Exam 2

#Exam 2 - Data Science for the Social World ##Laura Morales ###June 28, 2021 **1. Please clear the environment in R.**

rm(list=ls(all=TRUE))

**2. Load the college\_scorecard dataset in R, and call it “college\_scorecard”.**

library(rio) #rio package to load data  
college\_scorecard = import("2021\_exam2\_data.xlsx", which=4) #load data

1. Provide summary statistics for the college\_scorecard dataset.

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

**4. students who graduated from four-year+ colleges and universities located in Texas (state\_abbr: “TX”) and Louisiana (state\_abbr: “LA”). Call the resulting data frame “small\_scorecard”.**

small\_scorecard = subset(college\_scorecard, college\_scorecard$state\_abbr == 'TX' | college\_scorecard$state\_abbr == 'LA') #subset the colleges based on the states abbreviation  
small\_scorecard = subset(small\_scorecard, small\_scorecard$year >= 2014)#subset the colleges based on the lower limit of the year  
small\_scorecard = subset(small\_scorecard, small\_scorecard$year < 2016)#subset the colleges based on the upper limit of the year

**5. Collapse the “small\_scorecard” data frame to get both (a) the average of number people working who graduated from universities in Texas and Lousiana; and (b) the total number of people working who graduated from universities in Texas and Lousiana. Call your resulting data frame “even\_smaller\_scorecard”. (Hint: Your resulting data frame should only have two observations–one for Texas, the other for Lousiana.)**

#load the dplyr library  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

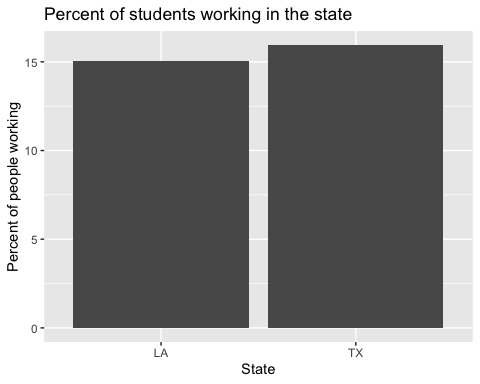
#pipe the small score card  
even\_smaller\_scorecard <- small\_scorecard %>%   
 na.omit(small\_scorecard, select= c("count\_not\_working", "count\_working")) %>% #drop the na values  
 group\_by(state\_abbr) %>% #group the states   
 summarize(across(where(is.numeric), sum)) #summarize count the amount of degrees

**6. Use the “even\_smaller\_scorecard” data frame to provide a bar graph detail- ing the percent of people working. Make sure to label the axes and provide an appropriate title for the graph. (Hint: you will need to create a new variable to answer this question.)**

even\_smaller\_scorecard$percent\_working <- (even\_smaller\_scorecard$count\_not\_working / (even\_smaller\_scorecard$count\_not\_working+even\_smaller\_scorecard$count\_working))\*100  
library(ggplot2)  
barplot = ggplot(data= even\_smaller\_scorecard, aes(even\_smaller\_scorecard$state\_abbr, even\_smaller\_scorecard$percent\_working))+  
 geom\_bar(stat="identity")+  
 labs(x= "State",   
 y= "Percent of people working",  
 title= "Percent of students working in the state")  
print(barplot)

## Warning: Use of `even\_smaller\_scorecard$state\_abbr` is discouraged. Use  
## `state\_abbr` instead.

## Warning: Use of `even\_smaller\_scorecard$percent\_working` is discouraged. Use  
## `percent\_working` instead.

 **7. On the basis of your graph, did people who graduated from four-year colleges/universities located in Texas or Louisiana have a better chance of being employed? More broadly, do you think that going to college/university in one state gives people a better chance at getting a job? (Hints: (a) you will want to take a look at the summary statistics of the “even\_smaller\_scorecard” data frame; and (b) you will want to take a look at the universities included in the “smaller\_scorecard”)** The people who graduated from colleges and universities located in Texas have a better chance of being employed.

summary(even\_smaller\_scorecard) #summary of the two states

## state\_abbr unitid pred\_degree\_awarded\_ipeds  
## Length:2 Min. : 23369594 Min. :144.0   
## Class :character 1st Qu.: 45647498 1st Qu.:259.8   
## Mode :character Median : 67925402 Median :375.5   
## Mean : 67925402 Mean :375.5   
## 3rd Qu.: 90203307 3rd Qu.:491.2   
## Max. :112481211 Max. :607.0   
## year earnings\_med count\_not\_working count\_working   
## Min. :167162 Min. : 2499300 Min. : 36477 Min. :205786   
## 1st Qu.:298072 1st Qu.: 4573125 1st Qu.: 63268 1st Qu.:343627   
## Median :428982 Median : 6646950 Median : 90058 Median :481468   
## Mean :428982 Mean : 6646950 Mean : 90058 Mean :481468   
## 3rd Qu.:559892 3rd Qu.: 8720775 3rd Qu.:116848 3rd Qu.:619310   
## Max. :690802 Max. :10794600 Max. :143639 Max. :757151   
## percent\_working  
## Min. :15.06   
## 1st Qu.:15.28   
## Median :15.50   
## Mean :15.50   
## 3rd Qu.:15.72   
## Max. :15.95

**8.**

library(rio)  
avocados = import("2021\_exam2\_data.xlsx", which=2)

**9.**

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ tibble 3.1.2 ✓ purrr 0.3.4  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.1

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

library(lubridate)

##   
## Attaching package: 'lubridate'

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

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

**10.**

#add World development indicators (WDI)  
library(WDI) #load WDI package  
deflator = WDI(country = "US", indicator = c("NY.GDP.DEFL.ZS"), start = 1960, end = 2018, extra = FALSE, cache = NULL) #create variable from the package  
library(data.table) #load data table package

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

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

setnames(deflator, "NY.GDP.DEFL.ZS", "deflator") #rename  
deflated\_data = left\_join(x=avocados,  
 y=deflator,  
 by='year') #left join the deflator and avocados dataframs  
deflated\_data$deflated\_amount = deflated\_data$average\_price/(deflated\_data$deflator/100) #adjust for inflation

**11.**

collapsed\_avocados <- deflated\_data %>%   
 group\_by(year) %>% #group the year   
 summarize(across(where(is.numeric), mean)) #summarize mean price

**12.**

wide\_avocados <- collapsed\_avocados %>% pivot\_wider(id\_cols = c ( "total\_volume", "average\_price", "year"), names\_from = "year", values\_from = "deflated\_amount" )#omake the year the columns  
head(wide\_avocados)#only see the first five values

## # A tibble: 4 x 6  
## total\_volume average\_price `2015` `2016` `2017` `2018`  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 5681498. 1.02 1.02 NA NA NA   
## 2 6105539. 1.05 NA 1.04 NA NA   
## 3 5834479. 1.25 NA NA 1.22 NA   
## 4 6786962. 1.08 NA NA NA 1.02

**13.**

library(data.table)#load library  
setnames(wide\_avocados, "total\_volume", "avocado\_total\_volume")#set names to be more descriptive

**14.**

training = import("2021\_exam2\_data.xlsx", which=3) #load data

**15.**

# long\_data\_frame <-   
# training %>%   
# pivot\_longer(cols = starts\_with('re\_74','re\_75','re\_78'),  
# names\_to = "training\_program",  
# names\_prefix = "re",  
# values\_to = c('re\_74','re\_75','re\_78'),  
# values\_drop\_na = FALSE) %>%   
# filter(!(c('re\_74','re\_75','re\_78')==0))

**16.**

titanic = import("2021\_exam2\_data.xlsx", which=1) #load data

**17.**

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

**18.**

library(gmodels)  
CrossTable(titanic$female, titanic$survived)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 2201   
##   
##   
## | titanic$survived   
## titanic$female | 0 | 1 | Row Total |   
## ---------------|-----------|-----------|-----------|  
## 0 | 126 | 344 | 470 |   
## | 116.071 | 243.243 | |   
## | 0.268 | 0.732 | 0.214 |   
## | 0.085 | 0.484 | |   
## | 0.057 | 0.156 | |   
## ---------------|-----------|-----------|-----------|  
## 1 | 1364 | 367 | 1731 |   
## | 31.515 | 66.045 | |   
## | 0.788 | 0.212 | 0.786 |   
## | 0.915 | 0.516 | |   
## | 0.620 | 0.167 | |   
## ---------------|-----------|-----------|-----------|  
## Column Total | 1490 | 711 | 2201 |   
## | 0.677 | 0.323 | |   
## ---------------|-----------|-----------|-----------|  
##   
##

Those who were female died at a rate almost ten times higher than those who were not female. **19.**

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

**19.**

titanic <- titanic %>% mutate(first\_class = ifelse(class == 1, "yes", "no"))  
table(titanic$first\_class, exclude=TRUE)

##   
## no yes   
## 1876 325