Introductory Course on Data Science and AI

Module one: Get Data

- *Introduction of the Data Science Life Cycle.*
- *Introduction to RStudio in Posit Connect*
- *Sources of data to include Open data.*
- *Different types of data formats: .csv files, excel files, urls, compressed data, Arrow-Parquet*
- *Rectangular data, vectors and data frames in R*
- *Methods for getting rectangular data from other sources or files into R data frames.*
- *Viewing data and getting summary statistics about data in R data frames*

Module two: Clean and Reshape data

- *Cleaning data*
- *Filtering rows using logical comparisons and %in%.*
- *Selecting columns using names and tidyselect.*
- *Reshaping data using pivots.*

Module three: Visualize Univariate Data

- *Variable types: continuous vs categorical*
- *Plots for continuous variables*
- *Plots for categorical variables*
- *Customizing plots*
- *Example of dynamic titles*

Module four: Visualizing Multi-Variate data

- *Create bivariate point plots and box plots.*
- *Use plot aesthetics to code by additional variables.*
- *Add and interpret linear and non-linear smoothers.*

Module five: Statistical tests and models

- *Why statistical tests and understanding a Null hypothesis.*
- *Interpreting a p-value.*
- *Using and interpreting the t.test function, aov/anova, and lm functions in R*

Module six: Classification Models

- *Binary Classification*
- *Logistic Regression*
- *Confusion Matrices and metrics*

Module seven: Overfitting and Bias-Variance Tradeoff

- *Degrees of Freedom and Restrictive versus flexible models*
- *Over-fitting*
- *Bias-Variance Trade off*
- *Validating and models by splitting data.*

Module eight: Evaluating and Tuning Models

- *Optimizing Loss Functions*
- *Evaluation Metrics*
- *Variable Selection and Partial F-Tests*
- *ROC Curves in Classification*

Module nine: Neural networks

- *Convergence of big data, big processing power (GPUs), and algorithmic advances.*
- *Neural Network architecture*
- *Role of Activation Functions and gradient descent and back-propagation.*
- *Challenges in Over fitting and convergence*

Module ten: Generative AI and prompt engineering

- *Generative AI compared to predictive models.*
- *Tokenization and Embedding*
- *Semantic similarity for prediction.*
- *Dangers of overfitting and the need for regularization.*
- *Elements of prompt engineering.*