1 Data Analysis : Week 2 Quiz

Question 1

In the text of the final write-up of a data analysis, how should the analyses be reported?

(i) Every analysis performed should be reported with a measure of uncertainty. (Not all data analyses contain an element of "uncertainty"- censuses)

(ii) Analyses should be reported in the order that they appear in the raw scripts files.

(iii) Analyses should be reported in an order to convey the story being told with the data analysis.

(iv) Analyses should be reported in chronological order of when they are performed. (useful but not a major concern)

Question 2

- Open a connection to the old version of my blog: 'http://simplystatistics.tumblr.com/',
- read the first 150 lines of the file and assign them to a vector 'simplyStats'.
- Apply the '`nchar()`' function to 'simplyStats' to count the characters in each element of 'simplyStats'.
- How many characters long are the lines 2, 45, and 122?

```
# open a connection to http://simplystatistics.tumblr.com/
# assign to vector 'simplyStats'
simplyStats <- readLines(
 url('http://simplystatistics.tumblr.com/'), 150)
```

How many characters are in each element of 'simplyStats'.

```
# apply 'nchar()'
simplyStatsChars <- nchar(simplyStats)
```

How many characters long are the lines 2, 45, and 122?

```
# how many characters long is line 2?
nchar(simplyStats)[2]

# how many characters long is line 45?
nchar(simplyStats)[45]

# how many characters long is line 122?
nchar(simplyStats)[122]
```

Question 3

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://dl.dropbox.com/u/7710864/data/csv_hid/ss06hid.csv`

`or here:`

`https://spark-public.s3.amazonaws.com/dataanalysis/ss06hid.csv`

and load the data into `R`. You will use this data for the next several questions.

Code Book

The code book, describing the variable names is here:

`https://dl.dropbox.com/u/7710864/data/PUMSDataDict06.pdf`

`or here:`

`https://spark-public.s3.amazonaws.com/dataanalysis/PUMSDataDict06.pdf`

How many housing units in this survey were worth more than $1,000,000?

```
# Download 2006 microdata survey
# re: housing for Idaho using download.file()
# setwd("~/DA")
download.file(
 'https://spark-public.s3.amazonaws.com/dataanalysis/ss06hid.csv',
              "ss06hid.csv", method="curl")

# Download the code book:
# download.file(
 'https://spark-public.s3.amazonaws.com/dataanalysis/PUMSDataDict06.pdf',
              "PUMSDataDict06.pdf", method="curl")

# load the data into R
idahoData <- read.csv("ss06hid.csv", header=TRUE)

# are we sure it's just Idaho data?
table(idahoData$ST)
#Check the PDF - what does 16 mean?

#any missing data?
```

```
summary(idahoData$ST)

# How many housing units [are] worth more than $1,000,000?
table(idahoData$TYPE,idahoData$VAL)
```

```
#from local files
idahoData <- read.csv("daquiz2.csv", header=TRUE)
```

Question 4

- Use the data you loaded from Question 3.
- Consider the variable FES.
- Which of the "tidy data" principles does this variable violate?

Revision

What are the three characteristics of tidy data?

- "***Tidy data***" by Hadley Wickham (RStudio)
- Submission to Journal of Statistical Software
- (http://vita.had.co.nz/papers/tidy-data.pdf)

Three Principles from Hadley Wickham's paper

1. Each variable forms a column,
2. Each observation forms a row,
3. Each table/file stores data about one kind of observation.

```
# let's look!
unique(idahoData$FES)
```

Options

(i) Each tidy data table contains information about only one type of observation.
(Not so)

(ii) Each variable in a tidy data set has been transformed to be interpretable.
(No)

(iii) Tidy data has no missing values.

(iv) Tidy data has one variable per column.

Question 5

Use the data you loaded from Question 3.

- How many households have 3 bedrooms and and 4 total rooms?
- How many households have 2 bedrooms and 5 total rooms?
- How many households have 2 bedrooms and 7 total rooms?

```
#USING TABLE
#Rooms on Rows , Bedrooms on Columns
#dnn adds dimension names

table(idahoData$RMS,idahoData$BDS,dnn=list("RMS","BDS"))
```

Another Way of Doing it

```
# How many households have 3 bedrooms and 4 total rooms?
nrow(idahoData[!is.na(idahoData$BDS) & idahoData$BDS==3 &
                 !is.na(idahoData$BDS) & idahoData$RMS==4,])
# How many households have 2 bedrooms and 5 total rooms?
nrow(idahoData[!is.na(idahoData$BDS) & idahoData$BDS==2 &
                 !is.na(idahoData$BDS) & idahoData$RMS==5,])
# How many households have 2 bedrooms and 7 total rooms?
nrow(idahoData[!is.na(idahoData$BDS) & idahoData$BDS==2 &
                 !is.na(idahoData$BDS) & idahoData$RMS==7,])
```

Question 6

- Use the data from Question 3.
- 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'.

```
# Like this (this wont run yet)
 which(agricultureLogical)
```

What are the first 3 values that result?

```
# Showing off a bit
q6cols <- c("ACR", "AGS")
which(names(idahoData) %in% q6cols)

# logical vector
agricultureLogical <- idahoData$ACR==3 & idahoData$AGS==6

# and:
 which(agricultureLogical)
```

1.1 Question 7

- Use the data from Question 3.
- 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 and assign it to the variable indexes.

```
indexes =  which(agricultureLogical)
```

If your data frame for the complete data is called `dataFrame` you can create a data frame with only the above subset with the command:

```
subsetDataFrame  = dataFrame[indexes,]
```

Note that we are subsetting this way because the NA values in the variables will cause problems if you subset directly with the logical statement.
How many households in the subsetDataFrame have a missing value for the mortgage status (MRGX) variable?

```
indexes <- which(agricultureLogical)
subsetIdahoData <- idahoData[indexes,]

# And then:
nrow(subsetIdahoData[is.na(subsetIdahoData$MRGX),])
```

Question 8

- Use the data from Question 3.
- Apply '`strsplit()`' to split all the names of the data frame on the characters "wgtp".
- What is the value of the 123 element of the resulting list?

```
List <- strsplit(names(idahoData), "wgtp")
List[123]
```

Question 9

What are the 0% and 100% quantiles of the variable `YBL`? Is there anything wrong with these values? *Hint: you may need to use the* ***`na.rm`*** *parameter.*

```
quantile(idahoData$YBL, na.rm=TRUE)
#  0%  25%  50%  75% 100%
#  -1    3    5    7   25
```

Question 10

In addition to the data from Question 3, the American Community Survey also collects data about populations. Using 'download.file()', download the population record data from:

`https://dl.dropbox.com/u/7710864/data/csv_hid/ss06pid.csv`

`or here:`

`https://spark-public.s3.amazonaws.com/dataanalysis/ss06pid.csv`

- Load the data into `R`. Assign the housing data from Question 3 to a data frame '`housingData`' and the population data from above to a data frame 'populationData'.
- Use the merge command to merge these data sets based only on the common identifier "SERIALNO".
- What is the dimension of the resulting data set?

```
download.file(
'https://spark-public.s3.amazonaws.com/dataanalysis/ss06pid.csv',
              'ss06pid.csv', method='curl')

rm(idahoData)
housingData <- read.csv("ss06hid.csv", header=TRUE)
populationData <- read.csv("ss06pid.csv", header=TRUE)

dim(merge(housingData,
 populationData, by="SERIALNO", all=TRUE))
```