

Improve Institutional Research Efficiency Using R:

Data Report, Data Freeze, Web Scraping, and Visualization

Find the Slides: <https://github.com/ZhouLinli>

[RCodes4DataAnalytics/ Data_Science_Ed/ NEAIRworkshops](#)

Linli Zhou, Ph.D.

2023-07-11



Agenda

- Set up: R and RStudio
- Multiple ways of using R for IR
 - Surveys Report and Visualization
 - Program Reviews
 - Data Freezing
 - Web Scraping
 - Annual Data Report (July 20 Workshop)
- Q&A



Understanding R for Institutional Research

About Me

- Lasell University: Small private university located in Newton, MA
- **2 person** IR office serving 2000 enrolled students and 200 faculty
- A little bit of everything:
 - Survey assessment of student experiences
 - Program evaluation/ review for accreditation and planning
 - External data reporting
- R can reproduce, and document IR tasks



Setup Tips

Set up - download both



- A programming language (a language used to talk to computer for data analysis related tasks)
- The "Engine"
- An integrated development environment (IDE) for writing and executing R code
- The "Dashboard"

RStudio Tips: Layout



Screenshot of RStudio showing the layout for an R Markdown file.

The left pane shows the R Markdown source code:

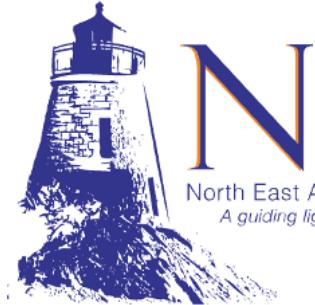
```
350 VERY SPACIAL (LONG) TO WRITE CODES| set...
351 | not...
352 | lm...
353 | D...
354 | L...
355 | L...
356 | `...
357 | Ag...
358 | Un...
359 | S...
360 | I...
361 | I...
362 | U...
363 | L...
364 | D...
365 | W...
366 | S...
367 | S...
368 | C...
369
```

The right pane shows the R console output:

```
R 4.3.0 · ~/Documents/Rprojects/IR-Projects/Data Science Ed/ →
> HERE IS THE CONSOLE TO RUN CODES WRITTEN IN THE RMD (THE LEFT) AND SHOW RESULTS
```

The bottom right pane shows the RStudio environment:

- Environment tab is selected.
- Data section lists datasets:
 - dta: 736944 obs. of 2 variables
 - ipeds.enroll: 2000 obs. of 7 variables
 - ug.t1: 6 obs. of 5 variables



RStudio Tips: Outline

.pull-left-large[

The screenshot shows an RStudio session with a dark theme. The code editor contains the following R script:

```
Untitled1* x
1
2 # Section A -----
3
4 add_one <- function(x) {
5   x + 1
6 }
7
8 # Section B -----
9
10 ad
11 ad
      Section A
      add_one(x)
```

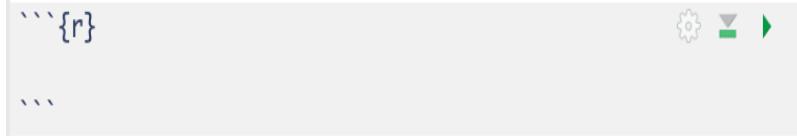
Two red arrows point to specific elements:

- A red arrow points from the bottom of the "Section A" code block to the "Section A" label in the outline panel.
- A red arrow points from the "Section B" label in the outline panel to the "Section B" code block in the editor.

The outline panel on the right side of the interface shows the following structure:

- Section A**
- add_one
- Section B**

RStudio Tips: Shortcuts



Create Code Chunk: cmd+opt+i

```

143 > ## RStudio Tips: Outline
155
156 > ---
157 > ## RStudio Tips: Shortcuts
183
184 > ---
185 > ## Starting to write codes: Types of Code Files
199
200 > ---
201 > ## Code File Setup Tip: Libraries
203
204 > ---
205 > ## Code File Setup Tip: Output
209
210 > ---
211 > # Multiple ways of Using R in the Context of IR
212
213 > ---
214 > ## Using R for Annual Data Reporting
240
241 > ---
242 > ## Leveraging R for Reviews and Surveys
268
269 > ---
270 > ## Data Freezing with R
296
297 > ---
298 > ## Web Scraping for IR Professionals
325
326 > ---
327 > ## Survey Visualization with R
352
353 > ---
354 > # Strategies to Improve Work Efficiency
378
379 > ---
380 > background-image: url("img/neairpdw.png")
381
382 > ## Conclusion and Q&A

```

Source: Posit Support, 2023

Collapse Sections: cmd+opt+o

The Source File

A file to always insert at the beginning of any R file

```
source("path/to/FrequentSetupfile")
```

What's in the Source File

Frequently use library/Rpackages

```
#common used library
#reading data
library(readxl)
#cleaning/wrangling data: include readr,tibble, stringr,forcats, dplyr, tidyR, purrr, ggplot2
library(tidyverse)
library(janitor)
library(pdftools) #read pdf
library(scales)
#text
library(tidytext)
library(rtweet)
#library(dataedu)
library(randomNames)
library(tidygraph)
library(ggraph)
#date
library(lubridate)
#save df
library(writexl)
```

What's in the Source File

Defined functions and themes

```
#defined customized theme
theme_lz <- function() {
  font <- "Helvetica" #assign font family up front
  theme_minimal() %+replace% #replace elements already strips axis lines,
  theme(
    plot.margin = margin(t = 20, r = 10, b = 40, l = 10, unit = "pt"),
    plot.margin=unit(c(0,0,
                      0,0),"cm"),#plot margin is how the whole (title,legend,viz all included,
    panel.grid.major = element_blank(), #no major gridlines
    panel.grid.minor = element_blank(), #no minor gridlines
    plot.title = element_text(family = font, size = 8, face = 'bold', hjust = 0, vjust = 0,
    plot.subtitle=element_text(size=8, hjust=0.5, face="italic", color="black"),
    axis.title = element_text(family = font, size = 9),
    axis.text = element_text(family = font, size = 9),
    axis.text.x = element_blank(),#element_text(family = font, size = 9, margin = margin(10,10,10,10)),
    axis.ticks = element_blank(), #strip axis ticks
    axis.text.y=element_blank(),
    legend.title = element_text(family = font, size=9),
    legend.margin=margin(t=-25),
```



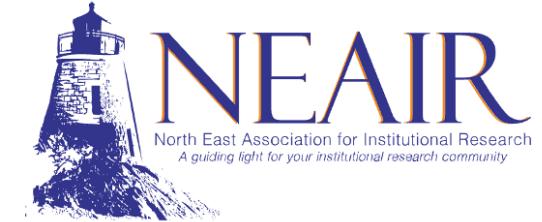
What's in the Source File

Global options

```
#Global options
knitr:::opts_chunk$set(echo = FALSE, include = FALSE, warning=FALSE, message=FALSE)
#show results only for specified chunks
#{r codechunkname, include=TRUE}

options(knitr.kable.NA = '') #in kable, show NA as blank

options(digits=1) # show 1 decimal point digits
</div>
```



Multiple ways of Using R: Survey Research



Multiple ways of Using R: **Survey Research**



Professional Survey Report



2022 Student Evaluation of Core Competencies Survey

Partial executive summary on page 1 of the survey report

Executive Summary

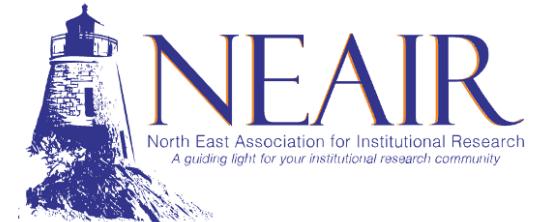
The Student Evaluation of Core Competencies Survey was collected from February 27 to March 13, 2023. Four email messages were sent to all enrolled matriculated students in 2023 Spring. We received 158 responses, representing a 14% response rate.

The survey asked for students' agreement with a few statements corresponding to each of the National Association of Colleges and Employers (NACE) competences, which include Career and Self-Development, Communication and Critical Thinking, Equity and Inclusion, Leadership/ Teamwork/ Professionalism, Technology, and Coursework.

The main findings are:

- Over 90% of students feel they have learned personally and professionally at Lasell. They can work well in teams with the necessary awareness and skills, and can use technology for tasks and goals. The most rewarding experiences for students are their internship experiences, and participation in courses and events that have a focus on real-world and career-applicable content.

Codes for PDF output: YAML



The screenshot shows an RStudio interface with the following code structure:

```
1 ---  
2 title: "BIS302 Variance Heterogeneity Practical"  
3 author: "Alex Douglas"  
4 date: "17/10/2019"  
5 output:  
6   pdf_document: default  
7   html_document: default  
8   fontsize: 11pt  
9 ---  
10  
11 Setup global options for knitr package. Normally I would not display these but I leave them here for your  
12 information. The arguments `width.cutoff` and `tidy = TRUE` keeps the displayed code within the code  
13 boxes (see what happens if you omit this).  
14 ```{r setup, include=TRUE}  
15 knitr::opts_chunk$set(echo=TRUE,tidy.opts=list(width.cutoff=55),tidy=TRUE)  
16  
17 ## Benthic Biodiversity experiment  
18 These data were obtained from a mesocosm experiment which aimed to examine the effect of benthic polychaete  
(*Nereis diversicolor*) biomass on sediment nutrient release (NH4~, NO3~ and PO3~). At the start of the  
experiment replicate mesocosms were filled with  
19 Import all the packages required for this exercise:  
20  
21 ```{r import data}  
22 nereis <- read.table("/Users/nhy163/Documents/Alex/tmp/Nereis2.txt", header = TRUE)  
23 nereis$fbiomass <- factor(nereis$biomass)  
24 str(nereis)  
25  
26  
27 3. How many replicates are there for each biomass and nutrient combination?  
28
```

Annotations on the left side of the code highlight specific sections:

- YAML header**: Points to the first 9 lines of code.
- formatted text**: Points to the explanatory text starting at line 11.
- code chunk**: Points to the code chunk starting at line 14.

Source: Douglas, et. al., 2023

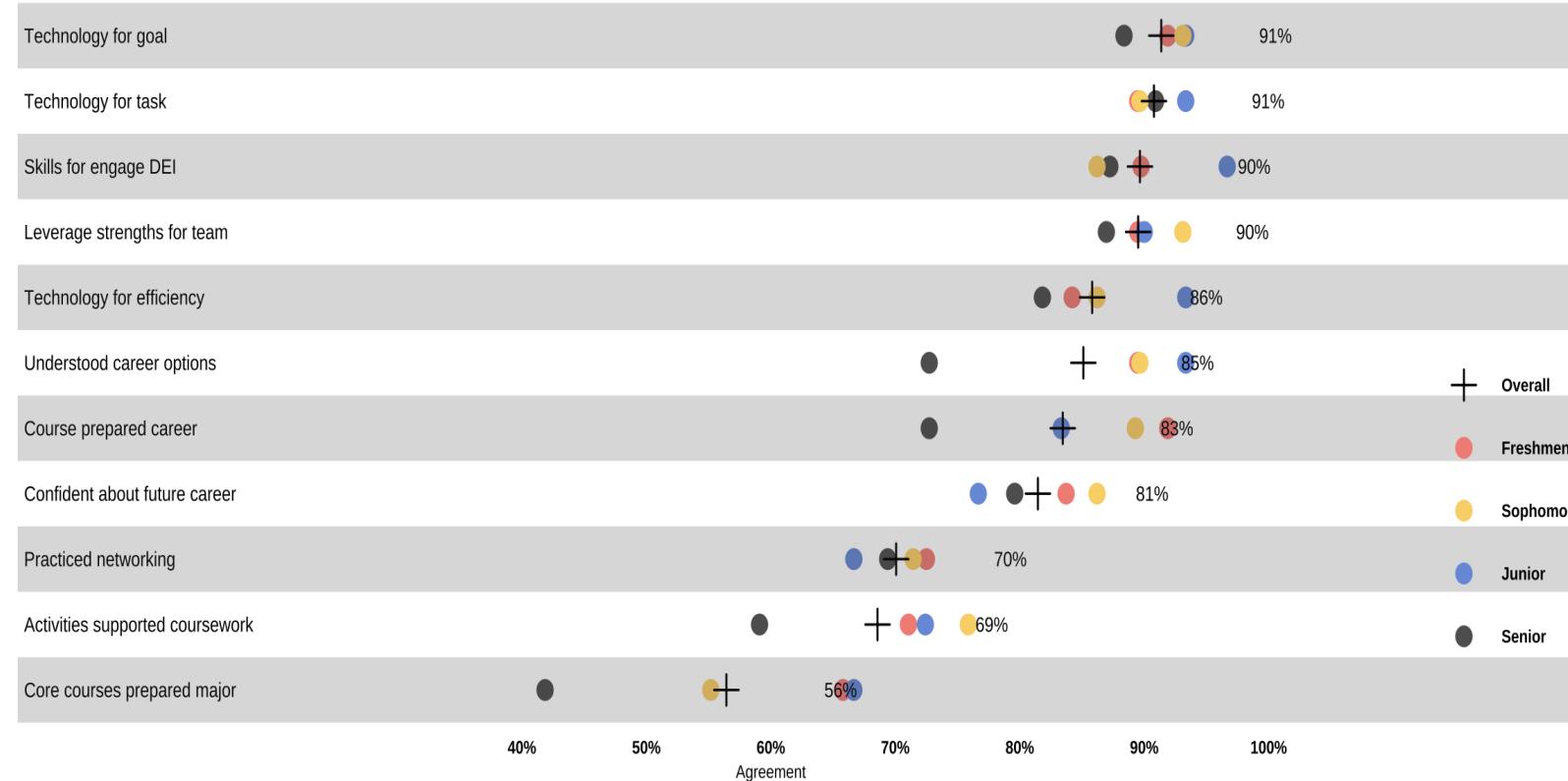
Codes for PDF output

```
title: "2022 Student Evaluation of Core Competencies Survey \\vspace{-2.9cm}"
format:
  pdf:
    pdf-engine: pdflatex
    prefer-html: true
    documentclass: article
    classoption: []
    fig-width: 8
    keep-md: true

  geometry:
    - top=25mm
    - bottom=20mm
    - left=15mm
    - right=15mm
    - textwidth=4.5in

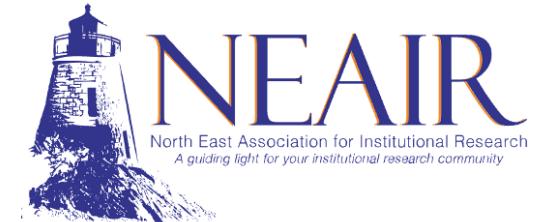
  include-in-header:
    - text: |
```

Compelling Visualization



Partial result viz
on page 2 of the
survey report

Codes for Visualization



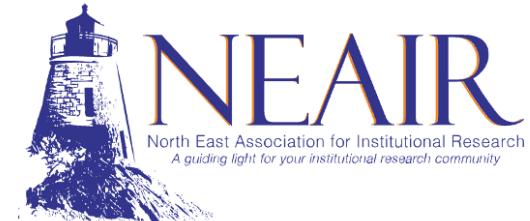
```
#create tables-----
overall.tbl<-df.l%>%filter(!is.na(gr.value))%>%
  group_by(gr.item, gr.value)%>%
  summarise(n=n())%>%mutate(prt=n/sum(n))%>%
  filter(gr.value=="Yes")%>%mutate(class_level="Overall")%>%
  select(gr.item,prt,class_level)

class.tbl<-df.l%>%filter(!is.na(gr.value))%>%
  group_by(gr.item, class_level,gr.value)%>%
  summarise(n=n())%>%mutate(prt=n/sum(n))%>%
  filter(gr.value=="Yes")%>%
  select(gr.item,prt,class_level)

tbl<-full_join(class.tbl,overall.tbl)%>%mutate(class_level=factor(class_level,levels=c("Overall","Yes")))

#set order-----
order<-tbl%>%filter(class_level=="Overall")%>%arrange(prt)
</div>
```

Codes for Visualization



```
#start building visualization-----
viz_stripped<-tbl%>%
  ggplot(aes(y=factor(agr.item,levels=order$agr.item),x=prt))+
  geom_point(aes(color=class_level,shape=class_level),
  stat = "identity",#position = position_dodge(width=0.9),
  size=2.5)+

  scale_color_manual(name="",values=c("black",color_redlight,color_yellow,color_blue_lasell,"c
  scale_shape_manual(name = "", values = c(3,16,16,16,16))+
  scale_x_continuous(limits = c(0,1.2), breaks=(seq(0.4, 1, .1)),labels=scales::percent_format

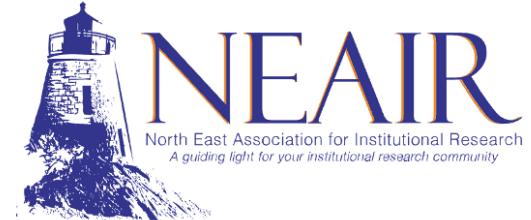
  geom_text(aes(label=if_else( class_level=="Overall", as.character(percent(prt,digits = 0)),
  annotate("text",x=rep(0,19),y=seq(1,19,1),label=order$agr.item,size=2,hjust = 0)+#labels : 

  theme_lz() + theme(
    axis.title.y = element_blank(),
    axis.title.x = element_text(size=5),
    axis.text.x = element_text(face ="bold",size=5),
    axis.text.y = element_blank(),
    legend.direction = "vertical"
```

Cleaning Comes First: Remove Columns

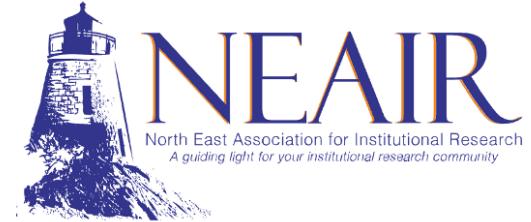
```
#load raw data
raw<-read_csv("SurveyExport.csv")

#rm unnecessary starter cols
df0<-raw%>%select(-(`Response ID`:Source),
                      -(`Invite Custom Field 2`:`Invite Custom Field 10`))%>%
janitor::clean_names()%>%unique()
df0<-df0[ , colSums(is.na(df0)) < nrow(df0)] #remove NA cols: sum each col's # of NA rows, i
</div>
```



Cleaning: Rename and Recode

```
#renaming and recoding
df.s<-df0%>%rename(agr.sec.question=long_raw_survey_question)%>%
  mutate(across(c(starts_with(c("agr")))),
    ~recode(.x, `Agree`="Yes",
      `Strongly Agree`="Yes",
      `Disagree`="No",
      `Strongly Disagree`="No")))%>%
  mutate(across(c(starts_with(c("agr"))), ~na_if(.x,"N/A")))#need to make "N/A" the real N/A
  #check: lapply(df.s%>%select(starts_with("agr")), unique)
</div>
```



Cleaning: Merge More Variables

```
# load a group variables
gp.df0<-read_excel("SP23 UG Backup Data Report.xlsx")
gp.df<-gp.df0%>%select(`People Code Id`, `Class level`, Degree, Curriculum, `College Attend`, `Ti
df.m<-left_join(df.s, gp.df, by=c("ppid"="people_code_id"))
</div>
```

Cleaning: Long Format

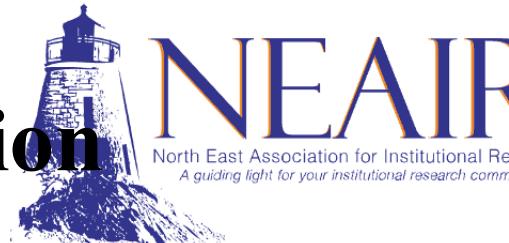
```
#longer df
df.l<-df.m%>%pivot_longer(cols=starts_with("agr"),
  names_to="agr.item",values_to="agr.value")

#easier all-at-once viz based on the long df
#df.l%>%group_by(agr.item, agr.value)%>%
#summarise(n=n())...
```

ppid	agr.career_prolearn	agr.career_strongweak	agr.career_confident
xxx54	Yes	Yes	Yes
xxx26	Yes	No	Yes
xxx66	Yes	Yes	No

ppid	agr.item	agr.value
xxx54	agr.career_prolearn	Yes
xxx54	agr.career_strongweak	Yes
xxx54	agr.career_confident	Yes
xxx26	agr.career_prolearn	Yes
xxx26	agr.career_strongweak	No
xxx26	agr.career_confident	Yes
xxx66	agr.career_prolearn	Yes
xxx66	agr.career_strongweak	Yes
xxx66	agr.career_confident	No

Cleaning: Long Format for Easier Aggregation



```
#longer df
df.l<-df.m%>%pivot_longer(cols=starts_with('agr.'))
names_to="agr.item",values_to="agr.value")

#easier all-at-once viz based on the long df
#df.l%>%group_by(agr.item, agr.value)%>%
#summarise(n=n())...
```

ppid	agr.career_prolearn	agr.career_strongweak	agr.career_confident
xxx54	Yes	Yes	Yes
xxx26	Yes	No	Yes
xxx66	Yes	Yes	No

agr.item	agr.value	n
agr.career_confident	No	1
agr.career_confident	Yes	2
agr.career_prolearn	Yes	3
agr.career_strongweak	No	1
agr.career_strongweak	Yes	2

ppid	agr.item	agr.value
xxx54	agr.career_prolearn	Yes
xxx54	agr.career_strongweak	Yes
xxx54	agr.career_confident	Yes
xxx26	agr.career_prolearn	Yes
xxx26	agr.career_strongweak	No
xxx26	agr.career_confident	Yes
xxx66	agr.career_prolearn	Yes
xxx66	agr.career_strongweak	Yes
xxx66	agr.career_confident	No

Cleaning Comes First: Save the Process

```
#save all process files: ExcelSheetName=RobjectName
write.xlsx(
  list("long"=df.l,
       "mutatedGroup"=df.m,
       "23sprRegistrarReport"=gp.df0,
       "simplified"=df.s),
  file="competencySurvey/cleandf.xlsx")
</div>
```

Extending the Survey Report: Regression

```
#different skills and characteristics
log.full<-polr(sum.tech~
  agr.career_prolearnt+agr.career_strongweak+agr.career_confident+
  agr.com_clear+agr.com_network+agr.com_seeneeds+
  agr.dei_engageskills+agr.dei_practice+agr.dei_appreciate+
  agr.team_leverage+agr.team_habit+agr.team_relation+
  agr.crs_careert+agr.crs_prep+agr.crs_core+agr.crs_activity+
  GENDER_CODE+ETHNICITY_REPORT_DESC+FA_Pell_ELIGIBLE_YN+HS_GPA+major+CLASS_LEVEL
  data=na.omit(df), Hess=TRUE)

#var selection
step.log.model <- log.full %>% stepAIC()#three left: agr.com_clear + agr.dei_engageskills +
# calculate p values and odds ratio
# p-value
coeftable<-coef(summary(step.log.model))
p <- pnorm(abs(coeftable[, "t value"]), lower.tail = FALSE) * 2
or<-exp(coef(step.log.model))
# show table
```

Program Review: Summary Tables

Table 1: AAC course enrollment by term

	Average	2018 Fall	2019 Spring	2019 Fall	2020 Spring	2020 Fall	2021 Spring	2021 Fall	2022 Spring
AAC102	56		80		43				46
AAC103	56	54	66	58	82	50	43	48	44
AAC104	50						50		

Codes for Creating the Table

```
tab.aacenroll<-aac.crs%>%group_by(coursecode,term)%>%summarise(n=n_distinct(people_code_id))
pivot_wider(names_from = term,values_from = n)

#Formatting the Table (PDF Output Only)
static_tab.aacenroll<-tab.aacenroll%>%
  kbl(align = "c",booktabs = T,label="tab.aacenroll",caption = "AAC course enrollment by term"
    kable_styling(full_width = F, font_size = 12, latex_options = c('hold_position'))%>%#set
      column_spec(1, bold = T, border_right = F, background = "white", width = "4em",color="black")
</div>
```

coursecode	term	n	Average
AAC102	2019 Spring	80	56
AAC102	2020 Spring	43	56
AAC102	2022 Spring	46	56
AAC103	2018 Fall	54	56
AAC103	2019 Fall	58	56
AAC103	2019 Spring	66	56
AAC103	2020 Fall	50	56
AAC103	2020 Spring	82	56
AAC103	2021 Fall	48	56
AAC103	2021 Spring	43	56

coursecode	Average	2018 Fall	2019 Spring	2019 Fall	2020 Spring	2020 Fall	2021 Spring	2021 Fall	2022 Spring
AAC102	56	NA	80	NA	43	NA	NA	NA	46
AAC103	56	54	66	58	82	50	43	48	44
AAC104	50	NA	NA	NA	NA	NA	50	NA	NA

Clean and Highlighted Table

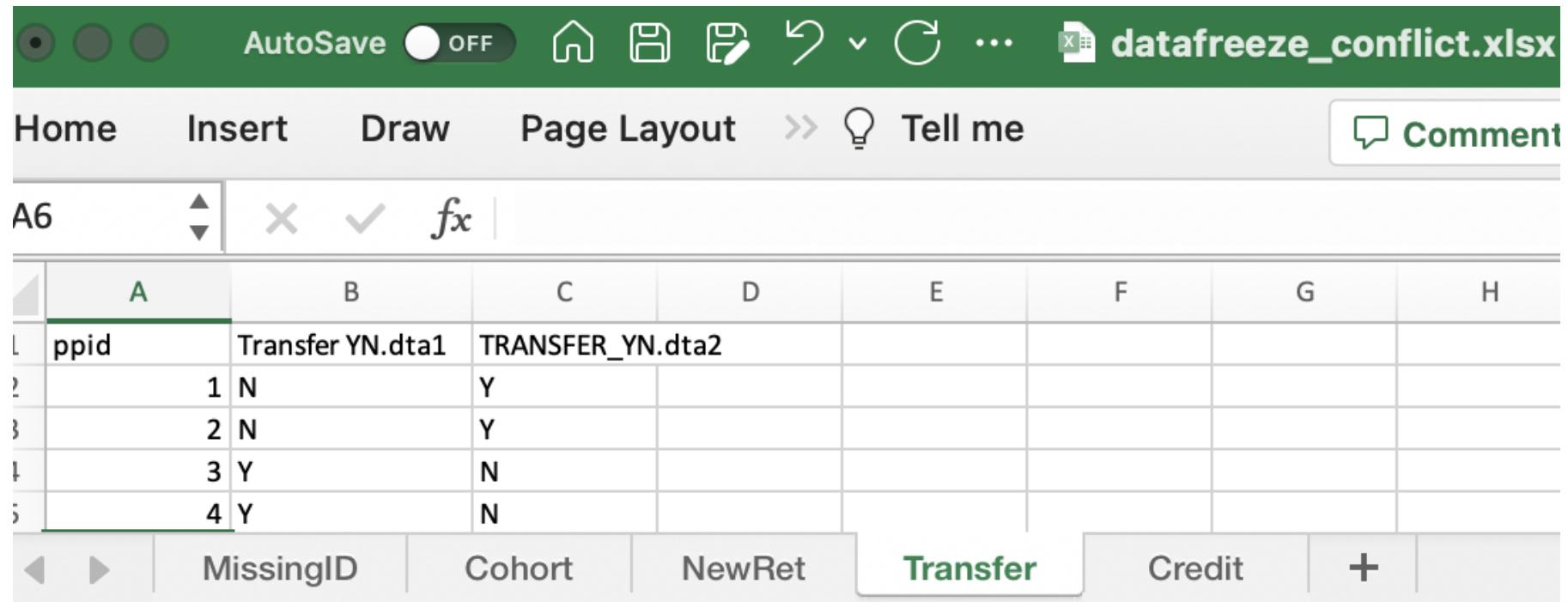
Table 1: AAC course enrollment by term

	Average	2018	2019	2019	2020	2020	2021	2021	2022
		Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
AAC102	56		80		43			46	
AAC103	56	54	66	58	82	50	43	48	44
AAC104	50					50			

```
#cf_color function defines kable condition
cf_color<-function(x,a=.9,b=.8,c=.5,d=.3,cc
  ifelse(is.na(x), "white", ifelse(
    x>a,col1,ifelse(#if not na and >90%, th
    x>b,col2,ifelse(#if <90% but >80%, th
      x>c, "white",ifelse(#if <80% but >50%
        x>d,col3,#if <50% but >30%, then
          col4))))#if <30%, then col4
```

```
#apply cf to multi columns
for (i in 3:ncol(tab.aacenroll)){
  static_tab.aacenroll<-
    column_spec(kable_input=static_tab.aacenroll,
    column=i, width = "3em",
    background =cf_color(tab.aacenroll[i],a=
    i=i+1
  static_tab.aacenroll}
```

Data Freezing: Review Consistency



The screenshot shows a Microsoft Excel spreadsheet titled "datafreeze_conflict.xlsx". The ribbon menu includes Home, Insert, Draw, Page Layout, Tell me, and Comment. The active cell is A6. The data is organized into columns A through H:

	A	B	C	D	E	F	G	H
L	ppid	Transfer YN.dta1	TRANSFER_YN.dta2					
1	1	N	Y					
2	2	N	Y					
3	3	Y	N					
4	4	Y	N					

Below the table, there are navigation buttons: MissingID, Cohort, NewRet, Transfer (highlighted in green), Credit, and a plus sign (+).

Comparing Two DataSets

```
#load data and consistency prep
dta1<-read_excel("dta1.xlsx")%>%
  rename_all(~paste0(.x,".dta1"))%>%#distinguish col from dta2
  mutate(..=recode(..))#value consistency
#dta2...

df.m<-full_join(dta1,dta2)

#find difference
conflict.major<-df.m%>%mutate(diff=ifelse(
  Curriculum.dta1==MAJOR_1.dta2, "matched","not match")) %>%
  filter(diff=="not match")%>%select(ppid,Name.rgs,Curriculum.dta1,MAJOR_1.dta2)

#save conflicts in different tab
write.xlsx(list("MissingID"=conflict.id,
  "Cohort"=conflict.cohort,
  "Level"=conflict.level,
  "Major"=conflict.major,
  "NewRet"=conflict.NewRet,
  "MajorFor"=conflict.majorFor
```



Web Scraping



Collect Web Data Efficiently

- Task: a list of 2021-2022 Courses for 43 Programs

Section	Course Title
BUSS104	Professional Development in Business
BUSS105	Excel for Business
BUSS205	Business Law



Step 1: Google a WebPage and Use "Inspect"

Screenshot of a web browser showing the Lasell University Academic Catalog page for the 2021-2022 academic year. The page displays the university's logo, navigation menu, and the title '2021 - 2022 ACADEMIC CATALOG'. Below this, there is a section for 'ACCOUNTING' with tabs for 'OVERVIEW', 'REQUIREMENTS' (which is selected), 'COURSE DESCRIPTIONS', and 'DEPARTMENT FACULTY'. At the bottom, there is a table with columns for 'COURSE CODE', 'COURSE TITLE', and 'CREDITS'. A context menu is open over the 'REQUIREMENTS' tab, showing options like 'Exit Full Screen', 'Back', 'Forward', 'Reload', 'Save As...', 'Print...', 'Cast...', 'Search Images with Google', 'Create QR Code for this Page', 'Translate to English', 'Get Image Descriptions from Google', 'View Page Source', and 'Inspect' (which is highlighted).

U Accounting – Lasell University

lasell.edu/academics/academic-catalog-and-calendar/21-22-academic-catalog/undergraduate-catalog-21-22/programs-of-study-/accounting-21-22.html#tab-2

Guest

Parents myLasell Interactive Map Apply Visit Request Info Give

LASSELL UNIVERSITY

Admissions > Discover Lasell > Tuition and Aid > Academics > Campus Life > Athletics > Alumni > Graduate Studies >

2021 - 2022 ACADEMIC CATALOG

Academics > Academic Catalog & Calendar > 21-22 Academic Catalog > Undergraduate Catalog 21-22 > Pr

ACCOUNTING

OVERVIEW REQUIREMENTS COURSE DESCRIPTIONS DEPARTMENT FACULTY

COURSE CODE	COURSE TITLE	CREDITS
School of Business Core	BUSS104	Professional Development in Business

Exit Full Screen
Back
Forward
Reload
Save As...
Print...
Cast...
Search Images with Google
Create QR Code for this Page
Translate to English
Get Image Descriptions from Google
View Page Source
Inspect

PROGRAMS OF STUDY

36 / 42



Step 2: Zoom in

Screenshot of a web browser showing the Lasell University Accounting program page. The page title is "ACCOUNTING". Below it are tabs for "OVERVIEW", "REQUIREMENTS" (which is active), "COURSE DESCRIPTIONS", and "DEPARTMENT FACULTY". The "REQUIREMENTS" section displays a table of core courses:

COURSE TITLE	
School of Business Core	
BUSS104	Professional Development in Business
BUSS105	Excel for Business
BUSS205	Business Law
BUSS220	Principles of Marketing
BUSS227	Managerial Accounting
BUSS440	Business Capstone

The browser's developer tools are open, showing the DOM structure and CSS styles for the table.

```
<nav> 600x1152
<main>
<div id="main-content">
<div class="container">
<div class="row">
<div class="col-sm-8">
<div id="tabss" class="ui-tabs ui-corner-all ui-widget ui-widget-content">
<ul role="tablist" class="ui-tabs-nav ui-corner-all ui-helper-reset ui-helper-clearfix ui-widget-header"><li><a href="#">OVERVIEW</a></li><li><a href="#">REQUIREMENTS</a></li><li><a href="#">COURSE DESCRIPTIONS</a></li><li><a href="#">DEPARTMENT FACULTY</a></li></ul>
<div id="tab-1" aria-labelledby="ui-id-1" role="tabpanel" class="ui-tabs-panel ui-corner-bottom ui-widget-content" aria-hidden="true" style="display: none;"></div>
<div id="tab-2" aria-labelledby="ui-id-2" role="tabpanel" class="ui-tabs-panel ui-corner-bottom ui-widget-content" aria-hidden="false">
<table>
<thead>
<tr><th>COURSE TITLE</th></tr>
</thead>
<tbody>
<tr><td>School of Business Core</td></tr>
<tr><td><a href="#">BUSS104</a></td><td>Professional Development in Business</td></tr>
<tr><td><a href="#">BUSS105</a></td><td>Excel for Business</td></tr>
<tr><td><a href="#">BUSS205</a></td><td>Business Law</td></tr>
<tr><td><a href="#">BUSS220</a></td><td>Principles of Marketing</td></tr>
<tr><td><a href="#">BUSS227</a></td><td>Managerial Accounting</td></tr>
<tr><td><a href="#">BUSS440</a></td><td>Business Capstone</td></tr>


Styles Computed Layout Event Listeners DOM Breakpoints Properties >  
Filter :hov .cls + □ □  
element.style {}  
div.table-wrap table { lasell.css:797  
line-height: 18px;  
min-width: 600px;  
}  
table { lasell.css:843  
border-collapse: collapse;


```



Step 3: Find the Class Name

Screenshot of a web browser showing the Lasell University Academic Catalog page for Accounting.

The URL is lasell.edu/academics/academic-catalog-and-calendar/21-22-academic-catalog/undergraduate-catalog-21-22/programs-of-study-/accounting-21-22.html#tab-2.

The page title is "ACCOUNTING".

Navigation tabs include: OVERVIEW (highlighted), REQUIREMENTS, COURSE DESCRIPTIONS, and DEPARTMENT FACULTY.

A table lists course codes and titles:

COURSE CODE	COURSE TITLE
BUSS104	Professional Development in Business
BUSS105	Excel for Business
BUSS205	Business Law
BUSS220	Principles of Marketing
BUSS227	Managerial Accounting
BUSS440	Business Capstone

The "BUSS104" link is highlighted with a blue box and a tooltip "a.mp-modal 63.73x23".

The right side of the image shows the browser's developer tools (Elements tab) with the DOM tree expanded to show the element corresponding to the "BUSS104" link.

```
<div class="row">
  <div class="col-sm-8 ">
    <div id="main" class="catalog">
      <h1 class="purple _text-centered">...</h1>
      <div class="body-copy detail">
        <div id="tabss" class="ui-tabs ui-corner-all ui-widget ui-widget-content">
          <ul role="tablist" class="ui-tabs-nav ui-corner-all ui-helper-reset ui-helper-clearfix ui-widget-header">...</ul>
          <div id="tab-1" aria-labelledby="ui-id-1" role="tabpanel" class="ui-tabs-panel ui-corner-bottom ui-widget-content" aria-hidden="true" style="display: none;">...</div>
          <div id="tab-2" aria-labelledby="ui-id-2" role="tabpanel" class="ui-tabs-panel ui-corner-bottom ui-widget-content" aria-hidden="false">
            <div class="table-wrap">
              <table>
                <thead>...</thead>
                <tbody>
                  <tr class="caption">...</tr>
                  <tr>
                    <td>== $0</td>
                    <a href="#modalx49742" class="mp-modal">BUSS104</a>
                  </td>
                  <td>Professional Development in Business</td>
                </tr>
              ...</tbody>
            </table>
          </div>
        </div>
      </div>
    </div>
  </div>

```

Styles tab shows the following CSS rule:

```
table th:first-child, table td:first-child { border-left: ▷ none; } lasell.css:907
```

Properties tab shows the following CSS rule:

```
table td { padding: ▷ 10px 15px; vertical-align: top; } lasell.css:924
```

Page footer: 38 / 42

Step 3: Find the Class Name or Full XPath

Screenshot of a web browser showing the 'Accounting – Lasell University' page at lasell.edu/academics/academic-catalog-and-calendar/21-22-academic-catalog/undergraduate-catalog-21-22/programs-of-study-/accounting-21-22.html#tab-2. The page displays a table of courses under the 'School of Business' category.

The table has two columns: 'COURSE CODE' and 'COURSE TITLE'. The 'COURSE TITLE' column contains several course names, with 'Professional Development in Business' highlighted in green. The 'School of Business' header is also highlighted in blue.

The 'Elements' tab of the developer tools is selected, showing the DOM structure. A context menu is open over the 'Professional Development in Business' cell, specifically over the text 'Professional Development in Business'. The menu includes options like 'Copy element', 'Copy full XPath', and 'Copy'. The 'Copy full XPath' option is highlighted.

The developer tools also show the CSS styles for the table and its cells, with 'lasell.css:924' and 'bootstrap.css:197' being referenced.

COURSE CODE	COURSE TITLE
School of Business	td 350x39
BUSS104	Professional Development in Business
BUSS105	Excel for Business
BUSS205	Business Law
BUSS220	Principles of Marketing
BUSS227	Managerial Accounting
BUSS440	Business Capstone
BUSS497	Business Internship & Seminar
DSCI202	Business Analytics
ECON101	Principles of Econ-Micro
MATH209	Business Statistics
Core Courses	
BUSS101	Fund of Bus in a Global Environment
BUSS203	Financial Management
BUSS226	Financial Accounting
BUSS301	Intermediate Accounting I
BUSS302	Intermediate Accounting II

R for Systematically Collecting Web Data

```
#read html and parse it into R readable contents
pg<-read_html("https://www.lasell.edu/academics/academic-catalog-21-22/undergraduate-catalog")

#collect data with the same class name
crscode<-pg %>% html_nodes(".mp-modal")%>%html_text()

#collect data with the ascending Xpath
crstitle<-lapply(1:(nrow(crs.df)+3), function(i) {pg %>% html_nodes(
xpath =paste0("/html/body/div[1]/main/div/div/div[1]/div/div/div[2]/div[1]/table/tbody/tr",
html_text()) })

#create a table with 2 columns of scrapped info
data.frame(crscode)%>%mutate(title=unlist(crstitle))%>%View()
```

```
> pg %>% html_nodes(".mp-modal")%>%
  html_text()
[1] "BUSS104"  "BUSS105"
[3] "BUSS205"  "BUSS220"
[5] "BUSS227"  "BUSS440"
[7] "BUSS497"  "DSCI102"
[9] "ECON101"   "MATH209"
[11] "BUSS101"  "BUSS203"
[13] "BUSS226"  "BUSS301"
[15] "BUSS302"  "BUSS306"
```

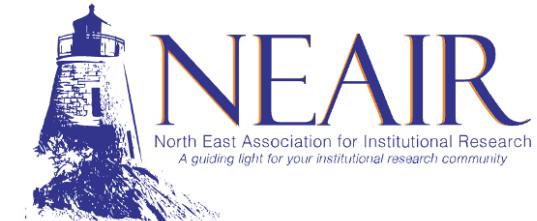
```
> lapply(1:(nrow(crs.df)+3), function(i) {
+   pg %>% html_nodes(xpath = paste0("/html/body/div[1]/main/div/div/div[1]/div/div/div[2]/div[1]/table/tbody/tr[", i, "]/td[2]")) %>%
+     html_text()
+ })
[[1]]
character(0)

[[2]]
[1] "Professional Development in Business"

[[3]]
[1] "Excel for Business"
```

Conclusion

R for IR: Strategies and Applications



- Tips for optimizing coding process
- Benefits of using R in institutional research
 - Survey Research: Professional PDF Report, Compelling Viz, Data Cleaning Process
 - Clean and highlighted summary tables
 - Discrepancy review across dataset
 - Web-scraping to efficiently collect web data
- Online resources and communities
 - youtube.com/@linlishareresearch
 - July 20 Workshop Focusing on IPEDS
- Q&A