ONERIO

AI & DATASCIENCE

FOUNDATION & INTERMEDIATE CURRICULUM

1.1 What is Data Science?

- Evolution and history of data science
- Importance and impact of data in business and research
- Real-world applications (healthcare, finance, e-commerce, etc.)

1.2 Lifecycle of a Data Science Project

- Data scientist: role, responsibilities, required skills
- data analyst: comparison with scientist; focus on business analysis
- data engineer: data pipelines, ETL, infrastructure
- collaboration within data teams

1.3 Roles in Data Science

- Problem definition and goal setting
- Data collection and data sources
- Data cleaning and preprocessing
- Exploratory data analysis (EDA)
- Model building and evaluation
- Deployment and monitoring

1.4 Overview of Tools

- Programming Languages: Python vs. R
- Jupyter Notebooks: Interactive coding environment
- Git & GitHub: Version control and collaboration
- IDEs and environments (Anaconda, VS Code)

1.2 Lifecycle of a Data Science Project

- Data scientist: role, responsibilities, required skills
- data analyst: comparison with scientist; focus on business analysis
- data engineer: data pipelines, ETL, infrastructure
- collaboration within data teams

2.2 Data Structures

- Data types (int, float, string, bool)
- Control structures (if-else, loops)
- functions (defining, arguments, return values)
- best practices (indentation, naming conventions)

2.3 File I/O and Error Handling

- Reading from and writing to text/CSV files
- Using with statement for file handling
- Exceptions: try-except blocks, raising custom errors

2.4 Using Jupyter Notebooks

- Markdown, code cells, and visualizations
- Magic commands and productivity tips
- Exporting and sharing notebooks

2.5 Introduction to Libraries

- NumPy: arrays, vectorized operations, broadcasting
- Pandas: Series, DataFrames, basic operations (head, tail, info)

Module 3: Data Wrangling and Analysis

3.1 Pandas Deep Dive

- Series vs DataFrame - Indexing, filtering, and selection
- Aggregations, groupby, and pivot tables

3.2 Data Cleaning Techniques

- Identifying and handling missing values
- Detecting and removing duplicates
- Handling incorrect data types
- Outlier detection basics

3.3 Data Munging

- Renaming columns, changing formats
- Encoding categorical variables (one-hot, label encoding)
- Data transformation (log, scaling)
- Date-time processing and parsing

3.4 Feature Engineering and Selection

- Creating new features from existing data
- binning, interaction terms, polynomial features
- feature selection techniques (correlation, variance threshold)
- dimensionality reduction (PCA - basic intro)

4.1 Data Visualization Principles

- Why visualize data?
- Choosing the right chart type
- Visual storytelling: clarity, color, scale, and interactivity
- Common mistakes and biases in visualization

4.2 Matplotlib and Seaborn

- Line, bar, scatter, and histogram plots
- Customization (titles, labels, legends, colors)
- Seaborn: correlation heatmaps, boxplots, violin plots
- Subplots and figure layouts

4.3 Plotly (Interactive Charts)

- Basic interactive plots
- Hover tools and tooltips
- Plotly express vs graph objects
- Use in dashboards and web apps

4.4 Dashboard Creation (Optional/Advanced)

- Introduction to streamlit: layout, widgets, deployment basics
- Power BI basics (for enterprise/business users)
- Connecting visualizations to live data

5.1 Descriptive Statistics

- Measures of central tendency: mean, median, mode
- Measures of spread: variance, standard deviation, IQR
- Skewness and kurtosis

4.3 Plotly (Interactive Charts)

- Basic interactive plots
- Hover tools and tooltips
- Plotly express vs graph objects
- Use in dashboards and web apps

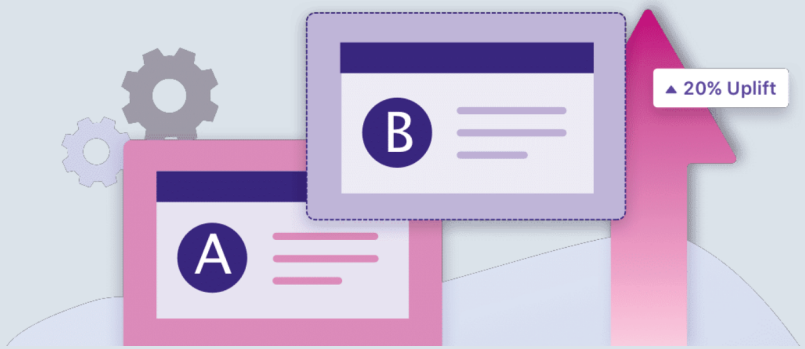
5.2 Inferential Statistics

- Population vs sample
- Confidence intervals
- Hypothesis testing: null vs alternative
- T-tests, chi-square tests, anova basics
- P-values and statistical significance

5.4 Central Limit Theorem

- Explanation and proof via simulation - Importance in sampling and hypothesis testing

5.5 A/B Testing



- Identifying and handling missing values
- Detecting and removing duplicates
- Handling incorrect data types
- Outlier detection basics

6.1 Supervised vs Unsupervised Learning

- Key differences and examples
- Problem types: regression, classification, clustering
- Typical workflow of ML projects

6.2 Scikit-learn Basics

- Data splitting: train/test
- Preprocessing pipelines (scaling, encoding)
- Model fitting and prediction
- Cross-validation

6.3 Regression Models

- Linear regression: interpretation, assumptions
- Logistic regression: sigmoid function, classification use-case
- Evaluation: RMSE, R^2 for regression

6.4 Classification Algorithms

- Decision trees: gini, entropy, pruning
- K-nearest neighbors: distance metrics, choosing k
- Support vector machines: margin, kernel trick
- Bias-variance trade-off

6.5 Clustering Algorithms

- K-means: inertia, elbow method
- Hierarchical clustering: dendrograms
- Silhouette score and evaluation of clusters

6.6 Model Evaluation

- Confusion matrix: TP, FP, TN, FN
- Precision, recall, f1-score
- ROC curve and AUC
- Overfitting and underfitting