CSC 465

Homework 2

Submit a PDF file with your answers.

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

Include written responses to questions and images of your visualizations (from screenshots or copying and pasting right from Tableau or RStudio into your document), you may use either, but some of these require techniques very difficult to achieve in Tableau or that are far simpler in RStudio.

1) **Reading (Not to turn in)** Read Cleveland sections 3.1-3.13.

2) **(5 pts, due Sunday April 21)** Submit the in-class participation for lecture 3 online. This participation consisted of analyzing three graphs for their visual inaccuracies or misrepresentations.

3) **(30 pts)** This problem continues 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. Create a dot chart of the median error by test. This is the same graph as the last homework part c), but instead of using a bar graph, you are using a dot-chart. Rather than coloring the dot chart categorically, color it so that the color emphasized positive and negative values with different colors. Describe in two or three sentences how the two compare in their communication. The plot should be clean and uncluttered.
   b. Build on your graph from the last homework, part e) by adding a jittered categorical scatterplot overlayed with the box plot to display, for each *Test* (don't distinguish between *Display* 1 & 2 or *Trial* B, C and D), 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? **(2 points of e.c. will be given if you use a normal distribution for the jitter. Make it clear if you have done this).**

c. **(3 points of e.c.)** The data shows quantization in the responses on multiple of 5% experiment with jittering along the direction of the response values (absoluteError) by a small amount … i.e. small enough that you aren't changing the quantized range of the responses. Write a short paragraph of analysis on what visual benefit/harm this provides as well as whether this is distorting the data or not.

d. Use a collection of violin plots to explore the Error field you calculated from last week. For which perception tests did people generally underestimated or overestimated the data? Would a jittered scatter plot overlay be helpful here in understanding the distributions? Analyze the results and explain in a short paragraph.

e. Create a visualization that compares 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). The visualization should have **two graphs**.

    i. One that compares overall results, not broken out by *Test*,
    ii. Another graph that shows each test divided into Display 1 and 2.

    Then in your analysis, answer the following: 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?

f. **Visualizing erroneous data:** An erroneous stimulus was used for the first *Display* of "vertical distance, non-aligned" for a small subset of the subjects. Imagine that you are trying to explain to your team that these responses are compromised and the responses from these subjects need to be removed from their analysis.

    You will have to find the erroneous responses by looking through the data (Excel?). They are an anomalous sequence of "1" responses across *Trials* B, C and D for **specific respondents (i.e. Subject ID numbers)**. Your first task is to look closely at the **original raw scores** and identify the sequence of subjects (hint: they are contiguous) and later in the dataset.

    Next, visualize the **raw scores** (not the errors) as a scatterplot or collection of scatterplots, with the *subject* ID as one of the axes and the *response* in the other axis. Filtering the data will be important. Use color and other visual features to clearly show that these values are different from the other responses. Your graph should make it clear not only they are outliers of with a very specific pattern but are most likely due to a bad stimulus. Some features that you might think about exploiting in the visualization:

    - they are **identical values** across all three *Trial*s, B, C and D, regardless of what the true values for the *Trial* is. This is why *response* is the proper field to plot
    - they are **only** for a small subset of subjects
    - they are **only** for display 1.

    Because of this, filtering (i.e. subsetting in R) will be key in building this visualization.

g. For the graph in b), recreate your visualization with the subjects from part e) removed. Explain whether and how the exclusion of these subjects changes the results.

The following problems requires the use and understanding of logarithmic scales from lecture. Notice that in R, the natural log is the function “log”, and it has two other functions, log2 and log10 for the other bases. In Tableau, the natural log is “ln(value)”, and it has another function called “log(value, base)”.

4) **(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 *Adj.Close*, both with and without a logarithmic scale. You may use the automatic log scale in Tableau or the scale_x_log ggplot scale in R. How does the logarithmic scale alter the visualization? Does it allow you to see any aspects of the data more clearly?
   b. Again, use a standard line graph and use the *Volume* field to alter the color of the line at each point. You may have to make the line thicker to see the result, and transparency (alpha in R, or the “Opacity” slider in the color properties in Tableau). Strike a balance between the visibility of the color and the definition of the line. Use a logarithmic scale for price.
   c. Create a calculated field for “logVolume” = ln(volume). Use the log of the volume for the color property in your line graph from b. Does this help or hinder the efforts to visualize the trading volume along the curve?
   d. Limit your graph to the single year of 1995 (filter on the date). Change the log scale to a natural log, and instead of having the adj.close numbers on the scale, display the natural log of the adj.close. To do this, you will have to create a lnClose = ln(adj.close) field for this. Then graph the ln of the adj.close for the year 1995. In your description for this part, identify three surges in price (from a local minimum to a local maximum) that are between 10 and 20% increases. For each, estimate as precisely as you can, just from the graph, the %-wise increase during the price surge.

5) **(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.

   Important note: 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. For distances, the scale should show the "order-of-magnitude" of the distance in light years (10, 100, 1000, etc.) clearly.

   a. Pick three of the variables in the data other than *Messier* Number. For each, plot the value for each of the objects against the *Messier Number* on the x-axis*,* one-by-one. 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? Submit the graph that exhibits a pattern. Remember, you should not have a large number of points lying along the axis, so if you do, how can you adjust for this?

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 *type* here, rather use *kind*. Sort the distribution displays in a way that makes the relationship clear. Make sure your distance axis is transformed in such a way as to not have most of these "Kinds" squashed down to the bottom of the graph along the axis.

c. Create a scatter plot with the distance to the Messier objects plotted against their *Apparent Magnitude* (this is a measure of how bright they are in the sky). Note that these values are backwards from what you might think. The **higher** the number, the **fainter** the object is in the sky, magnitude 1 = very bright, magnitude 9 = very dim. I**ncorporate that into your visualization to make the meaning of the Apparent Magnitude clear**. Again, pay attention to how you handle distance so that the full range is clearly displayed and the distance to all objects are clearly readable.

d. Finally, create a scatterplot or another type of visualization that that displays, for all four of the parameters: Distance, Kind, Apparent Magnitude and the angular Size of the objects readably in one scatterplot. Think closely about what the two axes should be and what would be better presented by a color, size, shape, etc. (One interesting idea: you might look up how these object "kinds" are presented on astronomical maps). Evaluate how easy it is to analyze all four aspects of the data from this graph.