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

The Rmd for making this presentation is also there 😉

Linli Zhou, Ph.D.

2023-07-13

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, Practices

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

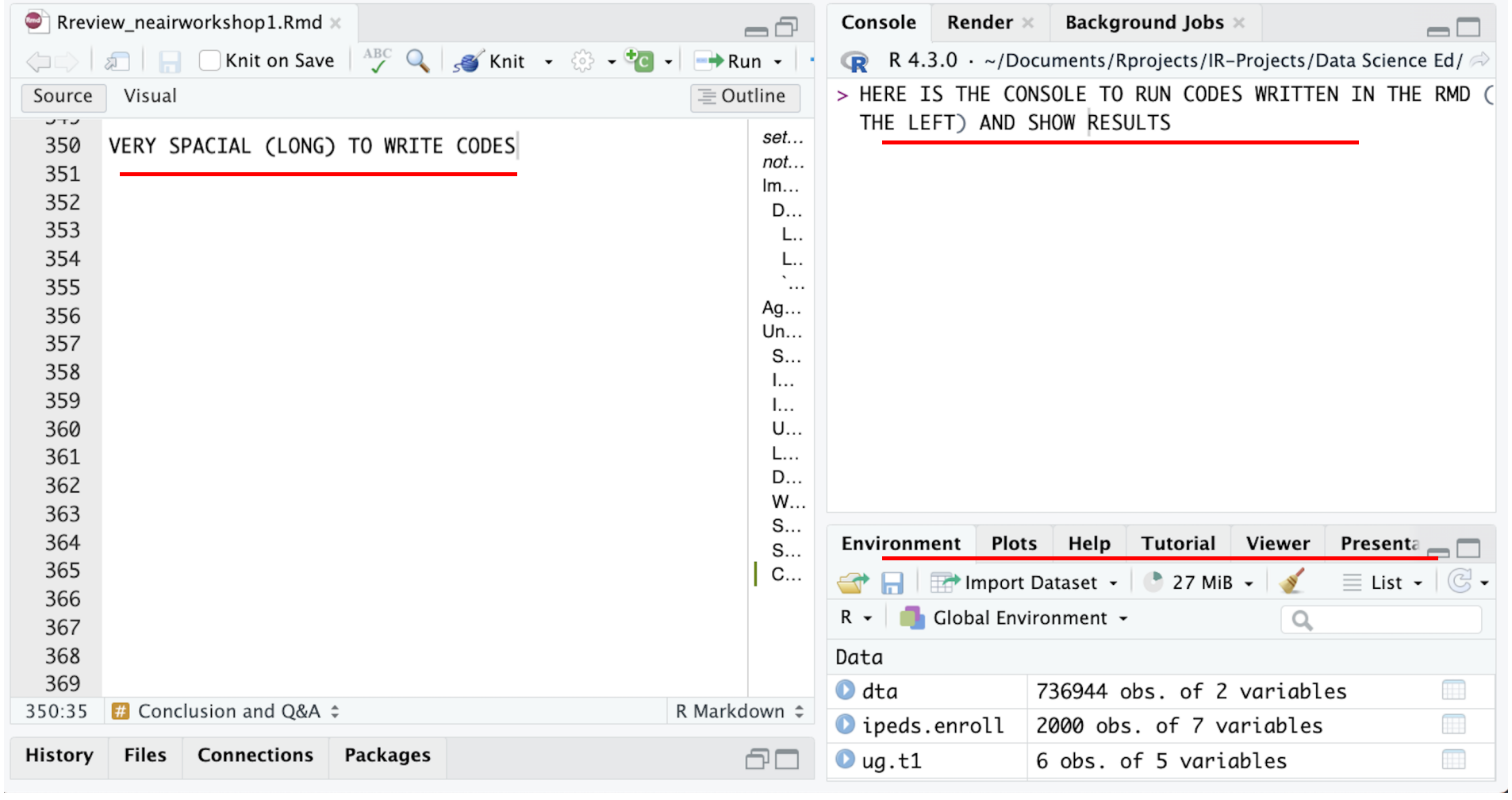
Setup Tips

Set up - download both

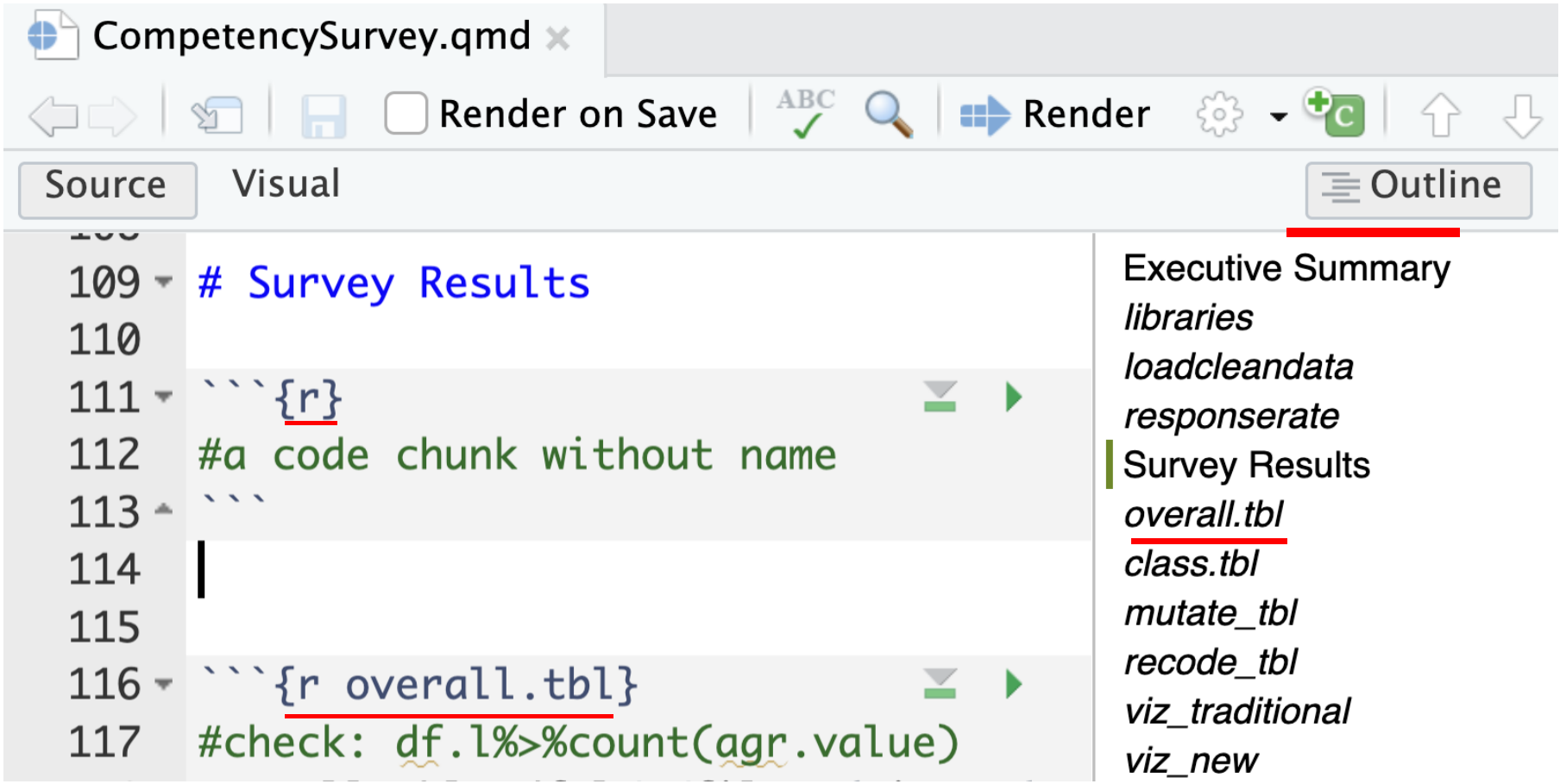
- A programming language (a language used to talk to computer for data analysis related tasks)
- The "Engine"

- An integrated development environment (IDE) for writing and executing R code
- The "Dashboard"

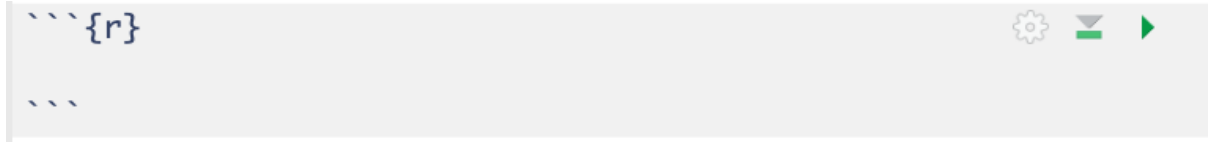
RStudio Tips: Layout



RStudio Tips: Outline



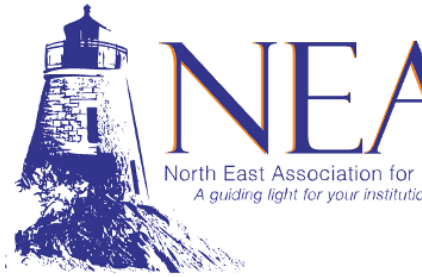
RStudio Tips: Code Chunk Creator



cmd+opt+i

Source: Posit Support, 2023

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
```

cmd+opt+o

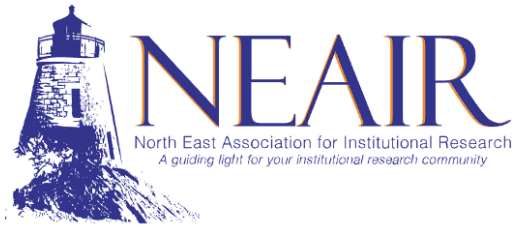
The Source File

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

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

Picture Source: Schafer, 2016

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, tidyr, purrr, ggplo
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
library(formattable)
library(ggrepel)
library(waffle)
library(patchwork)
library(wordcloud)
library(broom)
#viz table
library(gt)
library(kableExtra)#for good-looking static tables
library(DT)#for filtering/interactive tables
#web scrapping html
library(RCurl)
library(XML)
library(rvest)
```

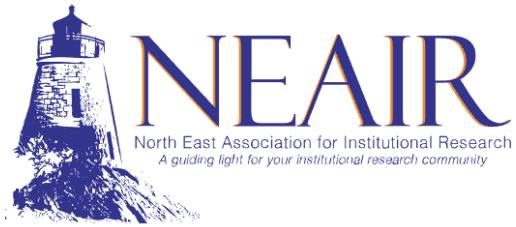
```
#for regression table
library(apaTables)#show cor/reg table
library(sjPlot)#show reg table
library(lme4)#multilevel modeling
library(performance)#calculate intra class correlation
#library(dummies)#lm already auto convert to dummy vars
library(caret)#creating predictive models: e.g. partitioning training vs test datasets
library(ranger)#random forest
library(tidylog)#see process of random forest
#data models
library(rstan)
#library(rethinking)#with the help of first installing: remotes::install_github("stan-dev/cr
library(GGally) #package to map cross-table of variables
library(leaps)#package for model selection
library(gvlma) #package to compare different model to select variables to include/exlude in
</div>
```

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
      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),
      legend.text = element_text(family = font, size=7),
      legend.position="top",
```

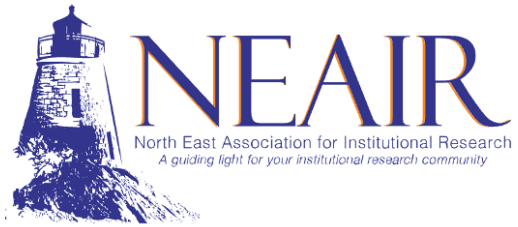
```
#continue code on the previous slide
      legend.margin=margin(t=-25),
      legend.text = element_text(family = font, size=7),
      legend.position="top",
      legend.direction = "horizontal",
      strip.text = element_text(family = font, size = 9, margin = margin(t=5,b=10)),#facet
      #strip.position = "bottom",
      strip.background = element_blank(),#facet background
      strip.text.y.left = element_text(family = font, size = 9, angle=0)#when facet label o
    )}
</div>
```

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>
```

Picture Source: Bjarnason, 2013



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.

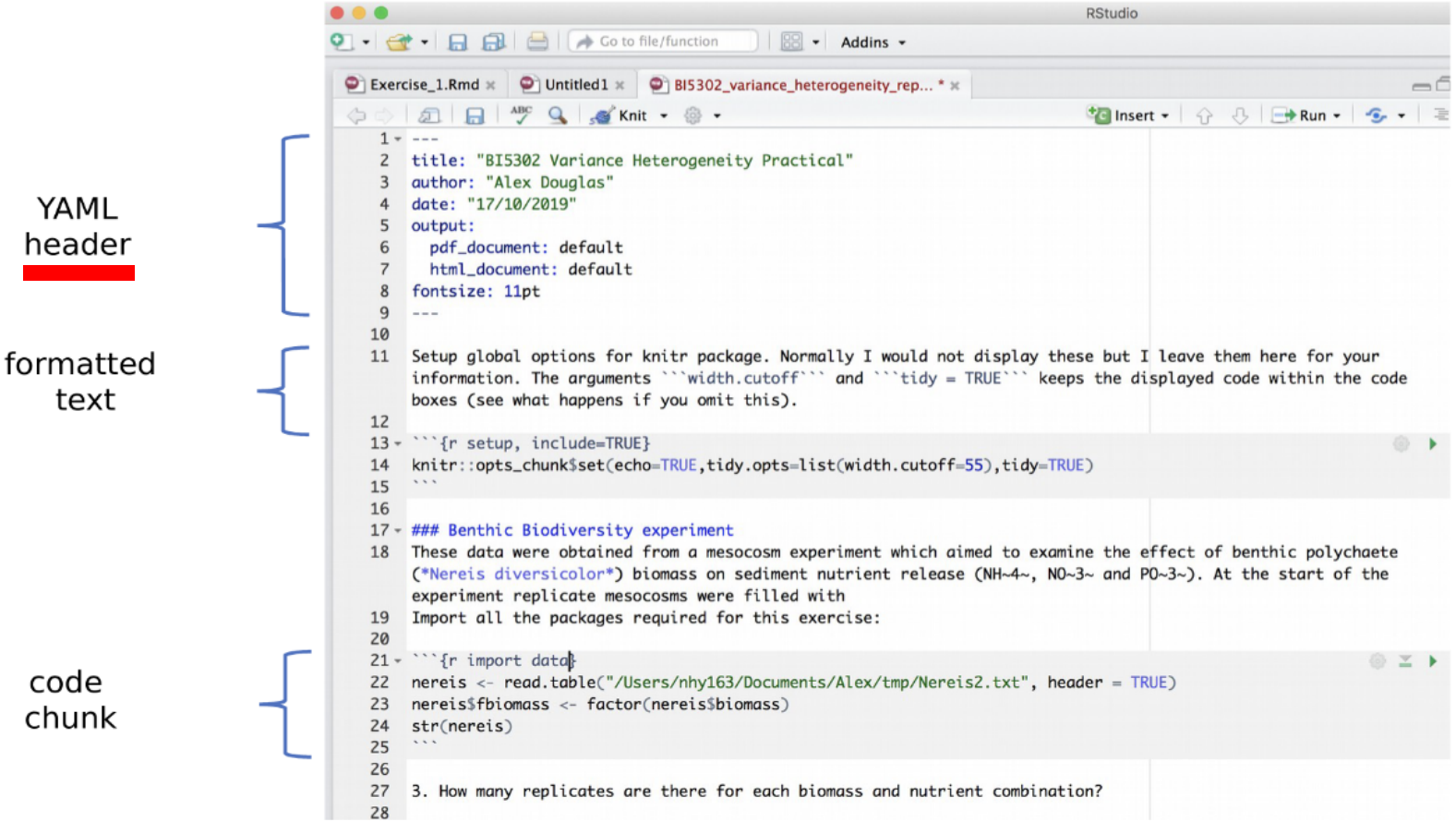
Codes for Setting PDF output

```
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
```

continue codes from the previous slide

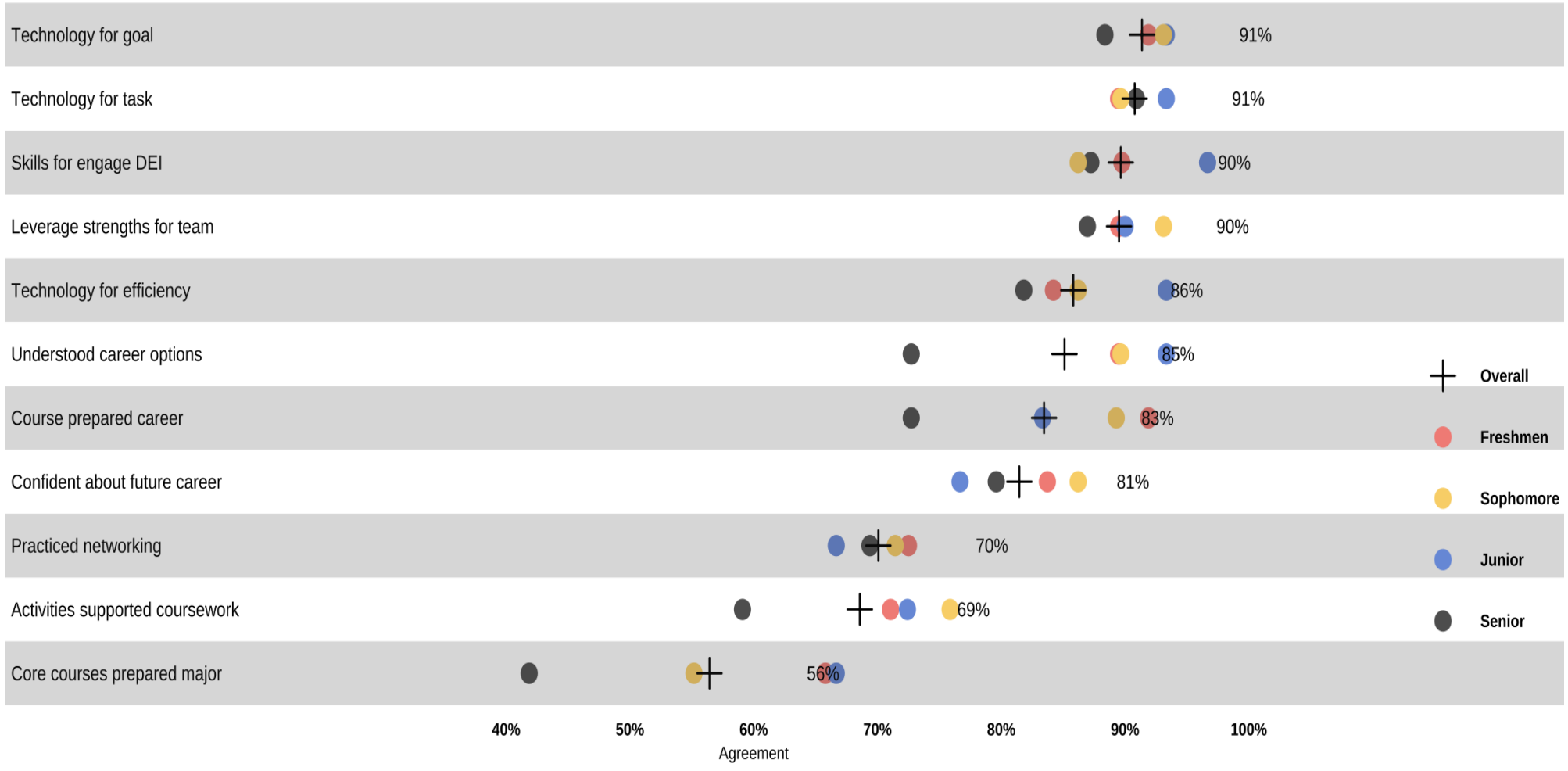
```
    include-in-header:
      - text: |
          \usepackage{titling}
          \setlength{\droptitle}{-1.5in}
          \pretitle{\begin{center}
          \includegraphics[height=1in]{/Users/linlizhou/Documents/Rprojects/IR.png}\LARGE\\
      - text: |
          \usepackage{fancyhdr}
          \addtolength{\headheight}{0.7cm}
          \pagestyle{fancy}
          \fancyhead[LO]{\includegraphics[height=1.3cm]{/Users/linlizhou/Documents/Rprojects
          \fancyhead[RE]{\vspace{4mm}\textbf{AAC Courses Review}}
          \fancyfoot[C]{\thepage}
          \renewcommand{\headrulewidth}{0pt}
</div>
```

YAML (Yet Another Markdown Language) Header



Source: Douglas, et. al., 2023

Compelling Visualization



Codes for Visualization: Creating Tables

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

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

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("Ove
```

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,"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: 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()%>%unique()%>%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 NA
  #check: lapply(df.s%>%select(starts_with("agr")), unique)
</div>
```

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`,`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")
```

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(
 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

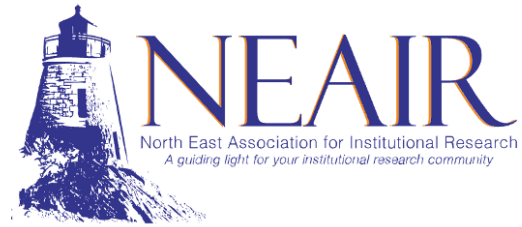
ppid	agr.career_prolearn	agr.career_strongweak	class_level
xxx54	Yes	Yes	Sophomore
xxx26	Yes	No	Freshmen
xxx66	Yes	Yes	Sophomore

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

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))%>%mutate(Average=round(n
  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(n
   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
```

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

Clean and Highlighted Table

Table 1: AAC course enrollment by term

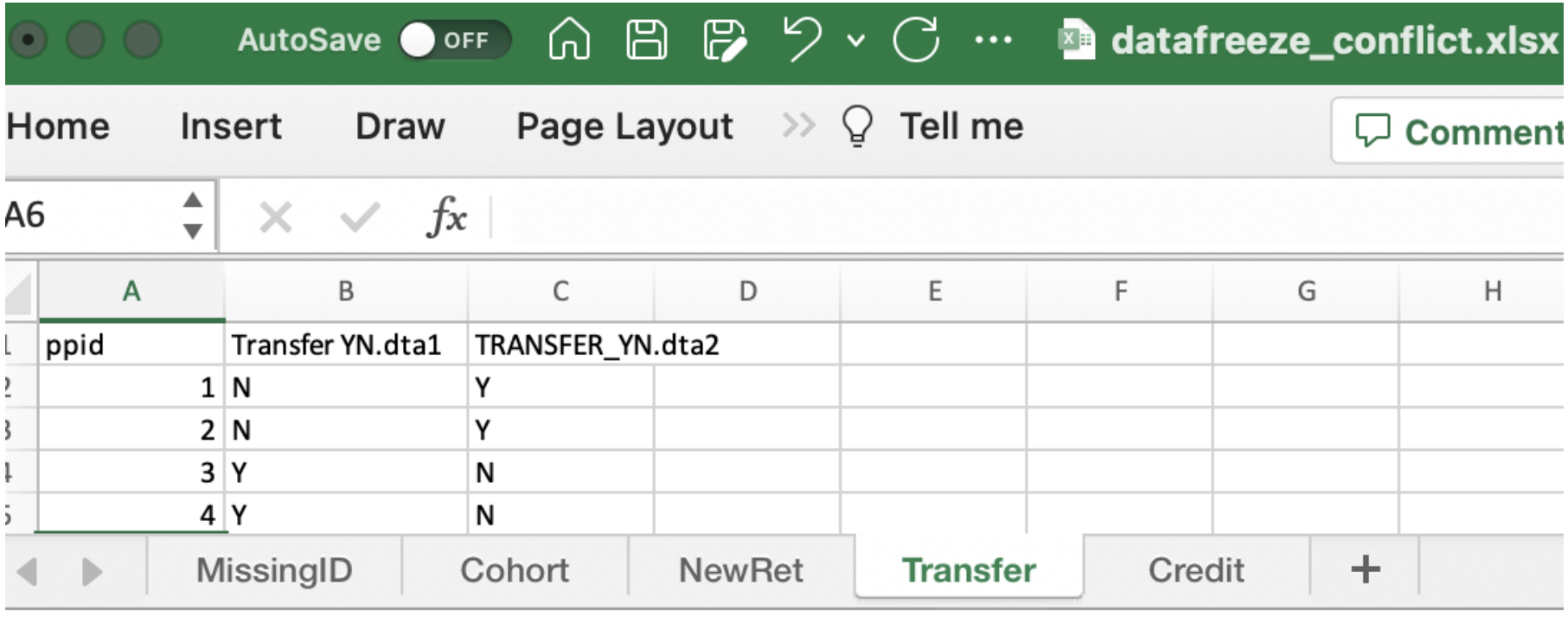
	Average	2018 Fall	2019 Spring	2019 Fall	2020 Spring	2020 Fall	2021 Spring	2021 Fall	2022 Spring
AAC102	56		80		43				46
AAC103	56	54	66	58	82	50	43	48	44
AAC104	50						50		

```
#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



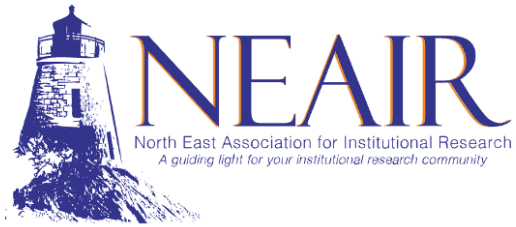
Comparing Two DataSets

```
#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)
```

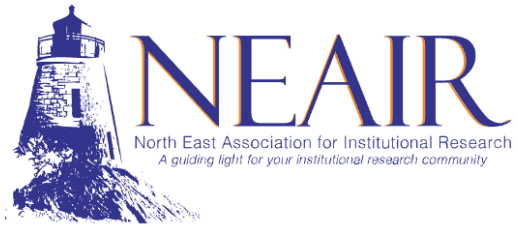
Comparing Two DataSets: If-else

```
#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

X Central repository of courses

X Individual reach out

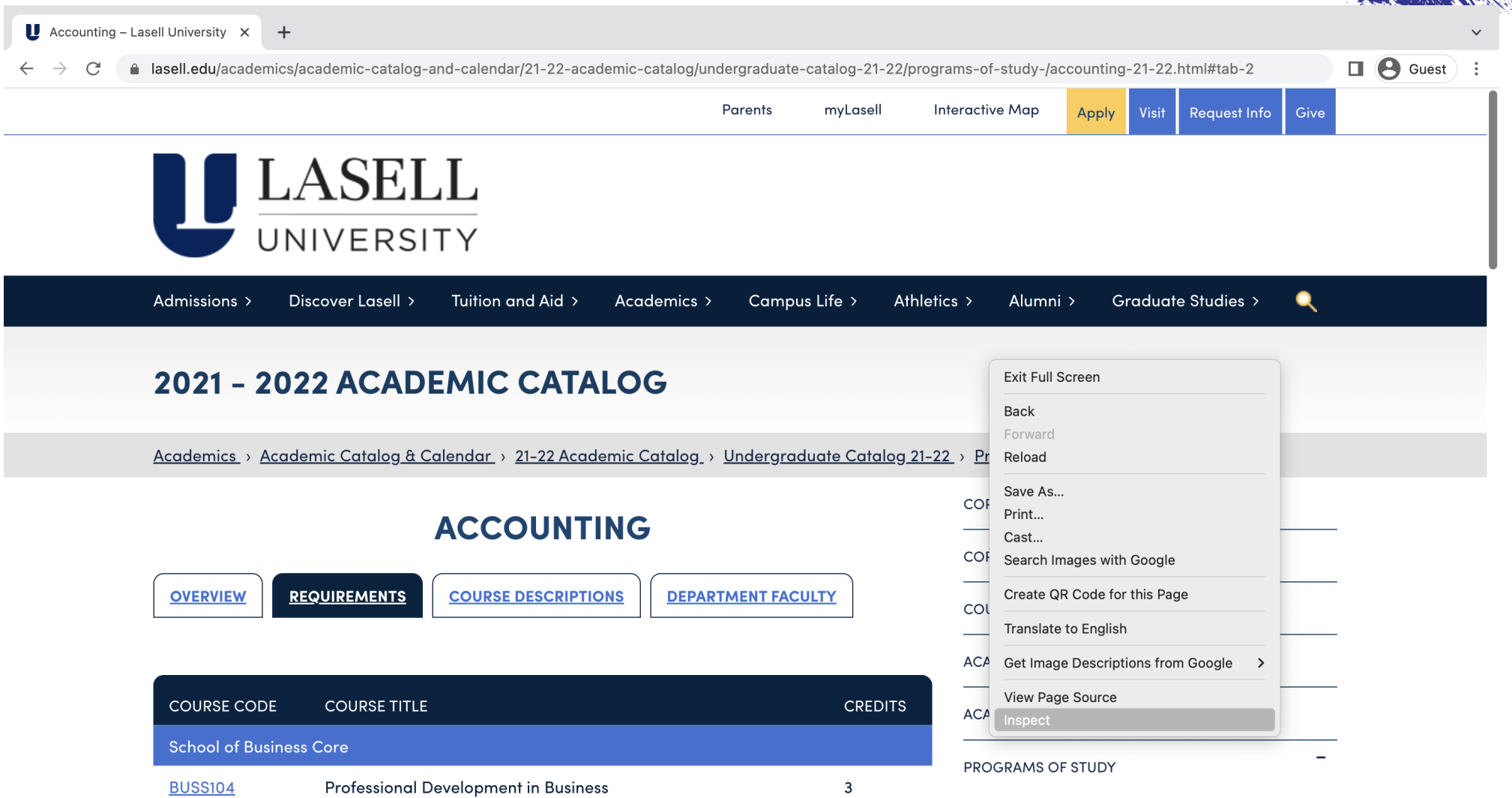
- Gather course information from the websites of each program

X Manual

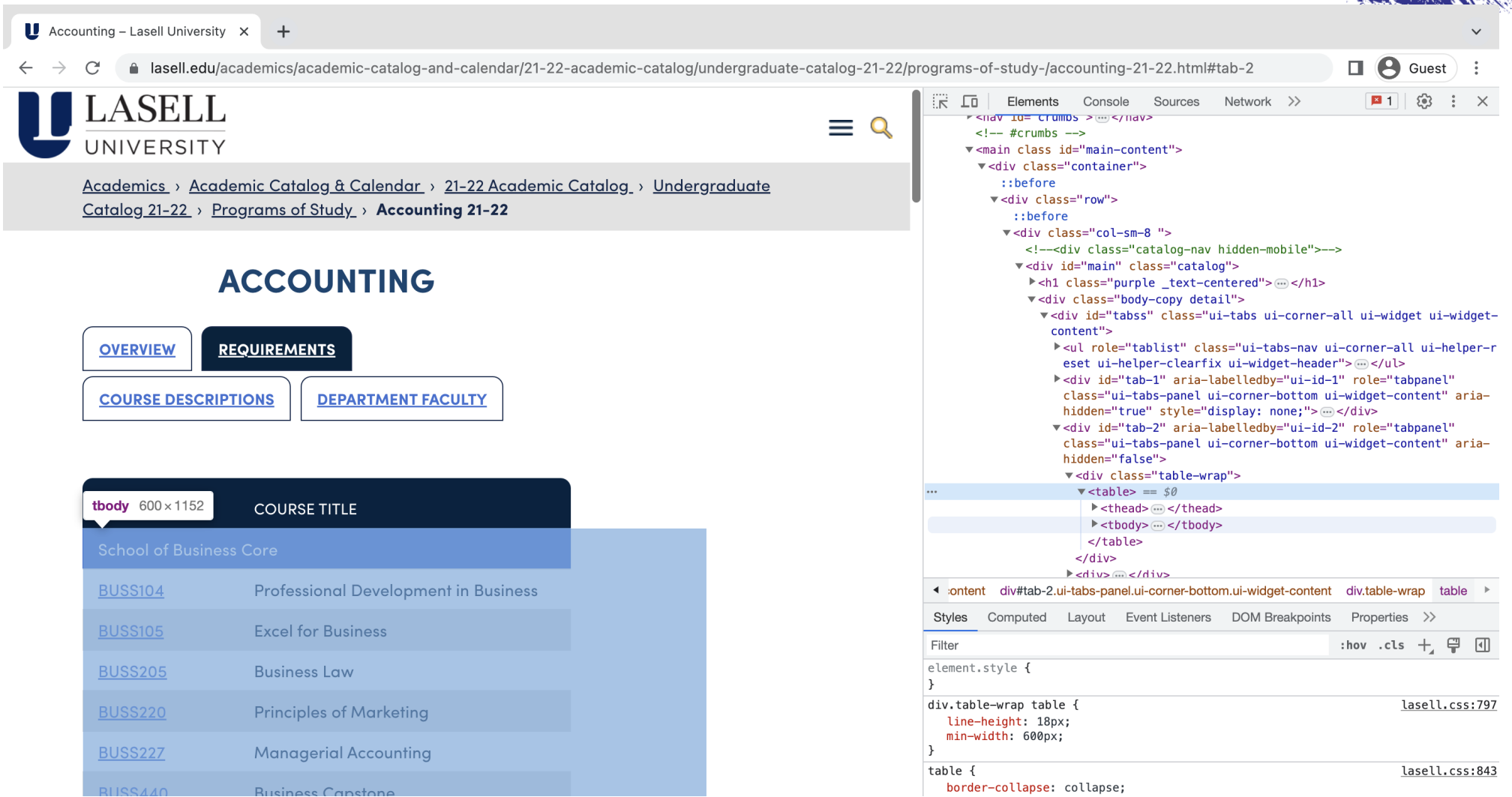
Picture source: Enos, 2022

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

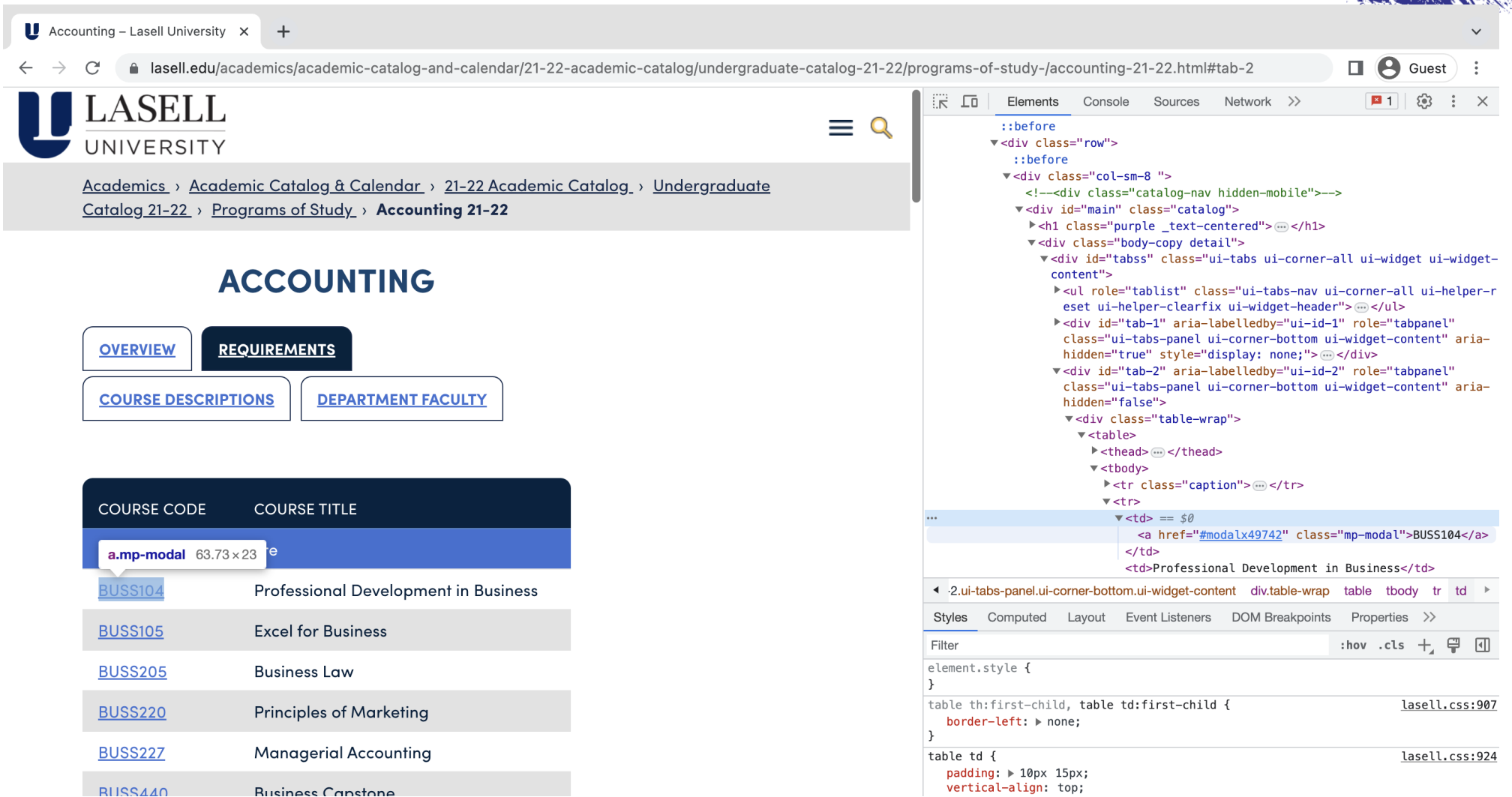
Step 1: "Inspect" in Chrome



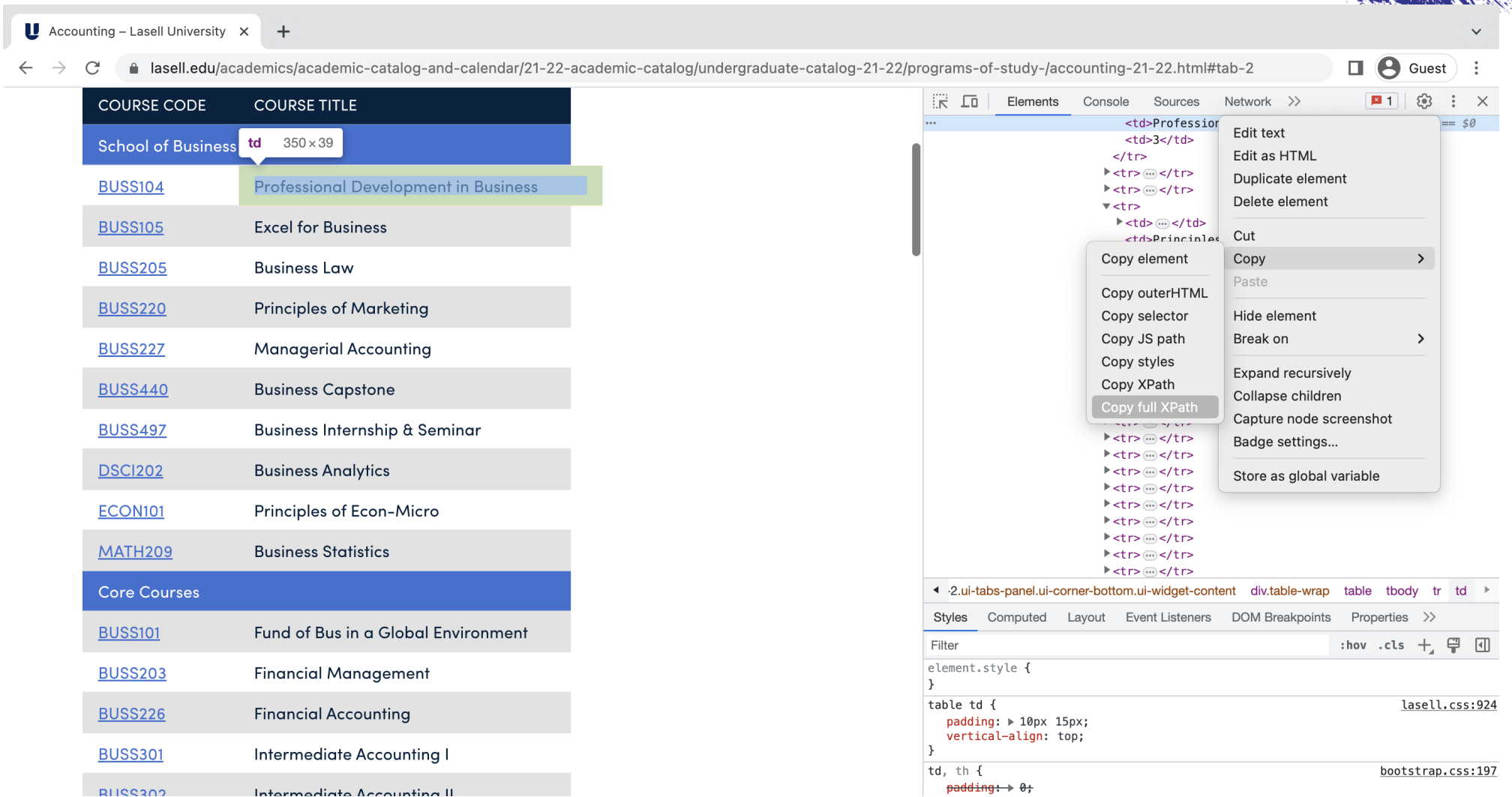
Step 2: Zoom in



Step 3: Find the Class Name



Step 3: Find the Class Name or Full XPath



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]

Find the rest of presentation here: https://youtu.be/cl3TB4B8inU

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
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/di
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-scrapping
- Online resources and communities
  - www.youtube.com/@linlishareresearch
  - July 20 Workshop Focusing on IPEDS
  - lzhou@lasell.edu

Practice!

Exercise questions: https://rb.gy/wiw14

Picture source: Avci, 2022

```
source("/Users/linlizhou/Documents/Rprojects/IR-Projects/theme_source.R")
#library(readxl)
#library(tidyverse)
#library(janitor)
#library(openxlsx)
#library(kableExtra)
```

Visualization

```
#read data
df<-read_excel("/Users/linlizhou/Downloads/sampledataset.xlsx")
#summarize table
tbl<-df%>%filter(!is.na(agr.value))%>%
  group_by(agr.item, agr.value)%>%
  summarise(n=n())%>%
  mutate(prt=n/sum(n))%>%filter(agr.value=="Yes")
#check the table
tbl%>%View()
```

```
#viz
tbl%>%ggplot(aes(y=agr.item,x=prt))+
  geom_point(stat = "identity")

#viz
tbl%>%ggplot(aes(y=agr.item,x=prt))+
  geom_bar(stat = "identity")
```

Conditional formatting

```
#function prep
cf_color<-function(x,a=.9,b=.8,c=.5,d=.3,col1=color_yellow,col2=yellow,col3=color_greylight,
  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

#read data
df<-read_excel("/Users/linlizhou/Downloads/sampledataset.xlsx")
#summarize table
tbl<-df%>%filter(!is.na(agr.value))%>%group_by(agr.item, agr.value)%>%summarise(n=n())%>%mut
```

```
#turn into a kbl object and do some initial formatting
tbl_kbl<-tbl%>%kbl(align = "c",booktabs = T) %>%
    kable_styling(full_width = F, font_size = 12, latex_options = c('hold_position'))%>%

#set conditional formatting to one column (the 4th column)
column_spec(column=4, width = "3em",
   background =cf_color(tbl[4],a=.7,b=.6,c=.5,d=.4))
```

Clean data

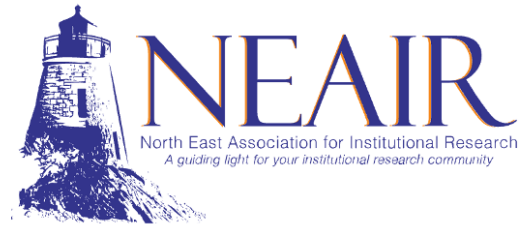
```
#load data
df<-read_excel("/Users/linlizhou/Downloads/messydata.xlsx")

#check column names: df%>%colnames()
#remove cols
df1<-df%>%select(-(`Response ID`:`Invite Status`))%>%
clean_names()%>%unique()%>%remove_empty()%>%
select(-which(colSums(is.na(.))==nrow(.)-1))%>%
filter(rowSums(is.na(.)) != ncol(.) - 1)
```

```
#rename and recode
#renaming
df2<-df1%>%rename(agr.learning=at_lasell_i_have_experienced_continued_personal_and_professio
agr.strength=at_lasell_i_have_gained_awareness_of_my_strengths_and_weaknesses_think_back_on_
agr.confident=i_feel_confident_about_my_future_career_think_back_on_your_experience_as_a_stu

#recoding
df3<-df2%>%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")))
```

Clean data: continue

```
#check sheet name: excel_sheets("/Users/linlizhou/Downloads/messydata.xlsx")
#load admin data
df.admin<-read_excel("/Users/linlizhou/Downloads/messydata.xlsx",sheet="administrativedata")

#merge to add vars
df4<-df3%>%left_join(df.admin,by=c("invite_custom_field_1"="studentID"))
```

Compare dataset




```
excel_sheets("/Users/linlizhou/Downloads/comparedata.xlsx")
#load and suffix
df1<-read_excel("/Users/linlizhou/Downloads/comparedata.xlsx",sheet="Sheet1")%>%
   rename_all(~paste0(.x,".dta1"))
df2<-read_excel("/Users/linlizhou/Downloads/comparedata.xlsx",sheet="Sheet2")%>%
   rename_all(~paste0(.x,".dta2"))
#merge
df.m<-full_join(df1,df2,by=c("ppid.dta1"="ppid.dta2"))
#diff
conflict.enroll<-df.m%>%mutate(diff=ifelse(
   enrollment_status.dta1==enrollment_status.dta2, "matched","not match")) %>%
  filter(diff=="not match")

conflict.transfer<-df.m%>%mutate(diff=ifelse(
   transfer.dta1==transfer.dta2, "matched","not match")) %>%
  filter(diff=="not match")
#save
write.xlsx(list("enrollment"=conflict.enroll,
                "transfer"=conflict.transfer),
           file="/Users/linlizhou/Downloads/Conflict_rd1.xlsx")
```