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

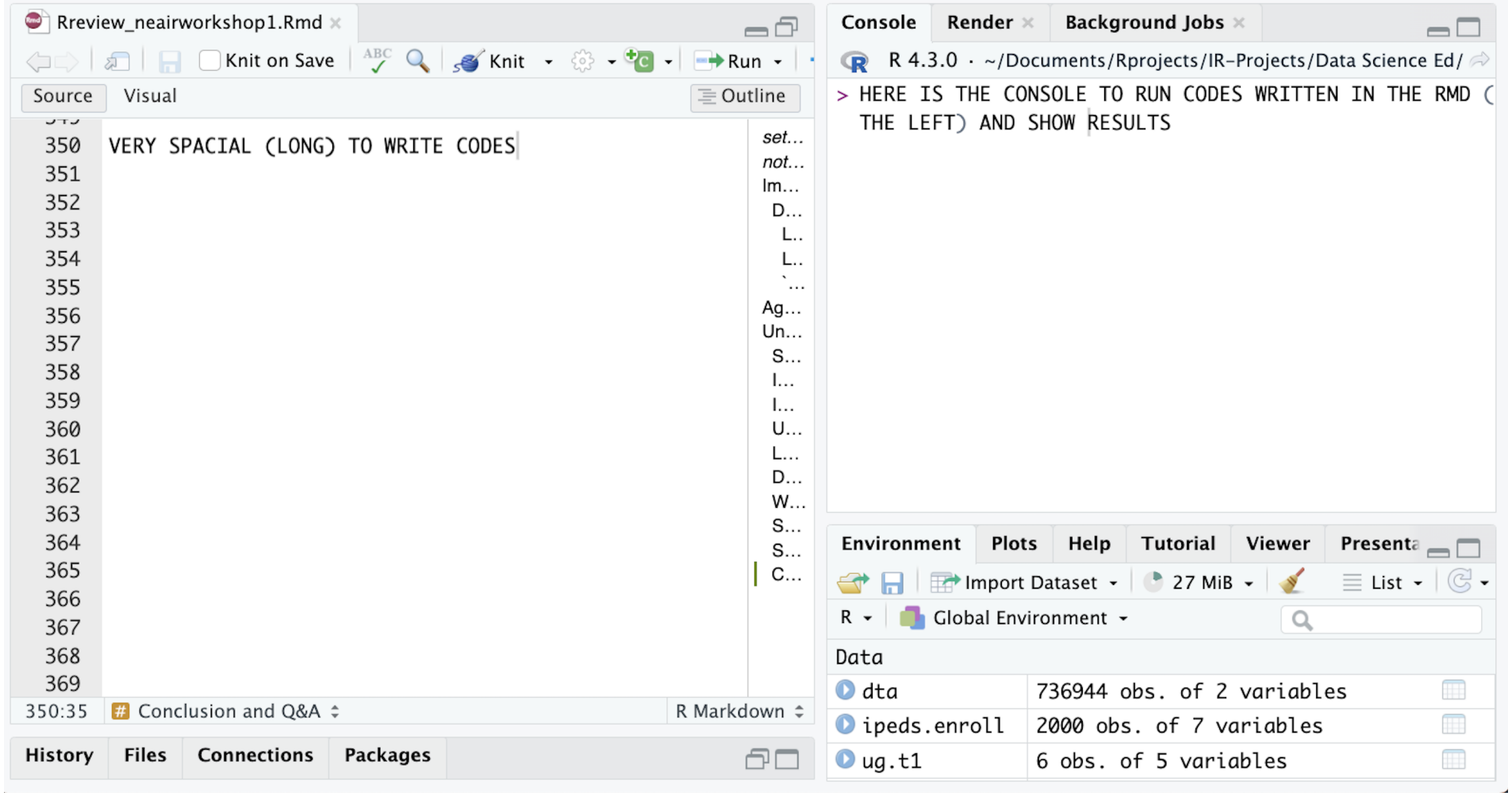
Set up - download both

- 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"

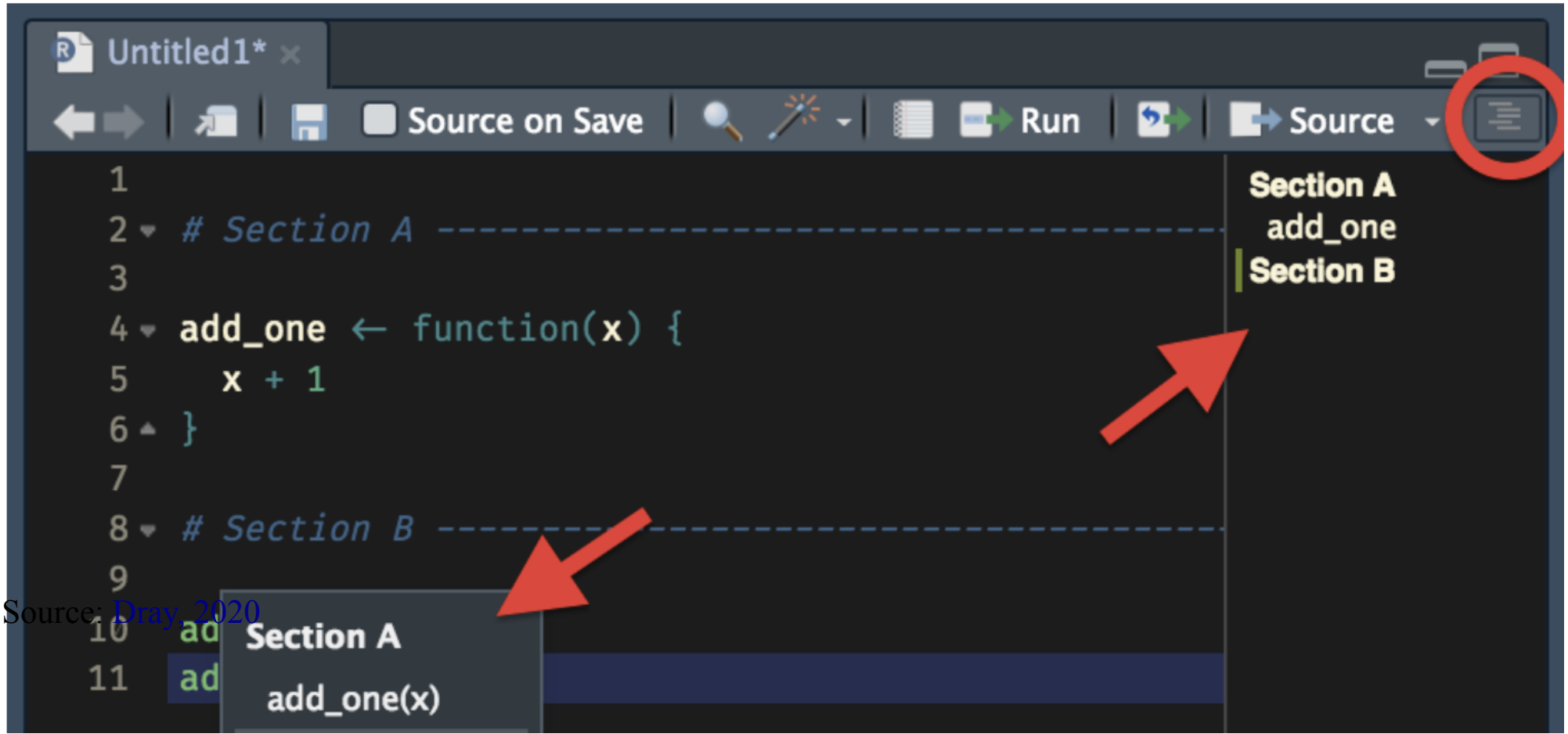
10:00

RStudio Tips: Layout



RStudio Tips: Outline

.pull-left-large[

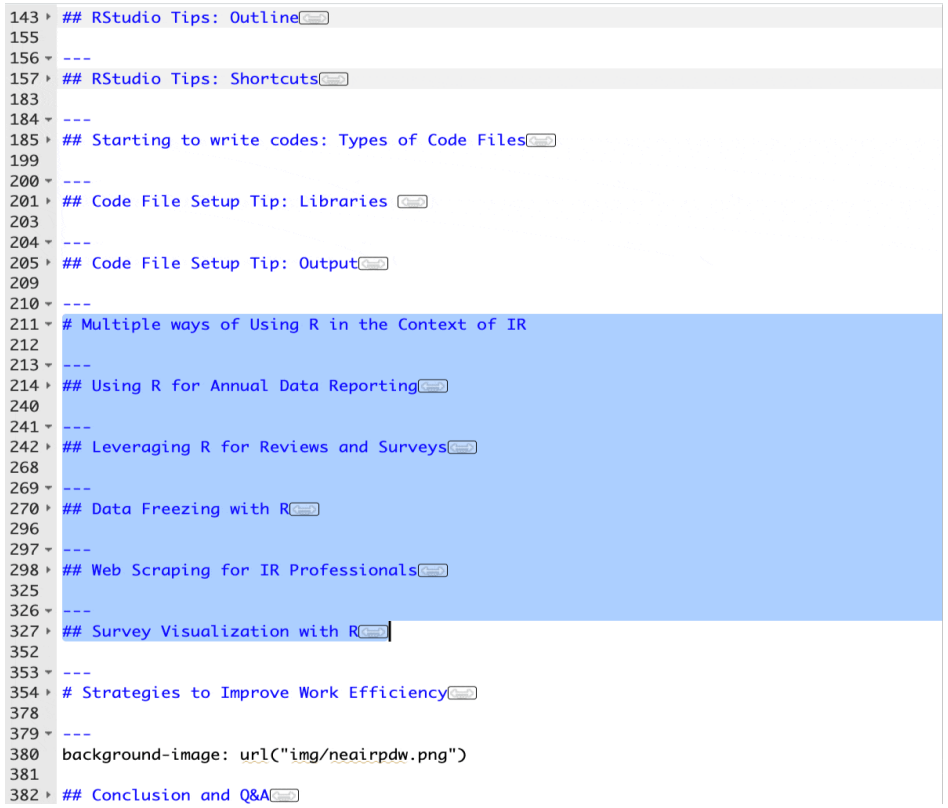


Source: Dray, 2020

RStudio Tips: Shortcuts

Create Code Chunk: cmd+opt+i

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, ggplo
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(t
      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

LASELL UNIVERSITY
INSTITUTIONAL RESEARCH

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.

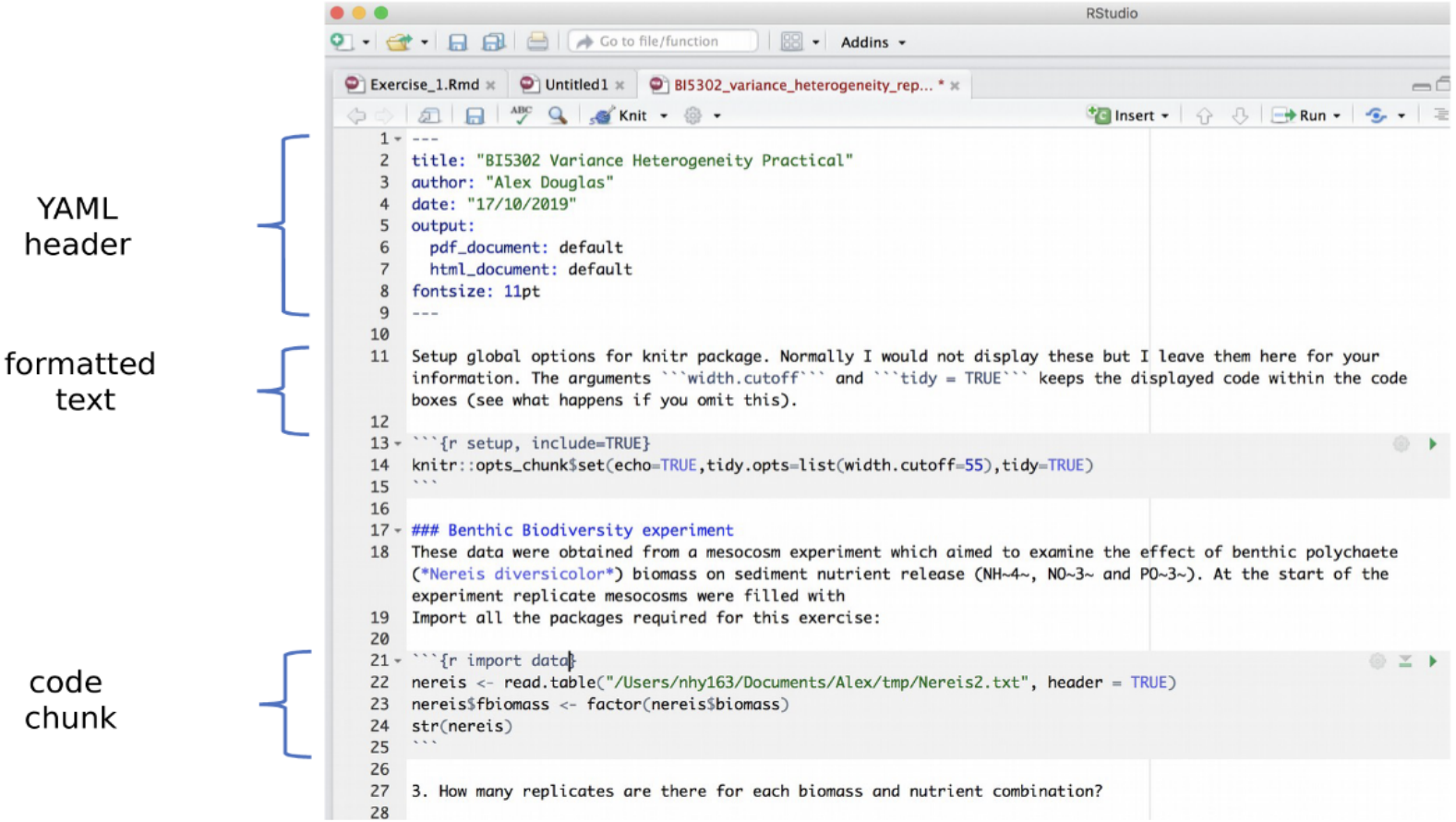
The survey asked for students' agreement with a few statements corresponding to each of the National Association of Colleges and Employers (NACE) competences, which include Career and Self-Development, Communication and Critical Thinking, Equity and Inclusion, Leadership/ Teamwork/ Professionalism, Technology,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

Partial executive summary on page 1 of the survey report

Codes for PDF output: YAML



Source: Douglas, et. al., 2023

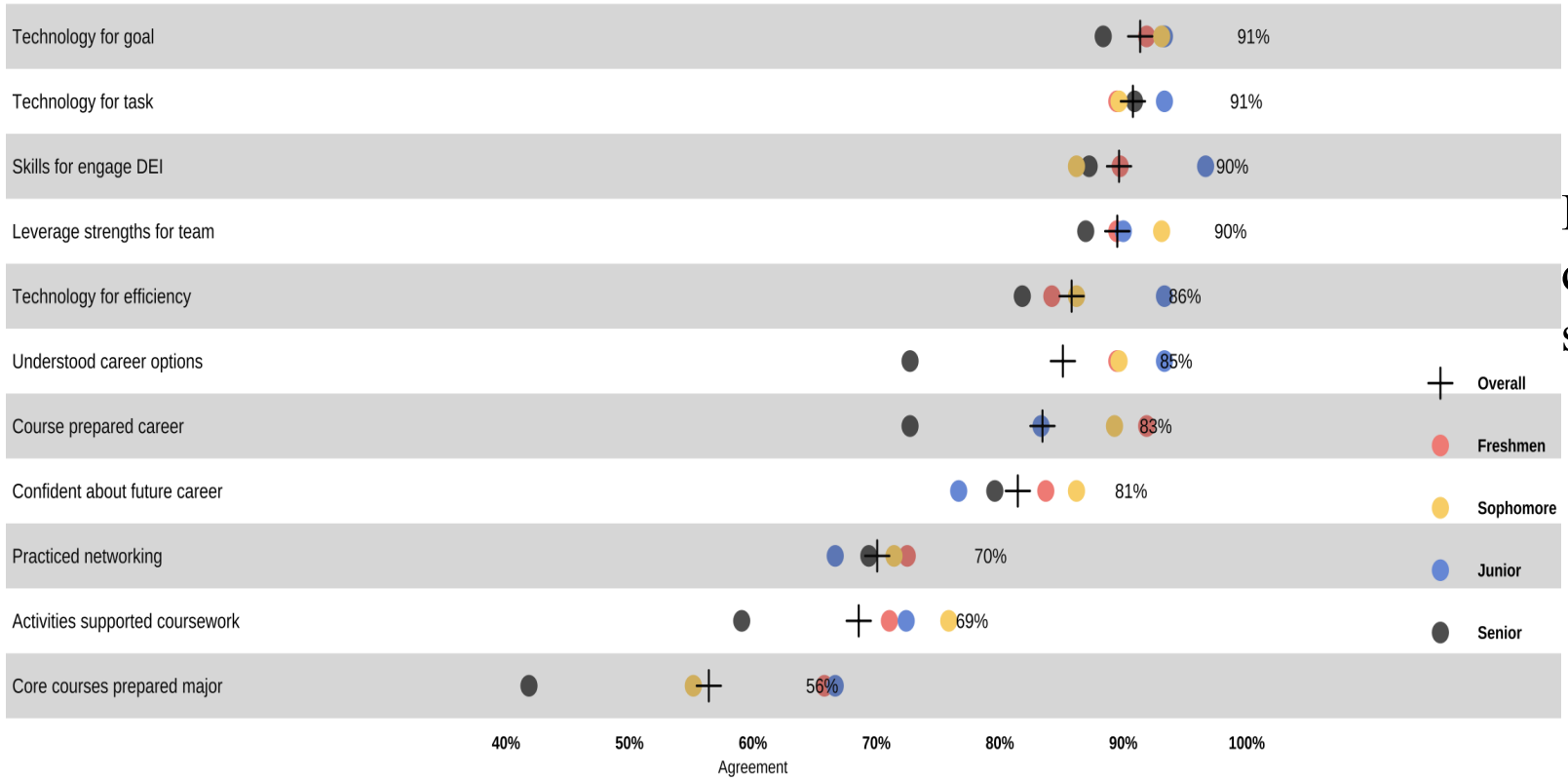
Codes for PDF output

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

Compelling Visualization



Partial result viz on page 2 of the survey report

Codes for Visualization

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

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

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

#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,"g
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,
</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 NA
  #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`,`Tr
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(
 names_to="agr.item",values_to="agr.value")

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

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

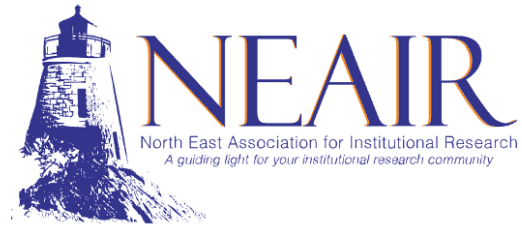
ppid	agr.career_prolearn	agr.career_strongweak	agr.career_confident
xxx54	Yes	Yes	Yes
xxx26	Yes	No	Yes
xxx66	Yes	Yes	No

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 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_prolearn+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_career+agr.crs_prep+agr.crs_core+agr.crs_activity+
                GENDER_CODE+ETHNICITY_REPORT_DESC+FA_Pell_ELIGIBLE_YN+HS_GPA+major+CLASS_LEV
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'))%>%#se
    column_spec(1, bold = T, border_right = F, background = "white", width = "4em",color="bl
</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

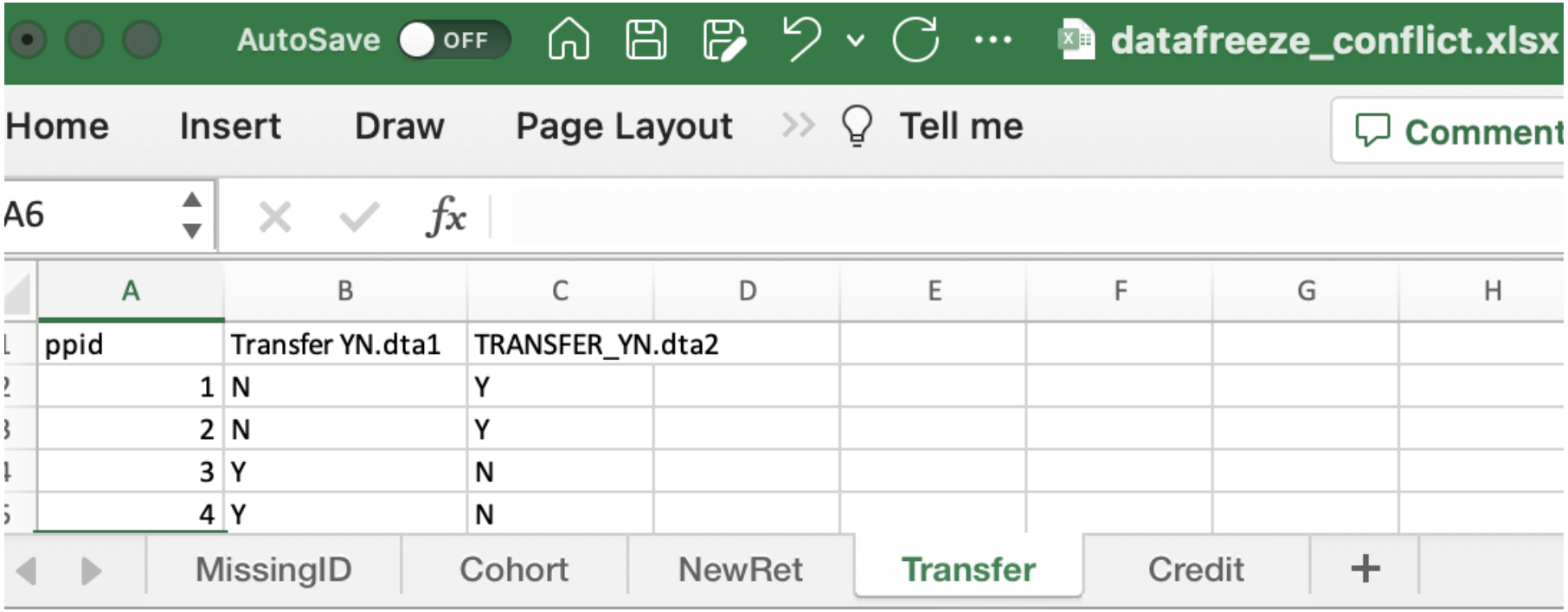
Table 1: AAC course enrollment by term

	Average	2018 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		

```
#cf_corlor function defines kable conditior
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.aacenr
  column=i, width = "3em",
   background =cf_color(tab.aacenroll[i],a=
i=i+1
static_tab.aacenroll}
```

Data Freezing: Review Consistency



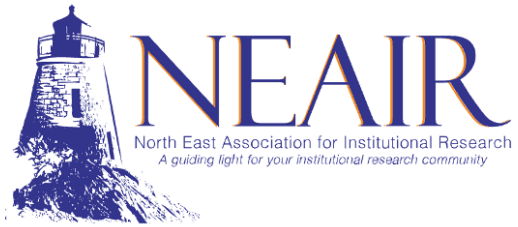
Comparing Two DataSets

```
#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,
                "Transfer"=conflict.transfer
```



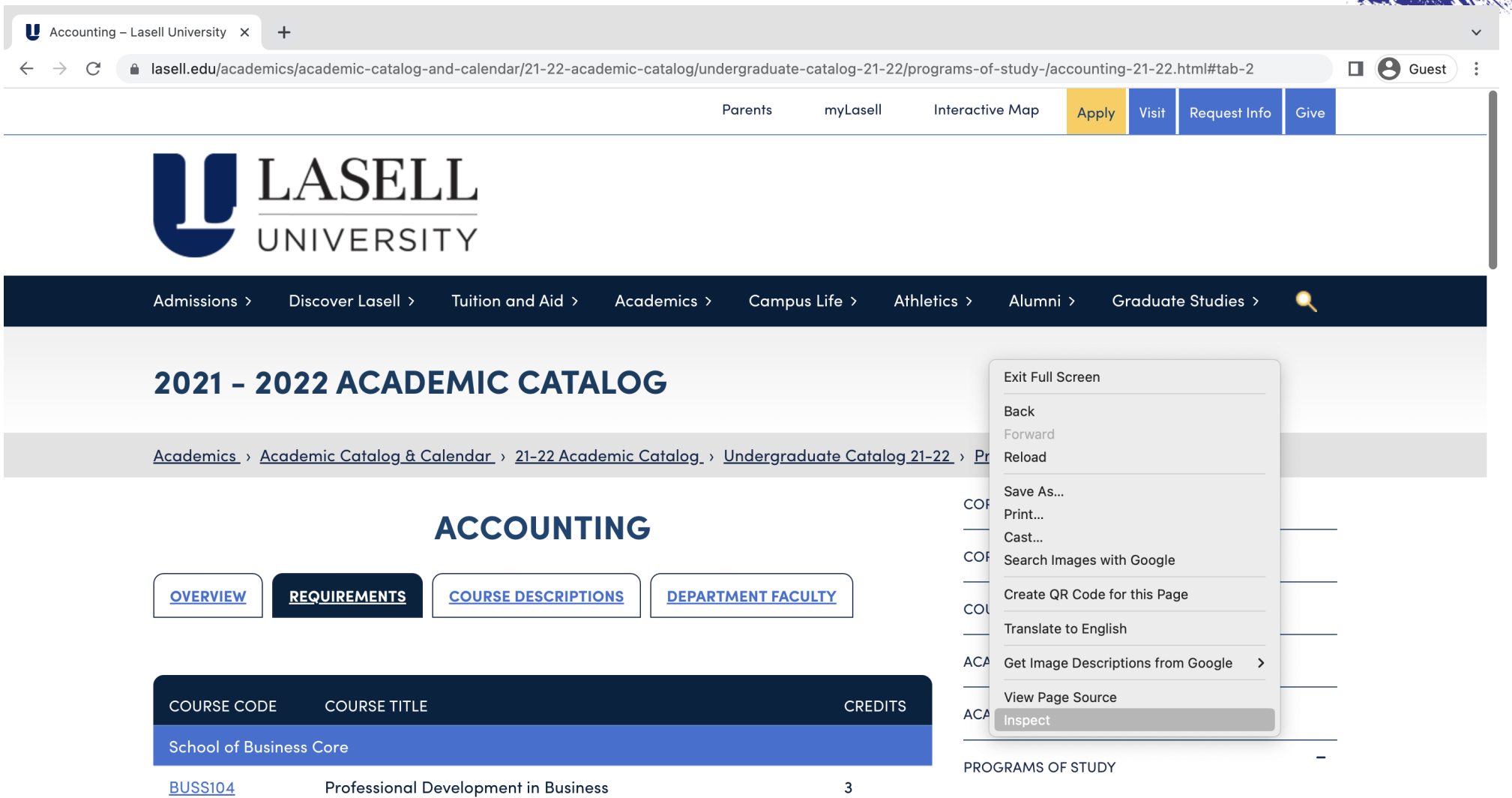
Web Scraping

Collect Web Data Efficiently

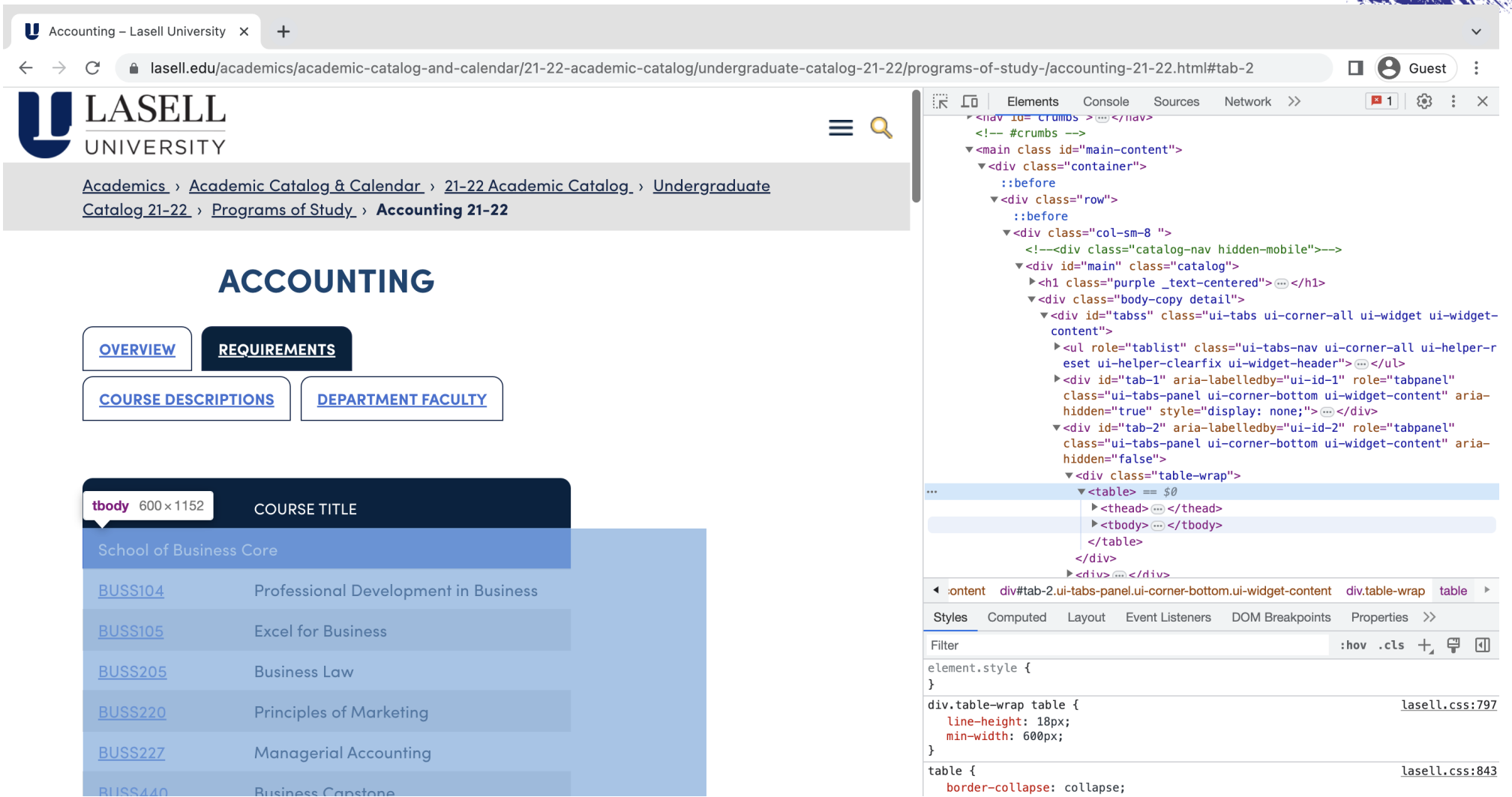
- Task: a list of 2021-2022 Courses for 43 Programs

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

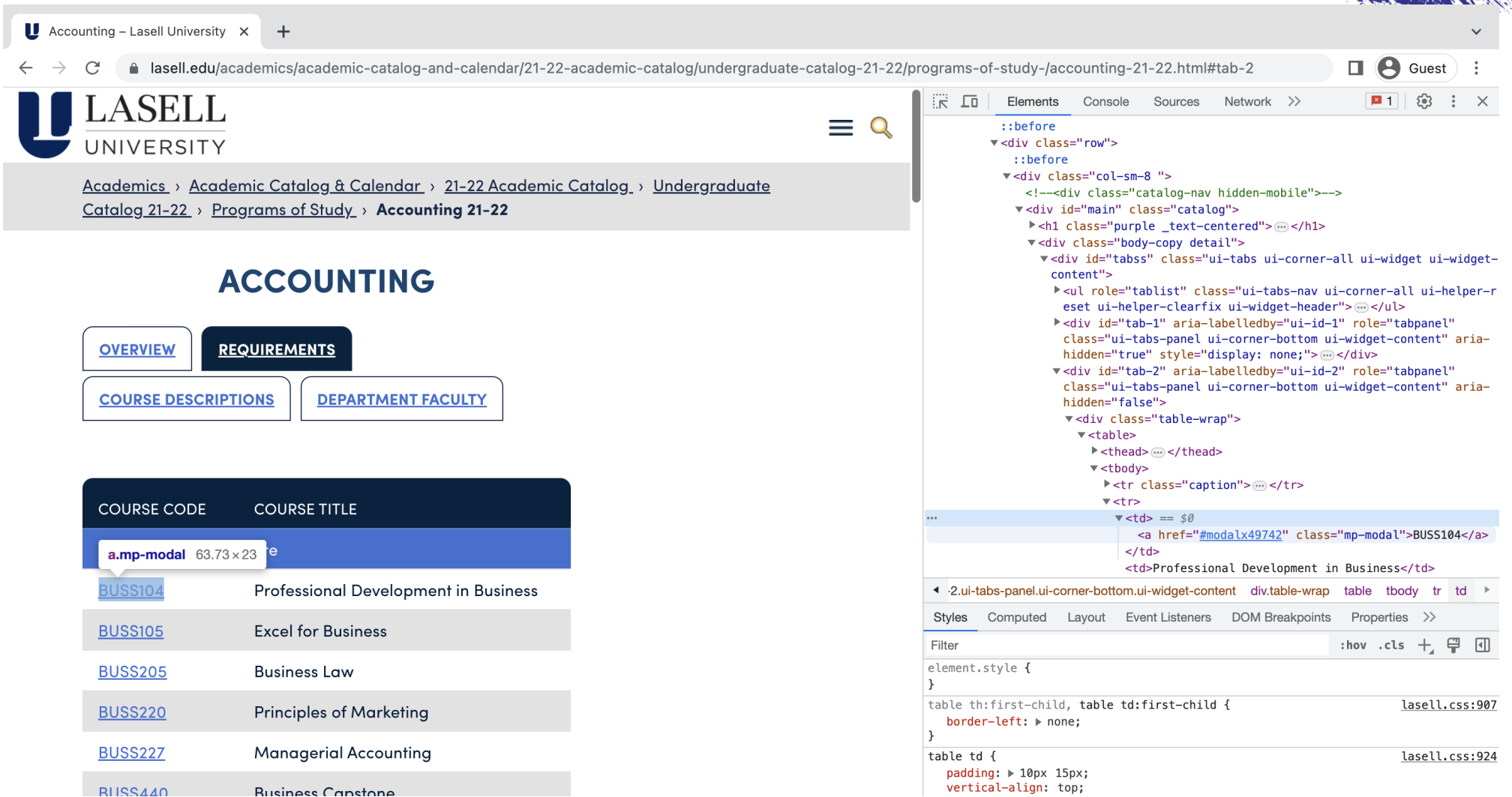
Step 1: Google a WebPage and Use "Inspect"



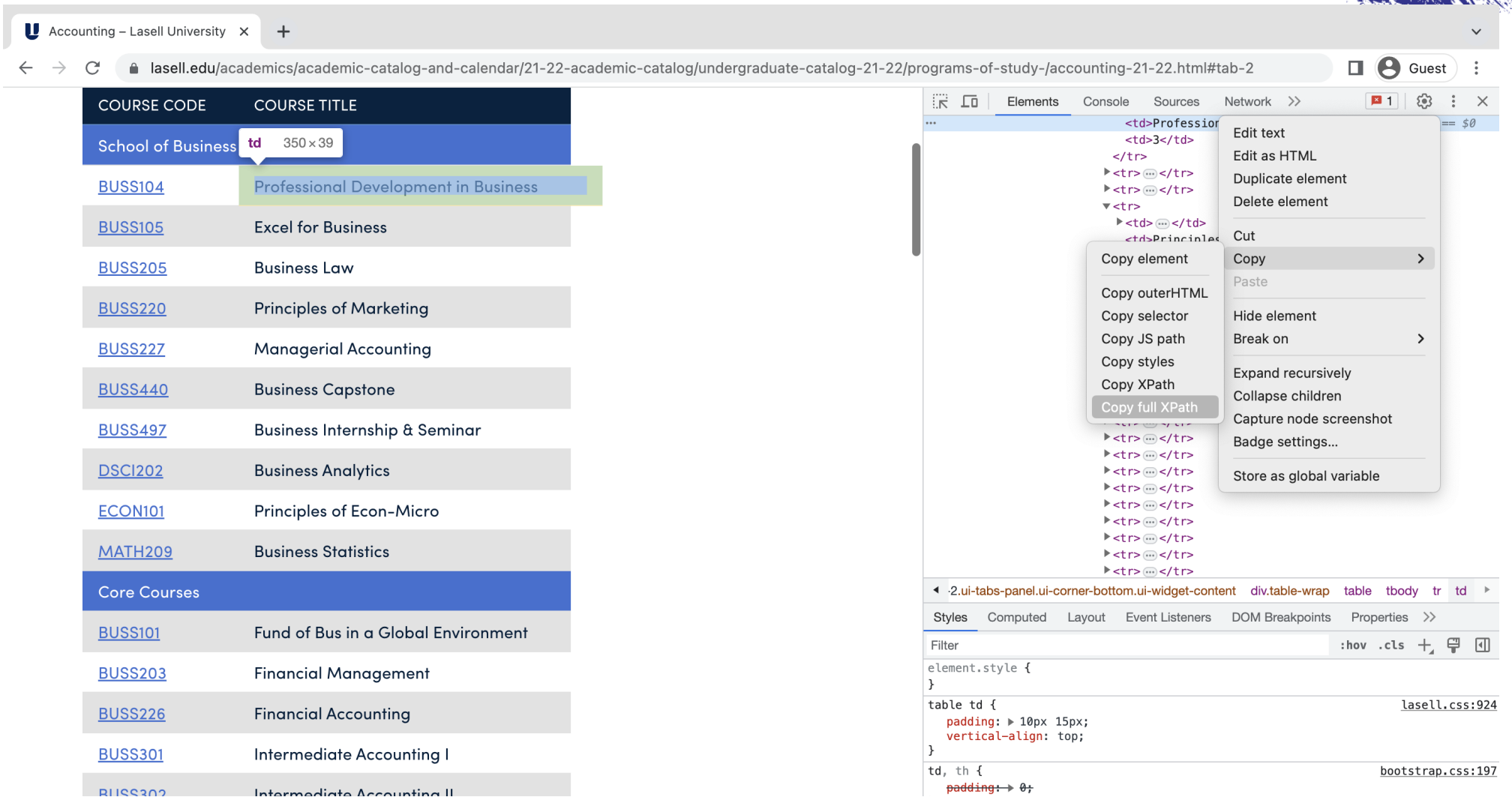
Step 2: Zoom in



Step 3: Find the Class Name



Step 3: Find the Class Name or Full XPath



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-catalo

#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/div[2]/div[1]/table/tbody/t
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"  "DSCI202"
 [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/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-scrapping to efficiently collect web data
- Online resources and communities
  - youtube.com/@linlishareresearch
  - July 20 Workshop Focusing on IPEDS
- Q&A