HW2: Data Visualization

Due: 19 February 2017, 11:59PM UTC-12:00 on T-Square

***Updated Feb 11**: Due date postponed to Feb 19; all significant changes highlighted.*

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.
- Choosing suitable 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).
- You need to turn in 2 files for this assignment:
  - Code: R script file named *exactly* **`hw2.R`** with code for the different problems.
  - Report: PDF file named *exactly* **`hw2_report.pdf`** with written responses.
- 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 top of both files (the one you use to log into T-Square or Passport), not in the filename.
- Total points in this assignment: 100.

Problems

1. Professional Education 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 education (represented by the column `'percprof'`), with a suitable graph that illustrates this relationship.
- Identify which state has the lowest and which has the highest percentage of adult population with a professional education. You may include another graph here if the previous one is not sufficient.

You can interpret this problem in one of two ways:

Interpretation A
Aggregate percprof for each state by combining the values from each county in that state, and generate an overall "percprof by state" plot. Note that since percprof is based on popadults, when aggregating it, you will have to use a formula like (s = state, c = county):

$$\forall s \ \ percprof_s = \frac{\Sigma_{c\in s} percprof_c \times popadults_c}{\Sigma_{c\in s} popadults_c}$$

Now describe which state has the lowest and which state has the highest percentage of adult population with a professional education, and display a graph supporting your conclusion.

Interpretation B
Treat the raw percprof value for each county as a metric. Plot percprof for each county grouped by state, using a suitable type of plot that can help show the properties of the distribution of percprof values found within that state.

Describe the distributions by pointing out relevant and interesting statistics that your plot illustrates, such as the mean or median value for some states, the spread of values for some states, states that have outlying counties, and so on.

Now can you point out which state has the lowest and highest percprof distribution, using some summary statistic? If not, explain why.

Note: Please mention which interpretation you are using, and any further assumptions you've made in answering this question. You may even use one interpretation for the first part of the problem and another for the second - in that case, specify the interpretation used for each part and include appropriate plots.

Also 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 education, states with lowest and highest percentage (or explanation why you cannot ascertain them), and a supporting graph. Mention which interpretation has been used: A or B.
- 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. Again, you may interpret this question in one of two ways, similar to Question 1 above:

Interpretation A
Aggregate perchsd and percollege for each state just like in Q1. Now illustrate the relationships between perchsd vs. state, percollege vs. state, as well as perchsd vs. percollege within each state, using suitable plots.

Are there any conclusions you can draw from these visualizations? If not, why?

Interpretation B
Treat the raw perchsd and percollege values per county as metrics. Now illustrate the distributions of perchsd grouped by state, percollege grouped by state, as well as perchsd vs. percollege values, using suitable plots (or a combined pair-wise plot, such as ggpairs).

Are there any conclusions you can draw about the distributions and relationships between variables? If not, why?

Submit

- Report: Your observation regarding the relationship between the percentage of people with a High School Diploma, percentage of College-educated population, and state, and a graph illustrating this relationship. Mention interpretation A or B.
- Code: R code you used to compute values and plot the graph, with brief comments.

3. Comparison of Visualization Techniques [20 points]

Describe the different elements of a Box Plot and how they depend on the statistical properties of a sample of 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 a sample of numbers?

Submit

- Report: Write-up describing the different elements of a Box Plot and how they depend on the sample of numbers (with an optional diagram). A comparison of pros and cons of a Box Plot and 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 values using the function `runif()` (read “random-uniform” not “run-if”), 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 vs. N, over a suitable range of N, for each format.

Note: To save a plot to disk, you may need to open a graphics device *before* plotting (for base graphics) or use ggsave() (for ggplots). If you have trouble saving the plots in any of these formats, you may pick some alternatives that are available on your system (such as *tiff*), but ensure that you plot file size vs. N for at least 3 different formats.

Submit

- Report: One or two sample scatterplots using randomly generated values, a suitable plot illustrating the relationship between file size and N for each of the given file formats, your observations on these relationships (asymptotic complexity is not compulsory, however, try to be as precise in your description of the relationships as evident from your plot).
- Code: R code you used to generate and save the scatterplots for different N, and to plot the graph showing the relationship between file size and N for each file format, 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 or bar charts for color, carat, and price, illustrating their distributions, and comment on the shapes of the distributions.

Investigate the three-way relationship between price, carat, and color. What are your conclusions? Provide a combined pair-wise plot and/or separate graphs that illustrate these relationships.

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