DSCI351-351m-451: Class 01a, (CWRU, Pitt, UCF, UTRGV)

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

27 September, 2022

Contents

1.1.1.1 Class Readings, Assignments, Syllabus Topics

1.1.1.1.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today:
  - For next class:
- Laboratory Exercises:
  - LE :
  - LE :
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Wednesday @ 4:00 PM to 5:00 PM, Will Oltjen
  - Saturday @ 3:00 PM to 4:00 PM, Kristen Hernandez
  - **Office Hours are on Zoom, and recorded**
- Semester Projects
  - DSCI 451 Students Biweekly Update 1 Due
  - DSCI 451 Students
    - Next **Report Out #1 is Due Friday September 30th**
  - All DSCI 351/351M/451 Students:
    - **Peer Grading of Report Out #1 is Due October 11th, 2022**
  - Exams
    - MidTerm: Tuesday October 18th, in class or remote, 11:30 - 12:45 PM

        - Final: Monday December 19, 2022, 12:00PM - 3:00PM, Nord 356 or remote

1.1.1.2 Textbooks

- Peng: R Programming for Data Science
- Peng: Exploratory Data Analysis with R
- Open Intro Stats, v4
- Wickham: R for Data Science
- Hastie: Intro to Statistical Learning with R, 2nd Ed.

Introduction to R and Data Science

- For R, Coding, Inferential Statistics
    - Peng: R Programming for Data Science
    - Peng: Exploratory Data Analysis with R

Textbooks for this class

- OIS = Diez, Barr, Çetinkaya-Runde: Open Intro Stat v4
- R4DS = Wickham, Grolemund: R for Data Science

Textbooks for DSCI353/353M/453, And in your Repo now

- ISLR2 = James, Witten, Hastie, Tibshirani: Intro to Statistical Learning with R 2nd Ed.
- ESL = Trevor Hastie, Tibshirani, Friedman: Elements of Statistical Learning
- DLwR = Chollet, Allaire: Deep Learning with R

Magazine Articles about Deep Learning

- DL1 to DL13 are "Deep Learning" articles in 3-readings/2-articles/

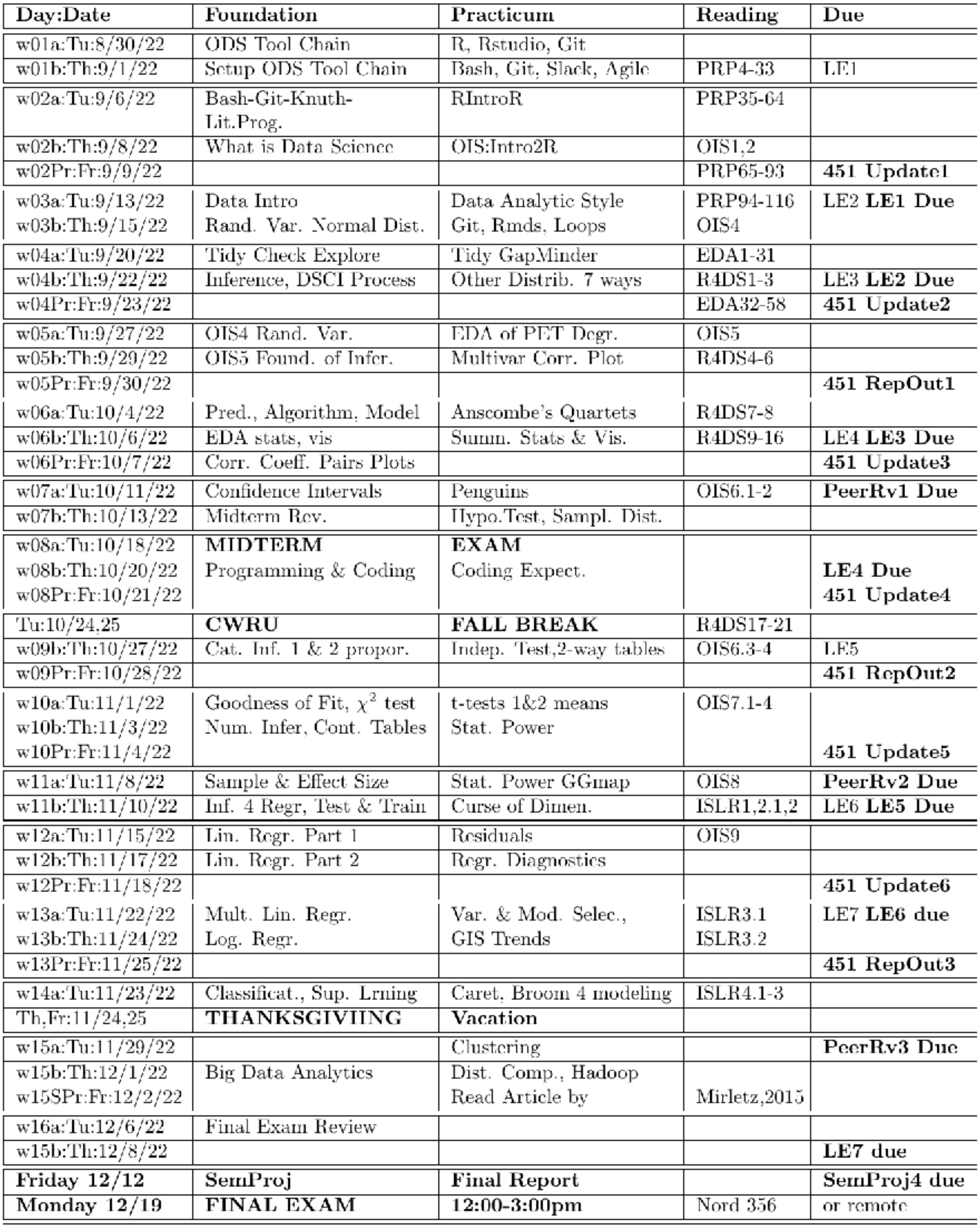
1.1.1.3 Syllabus

1.1.1.4 For the DSCI 451 students they have an EDA SemProj to do

- SemProjects:
    - SemProjects have a bi-weekly progress update
        - due Friday's at 11:59 pm (6 updates)
    - Each update should be made in the report template
        - found in the Repo with each update filled out
    - SemProj Report Out #1 Class W5, (recorded 10 min presentation)
        - Peer Grading by All DSCI 351/351m/451 students due a week later
    - SemProj Report Out #2 in Class W9 (recorded 10 min presentation)
        - Peer Grading by All DSCI 351/351m/451 students due a week later
    - SemProj Report Out #3 in Class W13 (recorded 10 min presentation)
        - Peer Grading by All DSCI 351/351m/451 students due a week later
    - SemProj Report is full comprehensive written project
        - (report template updated from each report)
        - **due Friday 12-11-2021**
- Assistance on SemProjects is done with DSCI352-352m-452 Class
    - SemProj's are taught by Prof. Laura Bruckman
    - SemProject office hours 9-10 am on Tuesdays

1.1.1.4.1 Care should be taken when choosing SemProj datasets.

- Report Out 1 focuses on
    - Explaining the 'why' of your research project
    - Describing your dataset
    - Presenting an analysis plan



Day:Date	Foundation	Practicum	Reading	Due
w01a:Tu:8/30/22	ODS Tool Chain	R, Rstudio, Git		
w01b:Th:9/1/22	Setup ODS Tool Chain	Bash, Git, Slack, Agile	PRP4-33	LE1
w02a:Tu:9/6/22	Bash-Git-Knuth-Lit.Prog.	RIntroR	PRP35-64	
w02b:Th:9/8/22	What is Data Science	OIS:Intro2R	OIS1,2	
w02Pr:Fr:9/9/22			PRP65-93	**451 Update1**
w03a:Tu:9/13/22	Data Intro	Data Analytic Style	PRP94-116	LE2 **LE1 Due**
w03b:Th:9/15/22	Rand. Var. Normal Dist.	Git, Rmds, Loops	OIS4	
w04a:Tu:9/20/22	Tidy Check Explore	Tidy GapMinder	EDA1-31	
w04b:Th:9/22/22	Inference, DSCI Process	Other Distrib. 7 ways	R4DS1-3	LE3 **LE2 Due**
w04Pr:Fr:9/23/22			EDA32-58	**451 Update2**
w05a:Tu:9/27/22	OIS4 Rand. Var.	EDA of PET Degr.	OIS5	
w05b:Th:9/29/22	OIS5 Found. of Infer.	Multivar Corr. Plot	R4DS4-6	
w05Pr:Fr:9/30/22				**451 RepOut1**
w06a:Tu:10/4/22	Pred., Algorithm, Model	Anscombe's Quartets	R4DS7-8	
w06b:Th:10/6/22	EDA stats, vis	Summ. Stats & Vis.	R4DS9-16	LE4 **LE3 Due**
w06Pr:Fr:10/7/22	Corr. Coeff. Pairs Plots			**451 Update3**
w07a:Tu:10/11/22	Confidence Intervals	Penguins	OIS6.1-2	**PeerRv1 Due**
w07b:Th:10/13/22	Midterm Rev.	Hypo.Test, Sampl. Dist.		
w08a:Tu:10/18/22	**MIDTERM**	**EXAM**		
w08b:Th:10/20/22	Programming & Coding	Coding Expect.		**LE4 Due**
w08Pr:Fr:10/21/22				**451 Update4**
Tu:10/24,25	**CWRU**	**FALL BREAK**	R4DS17-21	
w09b:Th:10/27/22	Cat. Inf. 1 & 2 propor.	Indep. Test,2-way tables	OIS6.3-4	LE5
w09Pr:Fr:10/28/22				**451 RepOut2**
w10a:Tu:11/1/22	Goodness of Fit, χ^2 test	t-tests 1&2 means	OIS7.1-4	
w10b:Th:11/3/22	Num. Infer, Cont. Tables	Stat. Power		
w10Pr:Fr:11/4/22				**451 Update5**
w11a:Tu:11/8/22	Sample & Effect Size	Stat. Power GGmap	OIS8	**PeerRv2 Due**
w11b:Th:11/10/22	Inf. 4 Regr, Test & Train	Curse of Dimen.	ISLR1,2.1,2	LE6 **LE5 Due**
w12a:Tu:11/15/22	Lin. Regr. Part 1	Residuals	OIS9	
w12b:Th:11/17/22	Lin. Regr. Part 2	Regr. Diagnostics		
w12Pr:Fr:11/18/22				**451 Update6**
w13a:Tu:11/22/22	Mult. Lin. Regr.	Var. & Mod. Selec.,	ISLR3.1	LE7 **LE6 due**
w13b:Th:11/24/22	Log. Regr.	GIS Trends	ISLR3.2	
w13Pr:Fr:11/25/22				**451 RepOut3**
w14a:Tu:11/23/22	Classificat., Sup. Lrning	Caret, Broom 4 modeling	ISLR4.1-3	
Th,Fr:11/24,25	**THANKSGIVIING**	**Vacation**		
w15a:Tu:11/29/22		Clustering		**PeerRv3 Due**
w15b:Th:12/1/22	Big Data Analytics	Dist. Comp., Hadoop		
w15SPr:Fr:12/2/22		Read Article by	Mirletz,2015	
w16a:Tu:12/6/22	Final Exam Review			
w15b:Th:12/8/22				**LE7 due**
Friday 12/12	**SemProj**	**Final Report**		**SemProj4 due**
Monday 12/19	**FINAL EXAM**	**12:00-3:00pm**	Nord 356	or remote

Figure 1: DSCI351-351M-451 Syllabus

    - Cleaning your data
- Report Out 2 focuses on:
    - EDA of your data
    - Visualizing your data
    - Further cleaning of your data
    - Reevaluation of your data analysis plan (Do you need more data?)
- Report Out 3:
    - More data visualization
    - Initial modeling
    - Conclusions about your data
    - Were you able to answer your why question?
    - What else would you need to do to get to understanding your data better?

1.1.1.5 Tidyverse Cheatsheets, Functions and Reading Your Code

- Look at the Tidyverse Cheatsheet
    - **Tidyverse For Beginners Cheatsheet**
        - In the Git/20s-dsci353-353m-453-prof/3-readings/3-CheatSheets/ folder
    - **Data Wrangling with dplyr and tidyr Cheatsheet**

  Tidyverse Functions & Conventions

    - The pipe operator `%>%`
    - Use `dplyr::filter()` to subset data row-wise.
    - Use `dplyr::arrange()` to sort the observations in a data frame
    - Use `dplyr::mutate()` to update or create new columns of a data frame
    - Use `dplyr::summarize()` to turn many observations into a single data point
    - Use `dplyr::arrange()` to change the ordering of the rows of a data frame
    - Use `dplyr::select()` to choose variables from a tibble,
        - keeps only variables you mention
    - Use `dplyr::rename()` keeps all the variables and renames variables
        - rename(iris, petal_length = Petal.Length)
    - These can be combined using `dplyr::group_by()`
        - which lets you perform operations "by group".
    - The `%in%` matches conditions provided by a vector using the c() function
    - The **forcats** package has tidyverse functions
        - for factors (categorical variables)
    - The **readr** package has tidyverse functions
        - to read_…, melt_… col_…, parse_… data and objects

Reading Your Code: Whenever you see

- The assignment operator `<-`, think **"gets"**
- The pipe operator, `%>%`, think **"then"**

1.1.1.6 Results and Observations from LEx, Exam x

1.1.1.6.1 The Median and a Standard Deviation,

- A visualization of the rank-ordered grades in points.

1.1.1.6.2 General Observations

1.1.1.6.3 Notable Solutions

1.1.1.6.4 Common Mistakes people made

1.1.1.6.5 Comments on Grading and Grading Philosophy

1.1.1.7 Topic

1.1.1.8 Links