#college\_scorecard data set

# clear environment

rm(list=ls(all=TRUE))

# Import college\_scprecard data set

library(readxl) college\_scorecard <- read\_excel(“~/Desktop/Summer 2021/2021\_exam2\_data.xlsx”, sheet = “college\_scorecard”) View(college\_scorecard)

##summary summary(college\_scorecard)

library (tidyverse)

##create smaller data set small\_scorecard <- subset (college\_scorecard, select= c(college\_scorecard, year== 2014, state\_abbr== “TX”, pred\_degree\_awarded\_ipeds== 3))

#AVOCADO DATA SET

##import avocado data library(readxl) avocados <- read\_excel(“~/Desktop/Summer 2021/2021\_exam2\_data.xlsx”, sheet = “avocados”) View(avocados)

##create variable year that only captures the year in which they were sold library (lubridate) year= avocados %>% dplyr::mutate(year = lubridate::year(avocados$date))

##deflate the average\_price library(WDI) Google “GDP deflator World Bank” <https://data.worldbank.org/indicator/NY.GDP.DEFL.ZS> deflator\_data = WDI(country = “all”, indicator = c(“NY.GDP.DEFL.ZS”), start = 1960, # start of foreign aid data end = 2018, # end of of foreign aid data extra = FALSE, cache = NULL)

##rename it using data.table package library(data.table) setnames(deflator\_data,“NY.GDP.DEFL.ZS”, “deflator”)

##gdp deflator for the US # select only the United States data usd\_deflator = subset(deflator\_data, country==“United States”)

##figure out the year base deflator subset(usd\_deflator, deflator==100) >>> 2015

#merge the deflator into our avocados frame. #drop deflator\_data first, you dont need it anymore # remove the rest of the deflator data rm(deflator\_data)

# drop unnecessary variables

usd\_deflatoriso2c <- NULL

#drop usd\_deflator remove(usd\_deflator)

#merge collapsed\_avocados= left\_join(year, usd\_deflator, by=c(“year”))

#deflate the data collapsed\_avocadosaverage\_price/ (collapsed\_avocados$deflator/100)

#collapse collapsed\_avocados %>% group\_by( year) %>% # tell R the unique IDs summarize(across(where(is.numeric), sum)) %>% # summarize numeric vars by sum select(c(“deflator”)) # drop transaction id (no longer relevant)

#show the new data frame head(collapsed\_avocados)

#reshape collapsed\_avocados wide

wide\_avocados <- collapsed\_avocados %>% pivot\_wider(id\_cols = c( “year”, “total\_volume”, “deflated\_price”), # unique IDs names\_from = “year”, # names for new wide vars values\_from = “year”, # data to put in new wide vars names\_prefix = “year\_” ) # prefix to add before years

head (wide\_avocados)

# label the variables

library(labelled) var\_label(wide\_avocados) <- list(total\_volume = “Volume of Sales”, year\_2015 = “2015”, deflated\_price = “Deflated Price”, year\_2016 = “2016”, year\_2017= “2017”, year\_2018 = “2018”)

TRAINING DATA SET

#import data set library(readxl) training <- read\_excel(“~/Desktop/Summer 2021/2021\_exam2\_data.xlsx”, sheet = “training”) View(training)

#reshape so that all earning are in a single column training$id = 1:nrow(training)

library(dplyr)

earnings= coalesce(trainingre\_75, training$re\_78)

training$earnings <- earnings

training\_reshape <- training %>% pivot\_longer(cols = starts\_with(“id”), # use columns starting with “year” names\_to =“idr”, # name of new column names\_prefix = “id\_”, # part of string to drop values\_to = “earnings”, # where to put numeric values values\_drop\_na = FALSE) %>% # don’t drop NAs

TITANIC DATA SET #import data set library(readxl) titanic <- read\_excel(“~/Desktop/Summer 2021/2021\_exam2\_data.xlsx”, sheet = “titanic”) View(titanic)

#summary statistics summary (titanic)

#cross-tabulation library(doBy) summaryBy(survived ~ female, data=titanic, FUN=c(mean,length))

According to these results, on average, more women survived in quantity (1731 vs 470), eventhough men have a higher mean of survival