Week 2 Homework

Noah Foilb

9/4/2020

## Week 2 Homework

### Working with Excel files

#### 1. Load the libraries to read .csv and .xlsx files (comma-separated variable and Excel files) into R

library(readr) #this library reads .csv files.   
library(readxl) #this library reads Excel files

#### 2. Read the Covid-19 Weekly Public Heatlh Report - Raw Data from mass.gov into R.

* This is an Excel file
* First, download the file from Blackboard to your computer. \*The original file is from mass.gov. If you’d like to view the source, see “Weekly Public Health Report - Raw Data - (Date)” at <https://www.mass.gov/info-details/covid-19-response-reporting#covid-19-daily-dashboard->
* Next, use the appropriate read command to load the Excel file into R. Include the following:
  + This file has multiple worksheets. Read in the sheet entitled “City\_Town\_Data”.
  + There ARE column names in the Excel file
  + We have to code spaces, \* and <5 entries from this spreadsheet as Not Available (NA) data. To do so, add na=c("“,”\*“,”<5") after the sheet= specifications (see the section called Data Import Using read\_excel() from the week\_2B code).
  + Use the assignment operator (<-) to name this dataframe covid\_data

# This is finding the location of the file and saving that location in Rstudios  
file\_covid <- "\\Users\\nofoi\\OneDrive\\Desktop\\R studios Data Science\\Homework\\Noah Foilb Week 2\\MA Covid.xlsx"  
  
# This is executing the loctation previously found and saving it as a data set  
covid\_data <- read\_excel(file\_covid,col\_names=TRUE, sheet="City\_Town\_Data",na=c("", "\*", "<5"))

### Data reformatting step

* Run this piece of code. It will work as long as your dataframe is named covid\_data. The case does matter, so if you named your dataframe Covid\_data, it won’t work
* You’re not expected to know what this code does, but comments are added if you’re curious

covid\_data <- within(covid\_data, {  
 Total\_case\_count <- as.numeric(Total\_case\_count) #codes Total\_case\_count as numeric data  
 Two\_Week\_Case\_Count <- as.numeric(Two\_Week\_Case\_Count) #codes Two\_Week\_Case\_Count as numeric data  
 Average\_Daily\_Incidence\_Rate\_per\_100000 <- as.numeric(Average\_Daily\_Incidence\_Rate\_per\_100000) #codes Average\_Daily\_Incidence\_Rate\_per\_100000 as numeric data  
 Fraction\_positive <- as.numeric(Fraction\_positive) #codes Fraction\_positive as numeric data  
 Relative\_Change\_in\_Case\_Count <- factor(Relative\_Change\_in\_Case\_Count, levels=c("Lower", "No Change", "Higher")) #codes Relative\_Change\_in\_Case\_Count as an ordered factor where in the order Lower<No Change<Higher  
 Change\_in\_fraction\_positivity <- factor(Change\_in\_fraction\_positivity, levels=c("Lower", "No Change", "Higher")) #codes Change\_in\_fraction\_positivity as an ordered factor where in the order Lower<No Change<Higher  
})  
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

covid\_data <- filter(covid\_data, City\_or\_Town!="Unknown"&City\_or\_Town!="State")

#### 3. Explore the data using the str() command and answer the following questions:

* How many numerical (num) variables are in the dataframe?
* How many character (chr) variables?
* How many categorical variables (Factor)?
* How many rows?
* How many columns?
  + Hint: tibble [503 × 18] would mean there are 503 rows and 18 columns

str(covid\_data) # This gives us the structure of the data set

## tibble [351 x 10] (S3: tbl\_df/tbl/data.frame)  
## $ City\_or\_Town : chr [1:351] "Abington" "Acton" "Acushnet" "Adams" ...  
## $ Total\_case\_count : num [1:351] 252 179 106 35 516 NA 197 121 345 NA ...  
## $ Two\_Week\_Case\_Count : num [1:351] NA 5 NA 0 9 0 NA NA 14 0 ...  
## $ Average\_Daily\_Incidence\_Rate\_per\_100000: num [1:351] 1.59 1.5 2.05 0 2.25 ...  
## $ Relative\_Change\_in\_Case\_Count : Factor w/ 3 levels "Lower","No Change",..: 1 3 3 2 1 2 1 1 1 2 ...  
## $ Total\_tests : num [1:351] 3858 5475 2047 1377 8841 ...  
## $ Total\_tests\_last\_14\_days : num [1:351] 525 1088 276 233 1927 ...  
## $ Positive\_Tests\_Last\_14\_days : num [1:351] 8 6 4 0 14 0 4 5 14 0 ...  
## $ Fraction\_positive : num [1:351] 0.01524 0.00551 0.01449 0 0.00727 ...  
## $ Change\_in\_fraction\_positivity : Factor w/ 3 levels "Lower","No Change",..: 1 3 3 2 1 2 1 1 1 2 ...

# There are seven "num" variables, one "chr" variable, and two "factor" variables.  
# There are 305 rows and 10 columns

#### 4. What are the first three cities/towns in the dataframe? What are the last five? Of these eight cities/towns, how many are missing data for the two week case count?

* For privacy reason, the state of MA won’t report case counts that are <5. Originally, these towns had this entry recorded as <5, but you recoded the entry to NA in your read\_excel() command.

# The frist three cities are Abington, Acton, and Acushnet.  
# The last five are Woburn, Worcester, Worthington, Wrentham, and Yarmouth  
# Four are counted as NA

### Exploring numerical variables

#### 5. Provide a numerical summary of the number of positive tests in the last 14 days (Positive\_Tests\_Last\_14\_days)

# By calling the function "summary" and by specifically calling the column heading of Postive\_Tests\_Last\_14\_days, we get one column summary  
summary(covid\_data$Positive\_Tests\_Last\_14\_days)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 3.00 13.19 9.00 744.00

#### 6. Provide the quantiles for the number of positive tests in the last 14 days (Positive\_Tests\_Last\_14\_days)

# By using the function Quantile it given us the quantiles of the data  
quantile(covid\_data$Positive\_Tests\_Last\_14\_days)

## 0% 25% 50% 75% 100%   
## 0 0 3 9 744

#### 7. What are the five highest two week case counts (Two\_Week\_Case\_Count)?

# By using head to display the first five data, and sort to get the data in order we are able to find the five highesr two week case counts  
head(sort(covid\_data$Two\_Week\_Case\_Count,decreasing = TRUE,),5)

## [1] 627 334 161 126 117

### Exploring categorical variables

* Before proceeding, recall what are the two categorical (Factor) variables (see question 2). Answer these questions using these categorical variables.

#### 8. How many towns had their fraction of positive test results (Change\_in\_fraction\_positivity) higher than last week? Lower? The same?

#By using summary we can find the change in a categorical graph  
summary(covid\_data$Change\_in\_fraction\_positivity)

## Lower No Change Higher   
## 148 131 72

# 148 lower, 131 the same, and 72 higher.

#### 9. How many towns had a decrease in the relative change in case counts (Relative\_Change\_in\_Case\_Count)? An increase? No change?

#By using summary we can find the change in a categorical graph  
summary(covid\_data$Relative\_Change\_in\_Case\_Count)

## Lower No Change Higher   
## 129 128 94

# 129 lower, 128 the same, and 94 higher

#### 10. Create a two-way table or a contingency table for the two categorical variables from questions 8 and 9 listing how many times each possible combination of the two values occurs. Do you notice any correlation between these two variables?

#By using table we can find the change in a categorical graph  
table(covid\_data$Change\_in\_fraction\_positivity,covid\_data$Relative\_Change\_in\_Case\_Count)

##   
## Lower No Change Higher  
## Lower 115 22 11  
## No Change 12 99 20  
## Higher 2 7 63

# there seems to be a strong corelation between the diagnol of this table (LowervsLower, NoChangevsNochange, and HighervsHigher)

### Working with .csv files

#### 11. The following website provides data on companies in the S&P 500. The data is in the form of a .csv file. Read the data into R.

* <https://raw.githubusercontent.com/datasets/s-and-p-500-companies-financials/master/data/constituents-financials.csv>
* First, assign the URL to a variable
* Use the proper read statement to read the file into R
  + The columns DO have column names
  + Make sure you assign the output from the read statement to a variable name

#assign the URL as a string to the variable  
companies\_url <- "https://raw.githubusercontent.com/datasets/s-and-p-500-companies-financials/master/data/constituents-financials.csv"  
  
#inpute the variable into read.csv to aquire the data  
companies\_data <- read.csv(companies\_url)

#### 12. What types of variables does the dataset contain (numerical, categorical, etc.), and how many of each type are there?

#Gives the structure of the dataset  
str(companies\_data)

## 'data.frame': 505 obs. of 14 variables:  
## $ Symbol : chr "MMM" "AOS" "ABT" "ABBV" ...  
## $ Name : chr "3M Company" "A.O. Smith Corp" "Abbott Laboratories" "AbbVie Inc." ...  
## $ Sector : chr "Industrials" "Industrials" "Health Care" "Health Care" ...  
## $ Price : num 222.9 60.2 56.3 108.5 150.5 ...  
## $ Price.Earnings: num 24.3 27.8 22.5 19.4 25.5 ...  
## $ Dividend.Yield: num 2.33 1.15 1.91 2.5 1.71 ...  
## $ Earnings.Share: num 7.92 1.7 0.26 3.29 5.44 1.28 7.43 3.39 6.19 0.03 ...  
## $ X52.Week.Low : num 259.8 68.4 64.6 125.9 162.6 ...  
## $ X52.Week.High : num 175.5 48.9 42.3 60 114.8 ...  
## $ Market.Cap : num 1.39e+11 1.08e+10 1.02e+11 1.81e+11 9.88e+10 ...  
## $ EBITDA : num 9.05e+09 6.01e+08 5.74e+09 1.03e+10 5.64e+09 ...  
## $ Price.Sales : num 4.39 3.58 3.74 6.29 2.6 ...  
## $ Price.Book : num 11.34 6.35 3.19 26.14 10.62 ...  
## $ SEC.Filings : chr "http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=MMM" "http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=AOS" "http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=ABT" "http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=ABBV" ...

#There are four "chr" and ten "num"

#### 13. How many companies are in the dataset (have R report the number of rows)? Does this surprise you?

#Use nrow to find the number of rows  
nrow(companies\_data)

## [1] 505

#There are 505 which is a lot more companies then I would have thought

#### 14. What are the first three and last three companies in the dataset?

#First three  
head(companies\_data,3)

## Symbol Name Sector Price Price.Earnings Dividend.Yield  
## 1 MMM 3M Company Industrials 222.89 24.31 2.332862  
## 2 AOS A.O. Smith Corp Industrials 60.24 27.76 1.147959  
## 3 ABT Abbott Laboratories Health Care 56.27 22.51 1.908982  
## Earnings.Share X52.Week.Low X52.Week.High Market.Cap EBITDA Price.Sales  
## 1 7.92 259.77 175.490 138721055226 9.048e+09 4.390271  
## 2 1.70 68.39 48.925 10783419933 6.010e+08 3.575483  
## 3 0.26 64.60 42.280 102121042306 5.744e+09 3.740480  
## Price.Book SEC.Filings  
## 1 11.34 http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=MMM  
## 2 6.35 http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=AOS  
## 3 3.19 http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=ABT

#Last three  
tail(companies\_data,3)

## Symbol Name Sector Price Price.Earnings  
## 503 ZBH Zimmer Biomet Holdings Health Care 115.53 14.32  
## 504 ZION Zions Bancorp Financials 50.71 17.73  
## 505 ZTS Zoetis Health Care 71.51 32.80  
## Dividend.Yield Earnings.Share X52.Week.Low X52.Week.High Market.Cap  
## 503 0.7948336 9.01 133.49 108.17 24454698119  
## 504 1.4809330 2.60 55.61 38.43 10670678640  
## 505 0.6823720 1.65 80.13 52.00 35991109776  
## EBITDA Price.Sales Price.Book  
## 503 2007400000 3.164895 2.39  
## 504 0 3.794579 1.42  
## 505 1734000000 9.280896 18.09  
## SEC.Filings  
## 503 http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=ZBH  
## 504 http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=ZION  
## 505 http://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=ZTS

#First Three: 3M Company, A.O. Smith Corp and Abbott Laboratories  
#Last Three: Zimmer Biomet Holdings, Zions Cancorp, and Zoetis

#### 15. Produce a numerical summary of the Market Cap.

* The column names aren’t in a good format for R at the moment. We’ll learn more about that later, but for now you’ll need to put back quotes around the column names, i.e., Market Cap
  + the backquote should be in upper left of your keyboard

# Use summary to find the summary  
summary(companies\_data$`Market.Cap`)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.626e+09 1.273e+10 2.140e+10 4.924e+10 4.512e+10 8.095e+11

#### 16. Produce a table showing the number of companies in each Sector in the S&P500 (the column name is Sector)

# use table to make the table  
table(companies\_data$Sector)

##   
## Consumer Discretionary Consumer Staples   
## 84 34   
## Energy Financials   
## 32 68   
## Health Care Industrials   
## 61 67   
## Information Technology Materials   
## 70 25   
## Real Estate Telecommunication Services   
## 33 3   
## Utilities   
## 28