Introduction to Data Analysis

Campus Labs

April 17, 2018

Contents

Overview

This is an introductory data analysis course. The primary goal of the course is to be able to answer elementary product data questions. We will be using R and SQL to accomplish this goal.

The three primary focci of the course are:

- Data Visualization
- Data Structuring
- Data Collection

Approach

I learned to play guitar in my late teens. I remember buying a book that had me start with scales and progressing through some kiddie folk songs like *Twinkle, Twinkle Little Star*, and finally ending with playing *Tom Dooley*. This was absolute torture. I never made it past page 3. Thankfully, the Internet was in it's infancy at this time but I was able to find music for Tom Petty's *Free Falling*. I started with a song I wanted to play and learned the chords to play it. It was a manageable song, not some Jimmy Page solo, but something meaningful & acheivable. I had success quickly which in turn motivated me to learn more. Eventually, I learned the Page solos too. I remember this experience when I am teaching others.

I take the educational stance that true learning happens around problems we're motivated to solve. We figure out what we want to do and then get the language, knowledge, and skills necessary to accomplish this goal. This is a top down approach. Often data analysis courses have the participants pass through a progressive series of steps, starting with a "Hello World" excerise and building up with a grand finale capstone project. The progression is boring and the final project is often unrelated to the problems the student actually wants

to solve. Sounds like a "scales and *Tom Dooley*" approach to me. I take the opposite approach. The course assumes that the participants have questions about product data that they would like to answer quickly. This is where the course will begin. Theory and skills will be woven into the process in a "just in time" fashion.

Secondly, I take the stance that struggle is a requiste for learning. Please don't confuse fustration and struggle. The former is not productive but the latter is essential. If you haven't struggled, I'd venture to say you haven't really learned anything of value. Learning something new is a series of inner dialogue feedback loops. We ask a question, learn about that for a bit. We are usually just learning what are the right words to Google. This first pass snowballs new questions, things we didn't know we needed to ask before. And this recursive process ends when we're satisfied that we've learned what we set out to learn originally (or we've determined it wasn't really worth the investment). This course will not give you all the language, knowledge, and skills you need to analyze your problems. Instead, it will give you the most important language, the most relevant knowledge, and the most generalizable skills. This will help to avoid fustration but not the struggle.

Required Texts

We will be using the following texts for the course.

1. Grolemund, G. & Wickham, H. (2016). *R for Data Science*. Retrieved from http://r4ds.had.co.nz
2. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. Retieved from https://github.com/hadley/ggplot2-book

Additionally, we will read from articles, blogs, & cheatsheets. All of the reading materials are provided.

Learning Outcomes

1. Query SQL databases.
2. Import common data file formats into R.
3. Combine data from multiple sources.
4. Manipulate data into a tidy format.
5. Transform data with mutations and summarizations.
6. Select appropriate graph forms and aesthetics to represent data.
7. Produce tabular and graphical summaries of data.
8. Export tabular and graphical summaries of data from R.

Course Schedule

This course is based around 7 modules.

1. Intro to R
2. Data Tranforming
3. Data Visualization
4. Exploratory Data Analysis (EDA)
5. Data Tidying
6. Data Import/Export (R & SQL)

Modules vary in size and scope and have a series of interactive lectures attached to them. This is the schedule for lectures.

EXPLAIN READINGS FOR NEXT TIME [FLIPPED-ISH] and ASSIGNMENTS FOR SELF MONITORING [will be obvious if not working outside]; typically assignment reinforce class and reading prime for next class

Module 1: Intro to R			
Day	**Topic**	**Reading**	**Assignment**
0	Install Course Tools	• Syllabus • Hadley's Caveat (4:06-4:51) • Grolemund & Wickham Ch. 3	• Install R, Rstudio, & SQL Ops
1	R, RStudio, & Visualization	• Grolemund & Wickham Ch. 5 • Hadley on 5 Verbs (8:25-51:10) • Hadley on Pipes (21:28-23:00) • R: Base R - Maths Functions Section	• Assignment 1
Module 2: Data Tranforming			
Day	**Topic**	**Reading**	**Assignment**
2	Select, Filter, Arrange, & Mutate	• Grolemund & Wickham Ch. 5 • R: Transformations Cheatsheet	• Assignment 2
3	Mutate, Summarize, & Group By	• Grolemund & Wickham Ch. 3	• Assignment 3
Module 3: Data Visualization			
Day	**Topic**	**Reading**	**Assignment**
4	Viz Grammar & Theory	• Grolemund & Wickham Ch. 7 • R: ggplot2 Viz Cheatsheet • Few's Graph Selection Matrix • FT's Graph Relationship Types • 50 ggplot2 Plots by Relationship Type	• Assignment 4
5	Viz Design [Relationships]	• Grolemund & Wickham Ch. 3 & 7 • R: ggplot2 Viz Cheatsheet	• Assignment 5
Module 4: Exploratory Data Analysis (EDA)			
Day	**Topic**	**Reading**	**Assignment**
6	Intro to EDA: Tables	• Grolemund & Wickham Ch. 7 • R: Transformations Cheatsheet	• Assignment 6
7	Finding Patterns: Viz	• Grolemund & Wickham Ch. 12 • Hadley on Tidy Data (6:18-9:57)	
Module 5: Data Tidying			
Day	**Topic**	**Reading**	**Assignment**
8	Reshaping	• Grolemund & Wickham Ch. 13 • R: Data Import Cheatsheet (p. 2)	• Assignment 8
9	Combining: Joins & Binds	• Grolemund & Wickham Ch. 11 • R: Transformations Cheatsheet (p. 2)	• Assignment 9
Module 6: Data Importing & Exporting			
Day	**Topic**	**Reading**	**Assignment**
10	Import/Export in R	• YouTube: Learn Basic SQL in 10 Minutes • R: Data Import Cheatsheet (p. 1)	• Assignment 10
11	SQL Querying: Select, From, & Where	• YouTube: Joins Tutorial for Beginners • SQL: Querying Cheatsheet	• Assignment 11
12	SQL Querying: Count, Group By, & Join		

Assignments

Assignment	Description
1	Data types & ggplot2
2	Transforms: select, filter, arrange, & mutate
3	Transforms: select, filter, arrange, mutate, summarize, & group_by
4	Visualization: grammar & theory

Assignment	Description
5	Visualization: design & pattern finding
6	Exploratory data analysis
7	-
8	Reshaping data
9	Combining data: joins & binds
10	Importing & exporting data: tabular & graphical
11	*

What You Won't Learn

This course will not teach advanced data science techniques. It will not cover statistcal modeling, machine learning, computer programming, big data, text analysis, reporting, or dashboarding. Additionally, R and SQL are capable of much more that this course uses them for. They are full featued and capable of integrating with a host of other technologies and products. We will utilize ~10% these tools' features to accomplish the primary goal and learning outcomes of this course.

Some of you may want to pursue R and/or SQL at a deeper level than covered in this course which will only further improve your workflow and produce an even richer analysis.

Required Technologies

We will use R, RStudio, and Install SQL Operations Studio. Sometimes people confuse R & RStudio. R is the backend while RStudio is an IDE for working with R.

1. Install R (Do this before installing RStudio)
   - Click on the download link corresponding to your computer's operating system
   - Download the installer and follow the directions.
2. Install RStudio
   - Scroll down to "Installers for Supported Platforms"
   - Click on the download link corresponding to your computer's operating system
3. Install SQL Operations Studio
   - Click on the download link corresponding to your computer's operating system
   - Download the installer (Windows) or .zip (Mac) and follow the directions

***Note**: If you need more detailed instructions on how to install R and RStudio, watch this DataCamp video (1m22s).*

Cheat Sheets

Cheat sheets provide nice references for utilizing the R & SQL tooling and selecting graph designs.

Link	Description
R: Base R	Covers many basic R functions
R: Data Transformations	Create new columns & summarize data (tidyverse)
R: Data Visualization	Visualization in ggplot2
Few's Graph Selection Matrix	Selecting & Designing Visualizations
FT's Graph Relationship Types	Selecting & Designing Visualizations
R: Data Import	Importing, cleaning, & restructuring (tidyverse)
R: Data Wrangling (old)	Wrangling (tidyverse)
SQL: Basic Querying	Basic SQL queries

Advanced

Resources for advanced topics.

Link	Description
R: The Complete ggplot2 Tutorial	Customizing ggplot2 plots
R: ggplot2 Extensions	Extra geoms and functionality not found in base ggplot2