NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp Fall 2015

Week 2 Days 1 + 2
Visualization

Due Date: Wednesday, September 30, 2015

Question 1: Dplyr Review

Download the Champion's League data, import it into R, and create a tbl_df object. The dataset records 100 Champion's League matches between different soccer clubs. Note that this dataset is generated from simulation (not the real match history).

```
CL = read.csv("Champions.csv")
library(dplyr)
CL = tbl_df(CL)
dim(CL)  # 100, 20
print(CL, n = 5)
```

The variables include the name of the club, number of goals, possession rate, number of yellow cards, etc., and each variable is recorded for home and away teams respectively.

1. Use filter to find out rows (games) that home team wins, i.e., HomeGoal > AwayGoal. These rows should be stored in a new tbl_df object. Also use filter to find out rows that the HomeTeam is either "Barcelona" or "Real Madrid".
2. Use select to create a new table which exactly includes all the variables about home team (and excludes variables about away team). Create another table which only includes 6 columns: HomeTeam, AwayTeam, HomeGoal, AwayGoal, HomeCorner, and AwayCorner. *Hint*: you may use the argument starts_with or contain in the function select.

3. (3) Use arrange to reorder the dataset by the number home goals, and display the following 6 columns of the reordered data: HomeTeam, AwayTeam, HomeGoal, AwayGoal, HomeCorner, and AwayCorner.
4. (4) For each HomeTeam, find out its average HomeGoal, average HomePossession (possession rate), and average HomeYellow (number of yellow cards). Summarise the results in a table.
5. (5) (Optional) Find out the top 5 frequent score (i.e., the combination of HomeGoal:AwayGoal). It is reported that 1:0 is the most frequent score in soccer games; does our dataset support this claim?

Question 2: Manipulating Data without Dplyr

Redo Question 2 using conventional method for data frame.

For example, the solution to part (1) can be:

```
CL[CL$HomeGoal > CL$AwayGoal, ]
```

Question 3: Scatterplot

The data frame cars in the datasets package records the speed (in mph) and stopping distance (in ft) for 50 cars.

(1) Use the plot function to create a scatterplot of dist (y-axis) vs. speed (x-axis).

(2) Refine the basic plot by labeling the x-axis with "Speed (mpg)" and the y-axis with "Stopping Distance (ft)". Also add a title to the plot.

(3) Revise the plot by changing the every point from the default open circles to red filled triangles (col="red", pch=17).

(4) Use ggplot2 to redo part (3).

Question 4: "Drawing Pictures" with R

The following function plots a house which is centered about the point (x, y). Note that we use the argument ... here.

```
house=function(x, y, ...) {
  lines( c(x - 1, x + 1, x + 1, x - 1, x - 1),
         c(y - 1, y - 1, y + 1, y + 1, y - 1), ... )
  lines( c(x - 1, x, x + 1), c(y + 1, y + 2, y + 1), ... )
  lines( c(x - 0.3, x + 0.3, x + 0.3, x - 0.3, x - 0.3),
         c(y - 1, y - 1, y + 0.4, y + 0.4, y - 1), ... )
}
```

1. Open a new plot window by the function plot.new. Using the plot.window function, specify that the horizontal and vertical coordinates both range from 0 to 10.
2. Draw three houses on the current plot window centered at the locations (1, 1), (4, 2), and (7, 6).
3. Please draw three additional houses on the current plot window.

   House 1: A "violet" house centered about (1,5), with line width as 2.

   House 2: A "slateblue" house centered about (5,6), with line type 2 as line width 3.

   Hint: By specifying the ... argument, one can pass parameters to sub-functions. Also, the file R color will be helpful.

4. Draw a boundary box about the current plot window using the box function.

Question 5: Density Curves

The Beta distribution is a distribution within the interval [0,1], which is usually applied to model the random behavior of a proportion. It is denoted as Beta(α, β), where α and β are shape parameters.

We can draw the density of Beta(5,2) by curve(dbeta(x, 5, 2), from=0, to=1).

1. Display the Beta(2, 6), Beta(4, 4), and Beta(6, 2) densities on a same plot. (*Hint*: specify the argument add=TRUE in the curve function.)

2. Use the following R command to title the plot with the equation of the beta density.title(expression(f(y)==frac(1,B(a,b))*y^{a-1}*(1-y)^{b-1}))
3. By the text function, label each density curve with its corresponding shape parameters a and b.
4. Redraw the graph using different colors/line types for the different curves.
5. Instead of using the text function, add a legend to the graph that shows the color or linetype for each of the beta density curves

Question 6: Boxplot and Density Curves

The dataset faithful contains the duration of the eruption (in minutes) eruptions and the waiting time until the next eruption waiting (in minutes) for the Old Faithful geyser.

1. In the faithful data frame, add a variable length that is "short" if the eruption is less than 3.2 minutes, and "long" otherwise.
2. Using the bwplot function in the lattice package, create parallel boxplots of the waiting times for the “short” and “long” eruptions.
3. Using the densityplot function, create overlapping density curves of the waiting times of the “short” and “long” eruptions.
4. Briefly describe your findings from the boxplots and the density curves,
5. Redo part (3) using the ggplot function in the ggplot2 package.
6. Redo part (4) using the ggplot function.

Question 7: NBA Data Visualization

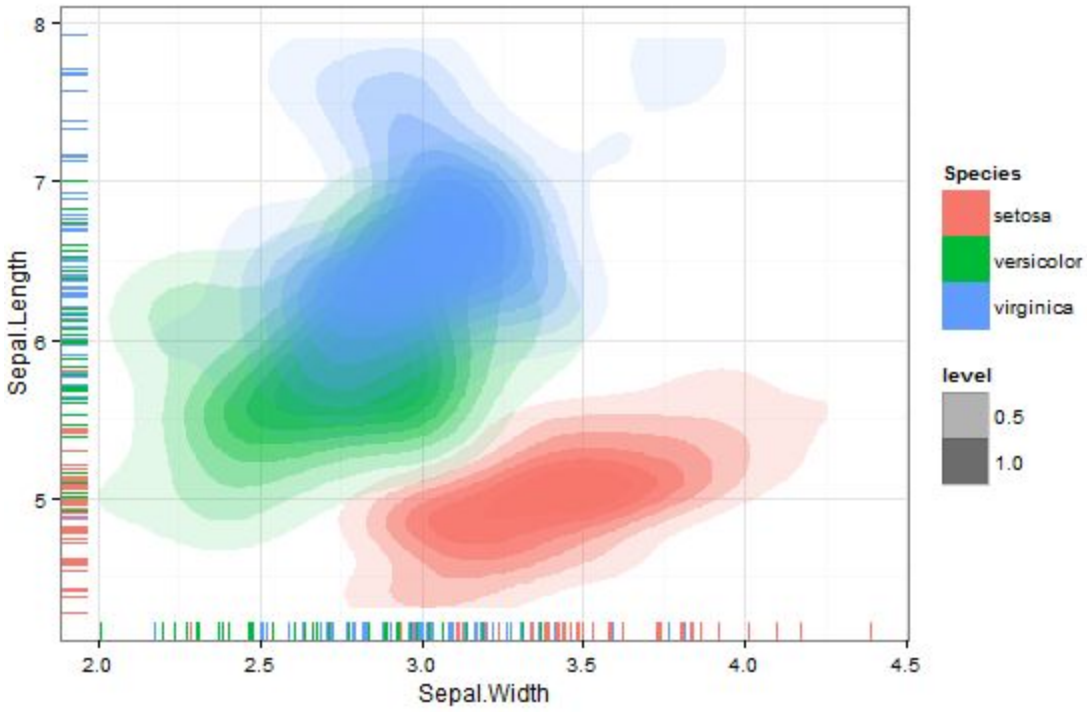
Load the New York Knicks NBA data.

1. Calculate the winning ratio of New York Knicks in different Seasons. **Visualize** how the winning ratio changes every year. (Barplot is the most appropriate here.)
2. Calculate the winning ratio both home and away. (The row labelled with visiting = 1 is an away game.) **Create bar-plots** to show home and away winning ratios for each season.
3. **Plot** five histograms to display the distribution of points in each season.

4. (Optional) Calculate the average winning ratio and the average point-difference (i.e.,points-opp) **by each opponent**. Create a scatter-plot to show winning ratio versus average point-difference. What pattern do you see in the graph?

Question 8: Density Plot

Recreate the following graph using the iris data.



You need to show the (joint) density of Sepal.Width and Sepal.Length in different Species. Also, please add the "rug".