DSCI351-351M-451 SemesterProject: Overview

Roger French

04 September, 2018

Contents

3.1.1.1 Reading, Homeworks, Projects, SemProjects

- Readings:
  - Open Intro Stats, 1-1.9
- Homeworks
  - HW2 due today
- Data Science Projects:
  - First Project given out 9/19, next Tuesday
  - Due Tuesday October 3rd
  - You have two weeks to do it
- 451 SemProjects:
  - 1st Report Out is week of Sept. 26/28, 2017
- Friday Comm. Hour
  - 451 students: Come discuss your SemProjects

3.1.1.2 Syllabus

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Figure 1: DSCI351-451 Syllabus

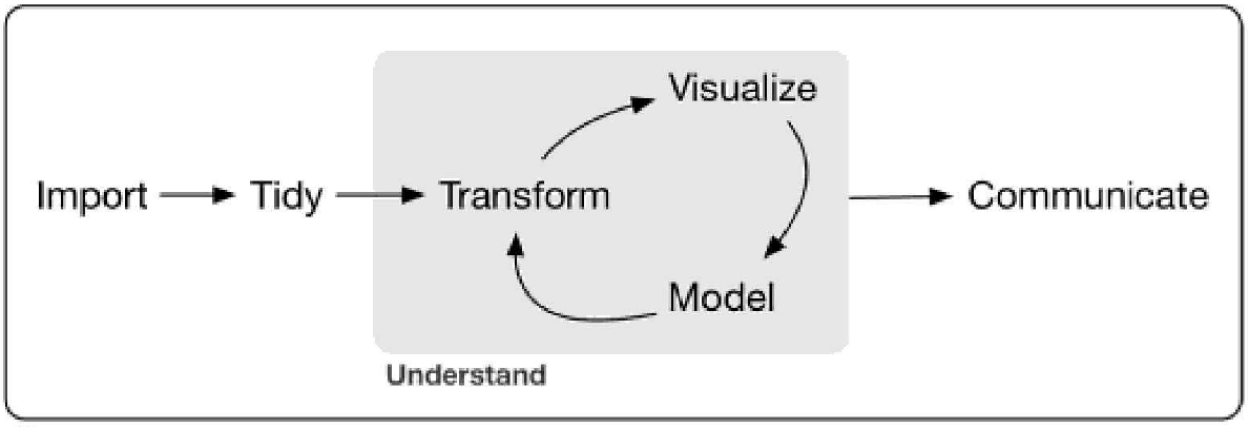

Figure 2: Data Analysis Workflow

3.1.1.3 DSCI451: Semester long Data Science Research Project

A Data Analysis/Prediction & Modeling Project

Same as what is done in DSCI 352-352M-452

3.1.1.3.1 Purpose of Semester Project Assignment

In the Semester Project, for students enrolled in DSCI451,

- you will take a four-part approach
  - to doing a data analysis through EDA and Insights
  - for a topic from your area of research interest.
- If you are able to do some modeling and statistical learning, all the better.

3.1.1.4 Basic steps we use to construct a data analysis.

Modified from Jeff Leek's slides

- (available in your repo in class/Leek)

And following Hadley Wickham's Tidyverse and R for Data Science approach

SemProj. Part a) Define Question

- Background on the research area and critical issues
- Define the question
- Define the ideal data set
- Determine what data you can access
- Define critical capabilities and identify packages you will draw upon
- Obtain the data, define you target data structure
- Clean and tidy the data

SemProj Part b) Cleaning and EDA

- Write you databook, defining variables, units and structures
- Data visualization and exploratory data analysis
- Observations of trends and functional forms
- Power transformations
- Validate with reference to domain knowledge
- Evaluate the types of Modeling Approaches to take

SemProj Part c) Modeling and Statistical Learning

- Types of modeling to try
- Statistical prediction/modeling
- Model selection
- Cross-validation, Predictive R2
- Interpret results
- Challenge results

SemProj Part d) Present your final models and learnings

- Present your results
- Present reproducible code
- Comparison to other modeling approaches in the literature

3.1.1.5 3 SemProj Report Outs and 1 Final Report

You will use .R scripts and

- do reports and presentations in .Rmd files,
- so that they are interactive, reproducible, open-science presentations.

Organize and store you code in your repo.

You will turn in all code (.R and .Rmd) and datasets as part of the project.

In Friday Community Hour we will have a SemProj discussion,

- to discuss your progress, experiences and questions.

3.1.1.6 SemProj. Report Outs

Your first SemProj. Rmd report and presentation will be

- the Week 5, September 25 and 27, 2017.

We will distribute the report out dates.

By now you've learned some R and Rmd,

- and now are either
  - defining your Semester Project topic
  - or been give access to your ODS DataPackage repo for your semester project.

So using the outlined structure in the syllabus,

- prepare a Rmarkdown document, with Rcode blocks,
- and demonstrate what types of things you have learned so far,
- and what you are thinking to pursue in your project.

3.1.1.6.1 In a data analysis this phase is

- defining good questions,
- defining the required data, and
- seeing if you have access to it.

- And what questions you might need to explore in this data.

You've also learned some R,

- from Peng's EDA with R materails,

- so you can demonstrate some of these elements also.

3.1.1.6.2 Presentation will be a total of 10 minutes

- 8 minutes
- with 2 minutes for questions from you mentor and classmates.
- Classmates must ask at least 2 questions.

3.1.1.6.3 It would be typical to present information on

- Background of the data and application
- Variables you will need to have in your dataset
- Units of these variables, and types (metadata, categorical, numerical)
- Assembly of your dataframe
- What would be the column variables in your dataframe
- What are the rows in your dataframe (the multiple simultaneous observations)
- What data cleaning and data validation do you need to consider doing
- What would be initial exploratory data analysis you should do

3.1.1.7 SemProj Report Out 2 and 3

Your second SemProj. Rmd report and presentation will be

- the Week 9b, October 25th, 2018.

Your third SemProj. Rmd report and presentation will be

- the Week 15, December 4th, and 6th, 2018.

3.1.1.8 The final SemProj. Rmd report will be

- the day of the final exam
- December 17th, 2017, by midnight.

3.1.1.9 Final Semester Project Report Structure and Format

For DSCI451, the final data science research report should be written like a scientific paper

3.1.1.9.1 Final Report Out has the following types of sections.

- Title
- Author
- Author Affilication
- License: ideally CC-BY-SA 4.0 (but a license choice is yours)
- Abstract
- IntroductionModeling
- Data Science Methods
- Exploratory Data Analysis
- Statistical Learning: Modeling & Prediction(if appropriate)
- Discussion
- Conclusions
- Acknowledgements
- References, Citations

3.1.1.9.2 Abstract

Summary of the nature, finding and meaning of your data analysis project.

3.1.1.9.3 Introduction

Background and motivation of the Data Science question

3.1.1.9.4 Data Science Methods

To be applied (such as image processing, time-series analysis, spectral analysis etc

3.1.1.9.5 Exploratory Data Analysis

Results and steps in the data analysis

3.1.1.9.6 Statistical Learning: Modeling & Prediction

If your analysis can accomplish some modeling, include it here.

3.1.1.9.7 Discussion

Discussion of the answers to the data science questions framed in the introduction

3.1.1.9.8 Conclusions

3.1.1.9.9 Acknowledgements

3.1.1.9.10 References

3.1.1.10 How to make your report

The report is done as an Rmarkdown document,

- which can be run/compiled to produce two versions of the report as a pdf.

One shows your R code and figures, and

- the other doesn't show R code, just your figures.

You'll then turn in a zip file (and leave a copy in your repo),

- with the dataset
  - (if its not to huge, if it is large, can you make a smaller dataset),
- Rmd file that works,
- and the two pdf reports.

Just do a pdf report as the final report,

- instead of a set of presentation slides.

The license choice of CC-BY-SA 4.0 is suggested

- so that others can use and build on your codes, in an open-source manner.

With more restrictive licenses,

- others won't be able to use your code in the future.