Exam 2 Submission

Echo Nattinger

6/28/2021

## College Scorecard Questions

# Clearing the environment and loading the dataset  
rm(list=ls(all=TRUE))  
library(rio)  
college\_scorecard = import("2021\_exam2\_data.xlsx", which=4)  
# Providing summary statistics  
summary(college\_scorecard)

## unitid inst\_name state\_abbr   
## Min. :100654 Length:48445 Length:48445   
## 1st Qu.:163532 Class :character Class :character   
## Median :212115 Mode :character Mode :character   
## Mean :260438   
## 3rd Qu.:409120   
## Max. :490009   
##   
## pred\_degree\_awarded\_ipeds year earnings\_med count\_not\_working  
## Min. :1.000 Min. :2007 Min. : 8400 Min. : 0.0   
## 1st Qu.:1.000 1st Qu.:2011 1st Qu.: 24700 1st Qu.: 46.0   
## Median :2.000 Median :2012 Median : 31600 Median : 115.0   
## Mean :1.913 Mean :2012 Mean : 33348 Mean : 369.4   
## 3rd Qu.:3.000 3rd Qu.:2014 3rd Qu.: 39800 3rd Qu.: 300.0   
## Max. :3.000 Max. :2016 Max. :186500 Max. :15960.0   
## NA's :15706 NA's :15801   
## count\_working   
## Min. : 8   
## 1st Qu.: 210   
## Median : 594   
## Mean : 2073   
## 3rd Qu.: 1477   
## Max. :94724   
## NA's :14772

# Creating smaller dataset  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

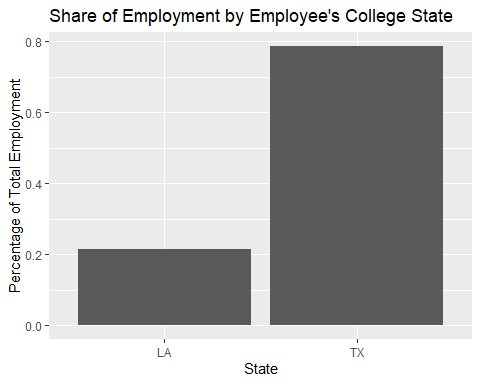
## v ggplot2 3.3.4 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.6  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

small\_scorecard = college\_scorecard%>%  
 filter(year == 2014 | year == 2015)%>%  
 filter(state\_abbr == "TX" | state\_abbr == "LA")

# Collapsing into smaller dataframe  
# First have to get rid of NAs in count\_working  
small\_scorecard\_no\_NA = small\_scorecard %>% na.omit(small\_scorecard$count\_working)  
even\_smaller\_scorecard = small\_scorecard\_no\_NA %>%  
 group\_by(state\_abbr)%>%  
 summarize(avwork = mean(count\_working),   
 totalworking = sum(count\_working))  
  
# Graphing  
library(ggplot2)  
  
ggplot(even\_smaller\_scorecard, aes(state\_abbr, totalworking/sum(totalworking)))+  
 geom\_col(stat="identify")+  
 labs(title="Share of Employment by Employee's College State",   
 x="State",   
 y="Percentage of Total Employment")

## Warning: Ignoring unknown parameters: stat



summary(even\_smaller\_scorecard)

## state\_abbr avwork totalworking   
## Length:2 Min. :2207 Min. :205786   
## Class :character 1st Qu.:2275 1st Qu.:343627   
## Mode :character Median :2343 Median :481469   
## Mean :2343 Mean :481469   
## 3rd Qu.:2411 3rd Qu.:619310   
## Max. :2479 Max. :757151

table(small\_scorecard\_no\_NA$state\_abbr)

##   
## LA TX   
## 83 343

On the basis of this graph, people who graduated from four-year colleges/universities in Texas had a better change of being employed. I do think going to college in certain states could give a student a better chance of employment, but I don’t think that’s what this graph proves. The average number of people working from each state is relatively close (2479 and 2207 for LA and TX respectively). Further, TX has way more schools in this dataset than LA.

## Avocado Sales

# Loading dataset  
avocados = import("2021\_exam2\_data.xlsx", which=2)

# Creating new variable  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

avocados = avocados%>%  
 mutate(year = lubridate::year(avocados$date))

# Deflating  
library(WDI)  
# loading deflator data and renaming it for ease  
deflatordata = WDI(country="US", indicator=c("NY.GDP.DEFL.ZS"),   
 start=1960, end=2018, extra=FALSE, cache=NULL)  
library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':  
##   
## hour, isoweek, mday, minute, month, quarter, second, wday, week,  
## yday, year

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

setnames(deflatordata, "NY.GDP.DEFL.ZS", "deflator")  
deflatordata$country = NULL  
deflatordata$iso2c = NULL  
# Merging  
deflatedavocados = left\_join(avocados, deflatordata, by=c("year"))  
  
# Creating deflated amounts  
deflatedavocados$deflated\_price\_2015 = deflatedavocados$average\_price/(deflatedavocados$deflator/100)

# Collapsing deflatedavocados  
collapsed\_avocados = deflatedavocados%>%  
 group\_by(year)%>%  
 summarize(avdeflatedprice = mean(deflated\_price\_2015))  
  
head(collapsed\_avocados)

## # A tibble: 4 x 2  
## year avdeflatedprice  
## <dbl> <dbl>  
## 1 2015 1.02  
## 2 2016 1.04  
## 3 2017 1.22  
## 4 2018 1.02

# Reshaping the data  
wide\_avocados = collapsed\_avocados%>%  
 pivot\_wider(id\_cols = "year",   
 names\_from = "year",   
 values\_from = "avdeflatedprice" )  
head(wide\_avocados)

## # A tibble: 1 x 4  
## `2015` `2016` `2017` `2018`  
## <dbl> <dbl> <dbl> <dbl>  
## 1 1.02 1.04 1.22 1.02

# Labeling variables in wide\_avocados  
library(labelled)  
var\_label(wide\_avocados) = list(`2015` = "Year 2015",   
 `2016` = "Year 2016",   
 `2017` = "Year 2017",   
 `2018` = "Year 2018")

## Training Dataset

# Loading dataset  
training = import("2021\_exam2\_data.xlsx", which=3)  
# Creating ID variable  
training$id = 1:nrow(training)  
# Reshaping  
traininglong = training%>%  
 pivot\_longer(cols = starts\_with("re"),   
 names\_to = NULL,   
 values\_to= "earnings",  
 values\_drop\_na = FALSE)  
# Summary  
summary(traininglong)

## training\_program age educ black   
## Min. :0.0000 Min. :17.00 Min. : 3.0 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:20.00 1st Qu.: 9.0 1st Qu.:1.0000   
## Median :0.0000 Median :24.00 Median :10.0 Median :1.0000   
## Mean :0.4157 Mean :25.37 Mean :10.2 Mean :0.8337   
## 3rd Qu.:1.0000 3rd Qu.:28.00 3rd Qu.:11.0 3rd Qu.:1.0000   
## Max. :1.0000 Max. :55.00 Max. :16.0 Max. :1.0000   
## hisp marr id earnings   
## Min. :0.00000 Min. :0.0000 Min. : 1 Min. : 0   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:112 1st Qu.: 0   
## Median :0.00000 Median :0.0000 Median :223 Median : 0   
## Mean :0.08764 Mean :0.1685 Mean :223 Mean : 2927   
## 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:334 3rd Qu.: 4045   
## Max. :1.00000 Max. :1.0000 Max. :445 Max. :60308

## Titanic Questions

# Loading dataset and providing summary statistics  
titanic = import("2021\_exam2\_data.xlsx", which=1)  
summary(titanic)

## class age female survived   
## Min. :1.000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:2.000 1st Qu.:1.0000 1st Qu.:1.0000 1st Qu.:0.000   
## Median :3.000 Median :1.0000 Median :1.0000 Median :0.000   
## Mean :2.977 Mean :0.9505 Mean :0.7865 Mean :0.323   
## 3rd Qu.:4.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :4.000 Max. :1.0000 Max. :1.0000 Max. :1.000

# Crosstabs  
library(doBy)

##   
## Attaching package: 'doBy'

## The following object is masked from 'package:dplyr':  
##   
## order\_by

summaryBy(survived ~ female, data = titanic, FUN=mean)

## female survived.mean  
## 1 0 0.7319149  
## 2 1 0.2120162

The results of this cross-tab show that those who are female (female=1) were less likely than their male counterparts to survive (with a mean of 0.21 compared to 0.73 for men).

# If else and frequency table  
titanic$first\_class = ifelse(titanic$class == 1, 1, 0)  
table(titanic$first\_class)

##   
## 0 1   
## 1876 325

Bonus: “My Heart Will Go On” by Celine Dion

Github link: <https://github.com/enatt/DataSciExam2>