1 Week 3 Quiz

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?

(1) 236, 238, 262

(2) 25, 36, 45

(3) 403, 756, 798

(4) 125, 238,262

```
> url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"
> file <- file.path(getwd(), "ss06hid.csv")
> download.file(url, file, method = "curl")
> dt <- data.table(read.csv(file))
> agricultureLogical <- (dt$ACR == 3) & (dt$AGS == 6)
> which(agricultureLogical)[1:3]
[1] 125 238 262
```

Question 2

Instructions

- 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)*

Options

(1) 10904118 and 10575416

(2) 15259150 and 10575416

(3) 16776430 and 15390165

(4) 14191406 and 10904118

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

```
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg"
file <- file.path(getwd(), "jeff.jpg")
download.file(url, file, mode = "wb", method = "curl")
img <- readJPEG(file, native = TRUE)
```

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? **Options**

1 189, St. Kitts and Nevis

2 234, Spain

3 190, Spain

4 189, Spain

5 190, St. Kitts and Nevis

6 234, St. Kitts and Nevis

Original data sources: http://data.worldbank.org/data-catalog/GDP-ranking-table
http://data.worldbank.org/data-catalog/ed-stats

Explanation

```
> url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
> file <- file.path(getwd(), "GDP.csv")
> download.file(url, file, method = "curl")
> dtGDP <- data.table(read.csv(file, 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"
> file <- file.path(getwd(), "EDSTATS_Country.csv")
> download.file(url, file, method = "curl")
> dtEd <- data.table(read.csv(file))
> 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
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?
Options

(i) 23, 45

(ii) 30, 37

(iii) 23, 30

(iv) 23.966667, 30.91304

(v) 32.96667, 91.91304

(vi) 133.72973, 32.96667

```
> dt[, mean(rankingGDP, na.rm = TRUE), by = Income.Group]
Income.Group        V1
1: High income: nonOECD  91.91304
2:           Low income 133.72973
3:  Lower middle income 107.70370
4:  Upper middle income  92.13333
5:    High income: OECD  32.96667
6:                   NA 131.00000
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?
Options

(1) 0

(2) 18

(3) 13

(4) 5

```
> 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
```