exam2

# Clearing the R environment

rm(list=ls(all=TRUE))

# Working with College\_Scorecard Data Sheet

## Loading the college\_scorecarddataset in R, naming it “college\_scorecard”

library(readxl)  
college\_scorecard <-  
 read\_excel("2021\_exam2\_data.xlsx", 4)

## Providing summary statistics for thecollege\_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

## Creating smaller dataset consisting only of data measured in 2014 and 2015 on formerstudents who graduated from four-year+ colleges and universities located inTexas (state\_abbr: “TX”) and Louisiana (state\_abbr: “LA”).

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

library(tidyverse)

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

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.1.2 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.3 ✓ forcats 0.5.1  
## ✓ readr 1.4.0

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

small\_scorecard1 <-   
 subset(college\_scorecard, year == c(2014, 2015))

## Warning in year == c(2014, 2015): longer object length is not a multiple of  
## shorter object length

#small\_scorecard <-   
 # subset(small\_scorecard1, state.abb == "LA|TX")  
 ##^^ I could not figure out how to subset for the two states, kept receiving the error "Input has size 6018 but subscript `r` has size 50"

# Working with Acovados Data Sheet

## Loading the college\_scorecarddataset in R, naming it “college\_scorecard”

library(readxl)  
avocados <-  
 read\_excel("2021\_exam2\_data.xlsx", 2)

## Creating new variable called year with the year avocadoes were sold

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

## Deflating average price variable using WDI data

library(WDI)  
  
#creating deflator set  
deflator\_data =WDI(country = "all", indicator =c("NY.GDP.DEFL.ZS"),  
 start = 1960, # start of foreign aid data  
 end = 2019, # end of of foreign aid data  
 extra = FALSE, cache = NULL)  
  
#renaming deflator variable  
library(data.table)

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

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

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

setnames(deflator\_data,"NY.GDP.DEFL.ZS", "deflator")  
  
#selectin only US data  
usd\_deflator =subset(deflator\_data, country=="United States")  
  
#getting base year  
subset(usd\_deflator, deflator == 100)

## iso2c country deflator year  
## 15185 US United States 100 2015

# drop unnecessary variable  
usd\_deflator$country <- NULL  
usd\_deflator$iso2c <- NULL  
  
#merging data  
deflated\_data =  
 left\_join(avocados,  
 usd\_deflator,  
 by=c("year"))  
  
#actually deflating the data  
deflated\_data$deflated\_amount = deflated\_data$average\_price/(deflated\_data$deflator/100)  
  
#renamed data with deflated price  
collapsed\_avocados <- deflated\_data  
  
  
#cleaning workspace  
rm(deflated\_data)  
rm(deflator\_data)  
rm(usd\_deflator)  
  
#showing output  
head(collapsed\_avocados)

## # A tibble: 6 x 6  
## date average\_price total\_volume year deflator deflated\_amount  
## <dttm> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2015-12-27 00:00:00 0.9 5040365. 2015 100 0.9   
## 2 2015-12-20 00:00:00 0.94 4695737. 2015 100 0.94  
## 3 2015-12-13 00:00:00 0.87 5259354. 2015 100 0.87  
## 4 2015-12-06 00:00:00 0.78 5775536. 2015 100 0.78  
## 5 2015-11-29 00:00:00 0.91 4575711. 2015 100 0.91  
## 6 2015-11-22 00:00:00 0.92 4804278. 2015 100 0.92

#pivotting data  
wide\_avocados <-  
 collapsed\_avocados %>%  
 pivot\_wider(id\_cols =c("year"),# unique IDs  
 names\_from = "date",# names for new wide vars  
 values\_from = "deflated\_amount")# data to put in new wide vars  
  
head(wide\_avocados)

## # A tibble: 4 x 170  
## year `2015-12-27` `2015-12-20` `2015-12-13` `2015-12-06` `2015-11-29`  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2015 0.9 0.94 0.87 0.78 0.91  
## 2 2016 NA NA NA NA NA   
## 3 2017 NA NA NA NA NA   
## 4 2018 NA NA NA NA NA   
## # … with 164 more variables: 2015-11-22 <dbl>, 2015-11-15 <dbl>,  
## # 2015-11-08 <dbl>, 2015-11-01 <dbl>, 2015-10-25 <dbl>, 2015-10-18 <dbl>,  
## # 2015-10-11 <dbl>, 2015-10-04 <dbl>, 2015-09-27 <dbl>, 2015-09-20 <dbl>,  
## # 2015-09-13 <dbl>, 2015-09-06 <dbl>, 2015-08-30 <dbl>, 2015-08-23 <dbl>,  
## # 2015-08-16 <dbl>, 2015-08-09 <dbl>, 2015-08-02 <dbl>, 2015-07-26 <dbl>,  
## # 2015-07-19 <dbl>, 2015-07-12 <dbl>, 2015-07-05 <dbl>, 2015-06-28 <dbl>,  
## # 2015-06-21 <dbl>, 2015-06-14 <dbl>, 2015-06-07 <dbl>, 2015-05-31 <dbl>,  
## # 2015-05-24 <dbl>, 2015-05-17 <dbl>, 2015-05-10 <dbl>, 2015-05-03 <dbl>,  
## # 2015-04-26 <dbl>, 2015-04-19 <dbl>, 2015-04-12 <dbl>, 2015-04-05 <dbl>,  
## # 2015-03-29 <dbl>, 2015-03-22 <dbl>, 2015-03-15 <dbl>, 2015-03-08 <dbl>,  
## # 2015-03-01 <dbl>, 2015-02-22 <dbl>, 2015-02-15 <dbl>, 2015-02-08 <dbl>,  
## # 2015-02-01 <dbl>, 2015-01-25 <dbl>, 2015-01-18 <dbl>, 2015-01-11 <dbl>,  
## # 2015-01-04 <dbl>, 2016-12-25 <dbl>, 2016-12-18 <dbl>, 2016-12-11 <dbl>,  
## # 2016-12-04 <dbl>, 2016-11-27 <dbl>, 2016-11-20 <dbl>, 2016-11-13 <dbl>,  
## # 2016-11-06 <dbl>, 2016-10-30 <dbl>, 2016-10-23 <dbl>, 2016-10-16 <dbl>,  
## # 2016-10-09 <dbl>, 2016-10-02 <dbl>, 2016-09-25 <dbl>, 2016-09-18 <dbl>,  
## # 2016-09-11 <dbl>, 2016-09-04 <dbl>, 2016-08-28 <dbl>, 2016-08-21 <dbl>,  
## # 2016-08-14 <dbl>, 2016-08-07 <dbl>, 2016-07-31 <dbl>, 2016-07-24 <dbl>,  
## # 2016-07-17 <dbl>, 2016-07-10 <dbl>, 2016-07-03 <dbl>, 2016-06-26 <dbl>,  
## # 2016-06-19 <dbl>, 2016-06-12 <dbl>, 2016-06-05 <dbl>, 2016-05-29 <dbl>,  
## # 2016-05-22 <dbl>, 2016-05-15 <dbl>, 2016-05-08 <dbl>, 2016-05-01 <dbl>,  
## # 2016-04-24 <dbl>, 2016-04-17 <dbl>, 2016-04-10 <dbl>, 2016-04-03 <dbl>,  
## # 2016-03-27 <dbl>, 2016-03-20 <dbl>, 2016-03-13 <dbl>, 2016-03-06 <dbl>,  
## # 2016-02-28 <dbl>, 2016-02-21 <dbl>, 2016-02-14 <dbl>, 2016-02-07 <dbl>,  
## # 2016-01-31 <dbl>, 2016-01-24 <dbl>, 2016-01-17 <dbl>, 2016-01-10 <dbl>,  
## # 2016-01-03 <dbl>, 2017-12-31 <dbl>, …

# Training Dataset Questions

#loading training dataset  
library(readxl)  
training <-  
 read\_excel("2021\_exam2\_data.xlsx", 3)

# Titanic Dataset Questions

library(readxl)  
titanic <-  
 read\_excel("2021\_exam2\_data.xlsx", 1)  
  
#providing summary statistics  
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

##creating cross tab of survivorship by gender  
male\_passengers <-   
 subset(titanic, female == 0)  
  
female\_passengers <-   
 subset(titanic, female == 1)  
  
#creating gender variable  
titanic$gender = NA  
titanic$gender[titanic$female ==1] = "Female"  
titanic$gender[titanic$female ==0] = "Male"  
  
#$checking table to esnure it all went through  
table(titanic$gender, exclude = TRUE)

##   
## Female Male   
## 1731 470

library(doBy)

##   
## Attaching package: 'doBy'

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

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

## # A tibble: 2 x 3  
## gender survived.mean survived.length  
## <chr> <dbl> <dbl>  
## 1 Female 0.212 1731  
## 2 Male 0.732 470

titanic$first\_class = NA  
#ifelse(test = class==1, yes = )

#Ran out of time, spent too much on the first set of Qs

##Thanks Professors Findley and Denly!