

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

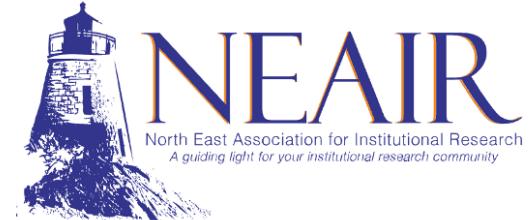
Linli Zhou, Ph.D.

2023-07-12



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



About Me

- Lasell University
- **2 person** IR office
 - Survey assessment of student experiences
 - Program evaluation/ review for accreditation and planning
 - External data reporting
- Reproducible codes



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



Review_neairworkshop1.Rmd x

Knit on Save ABC Knit Run Outline Source Visual

350 VERY SPACIAL (LONG) TO WRITE CODES
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369

set...
not...
lm...
D...
L..
L..
`...
Ag...
Un...
S...
I...
I...
U...
L...
D...
W...
S...
S...
C...

Console Render Background Jobs x

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

Environment Plots Help Tutorial Viewer Presentations

Import Dataset 27 MiB List C

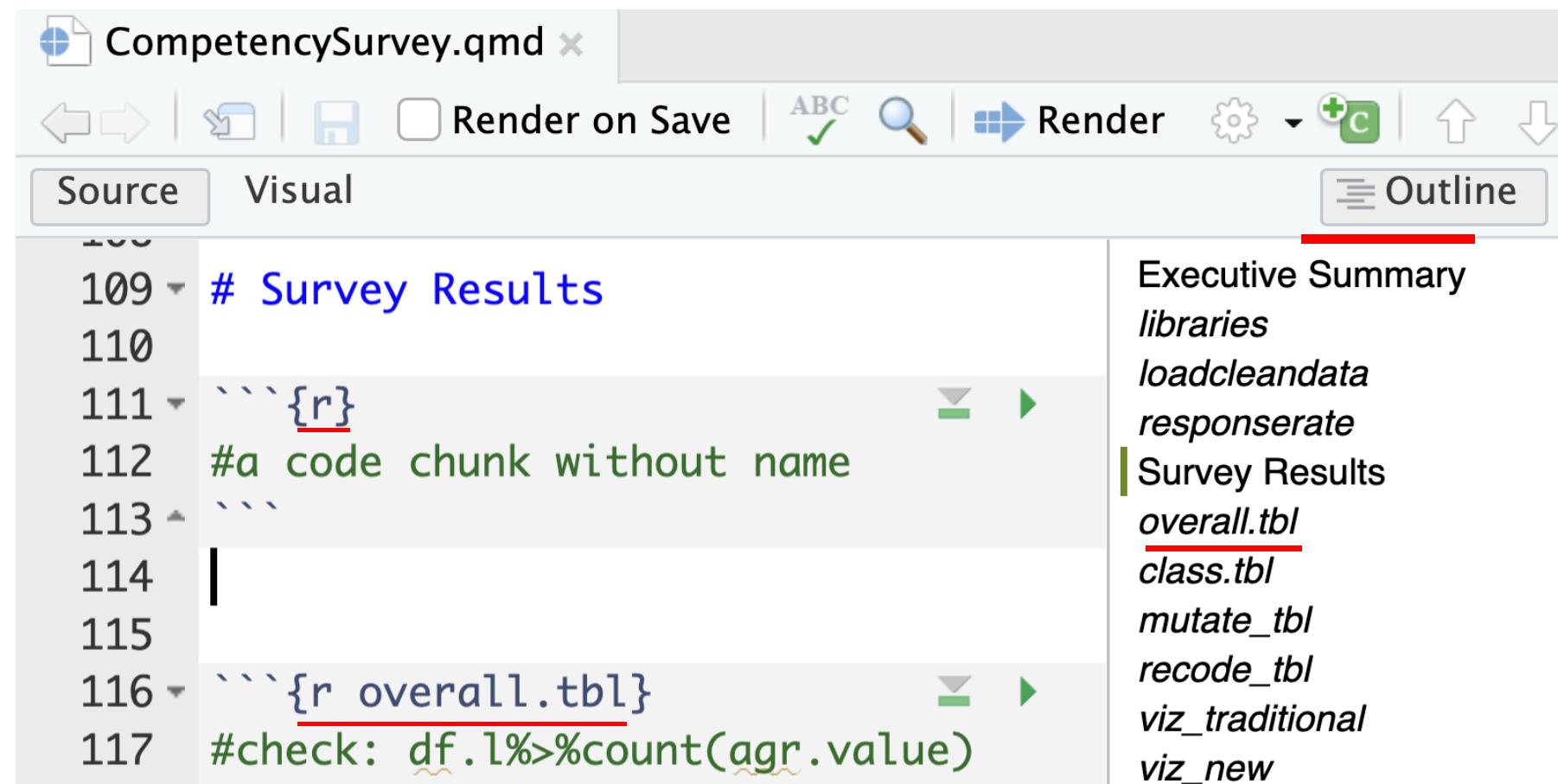
R Global Environment

Data

| | |
|--------------|----------------------------|
| dta | 736944 obs. of 2 variables |
| ipeds.enroll | 2000 obs. of 7 variables |
| ug.t1 | 6 obs. of 5 variables |

History Files Connections Packages

RStudio Tips: Outline



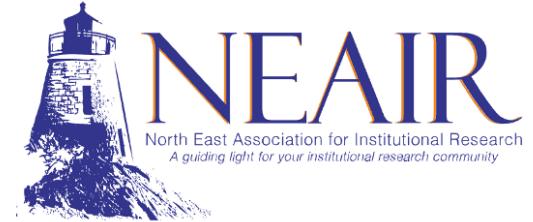
The screenshot shows the RStudio interface with the 'Outline' tab selected in the top navigation bar. The left pane displays R code, and the right pane shows the generated outline.

Code (Left Pane):

```
109 # Survey Results
110
111 ```{r}
112 #a code chunk without name
113 ```
114
115
116 ```{r overall.tbl}
117 #check: df.l%>%count(agr.value)
```

Outline (Right Pane):

- Executive Summary
- libraries*
- loadcleandata*
- responserate*
- Survey Results
- overall.tbl*
- class.tbl*
- mutate_tbl*
- recode_tbl*
- viz_traditional*
- viz_new*



RStudio Tips: Code Chunk Creator



cmd+opt+i

RStudio Tips: Collapsing and Expanding

```
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
```

cmd+opt+o

The Source File

- Frequently use library/Rpackages
- Defined functions
- Global settings

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

Programming Terms

DRY
Don't Repeat Yourself

Example: Frequently use library/Rpackages

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

Example: Defined functions

```
#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_text(family = font, size = 9, margin = margin(t=-25, r=0, b=0, l=0)),
    axis.ticks = element_blank(), #strip axis ticks
    axis.text.y=element_text(family = font, size=9),
    legend.title = element_text(family = font, size=9),
    legend.margin=margin(t=-25,
```

Example: 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}

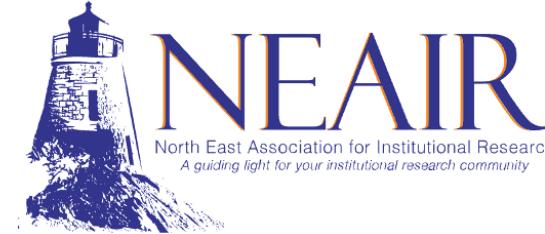
options(knitr.kable.NA = '')#in kable, show NA as blank

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



Multiple ways of Using R: Survey Research

Professional Survey Reports



2022 Student Evaluation of Core Competencies Survey

Executive Summary

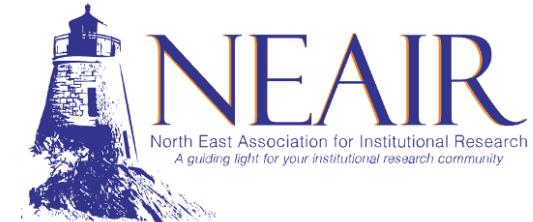
The Student Evaluation of Core Competencies Survey was collected from February 27 to March 13, 2023. Four email messages were sent to all enrolled matriculated students in 2023 Spring. We received 158 responses, representing a 14% response rate.

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.

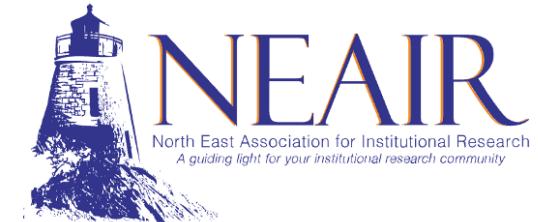
Codes for Setting PDF output



```
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: |
          \usepackage{titling}
```

YAML (Yet Another Markdown Language) Header



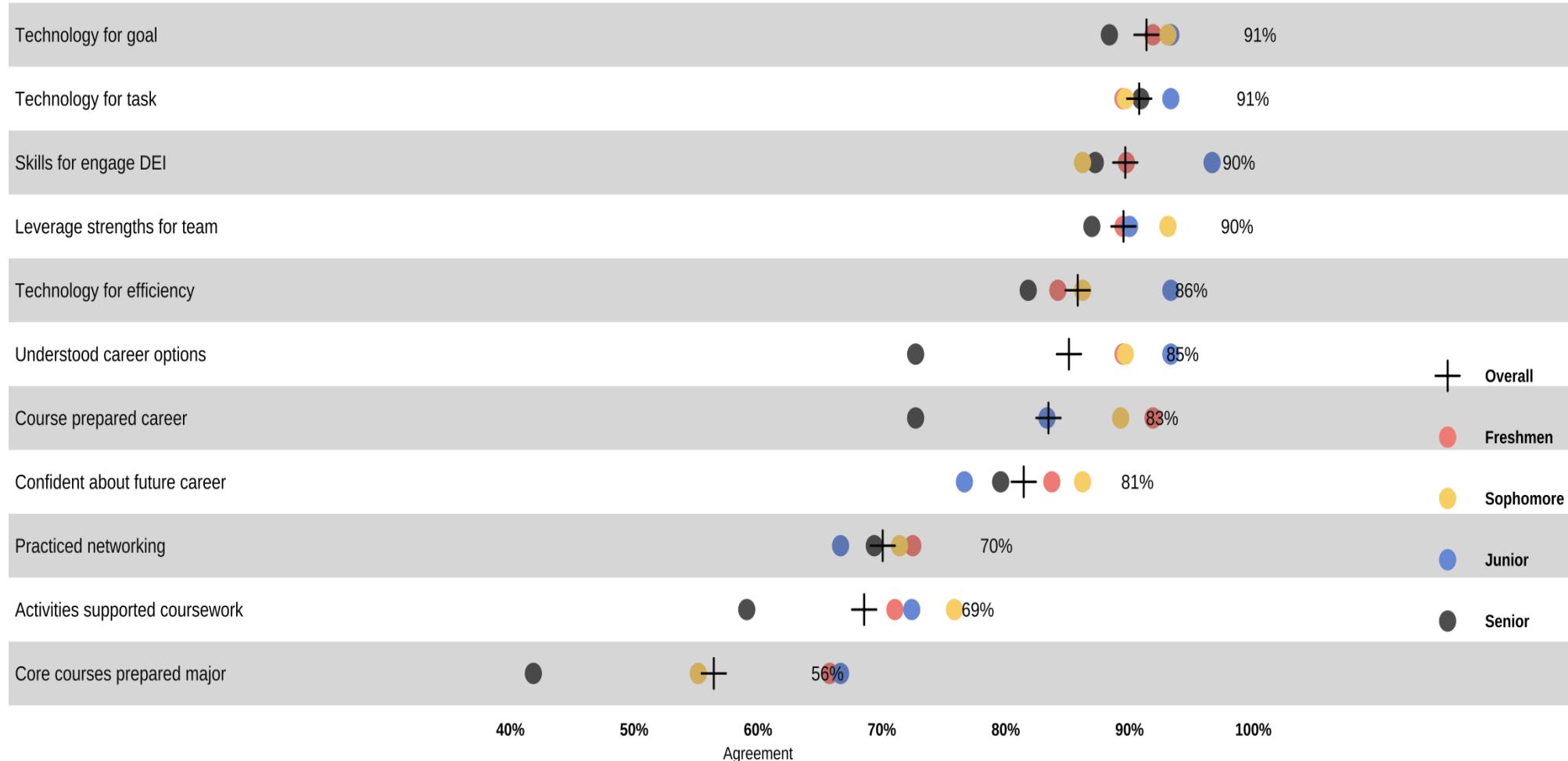
The screenshot shows an RStudio interface with an Rmd file open. The file contains the following content:

```
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 block identify specific sections:

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

Compelling Visualization



Codes for Visualization: Creating Tables

```
#students' percentage of agreement
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)
select(gr.item,prt,class_level)

#each class level's percentage of agreement
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)
```

| agr.item | prt | class_level |
|-----------------------|-----|-------------|
| agr.career_confident | 0.8 | Overall |
| agr.career_prolearn | 1.0 | Overall |
| agr.career_strongweak | 1.0 | Overall |
| agr.com_clear | 0.9 | Overall |

| agr.item | prt | class_level |
|----------------------|-----|-------------|
| agr.career_confident | 0.8 | Freshmen |
| agr.career_confident | 0.8 | Junior |
| agr.career_confident | 0.8 | Senior |
| agr.career_confident | 0.9 | Sophomore |

Codes for Visualization: Merging Tables

```
#merge into one table
tbl<-full_join(class.tbl,overall.tbl)%>%mutate(class_level=factor(class_level,levels=c("Over
```

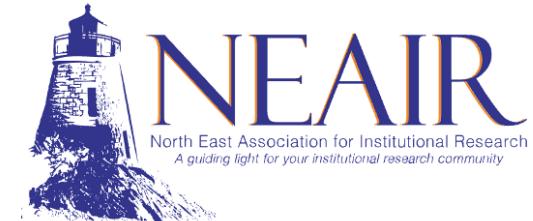
| agr.item | prt | class_level |
|----------------------|-----|-------------|
| agr.career_confident | 0.8 | Overall |
| agr.career_confident | 0.8 | Freshmen |
| agr.career_confident | 0.9 | Sophomore |
| agr.career_confident | 0.8 | Junior |

Codes for Visualization: Setting Order

```
#set order: the overall percentage of agreement
order<-tbl%>%filter(class_level=="Overall")%>%arrange(prt)
```

| agr.item | prt | class_level |
|----------------------|-----|-------------|
| agr.crs_core | 0.6 | Overall |
| agr.crs_activity | 0.7 | Overall |
| agr.com_network | 0.7 | Overall |
| agr.career_confident | 0.8 | Overall |
| agr.crs_prep | 0.8 | Overall |
| agr.crs_career | 0.9 | Overall |

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: Remove Columns

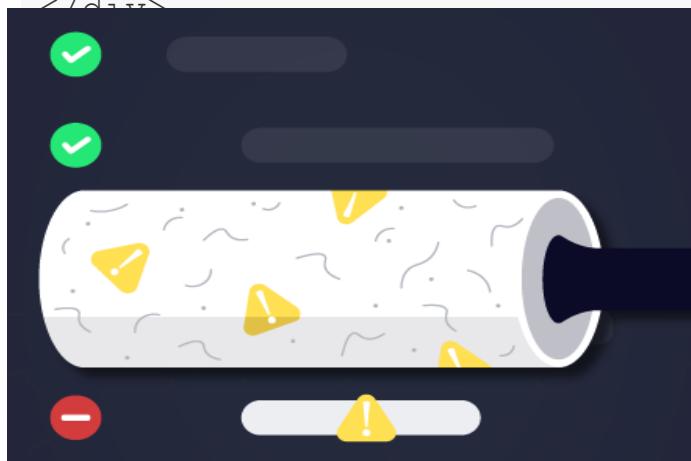
```
#load raw data
raw<-read_csv("SurveyExport.csv")

#rm unnecessary starter cols
df0<-raw%>%select(-(`Response ID`:Source),
                      -(`Invite Custom Field 2`:`Invite Custom Field 10`))%>%
  clean_names()%>%remove_empty()%>%#in the janitor package
  select(-which(colSums(is.na(.))==nrow(.)-1))%>%#remove col that has nrow-1 NA
  filter(rowSums(is.na(.)) != ncol(.) - 1) #remove row that has ncol-1 NA (keep those whose to
</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)

```



Picture source: Bellairs, 2019



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")
```

| ppid | agr.career_prolearn | agr.career_strongweak | class_level |
|-------|---------------------|-----------------------|-------------|
| xxx54 | Yes | Yes | Sophomore |
| xxx26 | Yes | No | Freshmen |
| xxx66 | Yes | Yes | Sophomore |

| ppid | class_level | agr.item | agr.value |
|-------|-------------|-----------------------|-----------|
| xxx54 | Sophomore | agr.career_prolearn | Yes |
| xxx54 | Sophomore | agr.career_strongweak | Yes |
| xxx26 | Freshmen | agr.career_prolearn | Yes |
| xxx26 | Freshmen | agr.career_strongweak | No |
| xxx66 | Sophomore | agr.career_prolearn | Yes |
| xxx66 | Sophomore | agr.career_strongweak | Yes |

Cleaning: Easier Aggregation

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

#summarize questions all-at-once
df.l%>%group_by(agr.item, agr.value)%>%summ
```

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

| ppid | agr.career_prolearn | agr.career_strongweak | class_level |
|-------|---------------------|-----------------------|-------------|
| xxx54 | Yes | Yes | Sophomore |
| xxx26 | Yes | No | Freshmen |
| xxx66 | Yes | Yes | Sophomore |

| ppid | class_level | agr.item | agr.value |
|-------|-------------|-----------------------|-----------|
| xxx54 | Sophomore | agr.career_prolearn | Yes |
| xxx54 | Sophomore | agr.career_strongweak | Yes |
| xxx26 | Freshmen | agr.career_prolearn | Yes |
| xxx26 | Freshmen | agr.career_strongweak | No |
| xxx66 | Sophomore | agr.career_prolearn | Yes |
| xxx66 | Sophomore | agr.career_strongweak | Yes |



Cleaning: Save the Process

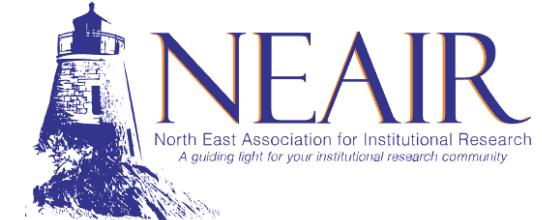
```
#save all process files: ExcelSheetName=RobjectName
write.xlsx(
  list("long"=df.l,
       "mutatedGroup"=df.m,
       "23sprRegistrarReport"=gp.df0,
       "simplified"=df.s),
  file="competencySurvey/cleandf.xlsx")
</div>
```

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
cbind(coeftable, "p value" = p, "odds ratio" = or)
```



Program Review: Summary Tables

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 | | |

Codes for Creating the Table

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

| 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 |
| AAC103 | 2022 Spring | 44 | 56 |
| AAC104 | 2021 Spring | 50 | 50 |

| coursecode | term | n | prt |
|------------|-------------|----|-----|
| AAC102 | 2019 Spring | 80 | 47% |
| AAC102 | 2020 Spring | 43 | 25% |
| AAC102 | 2022 Spring | 46 | 27% |

| term | coursecode | n | prt |
|-------------|------------|----|-----|
| 2022 Spring | AAC102 | 46 | 51% |
| 2022 Spring | AAC103 | 44 | 49% |

Codes for Creating the Table

```
tab.aacenroll<-aac.crs%>%
group_by(coursecode, term)%>%summarise(n=n_distinct(people_code_id))%>%mutate(Average=round(r
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="b
```

| 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 | | | |

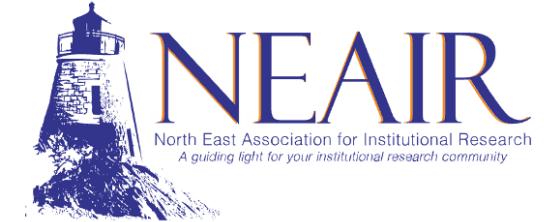
```
#cf_color function defines kable conditional formatting (colors based on condition function)
cf_color<-function(x,a=70,b=60,c=45,d=40,col1=color_yellow,col2=color_yellowlight,col3=color
ifelse(is.na(x),"white",ifelse(
  x>a,col1,ifelse(#if not na and >90%, then col1;
  x>b,col2,ifelse(#if <90% but >80%, then col2;
  x>c,"white",ifelse(#if <80% but >50%, then "white";
  x>d,col3,#if <50% but >30%, then col3;
  col4))))}#if <30%, then col4
```

Clean and Highlighted Table

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 | | | |

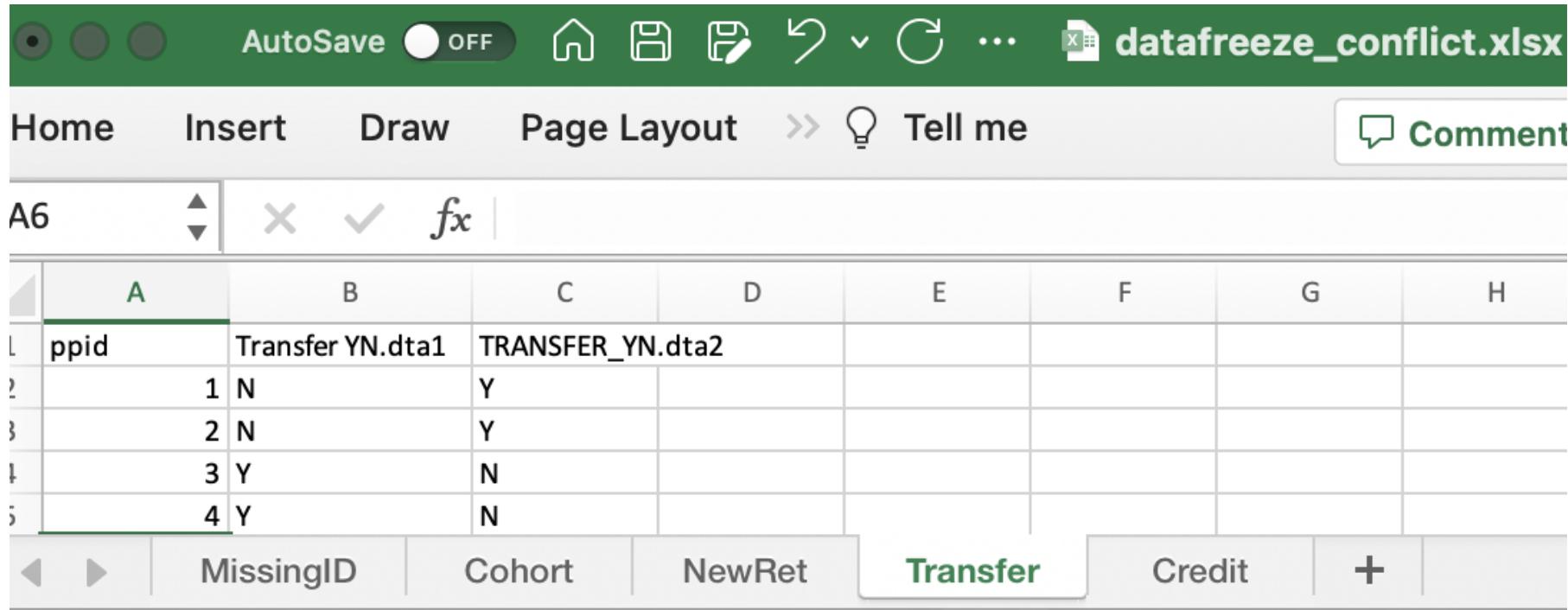
```
#apply cf to multi columns
for (i in 3:ncol(tab.aacenroll)){
  static_tab.aacenroll<-
    column_spec(kable_input=static_tab.aacenroll,
    column=i, width = "3em",
    background =cf_color(tab.aacenroll[i],a=70,b=60,c=45,d=40) )
  i=i+1
  static_tab.aacenroll}
```



Data Freezing: Review Consistency

Data freeze for Data Integrity

- Collaborative
 - IT
 - Registrar



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 three columns:

| | A | B | C | D | E | F | G | H |
|---|------|------------------|------------------|---|---|---|---|---|
| 1 | ppid | Transfer YN.dta1 | TRANSFER_YN.dta2 | | | | | |
| 2 | 1 | N | Y | | | | | |
| 3 | 2 | N | Y | | | | | |
| 4 | 3 | Y | N | | | | | |
| 5 | 4 | Y | N | | | | | |

Below the table, there are tabs for MissingID, Cohort, NewRet, Transfer (which is 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)
```

Comparing Two DataSets: If-else

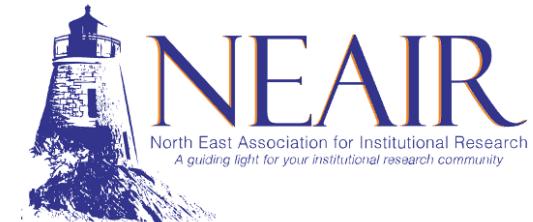
```
#find difference
conflict.major<-df.m%>%mutate(diff=ifelse(
  Curriculum.dta1==MAJOR_1.dta2, "matched", "not match")) %>%
  filter(diff=="not match")%>%select(ppid,Name.rgs,Curriculum.dta1,MAJOR_1.dta2)
```

Comparing Two DataSets: Saving

```
#save conflicts in different tab
write.xlsx(list("MissingID"=conflict.id,
                "Cohort"=conflict.cohort,
                "Level"=conflict.level,
                "Major"=conflict.major,#
                "NewRet"=conflict.NewRet,
                "Transfer"=conflict.transfer,
                "Credit"=conflict.credit),
           file="datafreezeConflict_rd1.xlsx")
```



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: "Inspect" in Chrome

Screenshot of a web browser showing the Lasell University Academic Catalog page for the 2021-2022 year. The page displays the university's logo, navigation menu, and a section for the Accounting program.

The browser's context menu is open over the "Inspect" option in the bottom right corner of the page content area.

Page Content (Visible Elements):

- Header: Accounting – Lasell University
- Header: lasell.edu/academics/academic-catalog-and-calendar/21-22-academic-catalog/undergraduate-catalog-21-22/programs-of-study-/accounting-21-22.html#tab-2
- Header Buttons: Parents, myLasell, Interactive Map, Apply, Visit, Request Info, Give
- Logo: LASSELL UNIVERSITY
- Navigation: Admissions > Discover Lasell > Tuition and Aid > Academics > Campus Life > Athletics > Alumni > Graduate Studies >
- Search: magnifying glass icon
- Title: 2021 - 2022 ACADEMIC CATALOG
- Breadcrumbs: Academics > Academic Catalog & Calendar > 21-22 Academic Catalog > Undergraduate Catalog 21-22 > Programs of Study > Accounting
- Section: ACCOUNTING
- Buttons: OVERVIEW (highlighted), REQUIREMENTS, COURSE DESCRIPTIONS, DEPARTMENT FACULTY
- Table Headers: COURSE CODE, COURSE TITLE, CREDITS
- Table Row: School of Business Core, BUSS104, Professional Development in Business

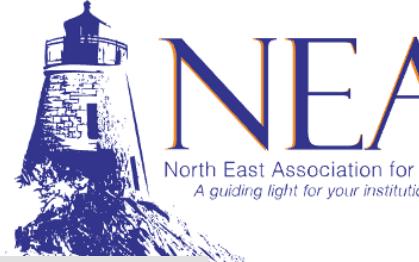
Context Menu (Open):

- 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

Page Footer:

PROGRAMS OF STUDY

44 / 51



Step 2: Zoom in

Screenshot of a web browser showing the Lasell University Accounting program page. The browser's developer tools (Elements tab) are open, highlighting the HTML structure of the page, specifically focusing on the course table.

The page title is "Accounting – Lasell University". 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 header includes the Lasell University logo and navigation links for the Academic Catalog & Calendar, Undergraduate Catalog 21-22, Programs of Study, and Accounting 21-22.

The main content area is titled "ACCOUNTING" and contains tabs for "OVERVIEW", "REQUIREMENTS", "COURSE DESCRIPTIONS", and "DEPARTMENT FACULTY". The "REQUIREMENTS" tab is active.

A table of courses is displayed under the "COURSE DESCRIPTIONS" tab:

| 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 developer tools show the HTML structure of the page, including the table element and its rows and columns. The CSS styles applied to the table are also visible, such as "line-height: 18px;" and "min-width: 600px;".

Step 3: Find the Class Name



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

The URL in the address bar 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 page is being viewed in a browser with developer tools open, specifically the Elements tab of the DevTools panel.

The DOM tree shows the structure of the page, including the main content area and the table of courses.

Styles tab of the DevTools panel shows the following CSS rule for the first table row:

```
table th:first-child, table td:first-child { border-left: none; }
```

Properties tab of the DevTools panel shows the following CSS rule for table cells:

```
table td { padding: 10px 15px; vertical-align: top; }
```

Page footer indicates "46 / 51".

Step 3: Find the Class Name or Full XPath

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

The page displays a table of courses under the "School of Business". The first course, "Professional Development in Business" (BUSS104), is highlighted with a green background.

The browser's developer tools (Elements tab) are open, showing the DOM structure. A context menu is displayed over the highlighted course cell, specifically over the "Professional Development in Business" text. The menu includes options like "Copy element", "Copy full XPath", and "Copy".

The "Copy full XPath" option is highlighted in the context menu.

The developer tools also show the CSS styles applied to the table cells:

```

table td {
    padding: 10px 15px;
    vertical-align: top;
}
td, th {
    padding: 0;
}

```

The URL in the browser address bar is: <https://lasell.edu/academics/academic-catalog-and-calendar/21-22-academic-catalog/undergraduate-catalog-21-22/programs-of-study-/accounting-21-22.html#tab-2>

Patterns of XPath

Xpath for the 1st course title:

/html/body/div[1]/main/div/div/div[1]/div/div/div/div[2]/div[1]/table/tbody/tr[**2**]/td[2]

Xpath for the 2nd course title:

/html/body/div[1]/main/div/div/div[1]/div/div/div/div[2]/div[1]/table/tbody/tr[**3**]/td[2]



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
col1<-pg %>% html_nodes(".mp-modal")%>%html_text()
crs.df<-data.frame(col1)
```

```
#collect data with the ascending Xpath
col2<-lapply(1:(nrow(crs.df)+3), function(i) {pg %>% html_nodes(xpath =paste0("/html/body/div[", i,
"]/td[2]")) %>
html_text()})

#merge
crs.df<-crs.df%>%mutate(title=unlist(col2))
```

```
> col1
```

```
[1] "BUSS104"  "BUSS105"  "BUSS205"  "BUSS220"
[5] "BUSS227"  "BUSS440"  "BUSS497"  "DSCI202"
[9] "ECON101"   "MATH209"  "BUSS101"  "BUSS203"
```

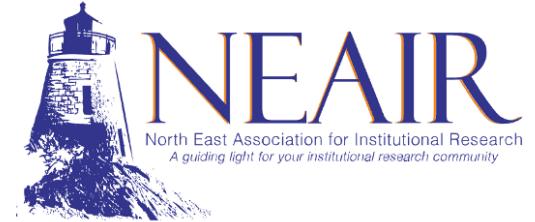
```
> col2
```

```
[[1]]
character(0)
```

```
[[2]]
```

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

```
[[3]]
```



Conclusion

R for IR



- Tips for optimizing coding process
 - RStudio Setup
 - Source File
- Applications
 - Survey Research: Reports, Viz, Cleaning, Reg
 - Summary tables
 - Compare datasets
 - Web-scrappling
- Online resources and communities
 - www.youtube.com/@linlishareresearch
 - July 20 Workshop Focusing on IPEDS
 - lzhou@lasell.edu

