Quiz 3

Xi Fang

4/17/2020

Q1

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?

```
# read the data into data.frame using read.csv
data <-read.csv("getdata_data_ss06hid.csv")
# set the criteria and assign it to a new variable
agricultureLogical <- data$ACR==3 & data$AGS ==6
# get the answer
head(which(agricultureLogical),3)
```

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

Q2

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?

```
# download and read the image
library(jpeg)
url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg"
z <- tempfile()
download.file(url,z,mode = "wb")
pic <- readJPEG(z, native = TRUE)
# get the quantile
quantile(pic, probs = c(0.3, 0.8))
```

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

Q3

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. 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

```
# read the data using fread into GDP2 and Country2
library("data.table")
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
GDP2 <- 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"))
Country2 <- data.table::fread('https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.cs
# merge the two dataset
mergedata2 = merge(GDP2, Country2, by = 'CountryCode')
# get the row number
nrow(mergedata2)
```

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

```
# arrange the merged data according to rank
mergedata2<-arrange(mergedata2, desc(Rank))
# get the 13th country name
mergedata2[13,]$Economy
```

```
## [1] "St. Kitts and Nevis"
```

Q4

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

```
tapply(mergedata2$Rank, mergedata2$`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
```

#Q5 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?

```
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':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
#cut the Rank colume into 5 quantiles
mergedata2$Rank=cut2(mergedata2$Rank, g=5)
#get dara summary according to rank and income group
table(mergedata2$Rank, mergedata2$`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
```