CSC 465

Homework 1

Submit a PDF file with your answers to the D2L submission box by the submission box deadline on D2L (but you're welcome to add a couple hours of grace period).

Clearly label which answer goes with which question. If it is not easy to find your answers, you may lose credit.

Include text answering questions and images of your visualizations (from screenshots or copying and pasting right from Tableau or RStudio into your document). For each question, explain very briefly how you created the visualizations and include R code files in your submission.

The idea behind this assignment is to get you using the tools we'll work with for this course and demonstrate your understanding of the first two weeks' material. You will make graphs with both R and Tableau. Follow the guidelines below for making quality graphs. Make each visualization and revise, making conscious decisions about the appearance, rather than using the default settings. It requires some fiddling with settings/code to get graphs the way you want them.

We'll learn more about guidelines for clear graphs, but here is a starting point:

- Each graph should be clean with easy-to-read graphical elements (not too thick, but not too thin either, not too much overlap of plot elements).
- The font size and weight should make labels easy to read, while not being intrusive.
- The defaults may be fine, but you are highly encouraged to experiment with different formatting options to try to improve the readability of the graphs. Plus, doing so helps you learn the software better!

1) (20 pts) For this problem, we'll look at data about Intel stock (Intel-1998 dataset from the website). The data covers stock market trading for the Intel corporation in 1998. Each row is a day, with the following columns: *Date*, *Trading Day* (integer day number, including skips), *Open* (price at market open), *High* (highest price of day), *Low* (lowest price of day), *Close* (price at market close), *Volume* (shares traded), and *Adj. Close* (adjusted closing price, meaning accounting for stock splits, which are not a problem in this data).

   Make the specified graphs in either R or Tableau:

a. Graph the closing price vs. the date with an ordinary line graph. If you use Tableau, you need to right-click on the *Date* and choose *Exact Date* from the dropdown menu so that it uses the full date with "day".
b. Graph the *Volume* vs. the exact *Date* as in the last part with a bar graph.
c. Create a scatterplot that graphs the *Volume* on the x-axis and the daily price range on the y-axis. You will need to create an additional column that contains the "range" of the prices for the day as the difference between the fields *High* and *Low*.

```
Range = High - Low
```

Tableau can do it with a *Calculated Field*. In R you can do it by making a new column equal to the result from subtracting the two columns. In Tableau, to get a scatter plot, you will need to right click on both the *Range* and *Volume* entries in graph and change them to "Dimensions".

2) (20 pts) Use Tableau for this question. Open the GM cars dataset included with this assignment (gmcar_price.txt). Each row represents a different car that was sold and includes information about features like the mileage and the price of sale. Hint: use the “Show Me” menu.

a. A treemap based on *Price* with a main subdivision for the *Make* of the car and a minor subdivision based on the *Model*. Because each row of the data file represents a single car but each box in the treemap represents all the cars with a given make and model, pay very close attention to what kind of aggregation is being used.
b. A packed bubble chart of the same type.
c. Write a short paragraph discussing the **differences** between the two plots. Describe for each something that displayed more clearly than with the other.
d. Create a contingency plot (Tableau calls it a *heat map* under *Show Me*) showing with color the number of cars (*Number of Records*) of each *Type* sold by each *Make*. Explain at least one observation about that data that this chart makes it easy to see.

3) (20 pts) This problem works with a dataset containing the population of Montana and of each of the 7 Native American reservations within it (reservation70-00.xlsx). There is a measurement for each decade between 1970 and 2000. Sheet1 has the original data.

We will use Tableau for this question, but Sheet1 has a header that confuses Tableau. If you’re interested, check out the “Data Interpreter” feature in Tableau to learn how to deal with this. Otherwise, use Sheet2, where I’ve removed the header. We need a few transformations to get the data ready to work with:

1. Renaming the 1970* field so it has no * and can be converted to a number

2. “Pivoting” the year fields in a similar manner to how it was demonstrated in the tutorial.
3. Changing the name of the pivot fields to *Year* and *Population*, and changing the type of the year field to “whole number”.
4. You can also hide the “Percent Change” field as it only contains information for change over the entire period, not per decade.
5. If you would like to have an actual Date field for the *Year*, so that it is treated by Tableau as a time instead of just a number, you need to create a “Calculated Field”. It should construct a Date using the *Year*, i.e. make a Date field that is on January 1 of the specified year:

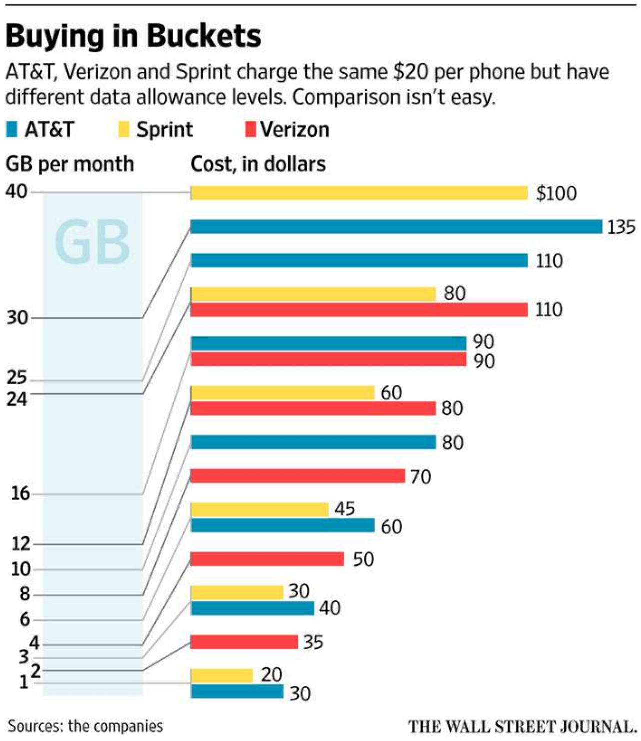
   ```
               makedate([Year], 1, 1)
   ```
6. We are not interested in the Montana population, only the reservation populations. When you have used Location on your graph, you can right mouse click (or click the down arrow within it) to apply filters. You can also use “Exclude” from the right click menu on the legend just below the “Marks” configuration.

Create graphs to show the following information, using appropriate graph types. Make sure that the graphs are properly labeled and that the axis scales properly reflect the type of data represented.

a. One chart that graphs the population growth over the years for the individual reservations.
b. One that graphs the total **reservation** population for each year, subdivided among the different reservations. The difference between this and (a) is that in (b) we are not looking only at each population individually but at the growth of the total population of all of them together, then subdivided by the reservations.

4) For this question, answer only with text. You may include an illustration if you would like, but you do not need to visualize data for this question.
   a. Explain what we mean by ‘pre-attentive’ attributes. Are these as effectively recognized by human perception when they are used in combinations?
   b. Use Weber’s Law to explain why it is important to include 0 in the numerical axis of a bar chart.

5) This graph of cell phone pricing plans is not very easy to use. Use R for this question and recreate this graph in two different ways of your choice. For each one, explain what you are trying to help the user see. For example, one might be to compare the cell phone companies to see what kind of plans they have. Another might be best for examining the trend of the relationship between price and data bandwidth. That relationship may hold overall, or you could look to see if it is different per company. You can decide what to visualize, i.e. what question to answer with your visualization, but make sure to explain what this visualization should be showing. To get full credit, you must produce a graph which makes the answer to your question immediately clear. It must also be well implemented, i.e. following the guidelines at the top for a clean graph.



Buying in Buckets

AT&T, Verizon and Sprint charge the same $20 per phone but have different data allowance levels. Comparison isn't easy.

Sources: the companies

THE WALL STREET JOURNAL.

You do not need to type in all the values by hand. Here is R code that makes a dataframe with these values in it:

```
cellPlans = data.frame(
    c("ATT", "Sprint", "Verizon", "ATT", "Sprint",
      "Verizon", "ATT", "Sprint", "Verizon", "ATT",
      "Verizon", "Sprint", "Verizon", "ATT",
      "Verizon", "Sprint", "ATT", "ATT", "Sprint"),
    c(1, 1, 2, 3, 3, 4, 6, 6, 8, 10, 12, 12, 16, 16,
      24, 24, 25, 30, 40),
    c(30, 20, 35, 40, 30, 50, 60, 45, 70, 80, 80, 60,
      90, 90, 110, 80, 110, 135, 100))

names(cellPlans) = c("Company", "DataGB", "Price")
```