Exam\_2\_official

Javvadi

6/28/2021

###strating

getwd()

## [1] "/Users/praveenajavvadi/College/Data\_Science/Exam-2"

rm(list =ls(all=TRUE))  
  
#importing data  
library(rio)  
college\_scorecard <- import("2021\_exam2\_data.xlsx", which = 4)  
  
#viewing summary statistics  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

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

#subseting data  
  
#only for LA and TX schools  
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

small\_scorecard <-  
 college\_scorecard %>%  
 dplyr::filter((state\_abbr=="LA"|  
 state\_abbr=="TX"))  
#filter for 4 years  
small\_scorecard<-  
 small\_scorecard %>%  
 dplyr::filter(pred\_degree\_awarded\_ipeds=="3")  
#filter for 2014 & 2015  
small\_scorecard<-  
 small\_scorecard %>%  
 dplyr::filter(year==2014|  
 year==2015)

#Starting avocados

#importing data for avocadoes  
avocados <- import("2021\_exam2\_data.xlsx", which = 2)  
  
#creating year variable  
#need these packages  
library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.1.0 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.3 ✓ forcats 0.5.1  
## ✓ readr 1.4.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x lubridate::as.difftime() masks base::as.difftime()  
## x lubridate::date() masks base::date()  
## x dplyr::filter() masks stats::filter()  
## x lubridate::intersect() masks base::intersect()  
## x dplyr::lag() masks stats::lag()  
## x lubridate::setdiff() masks base::setdiff()  
## x lubridate::union() masks base::union()

#creating year variable  
avocados <-   
avocados %>%   
 dplyr::mutate(year = lubridate::year(avocados$date))  
  
  
#deflating average price variable  
#load package  
library(WDI)  
deflator\_data = WDI(country ="all", indicator = c("NY.GDP.DEFL.ZS"),  
 start = 2015,  
 end = 2018,  
 extra = FALSE, cache = NULL)  
  
library(data.table)

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

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

## 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

setnames(deflator\_data,"NY.GDP.DEFL.ZS","deflator")  
  
  
#working in US dollars means we only need US data  
usd\_deflator = subset(deflator\_data, country == "United States")  
  
#the base of the deflator, where are we deflating or inflating from  
subset(usd\_deflator, deflator ==100)

## iso2c country deflator year  
## 1016 US United States 100 2015

#remove things we don't need anymore  
rm(deflator\_data)  
  
#merge the deflator data avocado data frame  
avocados = left\_join(avocados,  
 usd\_deflator,  
 by=c("year"))  
#now lets deflate the data  
avocados$deflatedprice\_2015 = avocados$average\_price/  
 (avocados$deflator/100)  
  
#collapsing data frame  
collapsed\_avocados <-  
 avocados%>%  
 group\_by(year)%>%  
 summarize(across(where(is.numeric), sum))%>%  
 select(-c("deflator","total\_volume", "average\_price"))  
  
  
#reshaping the collapse avocados dataframe   
wide\_avocados <-  
 collapsed\_avocados%>%  
 pivot\_wider(id\_cols =c("year"),  
 # unique IDs  
 names\_from = "year",  
 # names for new wide vars  
 values\_from = "deflatedprice\_2015")  
# data to put in new wide vars  
head(wide\_avocados)

## # A tibble: 1 x 4  
## `2015` `2016` `2017` `2018`  
## <dbl> <dbl> <dbl> <dbl>  
## 1 53.1 53.8 64.5 12.3

#Training Dataset

#training data set questions  
training <- import("2021\_exam2\_data.xlsx", which = 3)  
  
#creating id variable  
training$id =1:nrow(training)  
#reshaping long  
long\_data\_frame <-  
 training%>%  
 pivot\_longer(cols =starts\_with("re\_"),  
 # use columns starting with "re\_"  
 names\_to ="earnings",  
 # name of new column  
 names\_prefix = "re\_",  
 # part of string to drop  
 values\_to = "id",  
 # where to put numeric values  
 names\_repair = "minimal")  
  
#summary statistics for long data frame  
summary(long\_data\_frame)

## 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 Length:1335   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:112 Class :character   
## Median :0.00000 Median :0.0000 Median :223 Mode :character   
## Mean :0.08764 Mean :0.1685 Mean :223   
## 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:334   
## Max. :1.00000 Max. :1.0000 Max. :445   
## id   
## Min. : 0   
## 1st Qu.: 0   
## Median : 0   
## Mean : 2927   
## 3rd Qu.: 4045   
## Max. :60308

#titanic dataset

#titanic questions  
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

#cross tabulation   
library(doBy)

##   
## Attaching package: 'doBy'

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

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

## survived female.mean female.length  
## 1 0 0.9154362 1490  
## 2 1 0.5161744 711

women were more likely to survive than men

#new variable  
titanic$first\_class = ifelse(test = titanic$class==2, yes = "first class", no = "not first class")  
  
#frequency table  
table(titanic$first\_class)

##   
## first class not first class   
## 285 1916

#“My Heart Will Go On” by Celine Dion