covered	Python, Statistics, Machine Learning Basics, Data Visualization
Delivery Mode	Live instruction + Hands-on practice + Weekly projects
Target Outcome	Job-ready junior data scientists with practical ML skills and portfolio

Program Structure

- **Month 1 (Weeks 1-4):** Python Programming & Data Manipulation
- **Month 2 (Weeks 5-8):** Statistics & Exploratory Data Analysis
- **Month 3 (Weeks 9-11):** Machine Learning Fundamentals
- **Week 12:** Capstone Project & Final Presentation

MONTH 1: PYTHON PROGRAMMING & DATA MANIPULATION

Week 1: Python Foundations

Live Sessions:

Day 1: Introduction to Data Science & Python Basics

- What is Data Science and its applications
- Python vs other programming languages
- Installing Python and Jupyter Notebook/Google Colab
- Python syntax and structure
- Variables and data types (int, float, str, bool)
- Basic operators (+, -, *, /, %, **)
- Print statements and comments

Day 2: Control Flow & Functions

- Conditional statements (if, elif, else)
- Comparison and logical operators
- Loops (for and while)
- Range function
- Defining functions (def)
- Function parameters and return values
- Built-in functions (len, type, input)

Self-Study:

- Practice writing simple Python scripts
- Solve basic coding exercises
- Explore Jupyter Notebook interface

Weekly Assignment:

Project 1: "Simple Calculator & Number Analyzer"

- Build a calculator using functions
- Create a program that analyzes a list of numbers (min, max, average)
- Use loops and conditional statements
- Apply user input and output
- **Due:** Sunday, 11:59 PM

Tools: Python (Jupyter Notebook or Google Colab)

Week 2: Data Structures & File Handling

Live Sessions:

Day 1: Python Data Structures

- Lists (creation, indexing, slicing, methods)
- Tuples (immutable sequences)
- Dictionaries (key-value pairs)
- Sets (unique elements)
- List comprehensions basics
- Nested data structures

Day 2: Working with Files & Libraries

- Reading and writing text files
- CSV file handling
- Introduction to Python libraries
- Installing libraries with pip
- Importing modules (import, from...import)
- Introduction to NumPy basics
- NumPy arrays vs Python lists

Self-Study:

- Practice data structure operations
- Experiment with file I/O
- Explore NumPy documentation

Weekly Assignment:

Project 2: "Student Grade Management System"

- Read student data from CSV file
- Store data in appropriate data structures (lists, dictionaries)
- Calculate average grades, highest/lowest performers
- Write results to a new file
- Use NumPy for calculations
- **Due:** Sunday, 11:59 PM

Tools: Python, NumPy

Week 3: Data Manipulation with Pandas

Live Sessions:

Day 1: Introduction to Pandas

- What is Pandas and why it's essential
- Series and DataFrame objects
- Creating DataFrames from various sources
- Reading CSV, Excel files
- Basic DataFrame operations (head, tail, info, describe)
- Selecting columns and rows
- Filtering data with boolean indexing

Day 2: Data Cleaning with Pandas

- Handling missing values (isnull, fillna, dropna)
- Removing duplicates
- Renaming columns
- Changing data types
- String operations on text data
- Sorting and ranking data
- Basic data transformations

Self-Study:

- Practice Pandas operations
- Clean messy datasets
- Explore Pandas documentation

Weekly Assignment:

Project 3: "E-Commerce Data Cleaning Project"

- Load a messy e-commerce dataset
- Handle missing values appropriately
- Remove duplicates and fix data types
- Clean text columns (strip spaces, standardize)
- Create new calculated columns
- Export cleaned data
- Document cleaning steps
- **Due:** Sunday, 11:59 PM

Tools: Python, Pandas

Week 4: Data Aggregation & Visualization Basics

Live Sessions:

Day 1: Pandas Aggregation & Grouping

- GroupBy operations
- Aggregate functions (sum, mean, count, min, max)
- Multiple aggregations
- Pivot tables in Pandas
- Merging and joining DataFrames

- Concatenating data
- Reshaping data (melt, stack, unstack)

Day 2: Introduction to Data Visualization

- Why visualize data?
- Introduction to Matplotlib
- Basic plots (line, bar, scatter, histogram)
- Plot customization (titles, labels, colors)
- Introduction to Seaborn
- Statistical plots with Seaborn
- Saving figures

Self-Study:

- Practice grouping and aggregating data
- Create various chart types
- Explore visualization galleries

Weekly Assignment:

MAJOR PROJECT 1: Python Data Analysis Dashboard

"Retail Sales Analysis with Python"

- Load and clean 12 months of sales data
- Perform exploratory data analysis using Pandas
- Group and aggregate data:
    - Sales by month, product, region
    - Top customers and products
    - Revenue trends
- Create 8+ visualizations using Matplotlib/Seaborn:
    - Line charts for trends
    - Bar charts for comparisons
    - Histograms for distributions
    - Heatmaps for correlations
- Write analysis report (3-4 pages)
- Create Jupyter Notebook with narrative
- **Due:** Sunday, 11:59 PM

Tools: Python, Pandas, Matplotlib, Seaborn

Week 4 Checkpoint: Python & Pandas skills assessment + Portfolio review

🗓️ MONTH 2: STATISTICS & EXPLORATORY DATA ANALYSIS

Week 5: Descriptive Statistics

Live Sessions:

Day 1: Measures of Central Tendency & Dispersion

- Mean, median, mode
- When to use each measure
- Range, variance, standard deviation

- Quartiles and percentiles
- Interquartile range (IQR)
- Understanding data distribution
- Skewness and kurtosis

Day 2: Probability Fundamentals

- Basic probability concepts
- Probability distributions
- Normal distribution
- Binomial and Poisson distributions
- Z-scores and standardization
- Central Limit Theorem (introduction)

Self-Study:

- Calculate statistics manually and with Python
- Practice probability problems
- Explore scipy.stats library

Weekly Assignment:

Project 4: "Statistical Analysis of Test Scores"

- Analyze student test score dataset
- Calculate all measures of central tendency and dispersion
- Create distribution plots
- Identify outliers using IQR method
- Calculate probabilities using normal distribution
- Compare distributions across different groups
- Write statistical summary report
- **Due:** Sunday, 11:59 PM

Tools: Python, NumPy, SciPy, Pandas

Week 6: Inferential Statistics

Live Sessions:

Day 1: Sampling & Hypothesis Testing Basics

- Population vs sample
- Sampling methods
- Sampling distributions
- Confidence intervals
- Introduction to hypothesis testing
- Null and alternative hypotheses
- P-values and significance levels
- Type I and Type II errors

Day 2: Common Statistical Tests

- T-tests (one-sample, two-sample, paired)
- Chi-square test
- ANOVA basics

- Correlation vs causation
- Pearson correlation coefficient
- When to use which test
- Interpreting test results

Self-Study:

- Practice hypothesis testing problems
- Run statistical tests in Python
- Understand test assumptions

Weekly Assignment:

Project 5: "A/B Testing Analysis"

- Analyze A/B test results dataset
- Formulate hypotheses
- Perform appropriate statistical tests
- Calculate confidence intervals
- Interpret p-values
- Make data-driven recommendations
- Create visualization of results
- Write statistical report with conclusions
- **Due:** Sunday, 11:59 PM

Tools: Python, SciPy, Statsmodels

Week 7: Exploratory Data Analysis (EDA)

Live Sessions:

Day 1: Comprehensive EDA Techniques

- EDA framework and process
- Univariate analysis (single variables)
- Bivariate analysis (two variables)
- Multivariate analysis (multiple variables)
- Correlation analysis
- Identifying patterns and trends
- Detecting outliers and anomalies
- Feature relationships

Day 2: Advanced Visualization for EDA

- Pair plots and correlation matrices
- Box plots and violin plots
- Distribution plots (histograms, KDE)
- Scatter plot matrices
- Categorical data visualization
- Time series visualization
- Interactive plots with Plotly (introduction)

Self-Study:

- Perform EDA on different datasets

- Create comprehensive EDA notebooks
- Practice storytelling with data

Weekly Assignment:

Project 6: "Comprehensive EDA Report"

- Choose a dataset from provided options
- Perform complete exploratory data analysis:
    - Data overview and structure
    - Summary statistics
    - Missing value analysis
    - Distribution analysis for all variables
    - Correlation analysis
    - Outlier detection
    - Feature relationships
- Create 12+ visualizations
- Write narrative EDA report in Jupyter Notebook
- Provide insights and recommendations
- **Due:** Sunday, 11:59 PM

Tools: Python, Pandas, Matplotlib, Seaborn, Plotly

Week 8: Feature Engineering & Data Preprocessing

Live Sessions:

Day 1: Feature Engineering Fundamentals

- What is feature engineering?
- Creating new features from existing ones
- Feature extraction from dates (year, month, day, weekday)
- Binning continuous variables
- Encoding categorical variables (Label Encoding, One-Hot Encoding)
- Feature scaling and normalization
- Handling imbalanced data

Day 2: Data Preprocessing for Machine Learning

- Train-test split concept
- Cross-validation introduction
- Handling missing values for ML
- Outlier treatment strategies
- Feature selection basics
- Dimensionality reduction introduction
- Pipeline creation
- Preparing data for modeling

Self-Study:

- Practice feature engineering techniques
- Experiment with different encodings
- Understand preprocessing importance

Weekly Assignment:

MAJOR PROJECT 2: End-to-End EDA & Preprocessing

"Customer Churn Analysis & Preparation"

- Load customer churn dataset
- Perform comprehensive EDA
- Engineer new features:
  - Customer tenure categories
  - Spending ratios
  - Engagement scores
- Handle missing values and outliers
- Encode categorical variables
- Scale numerical features
- Split data into train and test sets
- Create preprocessing pipeline
- Document all transformations
- **Presentation:** 8-minute video walkthrough
- **Due:** Sunday, 11:59 PM

Tools: Python, Pandas, Scikit-learn

Week 8 Checkpoint: Statistics & EDA skills assessment + Portfolio review

🗓️ MONTH 3: MACHINE LEARNING FUNDAMENTALS

Week 9: Introduction to Machine Learning

Live Sessions:

Day 1: Machine Learning Foundations

- What is Machine Learning?
- Types of ML: Supervised, Unsupervised, Reinforcement
- Classification vs Regression
- ML workflow overview
- Introduction to Scikit-learn
- Model training and prediction
- Evaluating model performance
- Overfitting and underfitting

Day 2: Linear Regression

- Simple linear regression concept
- Multiple linear regression
- Assumptions of linear regression
- Training a regression model with Scikit-learn
- Making predictions
- Model coefficients and interpretation
- R-squared and MSE metrics
- Residual analysis

Self-Study:

- Understand ML concepts
- Practice linear regression examples
- Explore Scikit-learn documentation

Weekly Assignment:

Project 7: "House Price Prediction Model"

- Load housing dataset
- Perform EDA on features
- Engineer relevant features
- Train linear regression model
- Evaluate model performance
- Interpret coefficients
- Make predictions on new data
- Visualize actual vs predicted values
- Write model report
- **Due:** Sunday, 11:59 PM

Tools: Python, Scikit-learn, Pandas

Week 10: Classification Algorithms

Live Sessions:

Day 1: Logistic Regression & Decision Trees

- Binary classification concept
- Logistic regression fundamentals
- Training logistic regression model
- Decision Trees concept
- How Decision Trees work
- Training Decision Tree classifier
- Feature importance
- Visualizing Decision Trees

Day 2: Model Evaluation for Classification

- Confusion matrix
- Accuracy, Precision, Recall, F1-Score
- ROC curve and AUC
- Classification report
- Cross-validation
- Hyperparameter tuning introduction
- Grid Search basics
- Model comparison

Self-Study:

- Practice classification problems
- Understand evaluation metrics
- Compare different models

Weekly Assignment:

Project 8: "Customer Churn Prediction"

- Use preprocessed data from Week 8
- Train multiple classification models:
    - Logistic Regression
    - Decision Tree
- Evaluate each model with appropriate metrics
- Create confusion matrices
- Plot ROC curves
- Perform cross-validation
- Select best model with justification
- Make predictions on test set
- **Due:** Sunday, 11:59 PM

Tools: Python, Scikit-learn

Week 11: Ensemble Methods & Model Deployment

Live Sessions:

Day 1: Ensemble Learning

- What are ensemble methods?
- Random Forest concept
- Training Random Forest models
- Gradient Boosting introduction
- XGBoost basics (optional)
- Voting classifiers
- Feature importance from ensembles
- When to use ensemble methods

Day 2: Model Deployment Basics & MLOps Introduction

- Saving and loading models (pickle, joblib)
- Creating prediction functions
- Introduction to Flask for model serving
- Building simple ML API
- Model monitoring concepts
- ML project structure
- Best practices for ML projects
- Documentation importance

Self-Study:

- Experiment with ensemble models
- Practice saving/loading models
- Explore deployment options

Weekly Assignment:

MAJOR PROJECT 3: Complete ML Pipeline

"End-to-End Machine Learning Project"

- Choose from provided datasets or propose your own
- **Phase 1:** EDA and insights
- **Phase 2:** Feature engineering and preprocessing
- **Phase 3:** Model training and evaluation
    - Train at least 3 different models
    - Compare performance
    - Tune hyperparameters
    - Select best model
- **Phase 4:** Model deployment preparation
    - Save final model
    - Create prediction function
    - Build simple API (Flask - optional)
- Create comprehensive Jupyter Notebook
- Write technical documentation
- **Presentation:** 12-minute video walkthrough
- **Due:** Sunday, 11:59 PM

Tools: Python, Scikit-learn, Flask (optional)

Week 11 Checkpoint: Machine Learning skills assessment

WEEK 12: CAPSTONE PROJECT & FINAL PRESENTATION

Week 12: Capstone Week

Monday-Tuesday: Final Capstone Development

- Finalize all deliverables
- Polish code and notebooks
- Complete documentation
- Prepare presentation

Wednesday: Dress Rehearsal

- Practice presentations with peers
- Final mentor feedback
- Q&A preparation

Thursday: Buffer Day

- Last-minute refinements
- Technical checks
- Code cleanup

Friday: 🏆 CAPSTONE FINAL PRESENTATIONS

- **Format:** 25-minute presentation + 10-minute Q&A
- **Audience:** Mentors, instructors, DataVerse leadership, peers, industry guests
- **Evaluation Criteria:**
    - Technical execution & code quality (35%)
    - Data analysis & insights (25%)
    - Model performance & methodology (25%)
    - Presentation & communication (15%)