Getting and Cleaning Data - Data Science - Quiz 3 - Coursera

Getting and Cleaning Data Quiz 3

This is Quiz 3 from the Getting and Cleaning Data course within the Data Science Specialization on Coursera. Topics include sorting, matching, and aggregating data.

Questions

---

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**

```
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv",destfile = "quiz2data.csv")

data <- read.csv("quiz2data.csv")

logic <- data$ACR ==3 & data$AGS == 6

which(logic)
```

```
##  [1]  125  238  262  470  555  568  608  643  787  808  824  849  952  955
## [15] 1033 1265 1275 1315 1388 1607 1629 1651 1856 1919 2101 2194 2403 2443
## [29] 2539 2580 2655 2680 2740 2838 2965 3131 3133 3163 3291 3370 3402 3585
## [43] 3652 3852 3862 3912 4023 4045 4107 4113 4117 4185 4198 4310 4343 4354
## [57] 4448 4453 4461 4718 4817 4835 4910 5140 5199 5236 5326 5417 5531 5574
## [71] 5894 6033 6044 6089 6275 6376 6420
```

---

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)

- **-15259150 -10575416**

```
library(jpeg)


download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg", destfile = "quiz2jpeg.jpg")
jpgdat = readJPEG("quiz2jpeg.jpg", native = TRUE)
quantile(jpgdat, probs = c(0.3, 0.8))
```

```
##       30%       80%
## -15259150 -10575416
```

---

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**

```
library(data.table)
library(dplyr)
```

```
## ------------------------------------------------------------------------
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
## ------------------------------------------------------------------------
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:data.table':
##
##     between, last
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv", destfile
= "quiz3data.csv")
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv
", destfile = "quiz3data2.csv")


countries = fread("quiz3data.csv", skip=4, nrows = 190, select = c(1, 2, 4, 5), col.names
=c("CountryCode", "Rank", "Economy", "Total"))
education = fread("quiz3data2.csv")
all = merge(countries, education, by = "CountryCode")
dim(all)
```

```
## [1] 189  34
```

```
all <- arrange(all, desc(Rank))


head(all,13)[33]
```

```
##                         Table Name
## 1                            Tuvalu
## 2                          Kiribati
## 3                 Marshall Islands
## 4                             Palau
## 5     S\xe3o Tom\xe9 and Principe
## 6           Micronesia, Fed. Sts.
## 7                             Tonga
## 8                          Dominica
## 9                           Comoros
```

```
## 10                           Samoa
## 11 St. Vincent and the Grenadines
## 12                         Grenada
## 13             St. Kitts and Nevis
```

---

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

- **32.96667, 91.91304**

```
unique(all$`Income Group`)
## [1] "Lower middle income"  "Upper middle income"  "Low income"
## [4] "High income: nonOECD" "High income: OECD"
tapply(all$Rank, all$`Income Group`, mean)
## High income: nonOECD    High income: OECD           Low income
##              91.91304             32.96667            133.72973
##  Lower middle income  Upper middle income
##             107.70370             92.13333
```

---

5. Cut the GDP ranking into 5 separate quantile groups. Make a table versus Income.Group. How many countries

---

- **5**

```
library(Hmisc)
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
```

```
##
##     combine, src, summarize
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
all$RankGroups <- cut2(all$Rank, g=5)
table(all$RankGroups, all$`Income Group`)
```

```
##
##              High income: nonOECD High income: OECD Low income
##   [  1, 39)                    4                18          0
##   [ 39, 77)                    5                10          1
##   [ 77,115)                    8                 1          9
##   [115,154)                    5                 1         16
##   [154,190]                    1                 0         11
##
##              Lower middle income Upper middle income
##   [  1, 39)                   5                   11
##   [ 39, 77)                  13                    9
##   [ 77,115)                  12                    8
##   [115,154)                   8                    8
##   [154,190]                  16                    9
```

```
table(all$RankGroups, all$`Income Group`)[4]
```

```
## [1] 5
```
