Lab 2

Statistical Inference, Spring 97

In this lab, we focus on how data visualization is done in R. You will see a variety of charts and plots and will be asked to plot and discuss them. Pay attention to the legends, scales, labels, etc. and try to plot each chart in the requested format. There are lots of samples and examples on the net for each of the following exercises, so it is so suggested to explore the related topics yourself and increase your knowledge and skills of data visualization in R.

a) Simple Plots

Sequences can be created in R using "seq" function. Here is an example:

```
> x <- seq(-10, 10, length=41)
```

1- Plot the standard Normal distribution within the range [-3, 3] using the "plot" function. Note that the chart should be continuous and well shaped. Also, the axes should be labeled with sensible values, avoid the default labeling. (Do not use the "curve" function)

2- Try to plot standard Normal distribution using samples pulled from this distribution (hint: apply the "rnorm" function). Plot the distribution with sample sizes 5, 10, 100, and 1000 separately. Which one is the closest to a perfect standard Normal distribution? Interpret the results.

3- Plot Student's t distribution with degrees of freedom 1, 2, 5 and 10; all in one plot. Use different colors for each distribution in the plot. Create a legend on the top left corner of the canvas. Also, increase the line width of the curve for df = 1.

b) QQ-Plots, Histograms, Box-Plots and Pie-Charts

"precip" dataset contains annual precipitation of U.S. cities. It is a named numerical vector.

4- Plot the quantiles of "precip" against the standard normal distribution. Discuss on the distribution of "precip" and explain the result.

5- Draw four histograms all in one plot (2x2): a histogram with 7 bins, a histogram with 13 bins, a histogram with 30 bins, and a histogram of densities with 13 bins. Discuss each of the histograms and explain which one visualizes the data more appropriately.

6- Fit a smooth curve over the last histogram of exercise 5. Tune the colors, labels, etc. to make it beautiful enough to get a bonus for this exercise.

"rivers" dataset contains lengths of major North American rivers.

7- By plotting the boxplot of "rivers", discuss the skewness of this distribution.

8- Extract values of whiskers in exercise 7. Is there any outliers? What are the exact values? (Don't estimate the values by looking at the plot!)

9- We want to categorize rivers into "tiny" (<500), "short" (<1500), "medium" (<3000) and "long" (>=3000). Plot a pie chart that visualizes frequency of these four categories. Your chart should be colorized and the labels should contain each category with its percentage.

"mpg" dataset contains fuel economy data from 1999 to 2008 for 38 popular models of cars.

10- Explain the result of executing the following command. What is the meaning of "~"? How we model formulas in R? How do you justify the plot shown as the result of the following command?

```
> boxplot(mpg$displ ~ mpg$class)
```

c) ggplot2

"ggplot2" is a plotting system for R based on the "Grammar of Graphics". Grammar of Graphics is a general scheme for data visualization, which breaks up graphs into semantic components such as scales and layers.

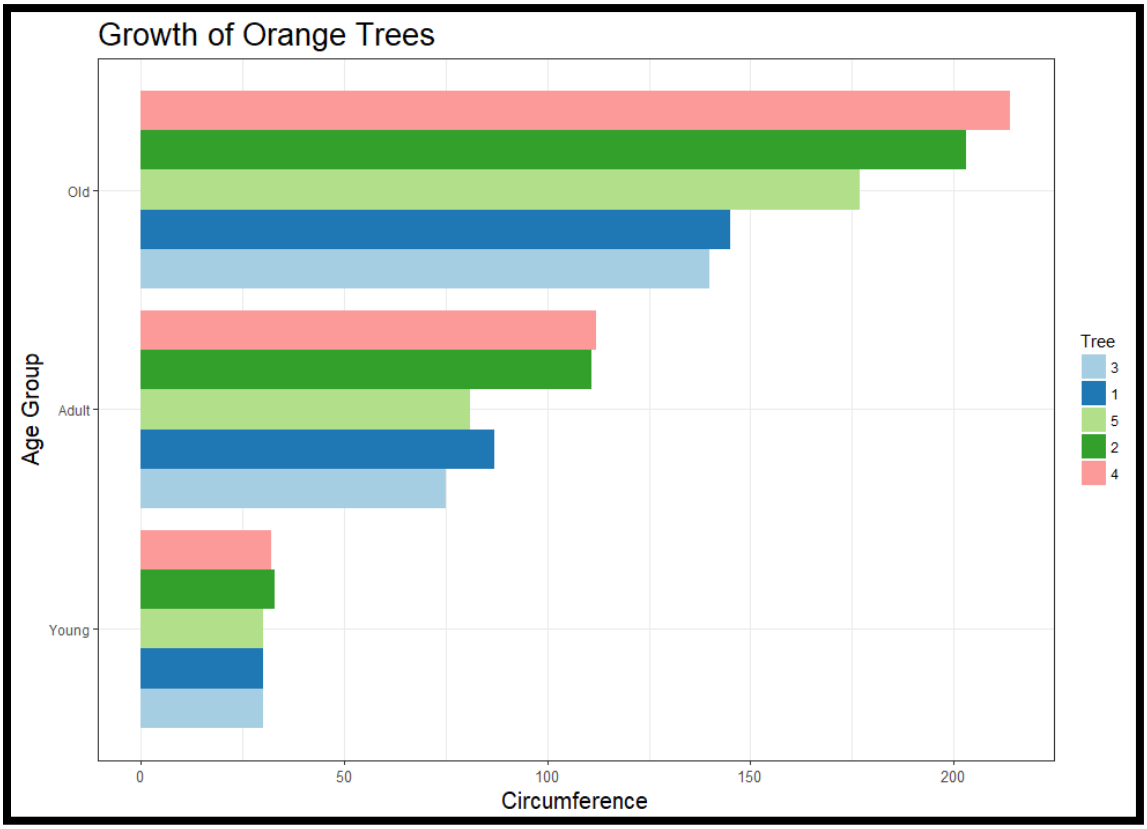
Consider the "Orange" dataset:

11- Create a simple scatterplot with ggplot2 that has the "age" variable on the X-axis and the "circumference" on the Y-axis.

12- Change the scatterplot of question 11 such that each point has a different according to the value of its "Tree" variable. Try to change the default ggplot2 coloring behavior, use more sharp colors like red, blue, orange, green, etc.

13- Plot a smooth graph (with its shadowed margin) which contains "age" and "circumference" variables on X and Y-axis. Discuss the "method" of smoothing which ggplot2 uses.

14- Create a new variable named "ageGroup". ageGroup takes the value "Young" for trees of the age 250 or less, "Adult" for trees between 250 than 900 years old and "Old" for trees older than 900. Plot a bar chart **exactly** like the one in the following figure. Tune the colors, labels, axes, etc., there would be about a dozen layers and controls needed to plot the chart.



d) ggmap

"ggmap" is a collection of functions to visualize spatial data and models on top of static maps from various online sources (e.g. Google Maps and Stamen Maps). It includes tools common to those tasks, including functions for geolocation and routing. "ggmap" is based on "ggplot2" principles and conventions, so it can be used easily in the concept of graphics layers.

"state" dataset contains some information about the states of the U.S. We can load this dataset into the R session environment with the following command:

```
> data(state)
```

"state.center" variable is a list that contains latitudes and longitudes of the centers of the states.

15- Convert "state.center" to a data frame named "centers_df".

16- Get the map of the "United States" with ggmap and plot the states' centers on the map. The result should be something like this: