CSC 465

Homework 3

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) **(20 pts)** This problem will not only give you practice creating visualizations, but requires you to follow carefully a somewhat complicated specification of experimental data and use visualization for problem solving. Recall the perception experiment from our first week. 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. On each slide the encodings (e.g. aligned bar, volume, etc.) changed, and each encoding was repeated. The data file for this problem, *PerceptionExperiment.csv*, contains the results from 92 previous students. (For those interested in experimental design, note that the order of the slides was changed for different classes.)

   Here is how the data are laid out in columns: each type of encoding is a *Test*, and each one got displayed with two separate slides. The individual PowerPoint slides are called *Display*s. Each individual *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).

   One way to help yourself understand this is to open the data up in RStudio (or Excel) and scroll through the rows. If you watch how the variable values change as you scroll, you will see what is happening. It is also helpful to use functions like *select, unique, filter, group_by* and *summarize* to get intuition. For example, use *select* to pick *Test* and then pipe to *unique* to find out how many encodings there were (*group_by Test* and then *summarise* accomplishes the same). Try *group_by* with *Test, Display, TestNumber* piped to *summarise* and then *arrange* to sort by *TestNumber.* See our earlier *tidyverse* tutorial for more information.

   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. Define *Error*:

   ```
   Error = Response - TrueValue
   ```

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 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.

a. 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.
b. 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* (absolute value of *Error*). 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?
c. Compare the data for *Display*s 1 and 2 for subjects 56-73 (you will need to filter the data in Tableau or R). 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.

2) **(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 encoded 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.

3) **(15pts)** Download and graph the Montana Population data set (different from the one we used previously). 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) **(20 pts)** We will look at data on air quality, captured from May to September in New York. This is actually built into R, but not as a data frame. There is a copy on the D2L site.
a. Use a scatterplot to look at the relationship between *Wind* and *Solar.R* (solar radiation). Show a fit line. Make sure to produce a clean visualization with emphasis on the trend. This provides one view of the relationship.

   For help doing this in R, see Tutorial 5. In Tableau, this is available from the *Analysis* tab. It is one of the tabs along with *Data* for the panel on the far left (i.e. look at the top of the panel from which you drag variables).
b. Use a plot that will show the distributions of *Wind* and *Solar.R* and allow you to compare with fine detail.

c. Finally, show these distributions in context of the rest of the variables by using a technique for comparing multiple distributions.

   Note: you will need to transform the data in a particular way that we have studied. I it showed in the Tableau tutorial and in an R tutorial. Hint – you need to collapse the current variables into two: (1) stores the original variable name, and (2) stores the corresponding original value.

d. For extra credit, compare *Wind* and *Solar.R* again with a QQ plot. What does this tell you?