Statistical & Spatial Data Analysis and Visualization using R Training
(Understanding and Analyzing Big Data)

Course Description

R is an open-source programming language that provides a wide variety of statistical, spatial and graphical techniques. R has "become the de-facto standard for writing statistical software among statisticians. This course will give you a solid introduction to creating statistical and spatial analysis solutions using the R language, and how to carry out a range of commonly used analytical processes.

Course Objectives:

The course will:

- Introduce participants to R as a quantitative data analysis tool and by the end of the training learners will have mastered the advanced use of R in data manipulation and cleaning, advanced statistical and spatial analysis, mapping and big data
- Enable learner master R software and R-studio as a user interface
- Enable learner import data from various sources
- Introduce basic to advanced statistics for exploratory data analysis including methods fordescribing and summarizing variable distributions
- Provide skills for data manipulation from introductory to advanced level
- Introduce the visual representation of variables in scatter graphs, bar charts, and histograms
- Learners will be able to use of R for advanced spatial data analysis and map making
- Learners will master big data handling and use learnt concepts in capstone project

Course Outline:

Day One:

Introduction to R/R-studio

- ❖ Overview of R-studio software
  - Installing, loading and updating R packages
  - Directory management
- ❖ Introduction to data structures in R
  - Introduction to vectors
  - Introduction to matrices
  - Introduction to lists
  - Introduction to dataframes

Day Two:

Data Quality Check and Exploration

- ❖ Introduction dlookr package
  - Data diagnosis, exploration and transformation
  - Calculating descriptive statistics, testing normalityin continuous data
  - Use of diagnosis function to run data quality checksand generate data diagnosis report
- ❖ Data Visualization using Line plots, box plots, scatter plots, Histograms and bar charts

Day Three: **Statistical Analysis in R**

- ❖ Correlation Tests in R
  - Pearson
  - Spearman test.
  - Visualization of the correlation matrix using plots
- ❖ Regression analysis in R
  - Linear regression
  - Multiple linear regression
- ❖ Descriptive statistical analysis
  - Measures of central tendency (mean, mode and median)
  - Measures of dispersion (variance, standard deviation, interquartile range, range, coefficient of variation),
  - Skewness and kurtosis
- ❖ Mean Comparison Tests in R
  - One Sample T Test
  - Independent Samples T Test
  - Paired Samples T Test
  - One-way analysis of variance (ANOVA)

Day Four: **An Introduction to Spatial Data Analysis in R**

- ❖ Spatial data types: vector, raster and beyond.
- ❖ Introduction to GIS with R
- ❖ Importing, exporting and plotting spatial data with R.
- ❖ Mapping spatial data and non-spatial attributes.
- ❖ Playing with map colors and symbols

Day Five: **Mapping and Visualization using R**

- ❖ Mapping Polygon Data in R
  - Loading shapefiles into R
  - Spatial join using R
  - Creating a quick map.
  - Creating maps with tmap package
  - Saving the shapefile
- ❖ Mapping Point Data in R
  - Loading point data into R
  - Mapping point data
    - Creating proportional symbol maps

Day Six: **Advanced Data Manipulation and Cleaning**

- **Advanced Data Structures and Manipulation**
  - Introduction to 'data.table' for efficient data manipulation
  - Advanced data wrangling with 'dplyr' and 'tidyr'
  - Handling missing data and outliers
- **Text Data Handling**
  - String manipulation with 'stringr'
  - Regular expressions for text processing
- **Time Series Data**
  - Introduction to handling time series with 'lubridate'
  - Time series decomposition and visualization

Day Seven: **Advanced Statistical Analysis and Multivariate Methods**

- **Advanced Regression Techniques**
  - Polynomial and spline regression
  - Logistic regression and Generalized Linear Models (GLM)
- **Model Diagnostics and Validation**
  - Residual analysis, multicollinearity, and heteroscedasticity
- **Multivariate Analysis**
  - Principal Component Analysis (PCA) and Factor Analysis
  - Cluster Analysis (K-means and Hierarchical Clustering)

Day Eight: **Advanced Spatial Data Analysis and GIS**

- **Spatial Data Handling**
  - Advanced spatial data types: spatial joins, buffers, and raster operations
  - Working with 'sf' and 'raster packages' for spatial data
- **Spatial Statistics**
  - Spatial autocorrelation and Moran's I
  - Spatial interpolation techniques: Kriging and Inverse Distance Weighting (IDW)
- **Introduction to GIS with R**
  - Geocoding, reverse geocoding, and spatial data visualization

Day Nine: **Interactive Visualizations and Reporting**

- **Interactive Plots and Dashboards**
  - Creating interactive graphics with 'plotly'
  - Building interactive web applications with Shiny
- **Dynamic Reporting with R Markdown**
  - Creating and customizing reports with R Markdown
  - Integrating code, text, and output into a single document
- **Reproducible Research**
  - Best practices for reproducibility and version control with Git and GitHub

Day Ten: **Advanced Mapping, Big Data, and Capstone Project**

❖ Advanced Map Making

- Creating complex thematic maps with 'ggplot2' and 'sf'
- Interactive maps with leaflet

❖ Big Data Handling

- Connecting to databases and handling big data with 'data.table', 'ff', and bigmemory
- Parallel processing and distributed computing with R

❖ Capstone Project

- Comprehensive project involving statistical analysis, spatial data, and visualization
- Application of techniques learned throughout the course
- Introduction to machine learning

❖ Review and Q&A

- Review of key concepts and techniques
- Open Q&A session to address any remaining questions