QBS 181: Data Wrangling

Fall 2020
Tuesday and Thursday, TBD
Course Instructor: Ramesh Yapalparvi

Data Wrangling:

This course introduces techniques which enables students to pull data from various sources and to work with different forms of data. Data wrangling requires a strong knowledge of the data structures that hold datasets. This course will cover the different types of data structures available in SQL and R, how they differ by dimensionality and how to create, add to, and subset the various data structures. You will also learn how to deal with missing values in data structures. Furthermore, since analysis is often a collaborative effort data scientist also need to know how to share their data. This course will cover the basics of importing tabular and spreadsheet data, scraping data stored online, and exporting data for sharing purposes. Writing a simple, replicable, and readable code is important to becoming an effective and efficient data scientist. Throughout this course, you will be introduced to the art of writing functions and using loop control statements to reduce redundancy in code. You will also learn how to simplify your code using various operators to make your code more readable. Consequently, this will help students to perform data wrangling tasks more effectively, efficiently, and with more clarity. Data wrangling is all about getting your data into the right form in order to feed it into the visualization and modeling stages. This typically requires a large amount of reshaping and transforming of your data. Throughout this course, you will learn the fundamental functions for “tidying” your data and for manipulating, sorting, summarizing, and joining your data. These tasks will help to significantly reduce the time you spend on the data wrangling process.

Lastly, data visualization is a key component that all data scientist needs to be fluent in. Students will become competent users of Tableau. A course of data visualization will help students spend plenty of time learning to properly visualize your data.

Pre-Requisites: Course work in Calculus, Algebra, and Programming. Intermediate programming experience in R.

Learning Objectives:

1. Gain a strong understanding of various data sources
2. Be able to do data cleaning using SQL and R
3. Become adept to industrial practices in data wrangling
4. Be able to use these skills in performing data visualization and analysis

Instructor's e-mail:

ramesh.yapalparvi@dartmouth.edu

Weekly topics

Week	Lecture	Topic	Slide Title
Week 1	**1**	Course Introduction and Specialization Introduction to Data Wrangling What is the role of data wrangling and Why Do We Do It? Data Wrangling Challenges Tools for Data Wrangling	**Introduction**
	2	What are the data sources? EMR, PRO (patient reported outcomes) File formats: JSON, XML, EXCEL, CSV, HTML, audio files. Cloud resources to databases: Amazon web services, Microsoft Azure cloud. Introduction to SQL	**Data Sources**
Week 2	**3**	***Guest Lecture by John Mecchella & Thirumal Aluka on Data Architecture and governance***	**Data Architecture & Governance**
	4	Connecting to a database using SQL server management studio (SSMS).SQL Joins, adding	**SQL**

		columns, renaming columns	
Week 3	**5**	Text mining. Using Excel to create SQL code. Data Wrangling techniques in SQL	**SQL**
	6	Dealing with Character strings	**R programming-StringR package**
Week 4	**7**	Introduction to regex (Regular expressions). Regex functions	**Regular expressions**
	8	Dealing with factors. Dealing with Dates	**Factors and dates**
Week 5	**9**	Dealing with Missing values	**Missing values**
	10	Rserve package-Integrating Tableau and R. Writing R code in Tableau	**Rserve**
Week 6	**11**	Importing data from various data formats. Scraping HTML text and table data using rvest	**Scraping HTML**
	12	Working with API's. Exporting data in R using	**API**

		readr,xlsx, r2excel packages	
Week 7	**13**	Importing data from various data formats. Scraping HTML text and table data using rvest	**Scraping HTML**
	14	Best practices in writing a R code. Loop control statement apply family functions	**Writing good code**
Week 8	**15**	Introduction to ODBC drives Setting up ODBC drivers Introduction to SQLDF package in R Running SQL queries in R	**ODBC drivers**
	16	Use of RODBC package to connect to a database in R	
Week 9	**17**	One hot encoding on categorical data	**One hot encoding**
	18	Additional topics	

Week 10	**19**	Project presentations	
	20	Project presentations continued..	

Class Climate and Inclusivity:

Everyone who meets the pre-requisites to take the class is welcome to take the course. A friendly learning environment will be fostered.

Location and Time:

TBD on Tuesday and Thursday from TBD

References:

1. Bradley C Boehmke. Data Wrangling with R, Springer, 1st ed. 2016
2. Hadley Wicham;Garrett Grolemund. R for data science, O'Reilly Media, Inc, Dec 2016
3. https://www.w3schools.com/sql/
4. R-Bloggers
5. Stack Overflow

Grades breakdown:

Exam I (midterm, take home exam) - 20%
Exam II (final, take home exam) - 20%
Homework - 20% (assigned on Friday via canvas, due next Wednesday, submitted electronically).
Team Project (presented at the last class meeting) - 30%
Quizzes-10%

Where class fits in terms of Data Science, Type and Applications?

Data Science		
Analytics	Algorithm	Inference
80	0	20
Course Type		
Theory	Methodology	Application
0	90	10
Application Area		
Application-driven	Specific	General
100	0	0

Academic Honor:

Academic integrity is at the core of our mission as mathematicians and educators, and we take it very seriously. We also believe in working and learning together.

Collaboration on homework is permitted and encouraged, but obviously it is a violation of the honor code for someone to provide the answers for you.

On written homework, you are encouraged to work together, and you may get help from others, but you must write up the answers yourself. If you are part of a group of students that produces an answer to a problem, you cannot then copy that group answer. You must write up the answer individually, in your own words.

On exams, you may not give or receive help from anyone.

Religious Observation:

Some students may wish to take part in religious observances that occur during this academic term. If you have a religious observance that conflicts with your participation in the course, please meet with me before the end of the second week of the term to discuss appropriate accommodations.

Student Accessibility and Accommodation:

Students with disabilities who may need disability-related academic adjustments and services for this course are encouraged to see me privately as early in the term as possible. Students requiring disability-related academic adjustments and services must consult the Student Accessibility

Services office in Carson Hall 125 or by phone: 646-9900 or email:
Student.Accessibility.Services@Dartmouth.edu.

Once SAS has authorized services, students must show the originally signed SAS Services and Consent Form and/or a letter on SAS letterhead to me. As a first step, if you have questions about whether you qualify to receive academic adjustments and services, you should contact the SAS office. All inquiries and discussions will remain confidential

Software:

Mac Users: You need to be able to use Microsoft SQL server management studio for working efficiently working in SQL and use of ODBC drivers which are not compatible with Mac. The alternate solution to install Parallels software and run Windows on it. The instructors are not responsible for any cost associated with software nor with any functionality issues with the software on your personal computers. It is the sole responsibility of students to have functional software on their PC. Please refer to online resources or get help from research computing