# CWRU DSCI351-351M-451: Exploratory Data Science

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

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

08 December, 2022

## Contents

#### 16.2.1.1 Reading, Homeworks, Projects, SemProjects

- Readings:
  - For last Class: Khalilnejad article, Khalilnejad et al_2020_Automated Pipeline Framework for Processing of Large-Scale Building Energy Time.pdf
  - For Thursday: Mirletz article in 3-readings/4-MatSci-And-SemProjReadings
    - *
- Lab Exercises:
  - LE7 due Thursday December 8th
- 451 SemProjects:
  - SemProj Peer Review 3 Due this past Tuesday
  - Final full SemProject Written Report Due Friday 12/11
- Final Exam
  - Final: Monday December 19, 2022, 12:00PM - 3:00PM, Nord 356 or remote

#### 16.2.1.2 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

#### 16.2.1.2.1 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”**

### 16.2.1.3 Syllabus

### 16.2.1.4 Final Exam ( worth 20 pts)

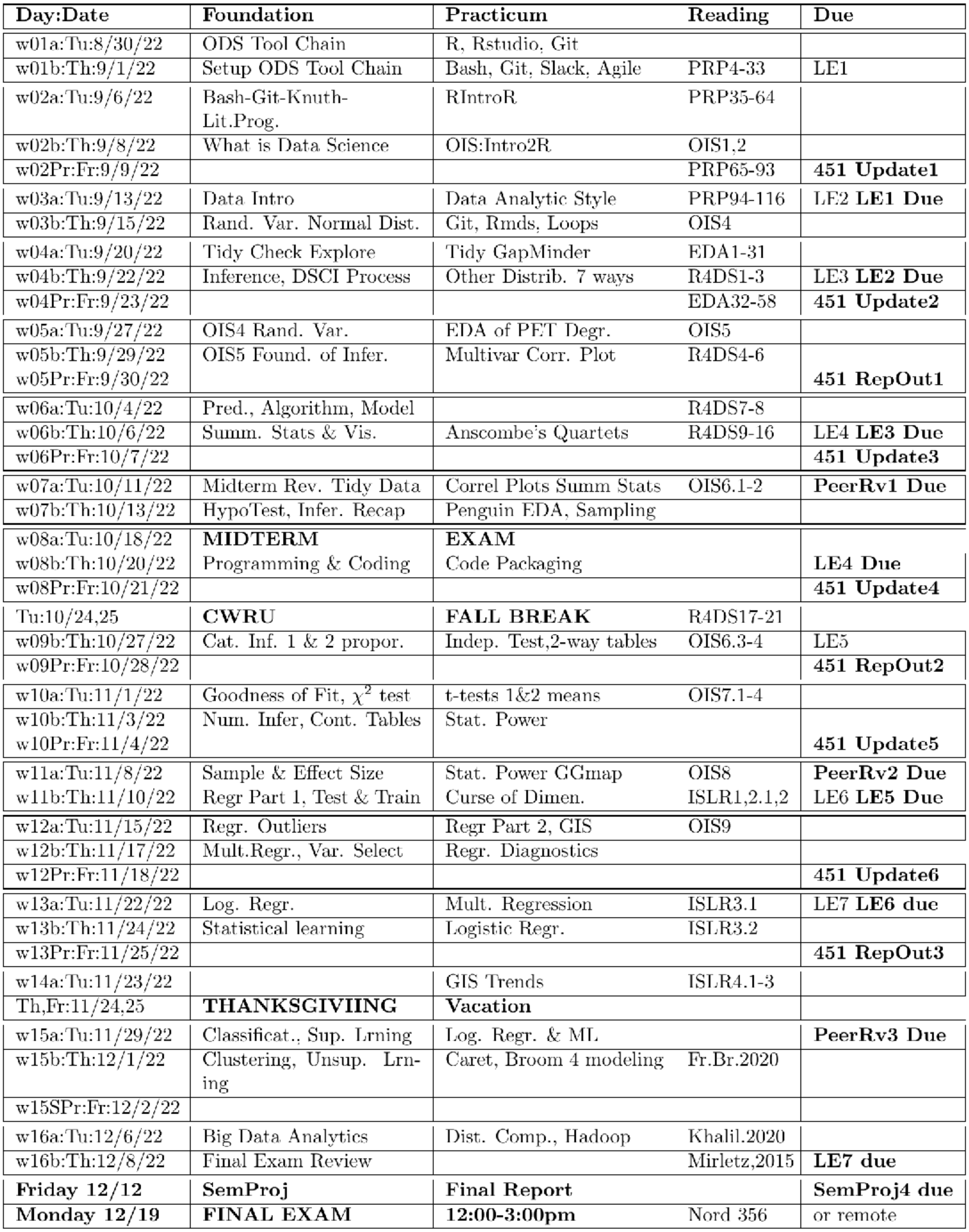
- Will be held Monday 12/13
  - From 12pm to 3pm
- Comprehensive overview of the course

#### 16.2.1.4.1 Before the final exam

- Confirm that you can
  - `git push` and `git pull` your class repo

So using the five commands on your fork of the git “...Prof” repository

- `git pull`
- `git status`
- `git add --all :/`
- `git status`
- `git commit -m 'my commit message'`



| **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 | | R4DS7-8 | |
| w06b:Th:10/6/22 | Summ. Stats & Vis. | Anscombe's Quartets | R4DS9-16 | LE4 **LE3 Due** |
| w06Pr:Fr:10/7/22 | | | | **451 Update3** |
| w07a:Tu:10/11/22 | Midterm Rev. Tidy Data | Correl Plots Summ Stats | OIS6.1-2 | **PeerRv1 Due** |
| w07b:Th:10/13/22 | HypoTest, Infer. Recap | Penguin EDA, Sampling | | |
| w08a:Tu:10/18/22 | **MIDTERM** | **EXAM** | | |
| w08b:Th:10/20/22 | Programming & Coding | Code Packaging | | **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, $\chi^2$ 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 | Regr Part 1, Test & Train | Curse of Dimen. | ISLR1,2.1,2 | LE6 **LE5 Due** |
| w12a:Tu:11/15/22 | Regr. Outliers | Regr Part 2, GIS | OIS9 | |
| w12b:Th:11/17/22 | Mult.Regr., Var. Select | Regr. Diagnostics | | |
| w12Pr:Fr:11/18/22 | | | | **451 Update6** |
| w13a:Tu:11/22/22 | Log. Regr. | Mult. Regression | ISLR3.1 | LE7 **LE6 due** |
| w13b:Th:11/24/22 | Statistical learning | Logistic Regr. | ISLR3.2 | |
| w13Pr:Fr:11/25/22 | | | | **451 RepOut3** |
| w14a:Tu:11/23/22 | | GIS Trends | ISLR4.1-3 | |
| Th,Fr:11/24,25 | **THANKSGIVIING** | **Vacation** | | |
| w15a:Tu:11/29/22 | Classificat., Sup. Lrning | Log. Regr. & ML | | **PeerRv3 Due** |
| w15b:Th:12/1/22 | Clustering, Unsup. Lrning | Caret, Broom 4 modeling | Fr.Br.2020 | |
| w15SPr:Fr:12/2/22 | | | | |
| w16a:Tu:12/6/22 | Big Data Analytics | Dist. Comp., Hadoop | Khalil.2020 | |
| w16b:Th:12/8/22 | Final Exam Review | | Mirletz,2015 | **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

- `git status`
- `git push`

### 16.2.1.4.2 Also confirm that you are running in Markov

- And confirm that you have this when you first launch your Rstudio-4.2.2 app
  - in your R console of Rstudio
- And the R version is now 4.2.2

---

initializing…

R lib path check: /home/rxf131/ondemand/ubuntu2004/r4

Time zone check: America/New_York

---

If you don't have "R lib path check:"

- With "/home/rxf131/ondemand/ubuntu2004/r4"
  - As the FIRST directory in the list
- Then you need to run the `source .....` command
  - That is in the "FixRstudioServer-R-libPaths.txt"
  - in the root directory of your class repo
- The command to run is
  - `source('/home/rxf131/ondemand/share/config/r-lib-path-fix.R')`

### 16.2.1.4.3 Final Exam Format

- The exam will appear in the prof repo
- In /assignments/finalexam folder
- Done as Rmd file to turn in as .pdf report
- Submit Final Exam .Rmd, .pdf to the Canvas Assignment Page

### 16.2.1.4.4 Types of Questions

- 8 questions total
- OI Stats questions to do
- Data Wrangling: Tidying, EDA
  - Read **Mirletz article**
- 5 Paragraph Essay Question with cites: about Data Science
  - Citations to literature supporting your discussion
    - These are done as footnotes
    - Format: Author, Title, Source:Journal,Magazine, Page, Year, URL link
- Data Analysis: Modeling using Linear Regression

### 16.2.1.4.5 Points per question

- 1. OIS 1 pt
- 2. OIS 1 pt
- 3. OIS 1 pt
- 4. Tidy data wrangling 2 pt
- 5. EDA, Summary Stats & Visualization 3 pts
- 6. 5 paragraph Essay 4 pts

- 7. EDA on Real Dataset problem 4 pts
- 8. Linear Regression on a dataset 4 pts

#### 16.2.1.5 Course Evaluations

- Please fill out and give feedback
  - On what works, what needs improvement
- Course Eval Form To Fill Out

We currently have 14% response rate

- So please go fill out the course evaluation

#### 16.2.1.6 Questions on Course

##### 16.2.1.6.1 Overarching Goal of Course

- Teach you how to do real data analysis projects
  - Using a modern data analysis tool chain
  - Using real-world and lab-based (messy) datasets
- Learn EDA to explore and discover insights from your data
  - And identify new data and metadata needed for data assembly

To achieve these goals

- What could be done better

##### 16.2.1.6.2 Utility of the 3 text books (R4DS, OIS, ISLR)

- Which did you find useful?
- Which were not useful?

##### 16.2.1.6.3 The 3 books we used

- (R4DS) R for Data Science
- (OIS) Open Intro Stats v3
- (ISLR) Introduction to Statistical Learning with Applications in R

##### 16.2.1.6.4 Git Class Repo structure to class

- This is a basic open-source collaboration method
  - did not use repo for turning in assignments
  - better by Git or by Blackboard/Canvas?

#### 16.2.1.7 Some CWRU alums in Computing

##### 16.2.1.7.1 Bill Gropp: National Center for Supercomputing Applications(NCSA)

##### 16.2.1.7.2 Donald Knuth: TeX, The Art of Computer Programming

**16.2.1.7.3 Peter Tippett: Norton Antivirus etc.** Things Tippett has done

- History & Development of Norton AntiVirus
- Verizon Data Breach Investigation Report
  - 2018 DBIR
- Veris: The Vocabulary for Event Recording and Incident Sharing
  - Veris DB, an open source database of data breaches