HW2: Data Visualization

Due: 12 February 2017, 11:59PM UTC-12:00 on T-Square

In this assignment, you will practice visualization techniques, and explore the relationships between different attributes of two separate datasets, *midwest* and *diamonds*.

Instructions

- You must complete this assignment on your own, not in a group.
- All your code must be written in the R programming language.
- You need to turn in 2 files for this assignment:
  - Code: R script file named **hw2.R** with code for the different problems.
  - Report: PDF file named **hw2_report.pdf** with responses to written portions.
- Each problem includes a Submit section that clearly mentions what you need to put in the Code and/or Report file.
- Make sure you include your GT account name at the beginning of both files (the one you use to log into T-Square or Passport).
- Total points in this assignment: 100.

Problems

1. Professional Employment by State [20 points]

Using the *midwest* dataset from *ggplot2*, describe the relationship between the states in the Midwest region and the percentage of people that have a professional employment (represented by the column 'percprof'). Describe which state has the lowest and which state has the highest percentage of total population with a professional employment, and display a graph supporting your conclusion.

Note: To use the midwest dataset, you must load the ggplot2 package first.

```
library(ggplot2)
data(midwest)
```

Submit

- Report: Your description of the relationship between states and professional employment, states with lowest and highest percentage, and a supporting graph.
- Code: R code you used to perform any computations and plot the graph, with brief comments clarifying different steps/components.

2. School and College Education by State [20 points]

Using the *midwest* dataset, explore the three-way relationship between the percentage of people with a High School diploma (represented by the column 'perchsd'), the percentage of college educated population ('percollege'), and the State. What are your observations? Display a graph supporting these observations.

Submit

- Report: Your observation regarding the relationship between the percentage of college education population and percentage of people with a High School Diploma, and how they vary by state, and a graph illustrating this relationship.
- Code: R code you used to compute values and plot the graph, with brief comments.

3. Comparison of Visualization Techniques [20 points]

Define the different elements of a Box Plot and how they depend on a sample of N numbers. What are the pros and cons of using a box plot or a histogram? When would you use a Box Plot, a Histogram, or a QQPlot to graphically summarize N points?

Submit

- Report: Writeup defining the different elements of a Box Plot and how they depend on the size of the dataset N (with an optional diagram). A comparison of pros and cons of a Box Plot and a Histogram. An explanation of when (i.e. with what kinds of data) would each of the following be most useful: Histogram, Box Plot, QQPlot.

4. Random Scatterplots [20 points]

Generate two sets of N random uniformly-distributed points using the function runif(), and display a corresponding scatterplot using one set as X-values and the other as Y-values. If you save the plot to disk, what is the resulting file size for the following file formats: *ps, pdf, jpeg, png*? How do these values scale with increasing N? Display your results by plotting file size by format over a suitable range of N.

Submit

- Report: One or two sample scatterplots, your observation regarding the relationship between File Size and File Format for the given file types , and a (single) plot illustrating these results.
- Code: R code you used to generate and save the scatterplots for different N, and to plot the graph showing the relationship between file format and file size over a range of N, with brief comments.

5. Diamonds [20 points]

The *diamonds* dataset within ggplot2 contains 10 columns (price, carat, cut, color, etc.) for 53940 different diamonds. Type help(diamonds) for more information. Plot histograms for color, carat, and price, and comment on their shapes. Investigate the three-way relationship between price, carat, and color. What are your conclusions?

Provide graphs that support your conclusions. Choosing suitable kinds of graphs for visualization is as much a part of this assignment as creating them. Please refrain from discussing what particular graph types you are using for specific problems (general discussion and best practices are fine).

If you encounter computational difficulties, consider using a smaller dataframe whose rows are sampled from the original *diamonds* dataframe. Use the function *sample* to create a subset of indices that may be used to create the smaller dataframe.

Note: To use the diamonds dataset, you must load the ggplot2 package first.

```
library(ggplot2)
data(diamonds)
```

Submit

- Report: Your observation on the distribution of values for color, carat, and price, and a histogram graph for each. A brief writeup explaining the three-way relationship between price, carat, and color, and included graph(s) to illustrate the relationship.
- Code: R code you used to plot the graphs.