Alex Teboul
DSC 465 Homework 1

Submit a single PDF file with your answers.

Clearly label which answer goes with which question or question part. If one graph answers multiple parts, you must clearly indicate how it does so. If it is not easy to find your answers, you may lose credit.

Include text answering each question with accompanying images of your visualizations (from screenshots or copying and pasting from your software into the document). For each question, explain very briefly how you created the visualization including any relevant code.

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, think about the following and spend some time learning how to alter each of these in both R and Tableau. Points will be deducted if you do not address the following:

- 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.
- Categorical axes should be sorted in a way that enhances the decoding of the graph.
- 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!

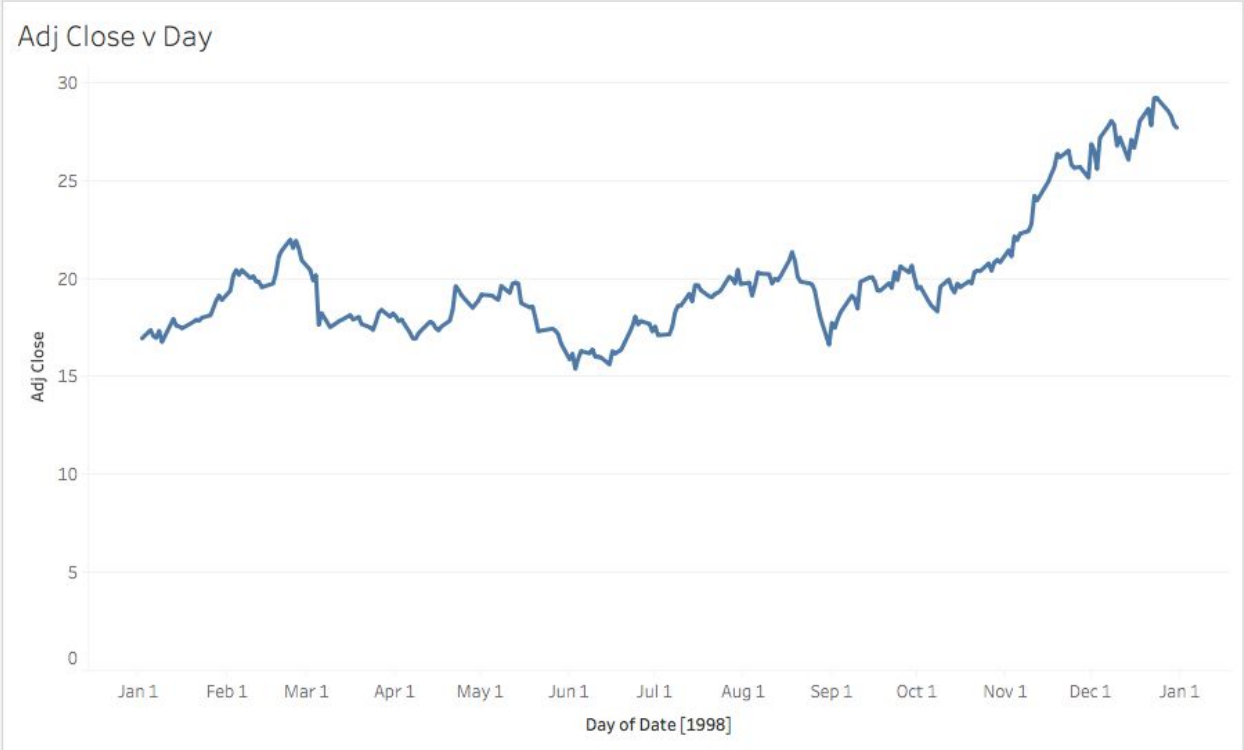
1) (Not Graded) Read sections 1.1-1.2 and 2.1-2.5 from the text, "The Elements of Graphing Data" by William Cleveland. In the second lecture, we will be covering specifically material from 2.2 through 2.5.

Tableau

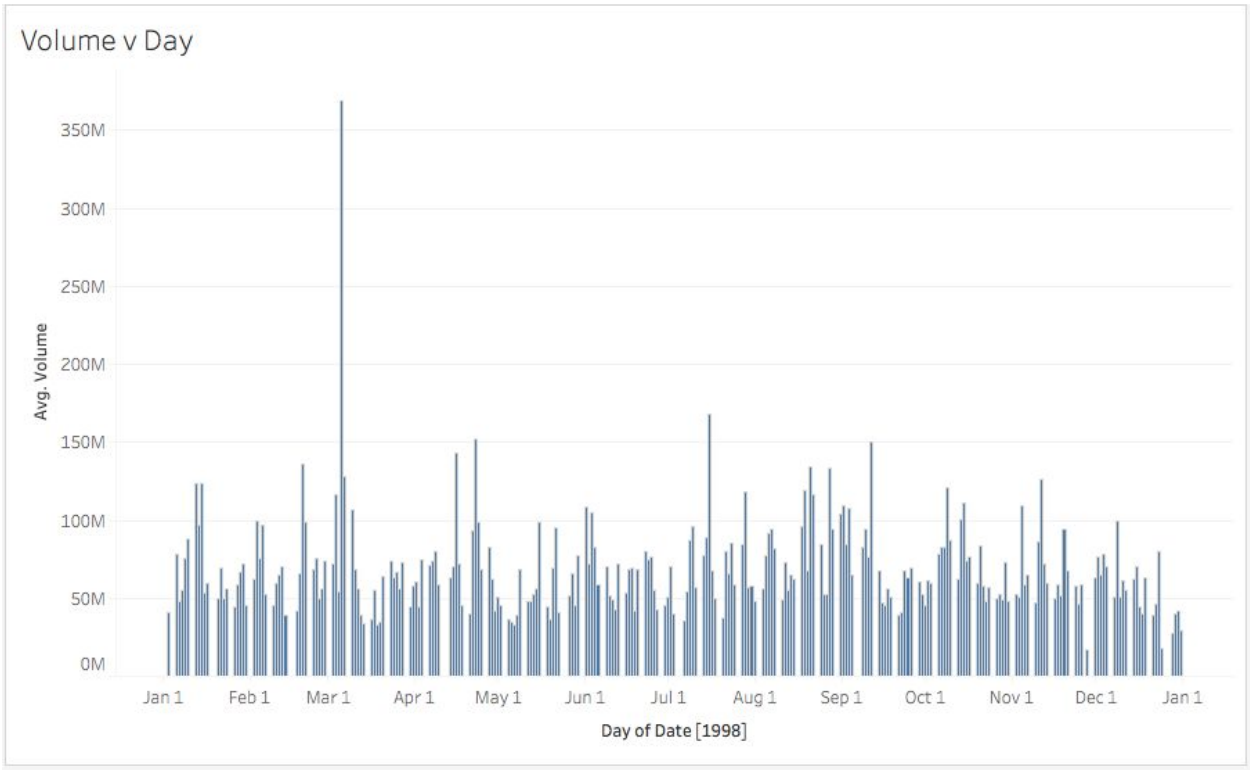
2) (20 pts) For this problem, use the Intel stock (Intel-1998 dataset from the zip file posted with this homework). The data covers stock market trading for the Intel corporation in 1998. Each row is a day, with the following columns:
-Date,
-Trading Day (integer day number, including skips),
-Open (price 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).
In the graphs below, "Price" will refer to the "Adj.Close".

Make the specified graphs in either R or Tableau:

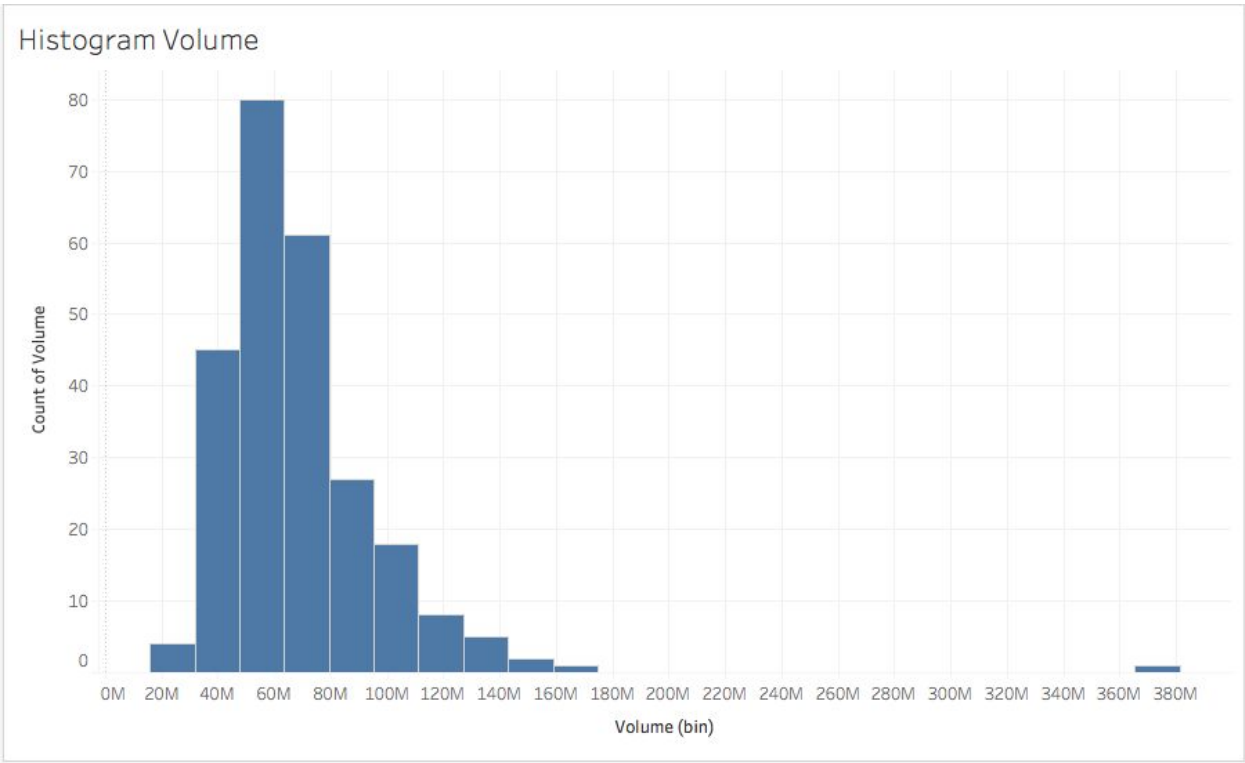
A. **Graph the Adj.Close vs. the date with a line graph. If you use Tableau, you need to right-click on the Date and choose Exact Date from the dropdown menu so that it uses the full date with "day". In R, you will need to convert the date field with as.date.**



B. Graph the Volume vs. the Date as in the last part with a bar graph. The graph should fit in a single display (don't need to scroll to see the rest of the graph) and the bars should be thin enough (use the "size" parameter) as to not overlap.



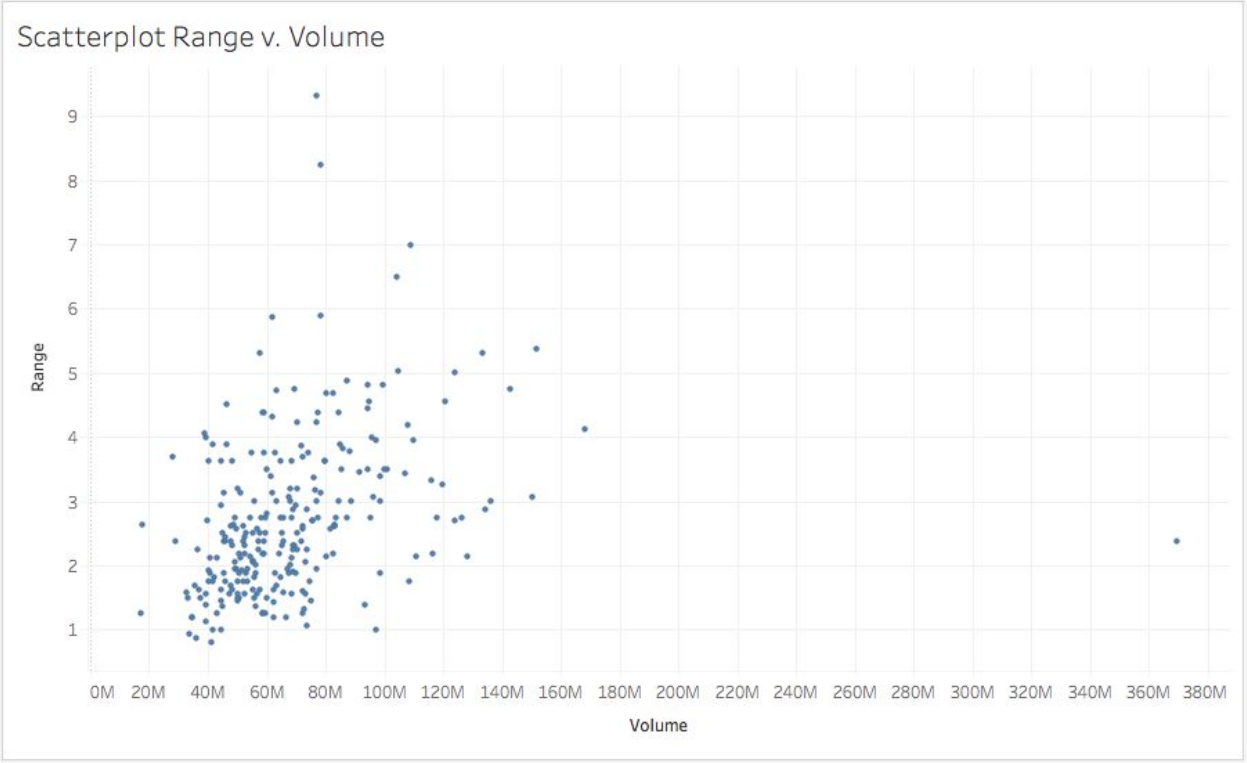
C. Create a histogram of the daily stock Volume. R has the hist command and a ggplot geom. In Tableau, the Histogram graph type in the Show Me box will be useful. Experiment with the bin size. It's an optional parameter in the R functions (e.g. breaks=20 for hist or bins=20 for geom_histogram). In Tableau, after you have the histogram, right click the "Volume (bin)" that is created for you in the "Dimensions" panel on the far left and select 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 in ggplot with the "binwidth" parameter.



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, which is accessible through the right click menu in the "Measures" panel (click in the area below the list of measures). 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".



Tableau

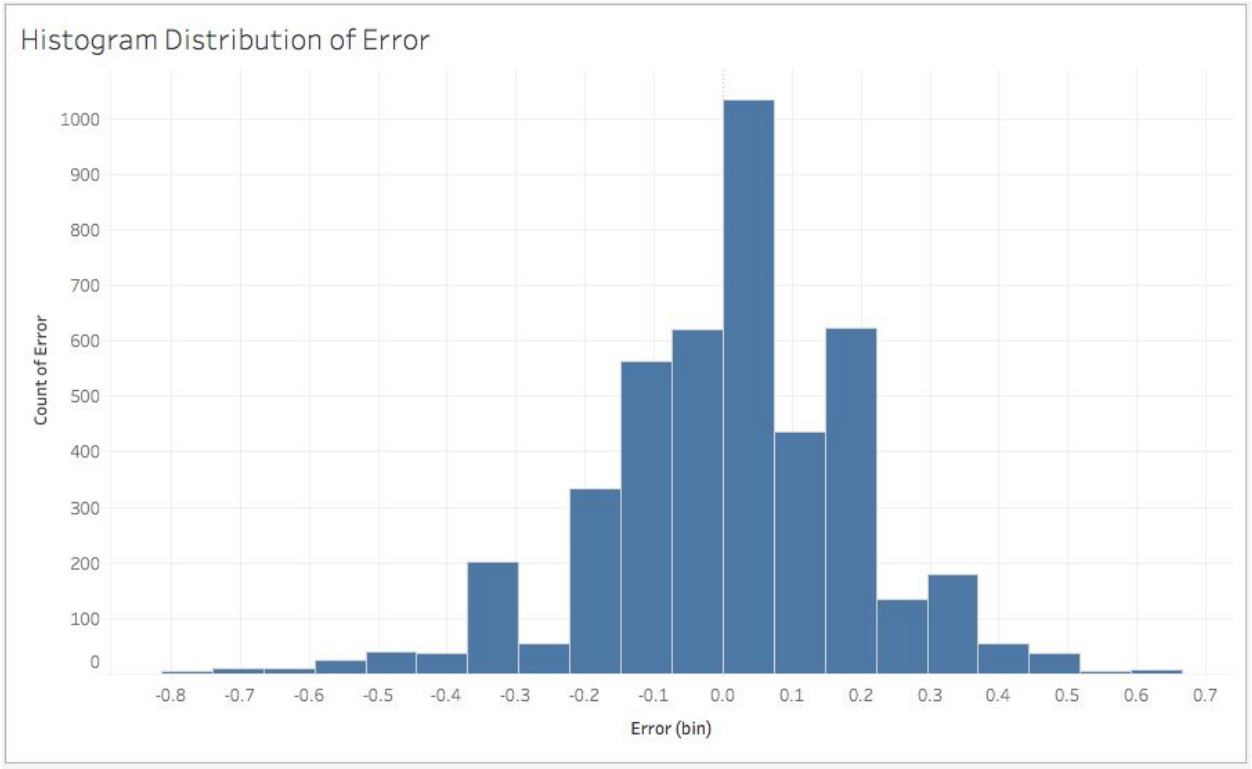
3) (20 pts) Analyze the perception data collected in class to see how accurate students were at perceiving values with different encodings (aligned bar vs. unaligned bars vs. volume, etc.). Use the PerceptionExperiment.csv data file, which has data from 92 students in previous years' classes. Remember that you saw a sequence of slides each with four encoded values, marked A, B, C and D. Each of the B, C and D are judged as a percentage 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 Displays.
-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 Subjects are the students and the estimates they made are the Responses.
-Each row has a copy of the TrueValue, i.e. the correct value that the Response is judged against.

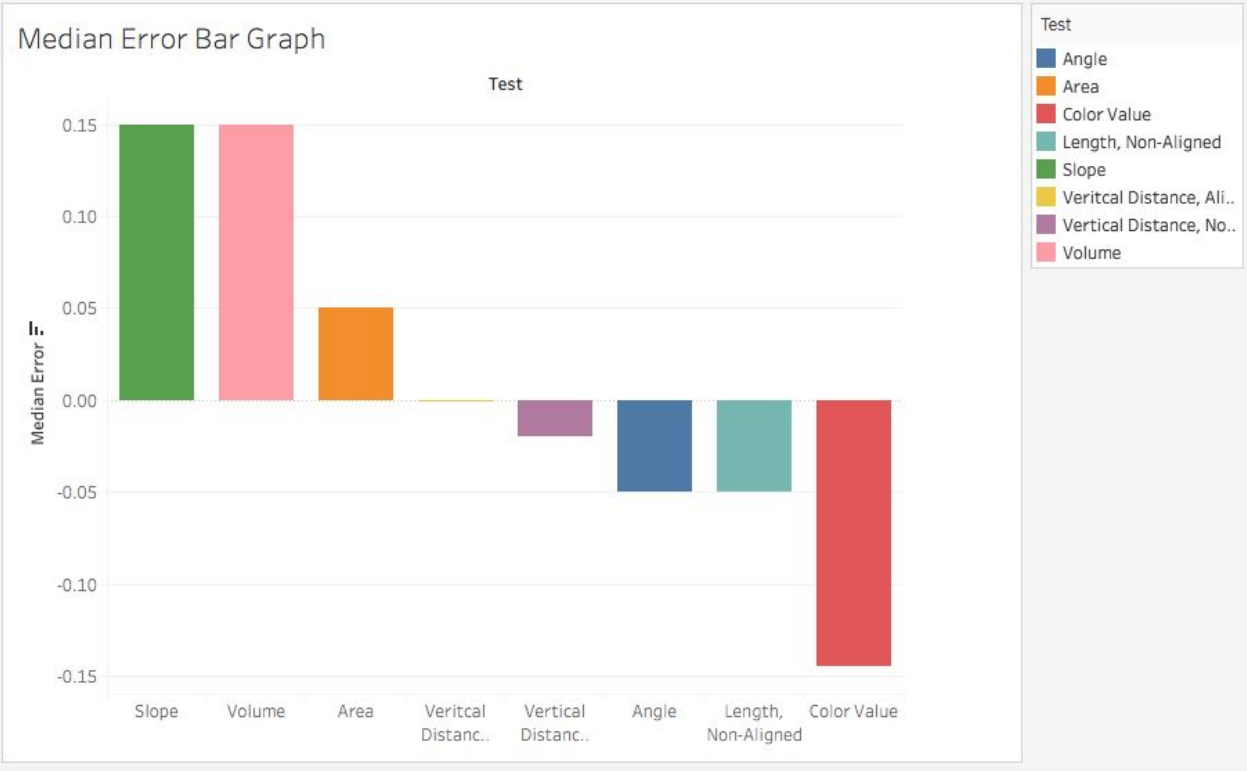
Perform the following to explore the data visually. For each, explain with a few sentences what the graph reveals.

A. **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 2D, define:**
Error = Response – TrueValue Using either Tableau or R, create the following graphs, and pay close attention to the ordering of categorical axes.

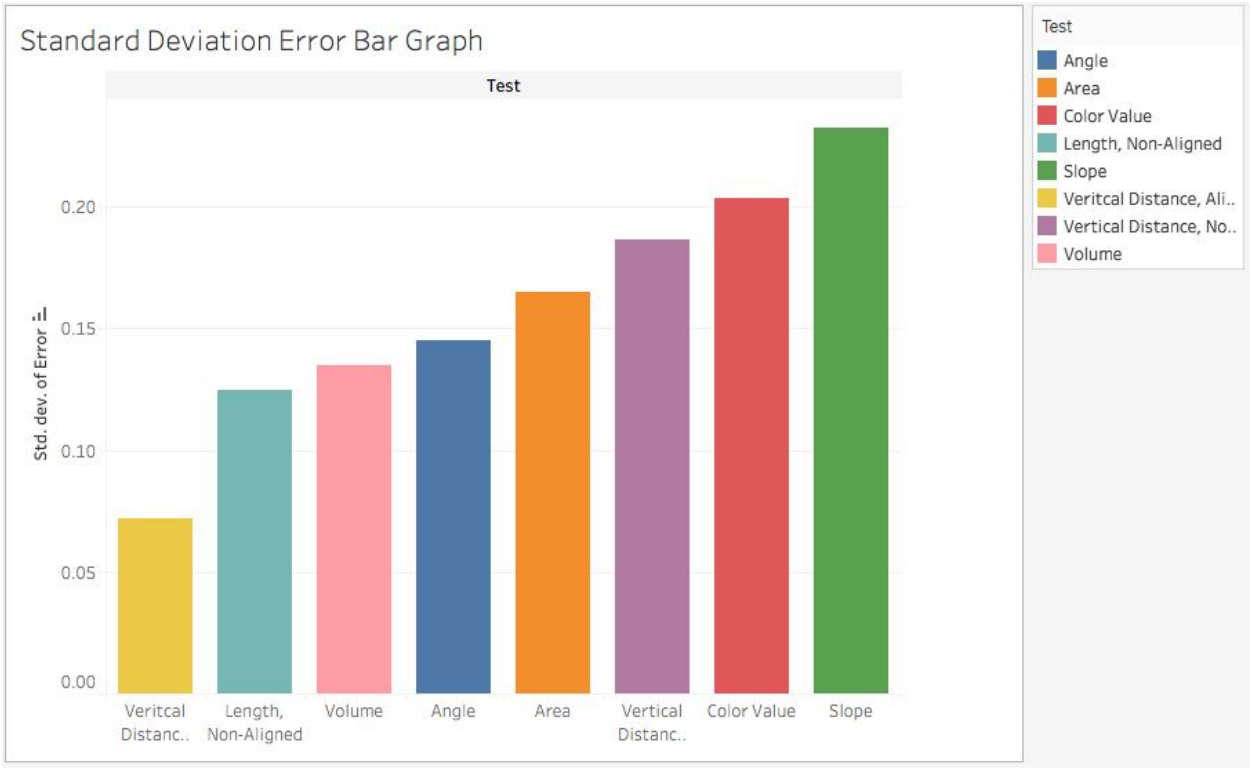
B. **A histogram of the overall distribution of Error.**



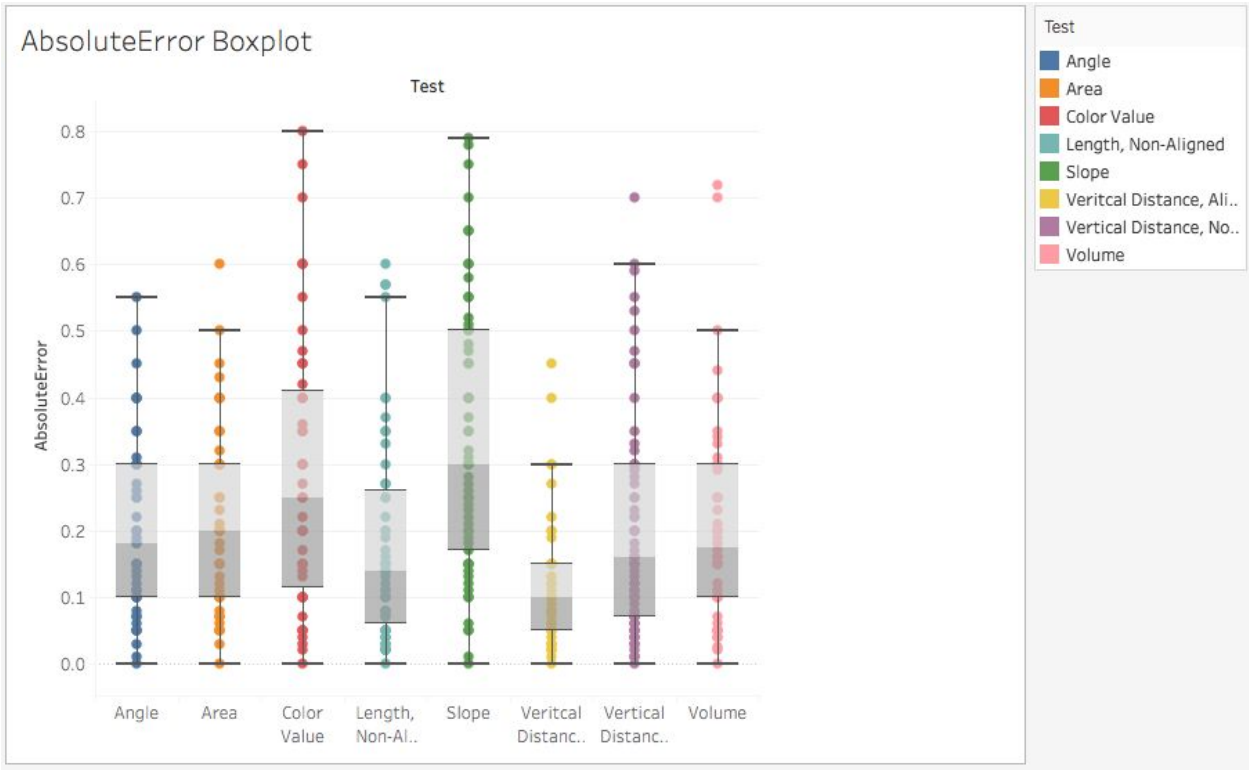
C. **A bar graph of the median Error vs. Test. Do not subdivide by Display or the Trial. Order the x-axis to make the graph as clear as possible. Remember, for bar graphs in general, do not necessarily keep the default order (e.g. alphabetical) of the x-axis.**



D. 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.



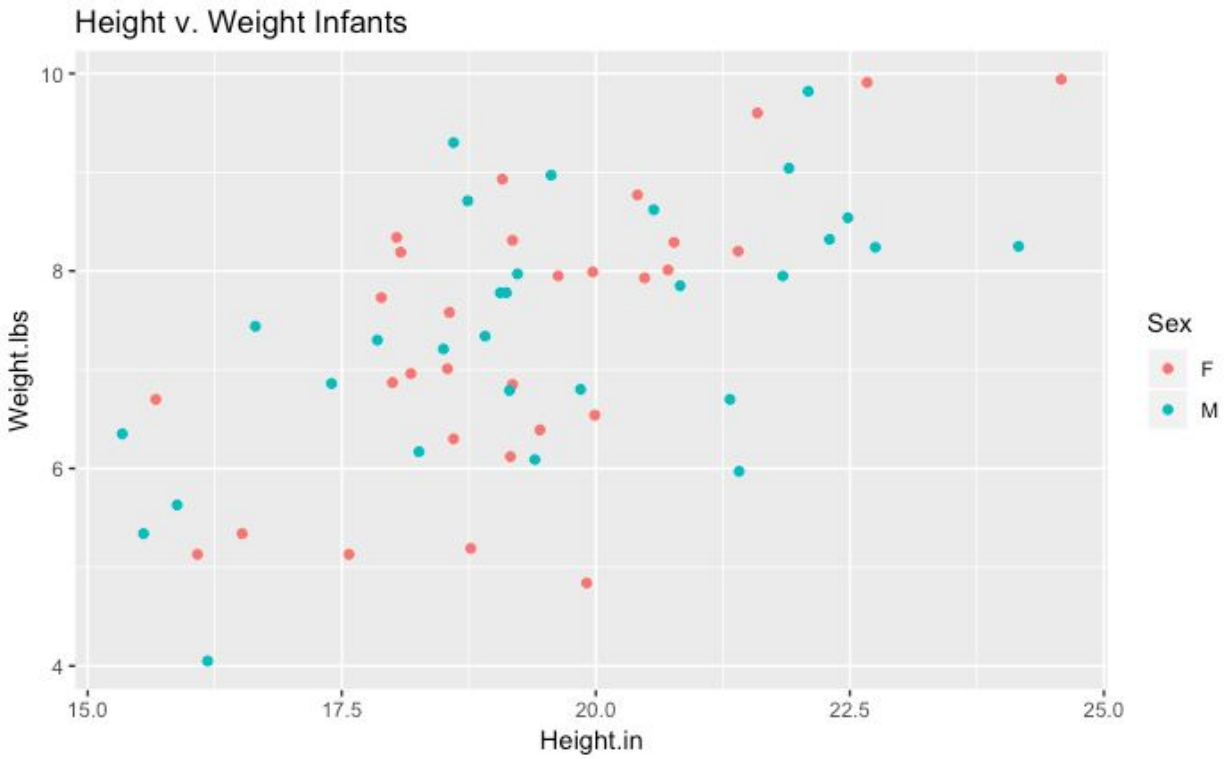
E. Create a new field called AbsoluteError by computing the absolute value of the Error field you created. Then create a box and whisker plot (boxplot) of AbsoluteError by Test.



R

4) (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. Color the plot points by M or F values for Sex.



P4 a)
==========

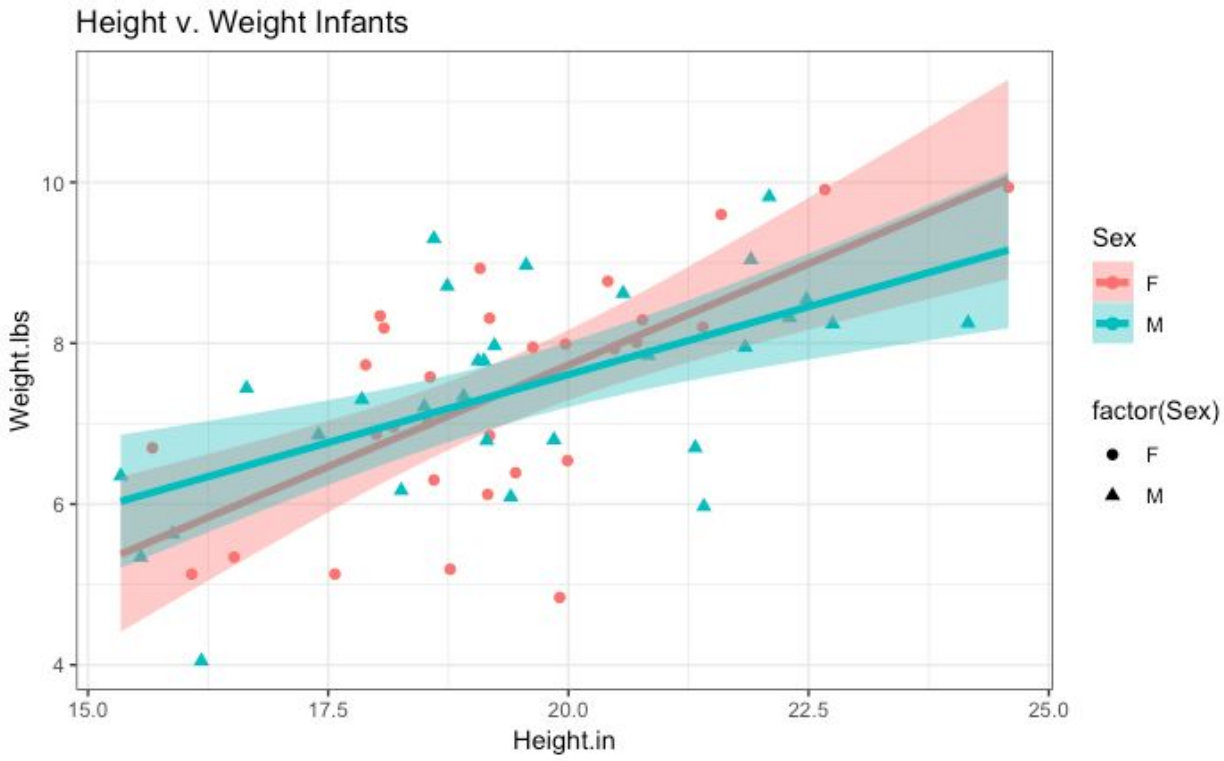
__a) Graph the data as a scatter plot of Height.in on the x-axis and Weight.lbs on the y-axis. Color the plot points by M or F values for Sex. __

```{r p4a}
# load in the data from file
library(ggplot2)
myd=read.csv("/Users/alexteboul/Desktop/Datasets/InfantData.csv", header=T)

#get the variables
Sex = myd$Sex
Height.in = myd$Height.in
Weight.lbs = myd$Weight.lbs

#scatterplot
#ggplot(myd, aes(x=Height.in, y=Weight.lbs, color=Sex)) + geom_point()
plot4a <- ggplot(myd,aes(x=Height.in, y=Weight.lbs, color=Sex))
plot4a + geom_point() + ggtitle("Height v. Weight Infants")
```

B. Create another graph that has the same data but with separate trend lines for the two populations on the graph plotted. Adjust both the line and datapoint size to make the scatter plot lighter but still clearly readable and to make the trend lines stand out.



```
P4 b)
============

__b) Create another graph that has the same data but with separate trend lines for the two populations on the graph plotted. Adjust both
the line and datapoint size to make the scatter plot lighter but still clearly readable and to make the trend lines stand out. __
```{r p4b}
# # Extend the regression lines beyond the domain of the data
plot4b <- ggplot(myd, aes(x=Height.in, y=Weight.lbs, color=Sex)) + theme_bw() +
  geom_point(aes(shape=factor(Sex)),size=2) +
  geom_smooth(method=lm, fullrange=TRUE,aes(fill=Sex), size=1.5) +
  ggtitle("Height v. Weight Infants")

plot4b

```
```

C. Explain in a short paragraph the decisions you made here and their impact on the graphs. See the R examples from the first two classes for reference.

a. Following the directions in part B, I adjusted the size of the points and trend lines slightly increasing them. I also switched to the black and white theme because part B mentioned wanting a 'lighter' graph. Colors can still be difficult to interpret for some viewers, so I also changed the shape of the points so that the M and F sexes are coded by both color and shape (circle v triangle). I kept the se=TRUE to show the error on the trendline and also coded it by the Sex colors. Overall, the graph is now easier to interpret than previously, in both color and extrapolating from the data with the trends. That said, these are pretty weak trends, produced by a small dataset.

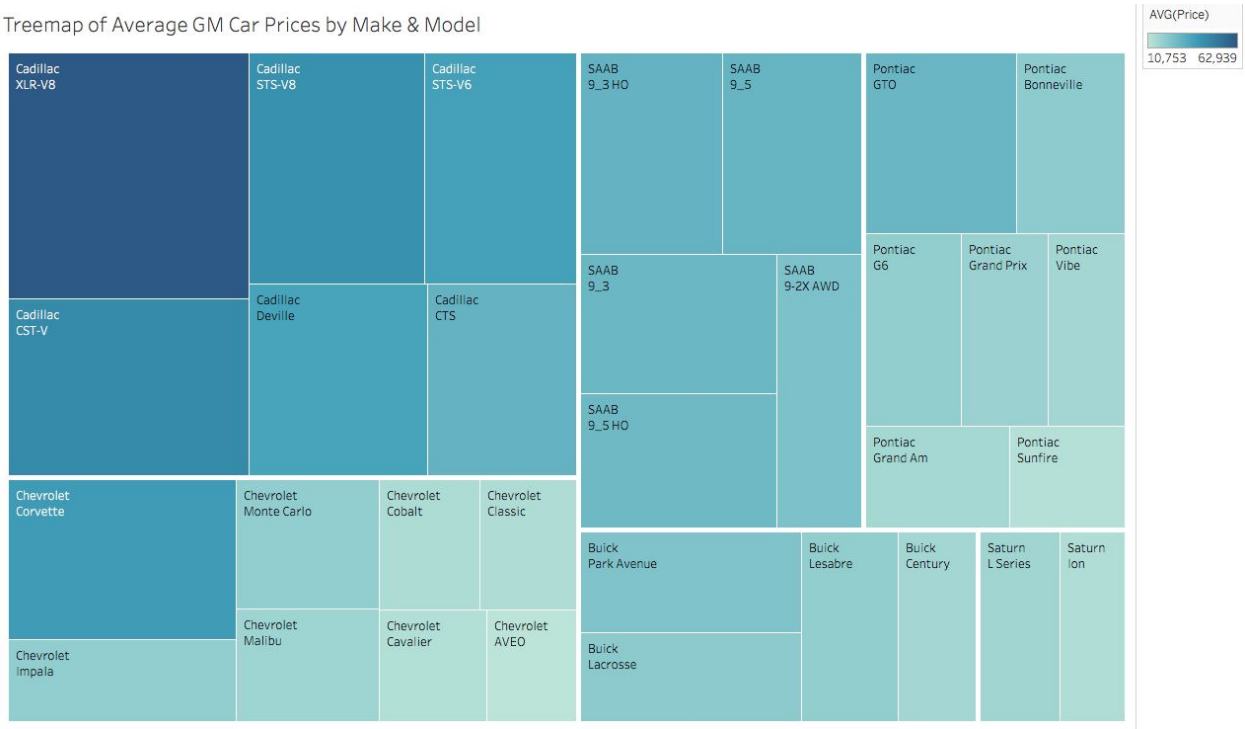
Tableau

5) (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 as demonstrated in class.

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 close attention aggregation type.

a. For aggregation type I used average(price) of each model, and then the aggregate average of all averages within a Make for the main divisions. These graphs are not very clear when copy pasted into this google doc, but for web viewing/powerpoints/etc they are quite clear.

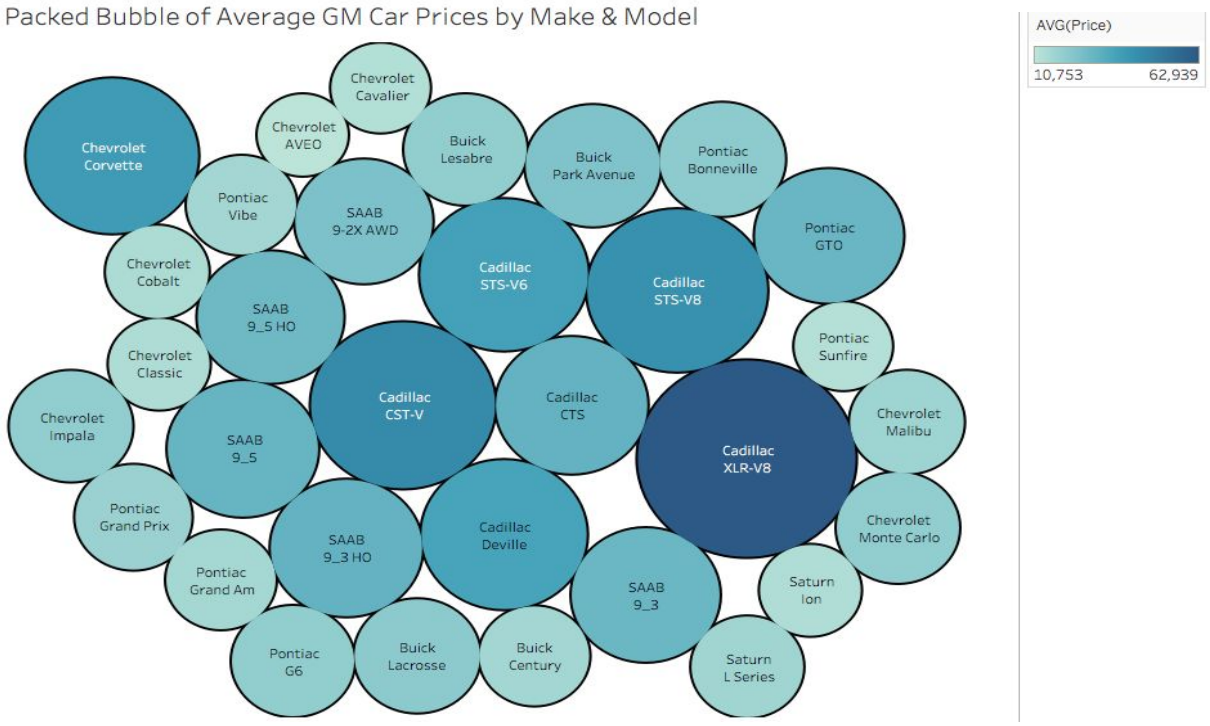
b. Below is the treemap based just on Price with the Make/Model Divisions.



c. I also like seeing the chart blown up with each Make as a different color, and while it may be more cluttered, I like having the average price on there as a label as well. I show this below but the original above is what my 'answer' to the question is.

B. A packed bubble chart of the same type.

a. The copy/paste from Tableau ends up blurry/small unless zoomed in on. Because color sorting is alphabetical and not based on those average prices, again I just show it based on the price like the question asked for.



b. For what it's worth I hate packed bubbles, grouping by Make makes it a little cleaner though.

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

a. The Treemap in Tableau presents the GM car data as a set of nested rectangles, with a main division for the Make of each car and subdivision for the different models of car with average Price being the aggregating feature. The Packed Bubble presents the data in the form of a cluster of sorted circles whose size is determined by the average price of each Make/Model. From the Treemap we can quickly determine that the Cadillac XLR-V8 Model has the highest average price of all cars, and that Cadillacs in general have the highest average prices across all Makes. The Chevy AVEO has the lowest average price, which you can tell from

the treemap fairly quickly and tell based on the color difference that it's less pricey than the Saturn Ion which would be the next best guess for lowest price Model.

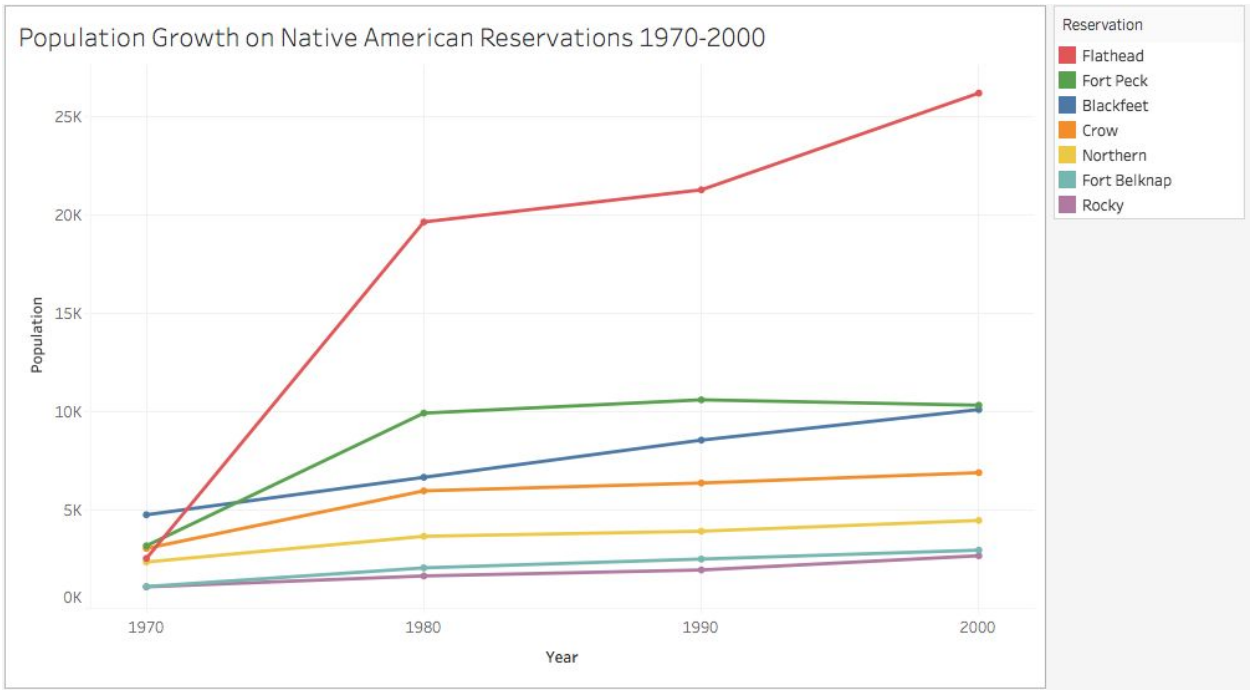
b. This is much more difficult in the Packed Bubble plot. First, telling the size on bubbles apart, especially at the lower end is difficult here. Additionally, the bubbles, while clustered together, do not have the same clear sorting that takes place in a treemap which would help to quickly find a relelevant datapoint. Even grouping by Make does not help make these packed bubbles any more clear. Treemap is better.

Tableau

6) (20 pts) This problem works with a dataset containing the population of each of 7 Native American reservations in Montana (reservation70-00.xlsx). There is a measurement for each decade between 1970 and 2000. Create graphs to show the following information, using appropriate graph types. Part of this problem is for you to discern, based on what we covered in class, what graph types are appropriate for each part.

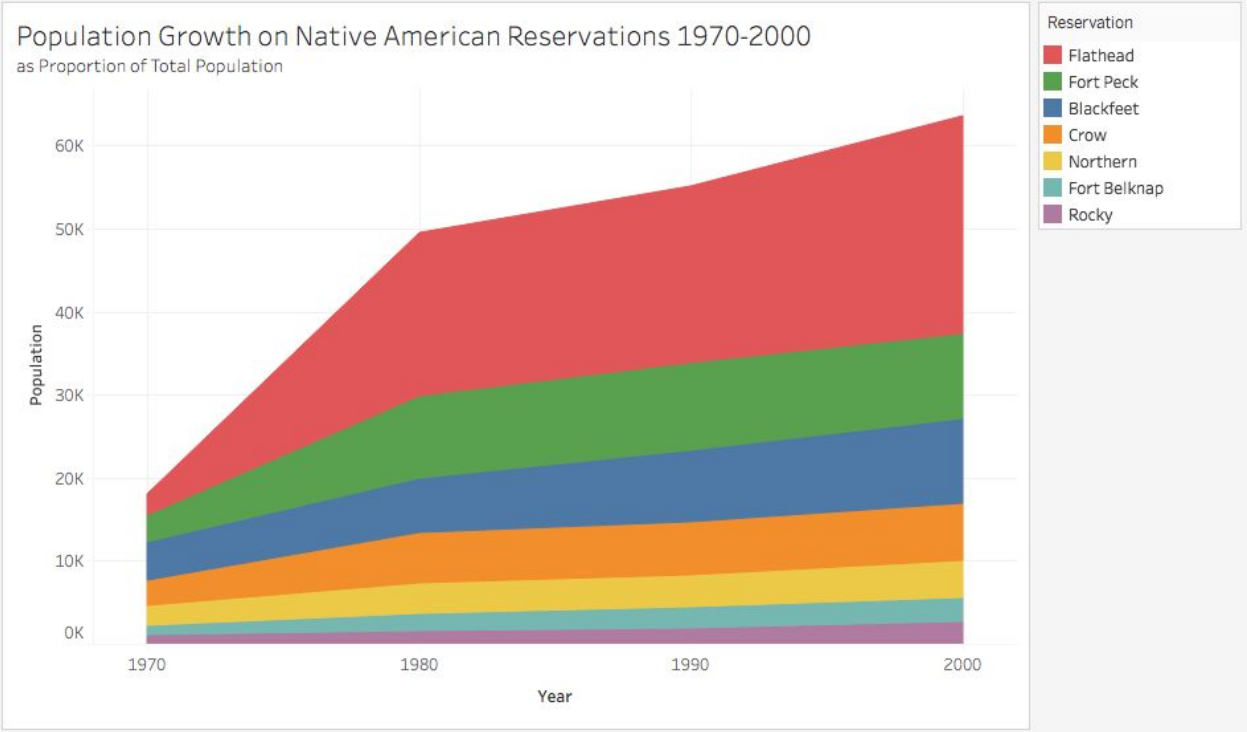
A. Create a graph that shows the continuous population growth over the years for each individual reservation. Think about the order of the reservations in the graph.

a. We don't really have that continuous of a dataset, only 4 data points per reservation. A line graph is sufficient to show these population changes over time.



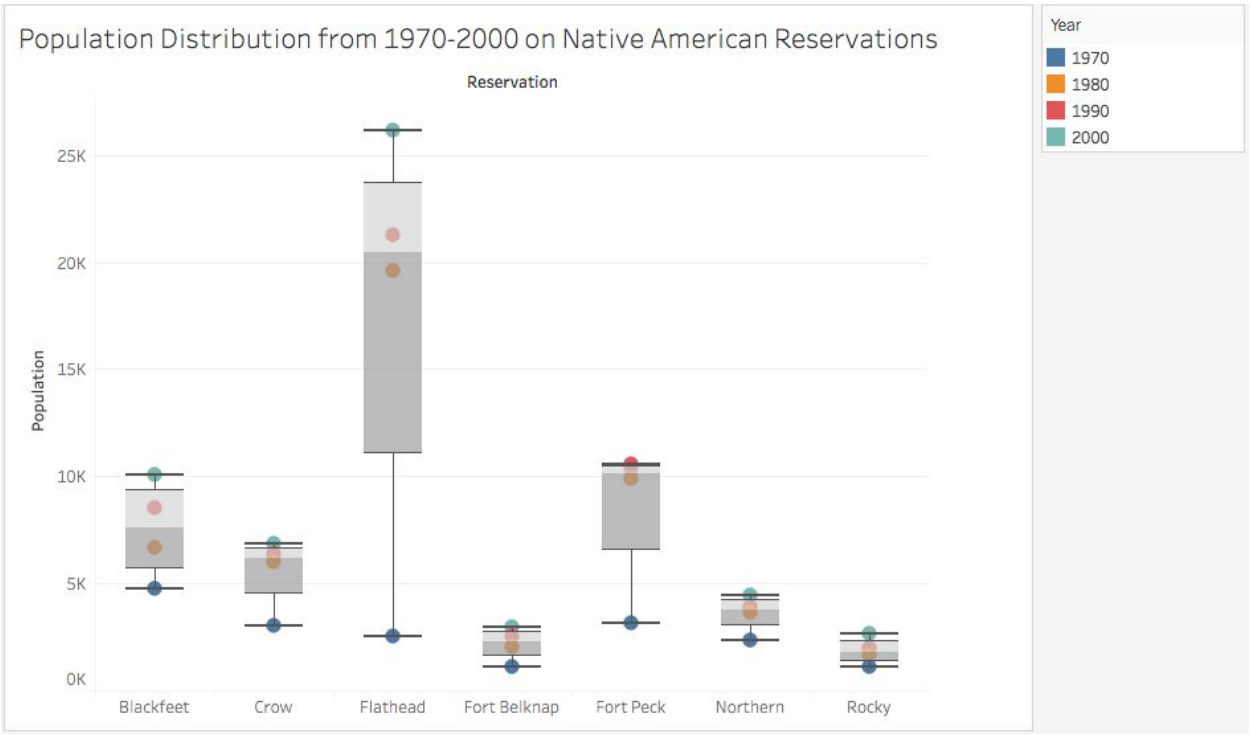
B. One that graphs the total reservation population for each year subdivided among the different reservations. The difference between this and (a) is that here we are not looking only at each population individually but at the total reservation population for each year, with each year subdivided into the reservation populations.

a. The subtle change here to a stacked area chart displays each reservation population and its growth in proportion to the other reservations.



C. One that graphs the population distribution vs. years for each reservation with a box-and-whisker plot. The x-axis should be the reservations, and the y-axis should be the reservation populations. Each reservation will have four values which will be summarized by the box-and-whisker plot. So, for c) we are showing a '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. The a box-and-whiskers for each reservation shows the distribution of population values at that location during this overall period.

a. Below is the box-and-whiskers plot for the population distribution. I didn't go with a gradient for the color of the dots to so they can be more easily compared across reservations when taking into consideration the transparency/opacity effect on the dots.

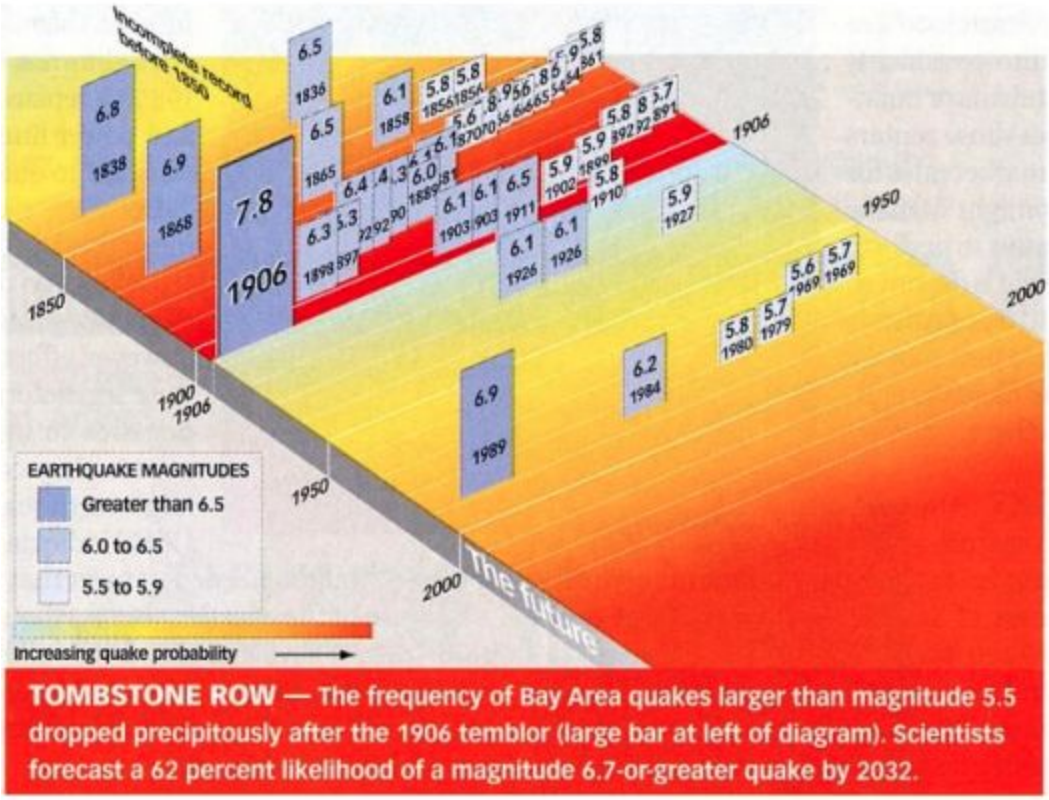


Make sure that the graphs are properly labeled and that the axis scales properly reflect the type of data represented.

Tableau

7) (10 pts) Analyze the following graph for its effectiveness and accuracy in displaying its data. Explain in a paragraph at least three (3) issues that the visualization has, and then use the criteria for clarity and accuracy presented in class to propose an alternative design for the graph that would better communicate the content of the display. Make sure though that all the data that is presented in the original is included in the new design but note that you do not need to organize the elements in the same way on the page and can even separate it into more than one display if it communicates better. If you do so, explain why. You may use paper and pencil or any software to draw the alternative design. It does not have to be 100% accurate (you do not need the numeric data) but it should clearly demonstrate the design changes proposed.

One thing to consider … how much bigger is a 7.0 vs a 6.0 magnitude earthquake?



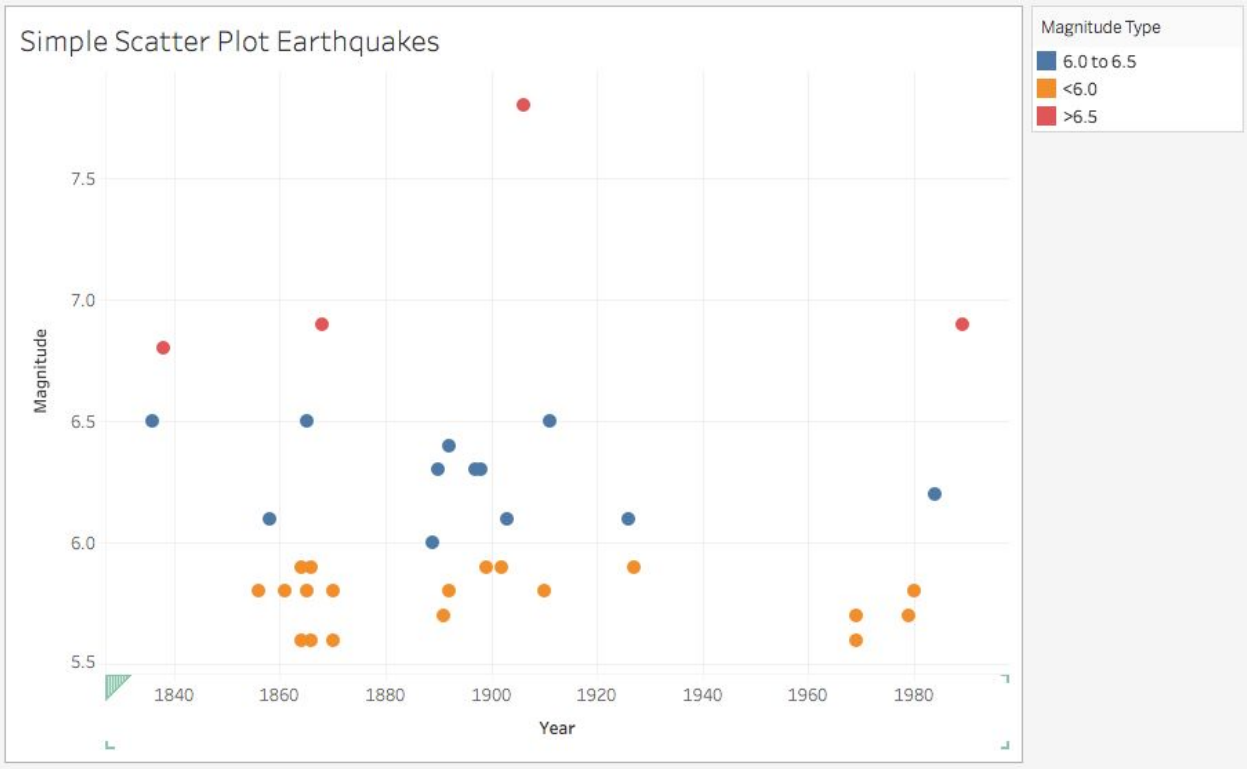
TOMBSTONE ROW — The frequency of Bay Area quakes larger than magnitude 5.5 dropped precipitously after the 1906 temblor (large bar at left of diagram). Scientists forecast a 62 percent likelihood of a magnitude 6.7-or-greater quake by 2032.

- A 7.0 is 10 times bigger than a 6.0 earthquake in terms of shaking amplitude.

Issues with this visualization:

1. **3D Effect** - The 3D effect does not add to the ability of the visualization to inform the audience about the magnitude and frequency of Bay Area earthquakes.
2. **Clutter** - The visualization is too cluttered to provide the audience with
3. **Area** - As discussed in class, area and comparing areas can sometimes be difficult. The variable area of each earthquake case based on magnitude is further confused by the 3D effect.
4. **'x-axis' (Time)** - The time axis does not have sufficiently clear labeling.
5. **Color Gradient** - The increasing quake probability gradient does not actual match the frequency of quakes in the dataset and the earthquake magnitudes categories do not have a sufficiently distinct difference in colors to tell them all apart.

Alternative Design:



- My proposed alternative design for the earthquake data is shown above. I took the values displayed on the given graph and made a basic scatter plot in Tableau with the different categories color coded. This offers a less cluttered, more succinct view of the data. Using this graph it appears that 2-3 <6.0 magnitude quakes and then 1-2 6.0 to 6.5 magnitude quakes generally precede a larger >6.5 magnitude quake. This is a small dataset, but it suggests nonetheless that it wouldn't hurt to be cautious if a ramp-up in <6.0 magnitude quakes were to be observed.