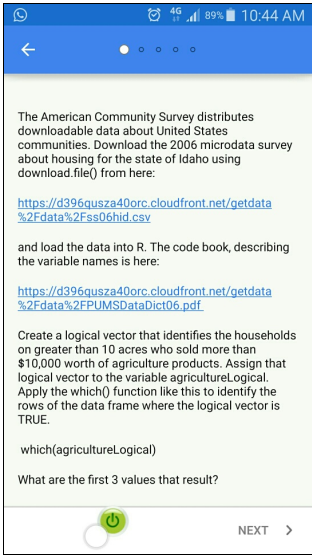
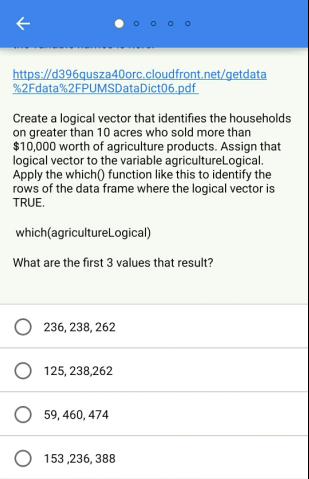
Question1





Answer : 125, 238, 262

**Solutions 1A:**

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?

# Question 1

# download file from server

#download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv", destfile = "q1.csv", method = "curl")

download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv", destfile = "q1.csv", mode = "wb")

# read csv file

q1 <- read.csv("q1.csv", header = TRUE)

# load library

library(plyr)

library(dplyr)

# create a logical vector

q1 <- mutate(q1, agricultureLogical=factor((ACR == 3 & AGS == 6), levels = c(TRUE, FALSE)))

# show the first 3 row names which the logical value are TRUE

head(row.names(q1[which(q1$agricultureLogical == TRUE),]), 3)

.names(q1[which(q1$agricultureLogical == TRUE),]), 3)

[1] "125" "238" "262"

**Solutions 1B:**

# write the file url and file destination to an object

file.url <- 'https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv'

file.dest <- 'ACS.csv'

# download from the URL

download.file(file.url, file.dest, mode = "wb")

# read the data

ACS <- read.csv('ACS.csv')

# create vector

ACS$agricultureLogical <- ifelse(ACS$ACR==3 & ACS$AGS==6,TRUE,FALSE)

# read lines

which(ACS$agricultureLogical)

> which(ACS$agricultureLogical)

[1] 125 238 262 470 555 568 608 643 787 808 824 849 952 955 1033 1265

[17] 1275 1315 1388 1607 1629 1651 1856 1919 2101 2194 2403 2443 2539 2580 2655 2680

[33] 2740 2838 2965 3131 3133 3163 3291 3370 3402 3585 3652 3852 3862 3912 4023 4045

[49] 4107 4113 4117 4185 4198 4310 4343 4354 4448 4453 4461 4718 4817 4835 4910 5140

[65] 5199 5236 5326 5417 5531 5574 5894 6033 6044 6089 6275 6376 6420

>

**Solutions 1C:**

#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?

library(httr)

direccion <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"

archivo <- "ss06hid.csv"

download.file(direccion, archivo, mode = "wb")

data <- read.csv("ss06hid.csv")

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

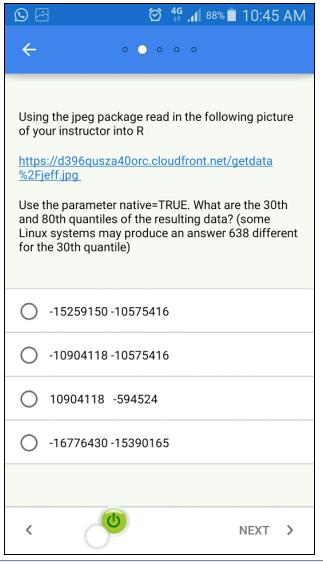
first3 <- which(logicalvector)[1:3] # which() treats NAs as FALSE

first3

> first3

[1] 125 238 262

Question2



Answer : -15259150 -10575416

**Solutions 2A:**

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)

# Question 2

# download jpeg file from server

#download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg", destfile = "q2.jpg", method = "curl")

download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg", destfile = "q2.jpg", mode = "wb")

# load library

library(jpeg)

# read jpeg file

q2 <- readJPEG(source = "q2.jpg", native = TRUE)

# show the 30th and 80th quantiles

quantile(q2, c(0.3, 0.8))

> quantile(q2, c(0.3, 0.8))

30% 80%

-15259150 -10575416

**Solutions 2B:**

# write the file url and file destination to an object

file.url <- 'https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg'

file.dest <- 'jeff.jpg'

# download from the URL

download.file(file.url, file.dest, mode='wb' )

# load package

library(jpeg)

# load the data

picture <- readJPEG('jeff.jpg', native=TRUE)

# get the quantile info

quantile(picture, probs = c(0.3, 0.8) )

> quantile(picture, probs = c(0.3, 0.8) )

30% 80%

-15259150 -10575416

**Solutions 2C:**

#QUESTION2

#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)

direccion2 <- "https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg"

archivo2 <- "jeff.jpg"

download.file(direccion2, archivo2, mode='wb' )

foto <- readJPEG("jeff.jpg", native = TRUE)

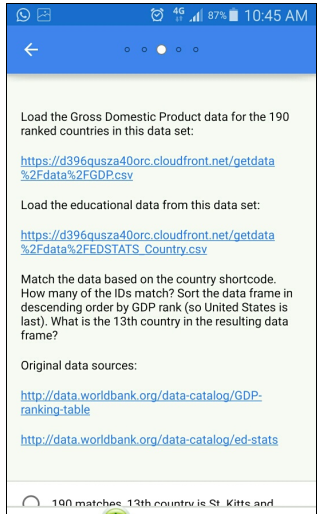
quantile(foto, c(0.3, 0.8))

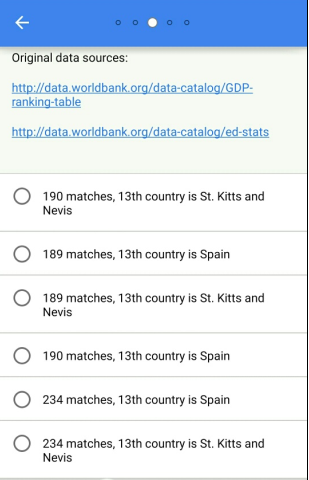
> quantile(foto, c(0.3, 0.8))

30% 80%

-15259150 -10575416

Question3





Answer : 189, St. Kitts and Nevis

**Solutions 3A:**

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

# Question 3

# download file from server

#download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv", destfile = "q3\_1.csv", method = "curl")

#download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv", destfile = "q3\_2.csv", method = "curl")

download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv", destfile = "q3\_1.csv", mode = "wb")

download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv", destfile = "q3\_2.csv", mode = "wb")

# load the datasets

q3\_1 <- read.csv("q3\_1.csv", header = TRUE, skip = 3, sep = ",")

q3\_2 <- read.csv("q3\_2.csv", header = TRUE)

# reshaping data

q3\_1 <- q3\_1[2:191, c(1,2,4,5)]

rownames(q3\_1) <- NULL

q3\_1 <- rename(q3\_1, CountryCode = X)

# merge two datasets

q3\_merge <- join(q3\_1, q3\_2)

# show the number of matches

sum(!is.na(unique(q3\_merge$Ranking)))

> sum(!is.na(unique(q3\_merge$Ranking)))

[1] 189

# convert the data type of Ranking

q3\_merge$Ranking <- as.numeric(as.character(q3\_merge$Ranking))

# show the 13th country after sort decending

q3\_merge <- arrange(q3\_merge, desc(Ranking))

q3\_merge[13,3]

> q3\_merge[13,3]

[1] St. Kitts and Nevis

229 Levels: East Asia & Pacific Euro area ... Zimbabwe

>

**Solutions 3B:**

# write the file url and file destination to an object

file.url <- 'https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv'

file.dest <- 'GDP.csv'

# download from the URL

download.file(file.url, file.dest )

# specify the right lines

rowNames <- seq(10,200, 2)

# read the data

gdp <- read.csv('GDP.csv', header=F, skip=5, nrows=190)

View(gdp)

# second data file

file.url <- 'https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv'

file.dest <- 'GDP2.csv'

# download from the URL

download.file(file.url, file.dest )

# read second file

fed <- read.csv('GDP2.csv')

View(fed)

# merge datasets

combined <- merge(gdp, fed, by.x='V1', by.y='CountryCode', sort=TRUE)

View(combined)

# Q3.

# sort the data

combined[with(combined, order(-V2) )]

> combined[with(combined, order(-V2) )]

Error in `[.data.frame`(combined, with(combined, order(-V2))) :

undefined columns selected

>

**Solutions 3C:**

QUESTION3

#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

library(data.table)

direccion3 <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"

archivo3 <- "GDP.csv"

download.file(direccion3, archivo3, mode="wb")

GDP <- data.table(read.csv("GDP.csv", skip = 4, nrows = 191))

GDP <- GDP[X != ""]

GDP <- GDP[, list(X, X.1, X.3, X.4)]

setnames(GDP, c("X", "X.1", "X.3", "X.4"), c("CountryCode", "rankingGDP", "Long.Name", "GDP"))

direccion4 <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv"

archivo4 <- "EDSTATS\_Country.csv"

download.file(direccion4, archivo4, mode="wb")

EDSTATS <- data.table(read.csv("EDSTATS\_Country.csv"))

data2 <- merge(GDP, EDSTATS, all = TRUE, by = c("CountryCode"))

sum(!is.na(unique(data2$rankingGDP)))

data2[order(rankingGDP, decreasing = TRUE), list(CountryCode, Long.Name.x, Long.Name.y, rankingGDP, GDP)][13]

> sum(!is.na(unique(data2$rankingGDP)))

[1] 189

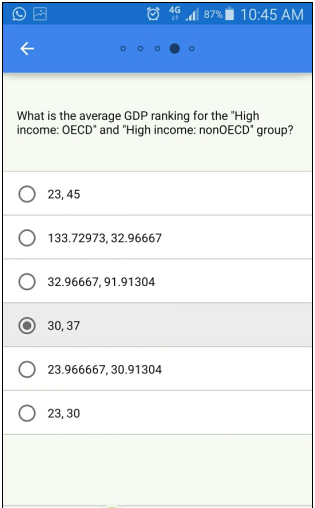
> data2[order(rankingGDP, decreasing = TRUE), list(CountryCode, Long.Name.x, Long.Name.y, rankingGDP, GDP)][13]

CountryCode Long.Name.x Long.Name.y rankingGDP GDP

1: KNA St. Kitts and Nevis St. Kitts and Nevis 178 767

>

Question4



Answer : High income: OECD 32.96667; High income: nonOECD 91.91304

**Solutions 4A:**

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

# Question 4

# Group q3\_merge by Income.Group

income\_group <- group\_by(q3\_merge, Income.Group)

summarise(income\_group, avg = mean(Ranking, na.rm = TRUE))

> summarise(income\_group, avg = mean(Ranking, na.rm = TRUE))

Source: local data frame [6 x 2]

Income.Group avg

(fctr) (dbl)

1 High income: nonOECD 91.91304

2 High income: OECD 32.96667

3 Low income 133.72973

4 Lower middle income 107.70370

5 Upper middle income 92.13333

6 NA 131.00000

**Solutions 4B:**

#QUESTION4

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

data2[, mean(rankingGDP, na.rm = TRUE), by = Income.Group]

> data2[, 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

>

**Solutions 4C:**

# Q4.

# OECD

mean(combined[combined$Income.Group=='High income: OECD',]$V2)

# non OECD

mean(combined[combined$Income.Group=='High income: nonOECD',]$V2)

> mean(combined[combined$Income.Group=='High income: OECD',]$V2)

[1] 32.96667

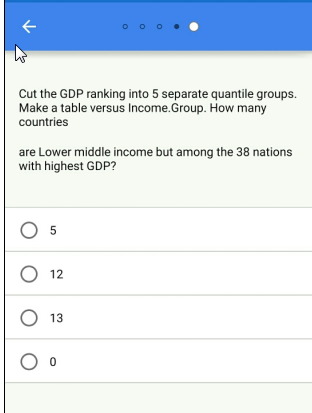
> # non OECD

> mean(combined[combined$Income.Group=='High income: nonOECD',]$V2)

[1] 91.91304

>

Question5



Answer : Lower middle income (1,38.8] 5

**Solutions 5A:**

ut 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?

# Question 5

# cut Ranking into 5 quantile groups

q3\_merge$RankingGroup <- cut(q3\_merge$Ranking, breaks = 5)

# make a table vs Income.Group

table(q3\_merge$RankingGroup, q3\_merge$Income.Group)

> table(q3\_merge$RankingGroup, q3\_merge$Income.Group)

High income: nonOECD High income: OECD Low income

(0.811,38.8] 0 4 18 0

(38.8,76.6] 0 5 10 1

(76.6,114] 0 8 1 9

(114,152] 0 4 1 16

(152,190] 0 2 0 11

Lower middle income Upper middle income

(0.811,38.8] 5 11

(38.8,76.6] 13 9

(76.6,114] 12 8

(114,152] 8 8

(152,190] 16 9

>

**Solutions 5B:**

#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(data2$rankingGDP, probs = seq(0, 1, 0.2), na.rm = TRUE)

data2$quantileGDP <- cut(data2$rankingGDP, breaks = breaks)

data2[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

**Solutions 5C:**

# Q5.

# assign quentile values

quentile <- c(0.2,0.4,0.6,0.8,1)

q <- quantile(combined$V2, quentile)

q1 <- combined$V2 <= 38

xtabs(q1 ~ combined$Income.Group)

> xtabs(q1 ~ combined$Income.Group)

combined$Income.Group

High income: nonOECD High income: OECD Low income

0 4 18 0

Lower middle income Upper middle income

5 11

>