Homework 4 Key

Jia-Ru Chung

October 26, 2017

**Note:**  
All answers have to be obtained using R code!

#### Max points = 120

### Question 1 {10 pts}.

Convert vector *v* to numeric. What is the 7th element in your new vector?

v <- factor(as.character(seq(1, 34, 3)))  
  
# R code  
v <- as.numeric(as.character(v))  
v[7]

## [1] 19

**Answer:**

### Question 2 {10 pts}.

Create your own *my.mean()* function that calculates arithmetic mean and ignores missing values, if any. Do not use the existing *mean()* function! What is the mean of vector *v*?

v <- seq(1, 13894, 34)  
v[seq(2, 391, 10)] <- NA  
  
# R code  
my.mean <- function(vector) {  
 average = sum(vector, na.rm = TRUE)/length(vector[!is.na(vector)])  
 return(average)  
}  
my.mean(v)

## [1] 6983.589

**Answer:**

### Question 3 {60 pts}.

Use the two **'Taxes'** datasets from Lecture 7 ('TC1400A1.txt' and 'TC1500A1.txt') and a new **'Population'** dataset ('PEP\_2015\_PEPANNRES.xls') to prepare data for analysis.

Note: you can convert excel file to csv prior to upload.

1. Import all three datasets {5 pts}.
2. Combine 2014 and 2015 **'Taxes'** data into one dataset so that each state has only two observations (one for year 2014 and another one for 2015). Instead of having *'GOVTAX\_TTL'* and *'AMOUNT'* column, create a separate column for each category of *'GOVTAX\_TTL'*. These columns should have *'AMOUNT'* as their values. The new column names should not have any space " ". {5 pts}.
3. In the **'Population'** data {20 pts}:
4. Keep only columns for state name, and years 2014 & 2015.
5. Remove any row that does not contain data.
6. Transform the Population data from 'wide' to 'long' format, so that each state has two observations (one for year 2014 and another one for 2015). (Hint: use melt() function). Your final **'Population'** data should have three columns: *GEO\_TTL*, *YEAR*, and *Population*.
7. Make sure the *population* and *YEAR* values are *'numeric'*.
8. Remove observations for "Puerto Rico" (since **Taxes** data doesn't have it).
9. Merge **'Taxes'** and **'Population'** datasets by *GEO\_TTL* and *YEAR*. Remove observations for the 'United States' {10 pts}.
10. Is it true the combined population in Alaska and Colorado was greater in 2014 than in 2015? {10 pts}

Data source: <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=STC_TC1500A1_FTP&prodType=document> <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=STC_TC1400A1_FTP&prodType=document> <http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=PEP_2015_PEPANNRES&src=pt>

# R code  
  
# Question 3 part 1- import data  
db2014 <- read.table("TC1400A1.dat", sep = "|", header = T)  
db2015 <- read.table("TC1500A1.dat", sep = "|", header = T)  
pop <- read.csv(file = "PEP\_2015\_PEPANNRES.csv", header = TRUE, sep = ",")  
  
# Question 3 part 2- restructure data  
require(reshape2)

## Loading required package: reshape2

d2014 = dcast(db2014, GEO\_TTL + YEAR ~ GOVTAX\_TTL, value.var = "AMOUNT")  
d2015 = dcast(db2015, GEO\_TTL + YEAR ~ GOVTAX\_TTL, value.var = "AMOUNT")  
Taxes <- rbind(d2014, d2015)  
colnames(Taxes) <- gsub(" ", "\_", colnames(Taxes)) # removing spaces  
  
# Question 3 part 3a  
pop <- pop[, -c(2:7)] # removing columns  
  
# Question 3 part 3b  
pop <- pop[rowSums(is.na(pop)) != ncol(pop), ] # removing rows not containing data  
  
# Question 3 part 3c  
colnames(pop)[2] = "2014"  
colnames(pop)[3] = "2015"  
pop\_long = melt(pop, id.vars = c("Geography"), measure.vars = c("2014", "2015"))

## Warning: attributes are not identical across measure variables; they will  
## be dropped

colnames(pop\_long)[1] = "GEO\_TTL"  
colnames(pop\_long)[2] = "YEAR"  
colnames(pop\_long)[3] = "Population"  
  
# Question 3 part 3d  
pop\_long[, 2] <- as.numeric(as.character(pop\_long[, 2]))  
pop\_long[, 3] <- as.numeric(gsub(",", "", pop\_long[, 3]))  
  
# Question 3 part 3e  
pop\_long = pop\_long[!(pop\_long[, 1]) == "Puerto Rico", ]  
  
# Question 3 part 4  
merged = merge(Taxes, pop\_long)  
merged = merged[!(merged[, 1]) == "United States", ]  
  
# Question 3 part 5  
AlaskaColorado2014 = sum(merged$Population[(merged$GEO\_TTL == "Alaska" | merged$GEO\_TTL ==   
 "Colorado") & merged$YEAR == 2014])  
AlaskaColorado2015 = sum(merged$Population[(merged$GEO\_TTL == "Alaska" | merged$GEO\_TTL ==   
 "Colorado") & merged$YEAR == 2015])  
isTRUE(AlaskaColorado2014 > AlaskaColorado2015)

## [1] FALSE

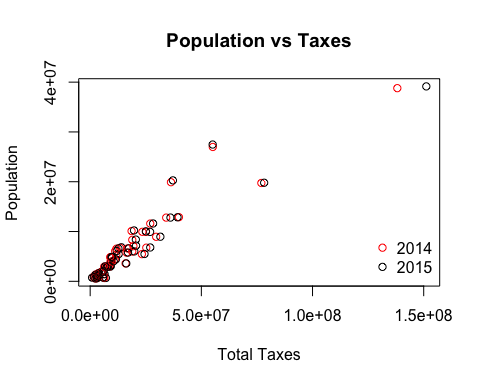
**Answer:**

### Question 4 {20 pts}.

Use final dataset created in Question 3.4.

1. Make a scatterplot of *'Total\_Taxes'* (x-axis) & *'Population'* (y-axis) colored by *'YEAR'* {10 pts}.
2. For year 2015, is there a correlation between *'Total\_Taxes'* & *'Population'*? Is it expected? Provide an estimate of the Pearson correlation coefficient and 95% confidence interval for the true correlation coefficient {10 pts}.

# R code  
  
# Question 4 part 1  
plot(merged$Total\_Taxes, merged$Population, xlab = "Total Taxes", ylab = "Population",   
 col = ifelse(merged$YEAR == 2014, "red", "black"), main = "Population vs Taxes")  
legend("bottomright", legend = c("2014", "2015"), col = c("red", "black"), pch = 1,   
 bty = "n")



# Question 4 part 2  
  
# The plot indicates high correlation between Total\_Taxes and Population;  
# this is expected because as the population increases, the taxes paid would  
# also increase  
  
# Estimated Correlation Coefficient  
Taxes2015 = merged$Total\_Taxes[merged$YEAR == "2015"]  
Population2015 = merged$Population[merged$YEAR == "2015"]  
correl = cor(Taxes2015, Population2015, method = "pearson")  
  
# 95% CI for True Correlation Coefficient  
correl

## [1] 0.9392414

cor.test(Taxes2015, Population2015, method = "pearson", conf.level = 0.95)$conf.int[1:2]

## [1] 0.8954306 0.9650356

**Answer:**

### Question 5 {20 pts}.

Use the final dataset created in Question 3.4 to create Geographic Maps of Population in 2014. Based on the map, which state had the highest population?

# credit https://gist.github.com/cdesante/4252133  
  
# install.packages('maps')  
require(maps)

## Loading required package: maps

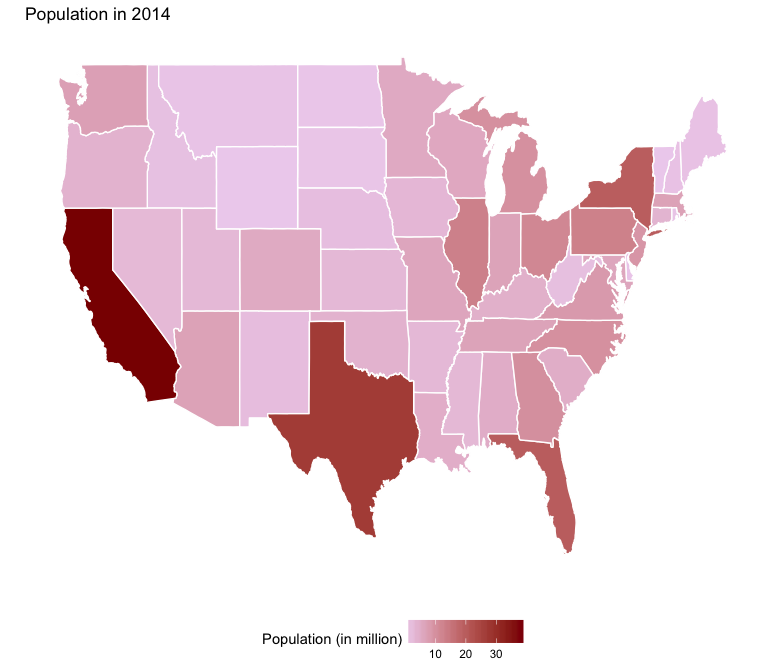
require(ggplot2)

## Loading required package: ggplot2

require(tools)

## Loading required package: tools

# getting US states information  
all\_states <- map\_data("state")  
all\_states$GEO\_TTL <- toTitleCase(all\_states$region) # converting state names to be in ProperCase   
  
# data for only 2014  
merged2014 = merged[merged$YEAR == "2014", c(1, ncol(merged))] # dataset to be used for creating the map  
  
# combining population and state data  
X <- merge(merged2014, all\_states, by = "GEO\_TTL")  
X$Population\_Scaled <- X$Population/1e+06  
  
# Making plot  
p <- ggplot() + geom\_polygon(data = X, aes(x = long, y = lat, group = group,   
 fill = Population\_Scaled), colour = "white") + scale\_fill\_continuous(low = "thistle2",   
 high = "darkred", guide = "colorbar")  
P1 <- p + theme\_bw() + labs(fill = "Population (in million)", title = "Population in 2014",   
 x = "", y = "")  
P1 + scale\_y\_continuous(breaks = c()) + scale\_x\_continuous(breaks = c()) + theme(legend.position = "bottom",   
 panel.border = element\_blank())



# Based on the map, California is the state with the highest population

**Answer:**