CSC 465

Homework 1

Submit a PDF file with your answers to the D2L dropbox by January 27.

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 Dropbox submission.

The idea behind this assignment is to get you using the tools we'll work with for this course. You will make graphs with both R and Tableau. It requires some fiddling with settings to get graphs the way you want them. Follow the criteria we've discussed in class for uncluttered graphs that clearly display the data to communicate some information. Recall we discussed clarity, lack of clutter, emphasizing the data and graphical integrity. Make each visualization and revise, making conscious decisions about the choices you make in the settings, rather than using the default settings.

You'll learn more about modifying graphs, and you'll usually get better graphs, if you think directly about how you want your graph to look. In particular:

- 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).
- Axis scales must adhere to the guidelines in the lectures (for example Lecture 2's material on tick marks and grid lines).
- You should have both horizontal and vertical grid-lines, but they should be a medium gray on white, or white on a medium gray background, and appropriate thickness, to keep them from competing with the data itself.
- 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. It 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 zip file posted with this homework). 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 histogram of the daily stock *Volume*. R has the *hist* command and a *ggplot* geom. In Tableau, the *Histogram* graph type in the *Show Me* box will be useful. Experiment with the bin size. It's an optional parameter in the R functions (e.g. n=20 for *hist* or bins=20 for *geom_histogram*). In Tableau, after you have the histogram, right click the "Volume (bin)" in the data bar on the far left and use *Edit.* In Tableau, it's not the number of bins, but their width (in terms of data). You can set them that way in R as well with different parameters.
d. 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) We will analyze the perception data collected in class to see how accurate students were at perceiving values with different encodings (aligned bar vs. unaligned bars vs. volume, etc.). Use the PerceptionExperiment.csv data file, which has data from 92 students in previous years' classes. Remember that you saw a sequence of slides each with four encoded values, marked A, B, C and D. You were supposed to write down the values for B, C and D as a proportion of A.

Here is what the column names in the data file mean: for each *Test*, i.e. for each type of visual encoding from angle to volume, there were two slides called *Display*s. Each individual slide, i.e. each *Display* of each *Test*, has a unique *TestNumber*. Each sample that you estimated a value for was labelled B, C or D as its *Trial*. The *Subject*s are the students and the estimates they made are the *Response*s. Each row has a copy of the *TrueValue*, i.e. the correct value that the student should have entered (if the whole point weren't how hard it is).

The *Responses* themselves are not very useful for initial visualizations because they will naturally cluster around each *True Value*. The first thing you will need to do is to

create a new column that contains the amount of error. Using the same procedure as in Question 1D, define:

```
Error = Response – TrueValue
```

Using either Tableau or R, create the following graphs:

a. A histogram of the overall distribution of *Error*.
b. A bar graph of the median *Error* by *Test* (aka *Error* vs. *Test*). Do not subdivide by *Display* or the *Trial*. Order the x-axis to make the graph as clear as possible. Remember, for bar graphs in general, do not necessarily keep the default order (e.g. alphabetical) of the x-axis.
c. A bar graph of the standard deviation of the *Error* by *Test*. Remember that this measures the spread of how widely subjects varied in their responses. Again, order the x-axis to make the graph clear.
d. Create a new field called *AbsoluteError* by computing the absolute value of the *Error* field you created. Then do the same as in (b) with the *AbsoluteError*.
e. For each of the above graphs, explain with a few sentences what the graph tells you.

3) (20 pts) Use R for this problem. We will look at data on infant sizes at birth (InfantData.xlsx). There are libraries to help you import the Excel file directly, but in my experience, they are finnicky. The easiest thing to do is open the file with Excel or other compatible software and save it as a CSV file.

Create the following graphs:

a. Graph the data as a scatter plot of *Height.in* on the x-axis and *Weight.lbs* on the y-axis. Differentiate in the plot between M or F values for *Sex*, but graph both on the same plot.
b. Then create another single graph that has separate trend lines for the two populations on the graph. Adjust both the line and data-point weight and color to make the scatter plot and trend lines stand out.
c. Explain in a short paragraph the decisions you made here and their impact on the visualization.

See the R examples from the first lab for reference.

4) (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. Create the following plots (we will look more closely at their meanings and design criteria later, but do the best you can to make them readable). 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.

Write a short paragraph discussing the **differences** between the two plots. Describe for each something that displayed more clearly than with the other.

5) (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.

Even with the header and footer removed, you will have to transform the data by

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 the example in the first lab.
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*, 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:

a. One chart that graphs the population growth over the years for the individual reservations.

b. One that graphs the total **reservation** population subdivided among the different reservations for each year. 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.

c. One that graphs the population distribution over years for each reservation with a box-and-whisker plot. The 'distribution over years' means we are

visualizing a distribution, i.e. multiple samples of something. In this case that is multiple samples of population value, one per year. For each reservation, we have four different year samples of population, so we will have a box-and-whiskers ‘column’ per reservation showing the distribution of population values at that location during this overall period.

Make sure that the graphs are properly labeled and that the axis scales properly reflect the type of data represented.