Potential Answers:

<https://github.com/bquast/Getting-Cleaning-Data/blob/master/Quiz3-Question3.R>

<https://github.com/benjamin-chan/GettingAndCleaningData/blob/master/Quiz3/quiz3.Rmd>

<https://github.com/theofpa/datascience/blob/master/getdata-quiz3.R>

**Question 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%2FPUMSDataDict06.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

153 ,236, 388

403, 756, 798

ACR Lot size

* b .N/A (GQ/not a one-family house or mobile home)
* 1 .House on less than one acre
* 2 .House on one to less than ten acres
* **3 .House on ten or more acres**

AGS Sales of Agriculture Products

* b .N/A (less than 1 acre/GQ/vacant/ 2 or more units in structure)
* 1 .None
* 2 .$ 1 - $ 999
* 3 .$ 1000 - $ 2499
* 4 .$ 2500 - $ 4999
* 5 .$ 5000 - $ 9999
* **6 .$10000+**

> url <- <https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv>  
> f <- file.path(getwd(), "ss06hid.csv") #Like this way to set path and file name  
> download.file(url, f)  
trying URL 'https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv'  
Content type 'text/csv' length 4246554 bytes (4.0 MB)  
opened URL  
downloaded 4.0 MB

> dt <- data.table(read.csv(f))  
> agricultureLogical <- dt$ACR == 3 & dt$AGS == 6  
> which(agricultureLogical)[1:3] #Finds the first three TRUE values.   
# Give the TRUE indices of a logical object, allowing for array indices  
[1] 125 238 262

**Question 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)

-16776430 -15390165

-15259150 -10575416

10904118 -594524

-14191406 -10904118

Install JPEG Package first.

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

> f <- file.path(getwd(), "jeff.jpg")

> download.file(url, f, mode="wb")

trying URL 'https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg'

Content type 'image/jpeg' length 23849 bytes (23 KB)

opened URL

downloaded 23 KB

> img <- readJPEG(f, native=TRUE)

> quantile(img, probs=c(0.3, 0.8))

30% 80%

-15259150 -10575416

**Question 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>   
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>

189 matches, 13th country is St. Kitts and Nevis

190 matches, 13th country is Spain

190 matches, 13th country is St. Kitts and Nevis

234 matches, 13th country is St. Kitts and Nevis

189 matches, 13th country is Spain

234 matches, 13th country is Spain

See Class3-Week3-Question3.R file in working directory



**Question 4**

What is the average GDP ranking for the "High income: OECD" and "High income: nonOECD" group?

133.72973, 32.96667

30, 37

23, 30

23, 45

23.966667, 30.91304

32.96667, 91.91304

**Question 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?

12

13

5

0

See file in working directory: Class 3 Quiz 3 Question 3-5

|  |  |
| --- | --- |
| cut2 {Hmisc} | R Documentation |

## **Cut a Numeric Variable into Intervals**

### Description

Function like cut but left endpoints are inclusive and labels are of the form [lower, upper), except that last interval is [lower,upper]. If cuts are given, will by default make sure that cuts include entire range ofx. Also, if cuts are not given, will cut x into quantile groups (g given) or groups with a given minimum number of observations (m). Whereas cut creates a category object, cut2 creates a factor object.



