# Week 3 Quiz

Vignesh

6/28/2020

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?

```
download.file(’https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv’
              , ’ACS.csv’
              , method=’curl’ )

# Read data into data.frame
ACS <- read.csv(’ACS.csv’)

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

```
## [1] 125 238 262
```

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)

```
library(jpeg)
# Download the file
download.file(’https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg’,’jeff.jpg’
              , mode=’wb’ )

# Read the image
picture <- jpeg::readJPEG(’jeff.jpg’, native=TRUE)

# Get Sample Quantiles corressponding to given prob
quantile(picture, probs = c(0.3, 0.8) )
```

```
##       30%       80%
## -15258512 -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

```
#install.packages("data.table") --> the reason for commenting this
#is because once you have installed the package you don’t have to
#do this operation again, uncomment this line only if you have not
#installed this package.
library(data.table)

#Download FGDP data
FGDP <- data.table::fread(’https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv’,
                          skip=5,
                          nrows = 190,
                          select = c(1, 2, 4, 5),
                          col.names=c("CountryCode", "Rank", "Economy","Total"))


# Download data and read FGDP data into data.table
FEDSTATS_Country <- data.table::fread(’https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_C


# Merging data from both the csv’s
mergedDT <- merge(FGDP, FEDSTATS_Country, by = ’CountryCode’)

# How many of the IDs match?
nrow(mergedDT)
```

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

```
# 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?
mergedDT[order(-Rank)][13,.(Economy)]
```

```
##                Economy
## 1: St. Kitts and Nevis
```

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

```
# "High income: OECD"
mergedDT[‘Income Group‘ == "High income: OECD"
         , lapply(.SD, mean)
         , .SDcols = c("Rank")
         , by = "Income Group"]
```

```
##         Income Group     Rank
## 1: High income: OECD 32.96667
```

```
# "High income: nonOECD"
mergedDT[‘Income Group‘ == "High income: nonOECD"
         , lapply(.SD, mean)
         , .SDcols = c("Rank")
         , by = "Income Group"]
```

```
##            Income Group     Rank
## 1: High income: nonOECD 91.91304
```

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?

```
# install.packages(’dplyr’)
library(’dplyr’)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:data.table’:
##
##     between, first, last

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

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

```
##           Income Group quantileGDP  N
## 1: Lower middle income (38.6,76.2] 13
## 2: Lower middle income   (114,152]  9
## 3: Lower middle income   (152,190] 16
## 4: Lower middle income  (76.2,114] 11
## 5: Lower middle income    (1,38.6]  5
```