

# Under the Wood

## Getting and Cleaning Data: Quiz 3

📅 2016-03-09

### Quiz

1. 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>

Create a logical vector that identifies the households on greater than 10 acres who sold more than \$10,000 worth of agriculture products. Assign that logical vector to the variable `agricultureLogical`. Apply the `which()` function like this to identify the rows of the data frame where the logical vector is TRUE.

```
which(agricultureLogical)
```

What are the first 3 values that result?

- 125, 238, 262
- 25, 36, 45
- 59, 460, 474
- 403, 756, 798

2. Using the `jpeg` package read in the following picture of your instructor into R

<https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg>

Use the parameter `native=TRUE`. What are the 30th and 80th quantiles of the resulting data? (some Linux systems may produce an answer 638 different for the 30th quantile)

- 10904118 -594524
- -15259150 -10575416
- -10904118 -10575416
- -16776430 -15390165

3. Load the Gross Domestic Product data for the 190 ranked countries in this data set:

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

Load the educational data from this data set:

[https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\\_Country.csv](https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv)

Match the data based on the country shortcode. How many of the IDs match? Sort the data frame in descending order by GDP rank (so United States is last). What is the 13th country in the resulting data frame?

Original data sources:

<http://data.worldbank.org/data-catalog/GDP-ranking-table>

<http://data.worldbank.org/data-catalog/ed-stats>

- 190 matches, 13th country is St. Kitts and Nevis
- 190 matches, 13th country is Spain
- 189 matches, 13th country is Spain
- 189 matches, 13th country is St. Kitts and Nevis
- 234 matches, 13th country is Spain
- 234 matches, 13th country is St. Kitts and Nevis

4. What is the average GDP ranking for the “High income: OECD” and “High income: nonOECD” group?

- 23, 45
- 23.966667, 30.91304
- 133.72973, 32.96667

- 32.96667, 91.91304
- 23, 30
- 30, 37

5. Cut the GDP ranking into 5 separate quantile groups. Make a table versus Income.Group. How many countries are Lower middle income but among the 38 nations with highest GDP?

- 5
- 13
- 12
- 0

## My solution

1. Download the file from the Internet, use browser or commands:

```
1  fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"
2  download.file(url = fileUrl, destfile = "getdata.csv", method = "curl")
3  data <- read.table("getdata.csv", header = TRUE, sep = ",")
4  head(data)
```

then analysis the data:

```
1  agricultureLogical <- data$ACR == 3 & data$AGS == 6
2  head(which(agricultureLogical), 3)
```

the result is 125 238 262 .


2. After install the “jpeg” package of R, download the file from the Internet, use command line:

```
1  library(jpeg)
2  fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg"
3  download.file(url = fileUrl, destfile = "jeff.jpg", method = "curl")
4  jpg <- readJPEG("jeff.jpg", native = TRUE)
5  quantile(jpg, probs = c(0.3, 0.8))
```

the result is -15259150 -10575416 .

3. Download the file from the Internet, use command line:

```
1  fileUrl1 <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
2  fileUrl2 <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country
3  download.file(fileUrl1, destfile = "GDP.csv", method = "curl")
4  download.file(fileUrl2, destfile = "country.csv", method = "curl")
```



load the file:

```
1  gdp <- read.csv("FGDP.csv", header = TRUE, skip = 3, sep = ",")
2  edu <- read.csv("country.csv", header = TRUE)
```

analysis the data:

```
1  library(data.table)
2  library(dplyr)
3  gdp <- fread("GDP.csv", skip = 4, nrow = 190, select = c(1, 2, 4, 5), col.names =
4  edu <- fread("country.csv")
5  View(gdp)
6  View(edu)
7  merge <- merge(gdp, edu, by = 'CountryCode')
```



show the result:

```
1  nrow(merge)
```

result is 189 .

```
1  arrange(merge, desc(Rank))[13, Economy]
```

result is "St. Kitts and Nevis".

4. analysis with the command line:

```
1 tapply(merge$Rank, merge$`Income Group`, mean)
```

shows:

1	High income: nonOECD	High income: OECD	Low income	Lower middle income
2	91.91304	32.96667	133.72973	107.70376
3	Upper middle income			
4	92.13333			

the result is 32.96667 91.91304 .

5. analysis with the command line:

```
1 merge$RankGroups <- cut(merge$Rank, breaks = 5)
2 table(merge$RankGroups, merge$`Income Group`)
```

the result is 5 .

The END.

◀ Getting and Cleaning Data: Quiz 2

Getting and Cleaning data: Course Project ▶

© 2016 ♥ Frank

Powered by [Hexo](#) | Theme - [NexT.Mist](#)