STAT 4410/8416 Homework 2

lastName firstName

Due on Sep 29, 2018

1. The data set `tips` contains tip for different party size as well as total bill personal information about bill payer. We can get the data from reshape2 packages as follows;

```
library(reshape2)
tips.dat <- tips
```

Now answer the following questions.

   a. Compute tip rate dividing tip by total bill and create a new column called `tip.rate` in the dataframe `tips.dat`. Demosntrate your result by showing the head of `tips.dat`.

   b. Draw a side by side violin plot of tip rate for each party size. Order the party size based on the median tip rate. Provide your codes and the plot. Which party size is responsible for highest median tip rate?

   c. Generate the similar plot you did in question 2b for each day (instead of party size) and facet by sex and smoker. Is the shape of violin plot similar for each faceted condition?

2. We generate a nxk matrix M and a vector V of length k for some specific values of n and k as follows;

```
set.seed(123)
n <- 7
k <- 8
V <- sample(seq(5), size=k, replace=TRUE)
M <- matrix(rnorm(n*k), ncol=k)
```

   a. Now, carefully review the following for loop. Rewrite the code that does the same job but doesn't use a for loop.

```
X <- M
for(i in seq(n)){
  X[i,] <- round(M[i,]/V, 2)
}
```

   b. Now do the same experiment for $n = 600$ and $k = 900$. Which code runs faster, your code or the for loop? Demonstrate that using function `system.time()`.

3. For the following questions please use data frame `tips`

   a. Create a bar chart that shows average tip by day.

   b. Compute the average tip, total tip and average size grouped by smoker and day. i.e., For each combination of smoker and day you should have a row of these summaries. Report the result in a nice table.

   c. Create a bar chart that shows average tip by day and also faceted by smoker.

   d. In questions 4a and 4c we plotted the summary of data which does not show us the whole picture. In practice we like to see the whole data. What plot do you suggest to serve the same purpose similar to what we did in question 4c? In other words, what would be a better plot to show tips by day and facetted by smoker? Please produce that plot and include your codes.

4. We want to generate a plot of US arrest data (USArrests). Please provide the detailed codes to answer the following questions.

a. Obtain USA state boundary coordinates data for USA map using function `map_data()` and store the data in `mdat`. Display first few data from `mdat` and notice that there is a column called `order` that contains the true order of coordinates.

b. You will find USA crime data in the data frame called `USArrests`. Standardize the crime rates and create a new column called `state` so that all the state names are lower case. Store the new data in an object called `arrest` and report first few rows of `arrest`.

c. Merge the two data sets `mdat` and `arrest` by state name. Merging will change the order of coordinates data. So, order the data back to the original order and store the merged-ordered data in `odat`. Report first few data from `odat`.

d. All the columns of `odat` is not necessary for our analysis. So, subset by selecting only columns long, lat, group, region, Murder, Assault, UrbanPop, Rape. Store the data in `sdat` and report first few rows.

e. Melt the data frame `sdat` with id variables long, lat, group, region. Store the molten data in `msdat` and report first few rows of data.

f. The molten data frame `msdat` is now ready to be plotted. Create a plot showing USA state map, fill with value and `facet_wrap` with variable. Please don't add any legend and make sure that faceting labels are identified so that we can compare the facetted plots.

g. Now examine the plot you have generated in question (f) and answer the following questions based on what you see in the plot.

   i. For each of the crimes, name two states with the highest crime rate.

   ii. Do you think larger urban population is an indicative of larger murder rate? Why or why not?

h. In question (3b) we standardized the crime rates. Why do you think we did this? Explain what would happen if we would not do this.

i. In question (3c) we ordered the data after merging. Why do you think we have to order? Explain what would happen if we would not order.

5. Life expectancy data for four countries are obtained from the world bank database which you will find on github. It contains life expectancy in years for different genders. Now answer the following questions.

a. Read the data from the above link and display first few rows of the data.

b. Generate a plot showing trend line of life expectancy over different year. Color them by sex and facet by country. Include your code and the plot.

c. Explain what interesting features you notice in the plot of question 5b.

6. We have a data set as below;

```
myDat <- read.csv("http://mamajumder.github.io/data-science/data/reshape-source.csv")
kable(myDat)
```

player	track	walking	cycling
1	A	408	43
1	B	402	31
1	C	386	41
2	A	373	53
2	B	404	41
2	C	422	30
3	A	403	25
3	B	393	46
3	C	422	48

We want to reshape the data and produce the following output.

player	variable	A	B	C
1	walking	408	359	359
1	cycling	23	45	39
2	walking	406	386	401
2	cycling	43	30	46
3	walking	418	401	392
3	cycling	42	45	43

Provide your codes that will produce the desired output. Demonstrate your answer by displaying the output as well.

7. **Ordering the factor** In class, we have seen how to order the factors. Suppose we have the following data about some rate during particular days of the week;

```
day <- c("Saturday", "Wednesday", "Friday", "Tuesday", "Thursday")
rate <- c(42,75,11,35,40)
df <- data.frame(day,rate)
```

Now please answer the following questions.

a. Convert the day column of dataframe `df` into a factor column. Demonstrate that it is indeed converted into a factor column.

b. Now generate a bar chart showing the rate of different days.

c. Notice the order of the levels of day is not natural, instead the plot shows the dictionary order. Now, order the bars according to the natural order of the levels of the class (days of the week as they appear in order) and regenrate the bar graph.

8. **Bonus (2 points)** for undergraduates and mandatory for graduate students. Suppose we have a vector of data as follows:

```
myVector <- c(0,3,6,9,12,15,18,21)
```

a. Using function `tapply()` compute the mean of first two values, next four values and rest of the two values. Show your codes and your result should be 1.5, 10.5 and 19.5.

b. Now compute the sum of squares instead of mean that you have done in question 8a. Show your codes and your result should be 9, 486 and 765.