

Quiz1

Mihai M. Craiu

4/7/2020

Q1

The American Community Survey distributes downloadable data about United States communities. Download the 2006 microdata survey about housing for the state of Idaho using `download.file()` from here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv>

and load the data into R. The code book, describing the variable names is here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDict06.pdf>

How many properties are worth \$1,000,000 or more?

```
# URL for the data and de pdf code book
file_housing_Url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"
code_book_pdf_Url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDict06.pdf"

#downloading the code book
download.file(code_book_pdf_Url, destfile = "./codebook.pdf", method = "curl")
#downloading the data
download.file(file_housing_Url, destfile = "./microdata_survey_housing.csv")
#reading the data

#Downloading date
dateDownloaded_q1 <- date()
dateDownloaded_q1
```

```
## [1] "Sat Jul 04 21:09:32 2020"
```

```
housingdata <- read.csv("microdata_survey_housing.csv")
#head(housingdata)

# Answer for the question

# VAL attribute says how much property is worth, code book
sum(housingdata$VAL == 24, na.rm = TRUE)
```

```
## [1] 53
```

Q2.

Use the data you loaded from Question 1. Consider the variable FES in the code book. Which of the “tidy data” principles does this variable violate?

Answer:

Tidy data one variable per column

Q3.

Download the Excel spreadsheet on Natural Gas Aquisition Program here:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx

Read rows 18-23 and columns 7-15 into R and assign the result to a variable called:

dat

What is the value of: `>sum(dat$Zip * dat$Ext, na.rm=T)`

```
# Url for the data
file_NaturalGas_Url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx"

# downloading the data .xlsx
download.file(file_NaturalGas_Url, destfile = "./NaturalGas.xlsx", method = "curl")

#Downloading date
dateDownloaded_q3 <- date()
dateDownloaded_q3
```

```
## [1] "Sat Jul 04 21:09:33 2020"
```

```
#Read rows 18-23 and columns 7-15 into R and assign the result to a variable called: dat
library(xlsx)

col <- 7:15
row <- 18:23
dat <- read.xlsx("NaturalGas.xlsx", sheetIndex=1, colIndex = col, rowIndex = row)
dat
```

```
##      Zip CuCurrent PaCurrent PoCurrent      Contact Ext      Fax email
## 1 74136         0         1         0 918-491-6998  0 918-491-6659    NA
## 2 30329         1         0         0 404-321-5711  NA      <NA>    NA
## 3 74136         1         0         0 918-523-2516  0 918-523-2522    NA
## 4 80203         0         1         0 303-864-1919  0      <NA>    NA
## 5 80120         1         0         0 345-098-8890 456      <NA>    NA
##      Status
## 1         1
## 2         1
## 3         1
## 4         1
## 5         1
```

```
# Answer
```

```
sum(dat$Zip*dat$Ext, na.rm = T)
```

```
## [1] 36534720
```

Q4.

Read the XML data on Baltimore restaurants from here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml>

How many restaurants have zipcode 21231?

```
library(XML)
```