Workshop: Data Science Concepts for Pavement Data & Modeling

Learning Objectives

At the conclusion of this workshop, participants will be familiar with foundational data science concepts, including supervised and unsupervised machine learning. Participants will gain hands on experience cleaning and preparing data, creating data visualizations, and performing basic modeling tasks.

The workshop will be a mix of instruction and hands on experience using the R programming language. The lab sections will be based on real-world pavement condition data and will introduce participants to analytical tools that can be used with data form their own agencies.

During the hands-on exercises, participants will read in pavement condition data, join multiple datasets together, clean and filter the data to remove erroneous values and outliers, cluster pavement segments together based on attributes, connect individual observations into series for segments, estimate deterioration rates for asphalt and concrete pavements using a linear regression model, and visualize the results.

Workshop Agenda

Morning Session: Data Science Foundations & Intro to R

8 – 9 am: Instruction - Data Science Concepts & Terminology

- Supervised and Unsupervised Machine Learning
  - ◊ Clustering
  - ◊ Regression
  - ◊ Classification
- Specific Topic Areas
  - ◊ AI & Deep Learning
  - ◊ Computer Vision
  - ◊ Natural Language Processing
- Common Tools
  - ◊ R
  - ◊ Python
  - ◊ Other Proprietary Tools (e.g., SAS, Alteryx, Watson)

9 – 9:30 am: Instruction - Introduction to R

- Basic Syntax
- Data Types
- Packages

9:30 – 10:15 am: Lab – Reading, Viewing, and Joining Data in R

- Importing Data
- Joining Data

10:30 – 11:00 am: Instruction - Data Cleaning and Preparation

- Exploratory Data Analysis
  - ◊ Summary Statistics
  - ◊ Histograms
  - ◊ Correlation Plots
- Cleaning
- Transformations (long-to-wide and wide-to-long)
- Joining/Merging Datasets

11 am – 12 pm: Lab - Data Cleaning and Preparation

- Exploratory Data Analysis
  - ◊ Summary Statistics
  - ◊ Histograms
  - ◊ Correlation Matrix
- Transformations (long-to-wide and wide-to-long)
- Joining/Merging Datasets

Afternoon Session: Data Visualization and Deterioration Modeling Using R

1 – 2 pm: Instruction – Data Visualization & Common Plots for Exploratory Data Analysis

- ggplot2 syntax
- Choosing an appropriate plot type
- Scatter Plots
- Bar Charts
- Line Graphs
- Box & Whisker Plots

2 – 2:45 pm: Lab – Data Visualization

- ggplot2 syntax
- Choosing an appropriate plot type
- Scatter Plots
- Bar Charts
- Line Graphs
- Box & Whisker Plots

3 – 4 pm: Instruction – Modeling

- Linear Regression
- K-nearest neighbors classification
- K-means clustering

4 – 5 pm: Lab – Modeling

- K-means clustering of pavements
- Linear Regression to deterioration