CWRU DSCI351-351M-451: Exploratory Data Science

Roger H. French, JiQi Liu

29 September, 2018

Contents

5.1.1.1 Reading, Homeworks, Projects, SemProjects

- Homework:
  - HW3 Due this morning
  -
- Readings:
  - R4DS 1-3 for today
  - R4DS Chapters 4-6 for Thursday
- Projects: We will have four 2 week EDA projects
  - Project 1 due Tuesday October 2nd, before class
- 451/352/352M/452 SemProjects:
  - SemProject Report Out 1, This Tues/Thurs

5.1.1.2 Syllabus

5.1.1.3 Data Science Twitter

5.1.1.3.1 What is WDI

- Analyzing World Bank data with WDI, googleVis Motion Charts

5.1.1.3.2 What is the Tidy-verse?

- Who is Hadley Wickham
- [Hadley Wickham (@hadleywickham) | Twitter](https://twitter.com/hadleywickham)

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

5.1.1.3.3 Other Data Scientists to Follow

- [Roger D. Peng (@rdpeng) | Twitter](https://twitter.com/rdpeng)
  - Johns Hopkins BioStats
- [Jeff Leek (@jtleek) | Twitter](https://twitter.com/jtleek)
  - Johns Hopkins BioStats
- [Mine CetinkayaRundel (@minebocek) | Twitter](https://twitter.com/minebocek?lang=en)
  - Open Intro Stats author
- [Hilary Parker (@hspter) | Twitter](https://twitter.com/hspter)
  - Biostats PhD now working in e-commerce
- [Jenny Bryan (@JennyBryan) | Twitter](https://twitter.com/JennyBryan)
  - Stats Prof at Univ. of British Columbia
- [Mara Averick (@dataandme) | Twitter](https://twitter.com/dataandme)
- [Rbloggers (@Rbloggers) | Twitter](https://twitter.com/Rbloggers)
- [Python Programming (@python_programm) | Twitter](https://twitter.com/python_programm)

5.1.1.4 We're starting to look at Distributions

- Such as the Normal Distribution
  - On which much of our statistical inference is based.
- Probability Distributions built up from samples of a population

5.1.1.5 For this we want to have a feeling for a few concepts

- An Observation
  - And a sequence of observations
- Sample Size
- Fluctuations in outcomes
  - Humans are not good at appreciating Fluctuations
- Priors, are previous observations
  - These are typically uncorrelated to the next observation
  - This is the Frequentist perspective
  - If we believe knowledge of Priors can inform us of the future
  - We are Bayesians, ie followers of Bayes' theorem

5.1.1.6 Are you a

- Frequentist
  - This is safest
- Bayesian?
  - This is coolest, but can be complicated to do correctly

5.1.1.7 Open Intro Stats Chapter 3