Introduction to Data Science and Software Development

Introduction

This summer program is designed to provide students with an introduction to the current industry best practices for applying statistical analysis to large data sets. The course curriculum is designed and administered by currently practicing data and software engineers who also have previous teaching experience. The objective of the course is to provide students with a foundational toolkit that allows them to:

- Identify problems that can be productively investigated through statistical analysis;
- Gather and organize available information from traditional databases, internet pages and application interfaces to build useful datasets;
- Manipulate, analyze and visualize datasets to inform meaningful insights;
- Produce intuitive graphics and predictive models to communicate and exercise insights from statistical analysis.

The course consists of 20 two-hour sessions over the span of 10 weeks for a total of 40 hours of direct instruction. Students will also be provided with access to curated third-party online learning platforms, through which they will be expected to complete self-directed learning modules between classes.

Target Audience

This course is designed for high school or college-aged students with a general interest in mathematics, engineering and software development. Students are not required to have prior expertise in either statistics or programming; however, all applicants will be required to complete a set of self-directed pre-work designed to provide the necessary prerequisite skills for success in the class.

Instructors

Derek Kaknes is a University of Virginia engineering graduate currently working as a Data Scientist with Certilytics, a healthcare analytics startup, and as an Associate Director with Mathnasium of Manhattan. Derek's prior work experience includes investment banking, developing software for automated personal financial analysis, and management consulting. Outside of work, Derek enjoys spending time with his Little Brother from New York's Big Brothers Big Sisters Program and freelance writing about local New York politics.

[SECOND INSTRUCTOR] is a University of California – Davis engineering graduate currently working as a Senior Software Developer with [COMPANY], a workforce optimization software platform, and as a computer science instructor for City College of New York's Advanced Coding Concepts & Web Development Python course. Outside of work, [INSTRUCTOR] is the lead organizer for the New York Python MeetUp group and enjoys performing and engaging in New York's improv comedy circuit.

Hardware and Software Requirements

All students are expected to bring their own laptop computers to each class. Computers must run either a Linux or Mac OSX operating system (No Windows!). Computers should have at least 8GB of RAM, but preferably 16GB. Please note that if your existing computer does not meet this memory requirement, then we can help you adequately upgrade before the first class (16GB of RAM costs ~$60 on Amazon).

Class Structure and Cost

Class will take place on two nights each week from 7:00pm to 9:00pm. The course runs from the week of June 26th through the week of August 28th. The total class size is limited to 20 students (10 students per instructor) and costs a total of $2,000 ($50/hr).

Overview of the Curriculum

1. An Introduction to the Unix Shell, Command Line and Vim
2. Open Source: An Introduction to the Git Workflow and Github
3. Functional Programming: An Introduction to R and RStudio
4. Object Oriented Programming: An Introduction to Python
5. Databases: An Introduction to SQL and Postgresql
6. The Internet: An Introduction to HTTP in Python
7. Beautiful Soup: An Introduction to HTML and Screen Scraping in Python
8. Homebrew, Cran, Pip: An Introduction to Package Managers
9. Data Import: Importing Data in R
10. Data Wrangling: Managing Data in R
11. Data Manipulation: Manipulating Data with the dplyr and reshape2 packages
12. Data Visualization: Creating Plots and Graphics with ggplot2, ggmap and ggvis
13. Foundations of Data Analysis: Partitioning Data, Training, Validating and Testing Linear Models
14. Intermediate Data Analysis: Training and Comparing Models with the caret package
15. Advanced Data Analysis: Ensembling Models with caretEnsemble
16. Case Study I: Analyzing New York City Council Election Results
17. Case Study II: Analyzing New York City Campaign Finance Data

18. Case Study III: Geocoding and Mapping the New York Voter File
19. Case Study IV: Predictive Modeling for Likely NYC Voters
20. Case Study V: Independent Project