CSC 465

Homework 2

Submit a PDF file with your answers to D2L. Clearly label which answer and visualization 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). Explain very briefly how you created the visualization and include R code files in your Dropbox submission.

1) **(30 pts)** We'll continue analyzing the data from the perception test that we started in Homework 1. In this problem, we will dig deeper into the distributions of each perception test and look for patterns that reveal any strengths or weaknesses. First, I recommend re-reading the description of the data from HW1 as this data has some subtleties to it. For the problem, explore the data for the following features and display them as clearly as possible using any techniques that we have covered for displaying and comparing univariate distributions. You may do this either in R or Tableau, but be aware that R will give you more options for your visualization. In either case, be thorough in looking at what methods are appropriate. Focus on the clarity of the display, keeping in mind the criteria from the lectures on clarity and accuracy. You will re-use your calculated fields for error and absolute error from HW1.

   a. Use a univariate scatterplot or another technique that shows fine detail for a collection of distributions. For each *Test* (don't divide between *Display* 1 & 2 or *Trial* B, C and D) plot the *absoluteError* of the responses. Then write a short paragraph of analysis. How do the distributions of the data compare across the different methods our perception test studied for encoding numerical data visually? Is there any noticeable clumping of responses for any of the methods?

   b. Were there any tests where people generally underestimated or overestimated the data? Explain what field you can graph to test this, what graphical method reveals this clearly. Analyze the results and explain in a short paragraph.

   c. Compare the data for *Display*s 1 and 2 for subjects 56-73 (in Tableau, you will need to filter the data here, and in R you will need to subset). Create a visualization that shows any differences in the response patterns between the two. These subjects all saw the first set of *Display*s before the second set. Is there any difference in the values for *Display*s 1 and 2? Did the participants get better at judging after having done it once?

   d. An erroneous stimulus was used for the first *Display* of "vertical distance, non-aligned" for a small subset of the subjects. They manifest themselves as an anomalous sequence of "1" responses across *Trial* B, C and D. Look closely at the original raw scores and identify the sequence of subjects (hint: they are contiguous). Visualize the raw scores in a way that highlights these values and makes their anomalous nature clear. It should make it clear not only that they are outliers but should show any features that distinguish them from ordinary outliers. Some features that you might think about exploiting: they are identical values across all three *Trial*s, regardless of what the true values for the *Trial* is; they are only for a small subset of subjects.

e. Continue your analysis and find at least one other **distinct** way to display the data that shows something interesting about the data that is not included in the previous questions**. Clearly explain what you found and how your visualization reveals this feature of the data.**

The following problems requires the use and understanding of logarithmic scales from lecture.

2) **(20pts)** Download the stock data for Intel. This time the file contains data over a longer period, just up to the 2001 .com bust. Graph the data in the following ways with one graph on each page of the workbook

a. Create a standard line graph of the *Adjusted Close*, but use the *Volume* measure to alter the thickness along the curve.
b. Use the *Volume* to alter the color of the line at each point. You may have to make the line thicker to see the result. Strike a balance between the visibility of the color and the definition of the line. Which of these two graphs, (a) or (b), communicates the *Volume* data more effectively? Explain your answer.
c. Start with the graph in (a) and change the scale on the graph to a logarithmic scale. How does this change the shape of the graph? What shape does it become approximately and what does this mean?
d. Now, start with the graph in (c) and change the plot so that the data is presented by years (use the average value over the year for *Price*, and you can use the sum for *Volume*). What years had the greatest increase in value %-wise?
e. Revisit the technique in (a) for adjusting the thickness of the curve. Does it work any better when the curve has fewer data points?

3) **(10pts)** Download and graph the Montana Population data set (different from the one we used on HW1). Create visualizations using logarithmic scales, and intended for a technical audience, that **clearly** demonstrate **visually** the answers to the following questions. Viewers should be able to read the answers to these directly off the graph scales. Different logarithmic scale techniques may be appropriate for each part. If you use a single graph to answer multiple parts, make it clear that you are doing so.

a. How many times has the population doubled since 1890?
b. Has the percentage rate of change in the population increased or decreased over the years? What years had the greatest increase in population %-wise?
c. What years was the population percentage increase greater than 15%?

4) **(20pts)** Download the astronomical data for the Messier objects. These are objects that can be seen in a dark sky with binoculars or a telescope that Charles Messier cataloged in France in the 18th century so that they wouldn't be confused with comets. Some of these are clusters of stars or great clouds of gas in our galaxy, some are galaxies that are **much** farther away. The dataset contains a list of 100 deep sky objects along with their distances from the earth in light-years. Graph this data in the following ways to explore the information provided about these

interesting objects.

For this dataset, you will have to pick suitable scales to make the data readable in your graphs. You should **not** wind up with a majority of the points squashed down along the one axis. In particular, for distances, the scale should show the "order-of-magnitude" of the distance in light years (10, 100, 1000, etc.) clearly.

a. Start by trying to graph one or more properties of the objects against the *Messier Number*. Remember, there is nothing ‘intrinsic’ about this number, it is just the order of Messier’s list. Is there any property that exhibits a pattern with respect to the ordering in his list?
b. Create a visualization that compares the distributions of the distances to the objects in each *Kind*. Note that the *Type* variable is a very different category and is really a sub-category of *Kind*. Do not use that here. Sort the distribution displays in a way that makes the relationship clear.
c. Create a scatter plot with the distance to the Messier objects plotted against their *Apparent Magnitude* (it’s their visual magnitude, a measure of how bright they are in the sky). Note that these values may be… backwards from what you would think. The **higher** the number the **fainter** the object is in the sky. Try to **incorporate that into your visualization to make the relationship clear**.
d. Augment the visualization in (c) by adjusting the size of the points in the scatter-plot based on the angular *Size* of the objects in the sky. Evaluate how easy it is to analyze all four aspects of the data from this graph and give a suggestion on how you might modify the graph to display all this information more readably.

5) **(20 pts)** Download the Portland Water Level dataset and explore it by creating the following visualizations of the time series from the techniques described in lecture. Use both R and Tableau for at least one graph here. They should, of course, adhere to the design criteria that we’ve learned, and should clearly display the information described in each part.

Then write a single paragraph outlining the differences between the information that each graph communicates.

a. This data contains a year of data with water level (*WL*) measurements every hour as a function of *Time* (i.e. 365 x 24 data points!). Since there is a lot of data, clean it up by smoothing the data by calculating a moving average. Use a window approach with window size that covers a range of days (remember, the data is hourly) and graph the smoothed result. Work with the window to see what size window gives you the best view of the changes in the data while still smoothing the noise well.

Remember that the moving average is in the Quick-Table calculations inside of the right click menu on the data item in Tableau, and we can compute it in R quite easily as shown in class.

b. Graph the cycles that happen each day (because of tides). You might try overlapping many days' data as separate overlapping time series, using a level plot, a horizon graph, etc. The point of this exercise is to try to come up with a way of showing the progression of the tides over some period of time that is rich and detailed and which shows the pattern, but which is still readable and which doesn't clutter the graph.