NYC Data Science Bootcamp

Data Visualization with ggplot2

***** Save all your code to *yourname*.R and push it to the homework Github repository.

Note: you may need to pull from origin before you push it to Github.

Question 1: Dplyr Review

Load the Champion's League dataset, Champions.csv, from the homework folder. 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).

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 contains in the function select.

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. 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. (Optional) Find out the top 5 frequent score (i.e., the combination of HomeGoal:AwayGoal). Note that 1:0 should be treated the same as 0:1.

Question 2: Scatterplot

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

1. 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).

Question 3: 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. Label each density curve with its corresponding shape parameters a and b using text function.
4. 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 4: Boxplot and Density Curves

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

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. Create parallel boxplots of the waiting times for the “short” and “long” eruptions.
3. 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.

Question 5: NBA Data Visualization

Load the New York Knicks dataset, Knicks.rda, from the homework folder.

Note: Winning ratio=Win/Total

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?