DataCamp Take Home

Shonte Stephenson

5/1/2019

Track Proposal Curriculum Area

Overall: `Data Engineering`

Learner Focus: `Introduction to Data Engineering for Data Scientists in R`

This track is for data scientists who want to learn the fundamentals of data engineering.

Why This Track:

Data engineers manage data to make it available for analysis. Keeping on top of consistency and accuracy are the two underlying jobs a Data Engineer engages in everyday. Learn the foundational skills of profiling, validating and analyzing your data with confidence.

Courses within Track

This track contains the following 6 courses:

- Core Concepts in Relational Databases
- Creating a Data Pipeline in R
- Data Profiling in R
- Data Cleansing in R
- Data Validation in R
- Automating Data Quality Monitoring and Anomoly Detection

Course Descriptions and Measureable Learning Objectives

Core Concepts in Relational Databases

This course introduces students to the areas involved in understanding relational databases and the efficient usability of it including the understanding of database schemas used to structure a given data set.

Students will be able to:

- Understand relational databases (star schemas, primary and foreign keys and table structure)
- Learn foundations of database arhitecture and best practices for design
- Learn how to normalize data from disparate sources into unified schemas

Prerequisite: `Introduction to SQL`

Creating a Data Pipeline in R

This course provides an introduction to frameworks for structuring pipelines for data analysis. After an introduction to data workflows, we look at ways to connect to a database from within `R` and how to run `SQL` queries within the same `R` environment. All of these steps are translated into a SQL statement and processed inside the database but called from `R`. We do not need to import the tables into R memory at any time, we just use dplyr to get the results quickly.

Students will be able to:

- Understand the concept of data workflows and data pipeline more specifically
- Learn how to connect to a database using the `dbConnect()` function from R
- how to query directly from a relational database from inside R using `dplyr`
- Learn how to write dataframes from R to a database

Prerequisites: `dplyr` and `Introduction to SQL`

Data Profiling in R

This course covers the importance of profiling and explore interesting and useful forms of metadata that the profiling process generates. You'll use functions to aggregate, summarize, and characterize the data by understanding relationships between various columns.

Students will be able to:

- Understand the concept of metadata and how to extract meaning from it
- Summarize data identify any data format irregularities such as time stamp missing hours or names being both upper or lower case

Data Cleansing in R

In this course, students will learn how to identify and detect dirty data using the `dplyr` package. First you will learn how to handle missing data and next you will learn how to quickly identify duplicate values in rows and by string types. In the final section, the focus is on practicing how to detect major data quality violations and quick ways to correct for them.

Students will Learn How to:

- Identify and handle missing values and remove unwanted observations
- Detect and remove duplicate values
- Sanitize dirty data sets and summarize from the clean data

Prerequisites: `dplyr`

Data Validation in R

This course will cover the basics on data validation and will give you an overview of how to use "Validate" to define, investigate an manipulate rule sets in R. It will teach you how to formulate rule sets based on what to expect from the data, evaluate your validation rules against one or more data sets and finally investigate and visualize the results of a data validation step.

Students will Learn How to:

- Understand the importance of creating reproducible validation rule sets to test the data against.
- Define and maintain data validation indicators
- Evaluate data set(s) against predefined rules, either in-or-cross-data sets
- Investigate and visualize the results of a data validation step

Automating Data Quality Monitoring and Anomoly Detection in R

The ability to produce clean and validated data for analysis is an essential part of your skill set as a data engineer. This course introduces you to monitoring the reliability and accuracy of data given the historical expecations of the data. Here, you'll learn how to differentiate expected from unexpected anomolies that may occur on certain occasions or seasonally.

Students will Learn How to:

- Understand and discover positive and negative anomalies
- Detect anomolies for short and long term time series
- Detect anomolies for time series with seasonal effects