

# Week 3 Quiz

---

The due date for this quiz is Sun 27 Apr 2014 4:30 PM PDT.

Load packages.

```
packages <- c("data.table", "jpeg")
sapply(packages, require, character.only = TRUE, quietly = TRUE)
```

```
## data.table      jpeg
##      TRUE      TRUE
```

Fix URL reading for knitr. See [Stackoverflow](#).

```
setInternet2(TRUE)
```

## 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%2FPUMSDDataDict06.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?

```
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"
f <- file.path(getwd(), "ss06hid.csv")
download.file(url, f)
dt <- data.table(read.csv(f))
agricultureLogical <- dt$ACR == 3 & dt$AGS == 6
which(agricultureLogical)[1:3]
```

```
## [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?

```
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg"
f <- file.path(getwd(), "jeff.jpg")
download.file(url, f, mode = "wb")
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](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. 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>

```
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
f <- file.path(getwd(), "GDP.csv")
download.file(url, f)
dtGDP <- data.table(read.csv(f, skip = 4, nrows = 215))
dtGDP <- dtGDP[X != ""]
dtGDP <- dtGDP[, list(X, X.1, X.3, X.4)]
setnames(dtGDP, c("X", "X.1", "X.3", "X.4"), c("CountryCode", "rankingGDP",
  "Long.Name", "gdp"))
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv"
f <- file.path(getwd(), "EDSTATS_Country.csv")
download.file(url, f)
dtEd <- data.table(read.csv(f))
dt <- merge(dtGDP, dtEd, all = TRUE, by = c("CountryCode"))
sum(!is.na(unique(dt$rankingGDP)))
```

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

```
dt[order(rankingGDP, decreasing = TRUE), list(CountryCode, Long.Name.x, Long.Name.y,
rankingGDP, gdp)][13]
```

```
##      CountryCode      Long.Name.x      Long.Name.y rankingGDP   gdp
## 1:      KNA St. Kitts and Nevis St. Kitts and Nevis      178   767
```

## Question 4

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

```
dt[, mean(rankingGDP, na.rm = TRUE), by = Income.Group]
```

```
##      Income.Group      V1
## 1: High income: nonOECD  91.91
## 2:      Low income 133.73
## 3: Lower middle income 107.70
## 4: Upper middle income  92.13
## 5:   High income: OECD  32.97
## 6:      NA 131.00
## 7:      NaN
```

## 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?

```
breaks <- quantile(dt$rankingGDP, probs = seq(0, 1, 0.2), na.rm = TRUE)
dt$quantileGDP <- cut(dt$rankingGDP, breaks = breaks)
dt[Income.Group == "Lower middle income", .N, by = c("Income.Group", "quantileGDP")]
```

```
##      Income.Group quantileGDP  N
## 1: Lower middle income (38.8,76.6] 13
## 2: Lower middle income  (114,152]  8
## 3: Lower middle income  (152,190] 16
## 4: Lower middle income  (76.6,114] 12
## 5: Lower middle income   (1,38.8]  5
## 6: Lower middle income      NA    2
```