Lab 3 Data import and export key

Fang Fang

9/14/2020

Preps

- Get a new folder (with a sub folder called data) for lab assignment 3
- Set this new folder as the working directory for this assignment.
- Establish a new R script for the assignment

Collect the SPSS dataset

Let's try to download the data using the URL from National Household Travel Survey.

Put it under this data sub-folder. Your can name your output file as `spss_NHTS.zip`. Please finish the script below.

```
getwd()
fileURL <- "https://nhts.ornl.gov/assets/2016/download/Spss.zip"
#Finish the code below
download.file(fileURL, destfile="data/spss_NHTS.zip")# finish the command

date_dl <- date()
```

Import SPSS data and export it out.

Extract the file called hhpub.sav from your zip file. Again our data is derived from National Household Travel Survey (https://nhts.ornl.gov/). This web is conducted by the Federal Highway Administration (FHWA), and it is the authoritative source on the travel behavior of the American public. It is the only source, and quite popular source of data that allows us to analyze trends in personal and household travel by time.

The `hhpub.sav` is a SPSS supported file. No worries if you have no experiences in SPSS. Let's try to import it in R, and export it out as a csv file later. Import the sav file and name it travel

```
# Please modify the syntax below to import your .sav file.
library(haven)
travel <- read_sav("hhpub.sav")
```

Q1 (5 points)

How many rows and columns does this data (travel) have?

Answer: 129696 rows and 58 columns

Briefly examine the columns. Here is the more information. https://nhts.ornl.gov/assets/NHTS2017_UsersGuide_04232019_1.pdf

Q2 (5 points)

Use sapply to examine the data type for each column use the class function

```
# Let's try to examine the class for each column. Try sapply!
sapply(travel, class)
```

Sometimes, the codes for data are provided in a separate code book file and you have to apply labels to the data yourself. Check out the code book here. https://nhts.ornl.gov/assets/codebook_v1.1.pdf

Let'e try to apply labels For column Urbanisize,

01=50,000 - 199,999

02=200,000 - 499,999

03=500,000 - 999,999

04=1 million or more without heavy rail

05=1 million or more with heavy rail

06=Not in an urbanized area

You will use function called factor, to created a new factor variable, and assign it to the existing column, to replace the codes.

```
# Now you will assign a new factor to the column called Urbanisize
new_factor <- factor(travel$urbansize,
        labels = c("<200000", "200,000 - 499,999",
    "500,000 - 999,999","1 million or more without heavy rail",
    "1 million or more with heavy rail ","Not in an urbanized area"))

# Try to assign this new_factor to your dataset
travel$urbansize <- new_factor
```

Let's try to extract records only for travel on weekend!

Hint: based on the code book, https://nhts.ornl.gov/assets/codebook_v1.1.pdf 01=Sunday, 07=Saturday. So you need the rows where the column called `travday` equals to 01 or 07. To translate into R command, you need identify where `travel$travday=='01'|travel$travday=='07'` The `|` means `or` condition.

Based on the hint above, complete the commands below to extract rows on weekend only and save it as a new object called `weekend` .

```
# Use the column called travday to subset the dataset
weekend <- travel[(travel$travday=='01'|travel$travday=='07'),]
```

Q 3 (10 points)

For the weekend dataset, which group travel most among all the categories in column called urbansize (6 groups as mentioned above)?

Hint: you can use the `table` function here.

```
table(weekend$urbansize)
```

Answer: People live in not an urbanized area travel the most.

Q4 (10 points)

Export records for weekend only to a new csv file. Name it weekend_sub.csv. You need to turn in this csv file in compass as delivery.

```
# Export weekend out as a csv. Modify the syntax below if needed.
write_csv(weekend, "data/weekend_sub.csv")
```

Q5 (10 points)

Monitoring beach water quality:

Assume you are working at the City of Chicago Park District where basically working on facilities management. You are interested in monitoring the beach water quality (e.g. temperature, turbidity etc). It is important to check on water quality conditions to protect our citizens health and make decisions e.g. when the beach season starts/ends. There are automated sensors installed in the water at beaches along Chicago's Lake Michigan lakefront. Now let's try to explore the beach water quality data generated by these sensors. Click here to view the data.

Load this water quality data into R using the URL. You can use the `read_csv` function.

What is the average water temperature (column called Water Temperature) for this dataset?

```
library(tidyverse)
myURL <- "https://data.cityofchicago.org/api/views/qmqz-2xku/rows.csv?accessType=DOWNLOAD"
water <- read_csv(myURL)

mean(water$`Water Temperature`)
```

```
## [1] 19.55211
```

Answer: average temperature is ~19.5

Q 6 (10 points)

Currently by default, the `Beach Name` column is under the data type of chr (character). Use the `factor` function to convert this column into a factor.

```
water$`Beach Name` <- factor(water$`Beach Name`)
```

Then use one of the following function: tapply/sapply to calculate the average water temperature for each beach (based on the Beach Name column). Which beach has the highest avg. water temperature?

```
tapply(water$`Water Temperature`,water$`Beach Name`,mean)
```

```
## 63rd Street Beach     Calumet Beach    Montrose Beach Ohio Street Beach
##          18.45990          20.37310          18.64053          20.43293
##    Osterman Beach     Rainbow Beach
##          17.93362          18.74125
```

Answer: Ohio street beach has the highest avg. temperature