Kubra Iqbal

CSC 465 Homework 1

Question 1 2 3 4 5 6 7

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. 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.

Done

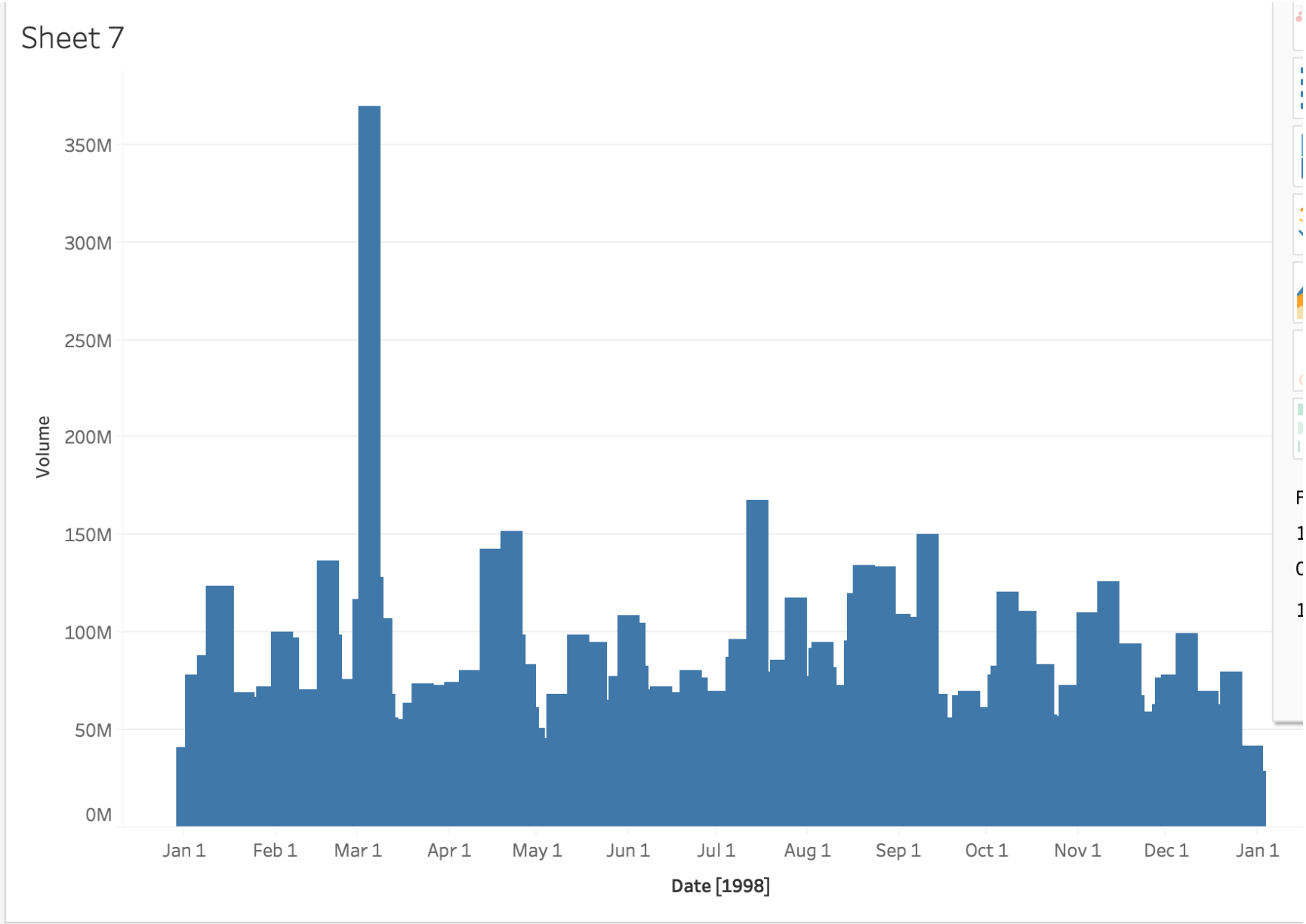
Question 2) (20 pts) For this problem, use the Intel stock (Intel-1998 dataset from the zip file posted with this homework). The data covers stock market trading for the Intel corporation in 1998. Each row is a day, with the following columns: *Date*, *Trading Day* (integer day number, including skips), *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:

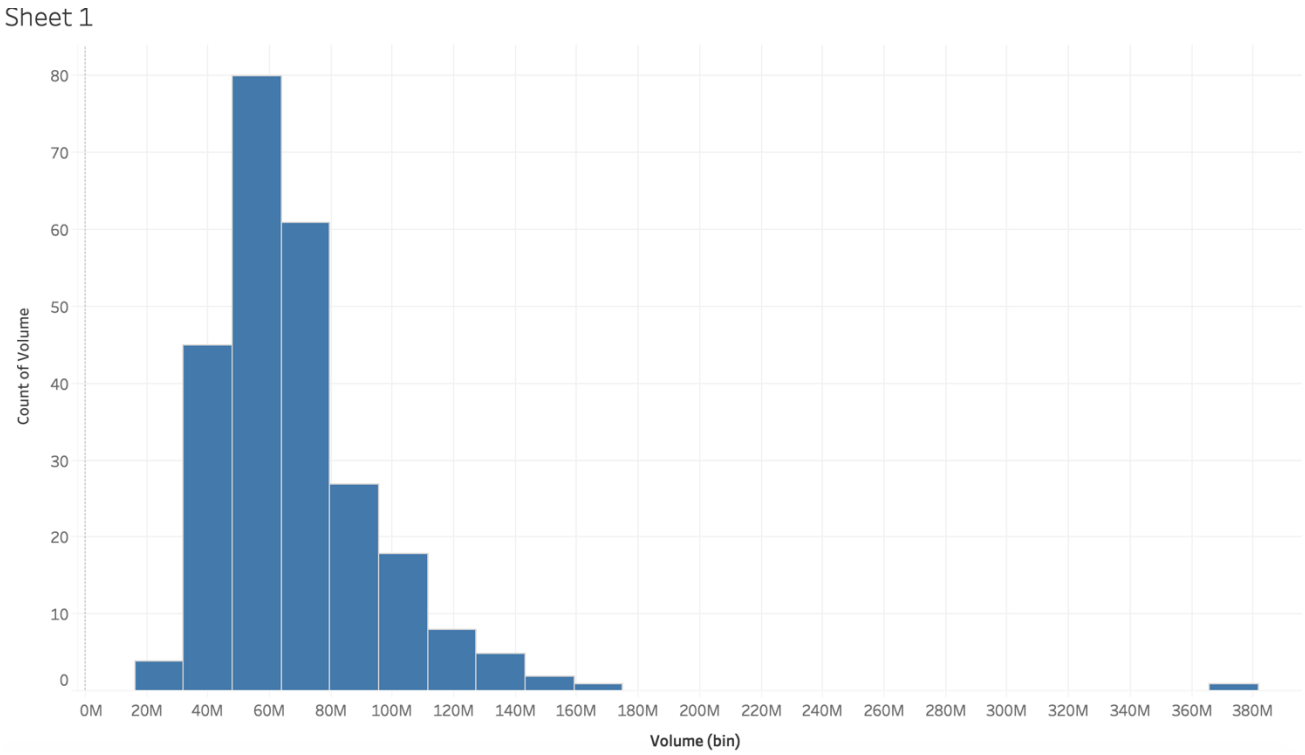
1. Graph the Adj.Close vs. the date with a line graph. If you use Tableau, you need to right-click on the *Date* and choose *Exact Date* from the dropdown menu so that it uses the full date with "day". In R, you will need to convert the date field with as.date.

Used Tableau to create this graph. ADJ close on the Y axis and Day of Data – X axis.

2. 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.


Sheet 7
350M
300M
250M
200M
150M
100M
50M
0M
Volume
Jan 1
Feb 1
Mar 1
Apr 1
May 1
Jun 1
Jul 1
Aug 1
Sep 1
Oct 1
Nov 1
Dec 1
Jan 1
Date [1998]

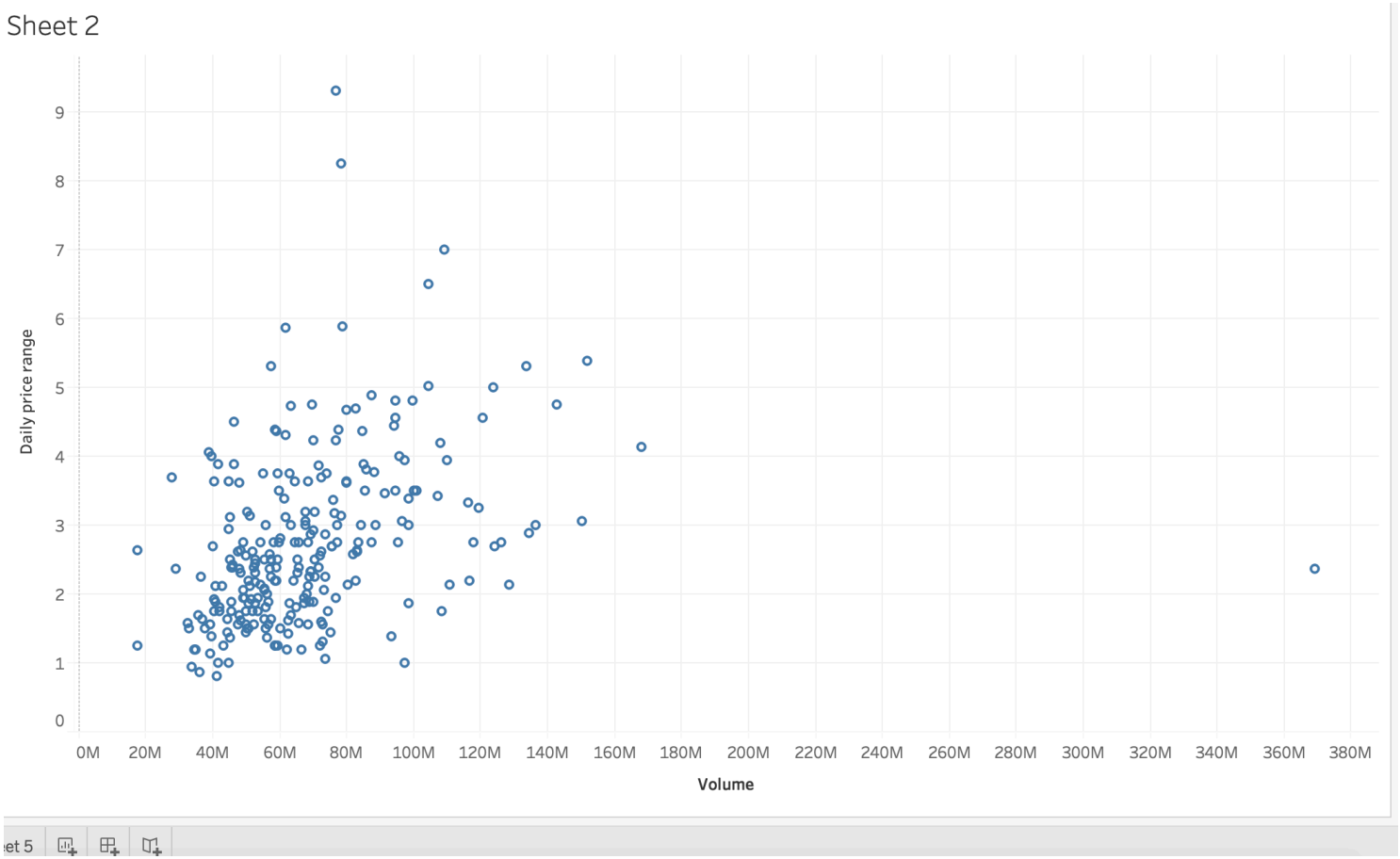
3. Create a histogram of the daily stock *Volume*. R has the *hist* command and a *ggplot* geom. In Tableau, the *Histogram* graph type in the *Show Me* box will be useful. Experiment with the bin size. It's an optional parameter in the R functions (e.g. breaks=20 for *hist* or bins=20 for *geom_histogram*). 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.



4. 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".


Sheet 2
Daily price range
0
1
2
3
4
5
6
7
8
9
0M
20M
40M
60M
80M
100M
120M
140M
160M
180M
200M
220M
240M
260M
280M
300M
320M
340M
360M
380M
Volume
et 5

Question 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 *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 Response is judged against.

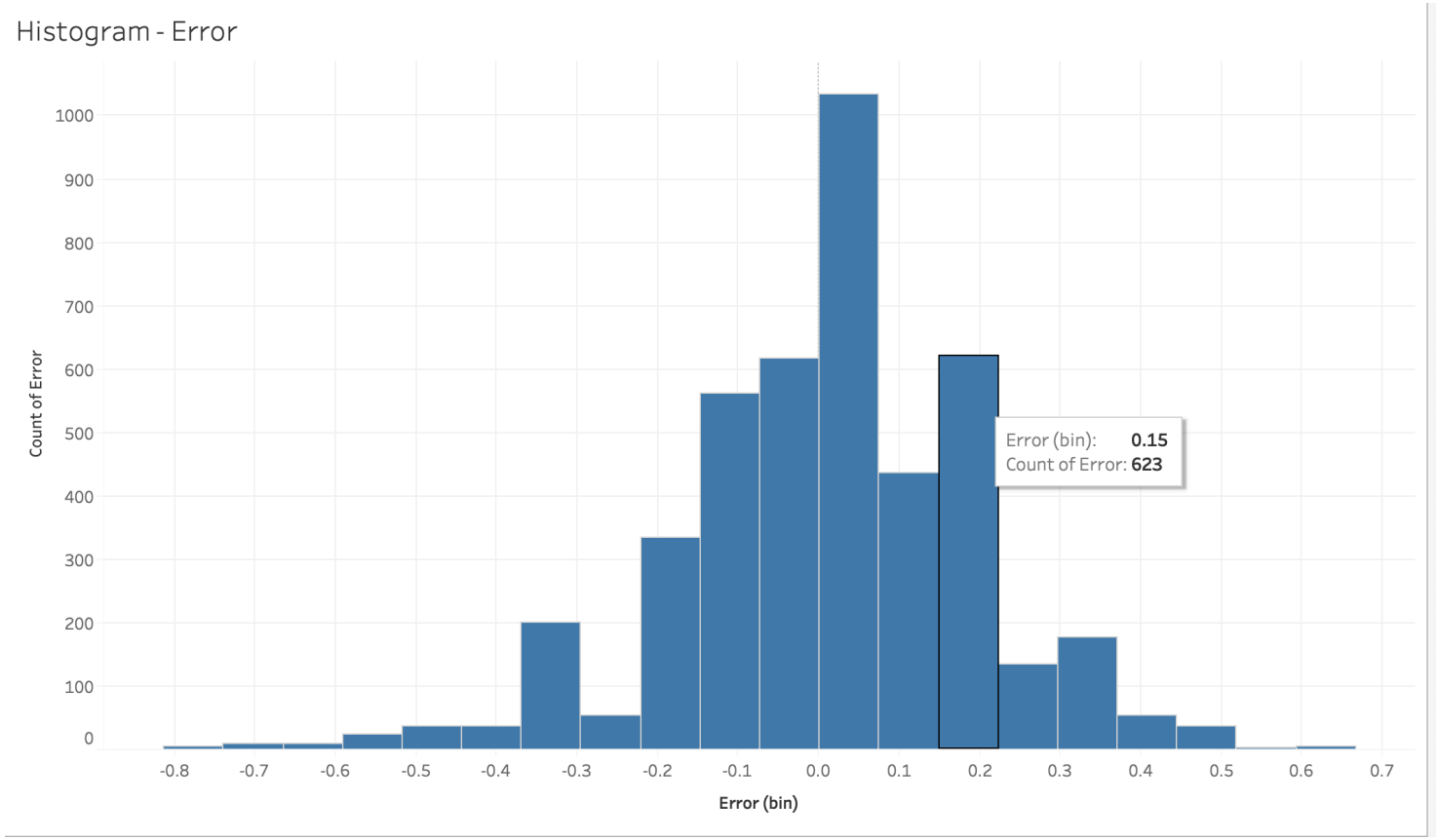
Perform the following to explore the data visually. For each, **explain** with a few sentences what the graph reveals.

1. 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

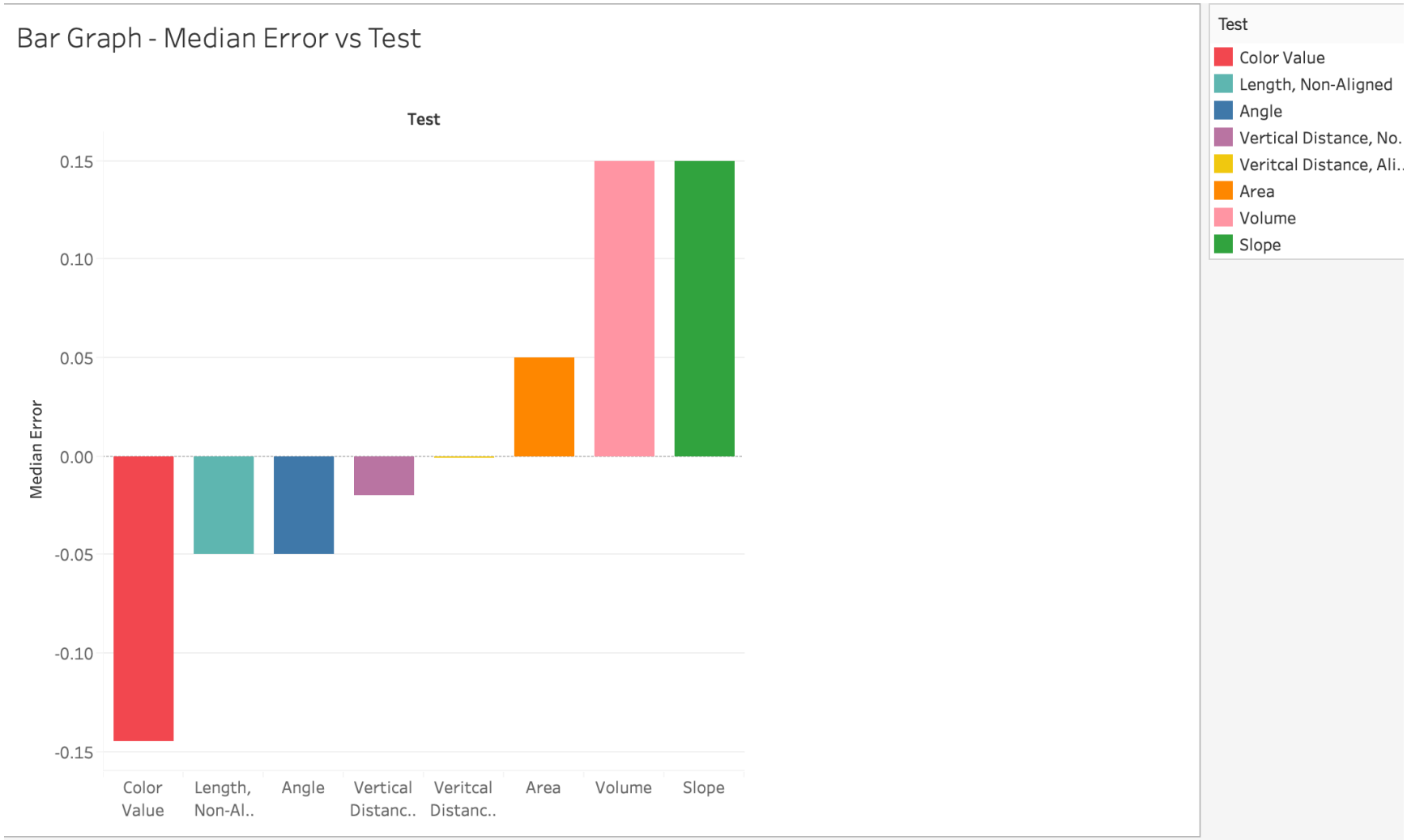
   Using either Tableau or R, create the following graphs, and pay close attention to the ordering of categorical axes.

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



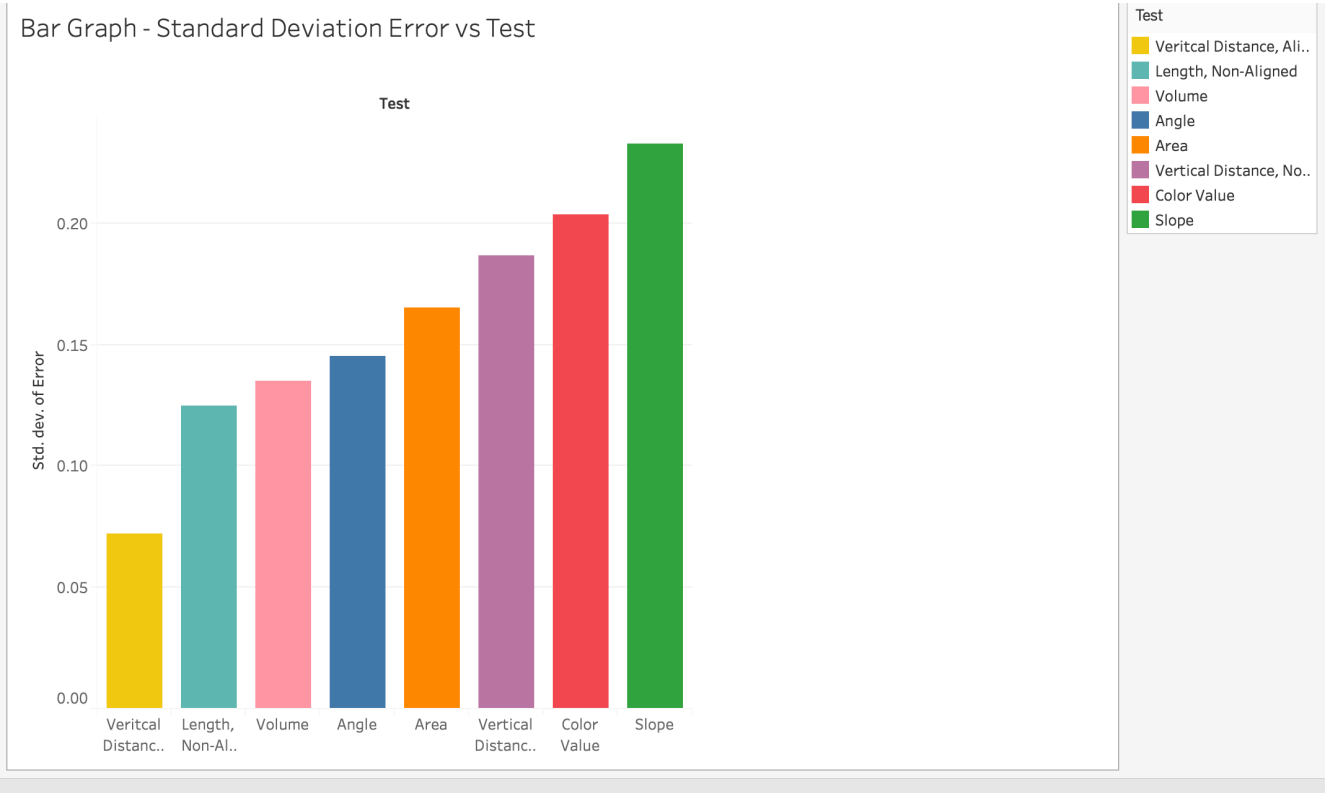
The histogram above shows the distribution of Error is skewed towards the left side. There are also some outliers that are present in this data.

3. 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.



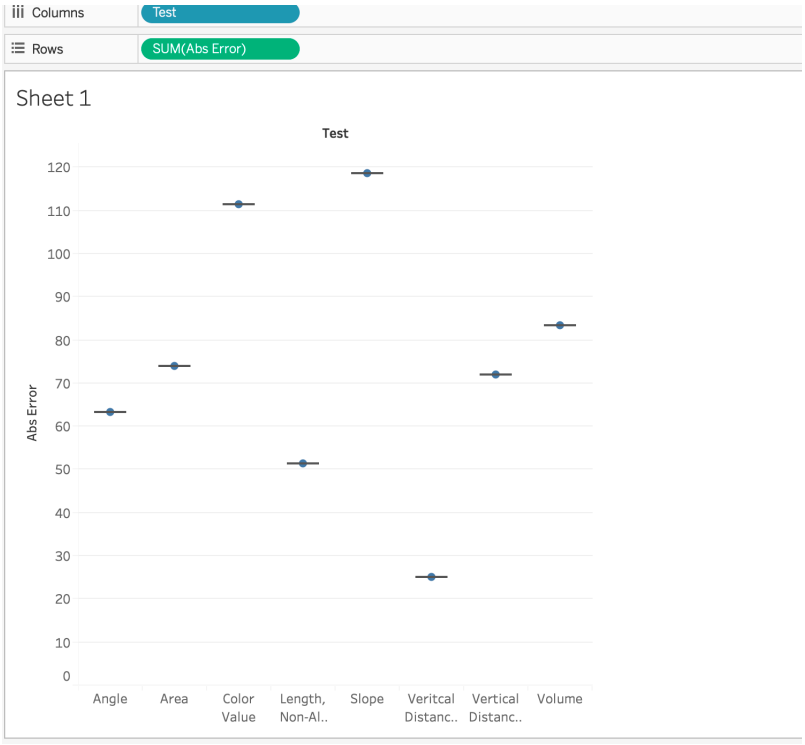
The bar chart above shows that the color value has the lowest median error and the slop has the highest median error. It can be noticed more in the bar chart below which has been sorted.

4. 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.



In this scenario – bars are very easy to compare. By sorting the chart to form the lowest to the highest – it is easy to visualize the range of the data that is being presented in the specific graph. The Bar Chart graph above shows that the Vertical Distance Aligned has the lowest standard deviation of error and the Slope has the highest standard deviation of error.

5. 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.

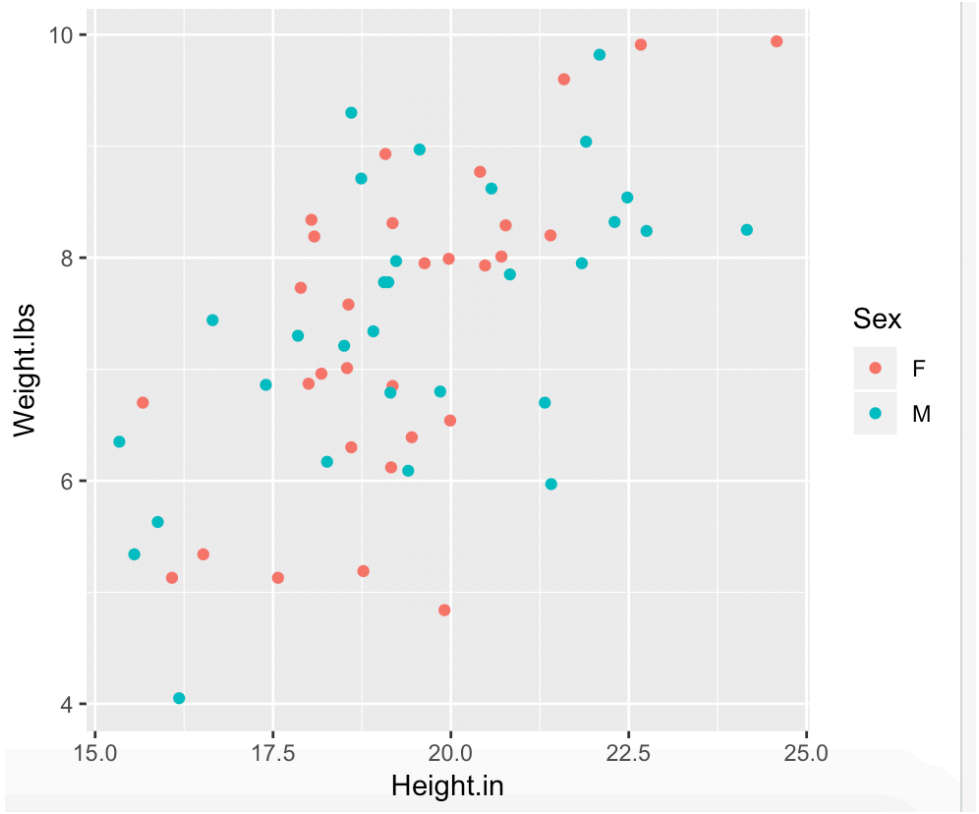


Question 4 (20 pts) Use R for this problem. We will look at data on infant sizes at birth (InfantData.xlsx). There are libraries to help you import the Excel file directly, but in my experience, they are finnicky. The easiest thing to do is open the file with Excel or other compatible software and save it as a CSV file.

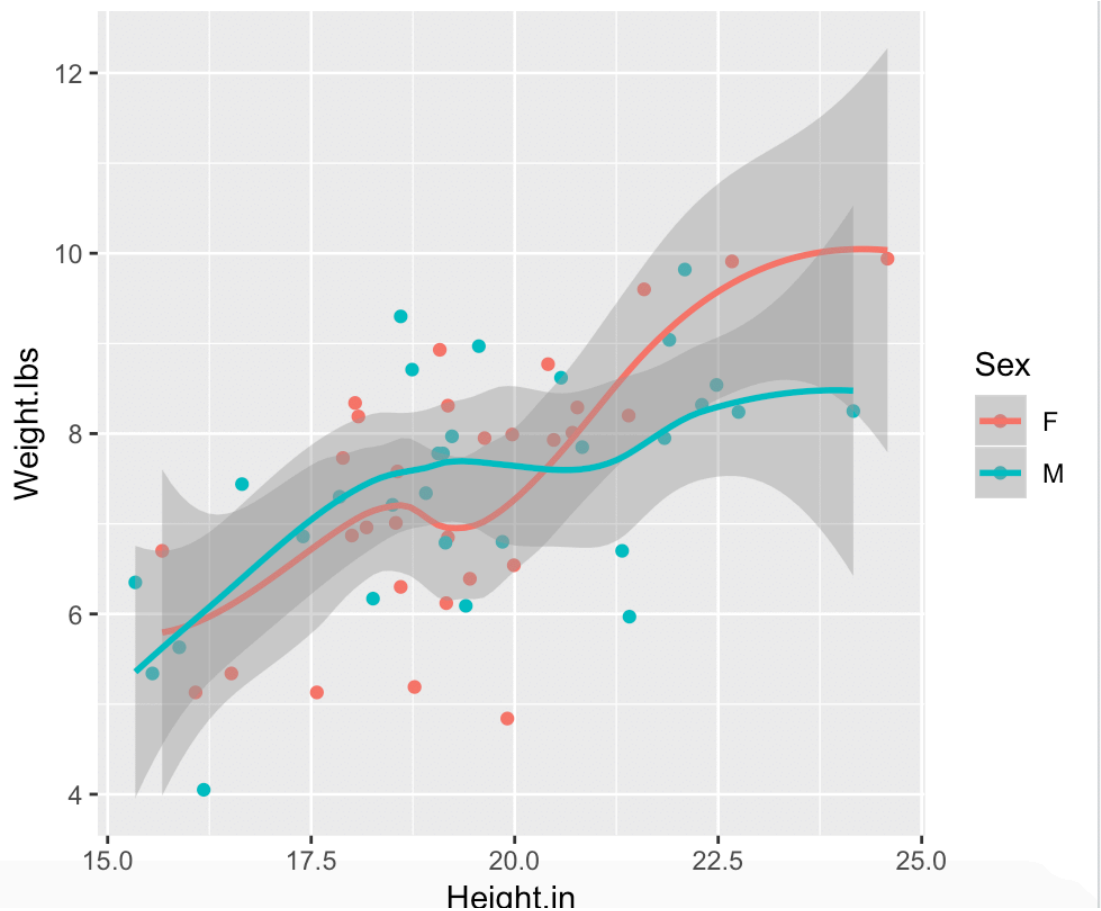
Create the following graphs:

1. 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*.

```
ggplot(data=ds, aes(x=Height.in, y=Weight.lbs, color=Sex)) + geom_point()
```



2. 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 data- point size to make the scatter plot lighter but still clearly readable and to make the trend lines stand out.



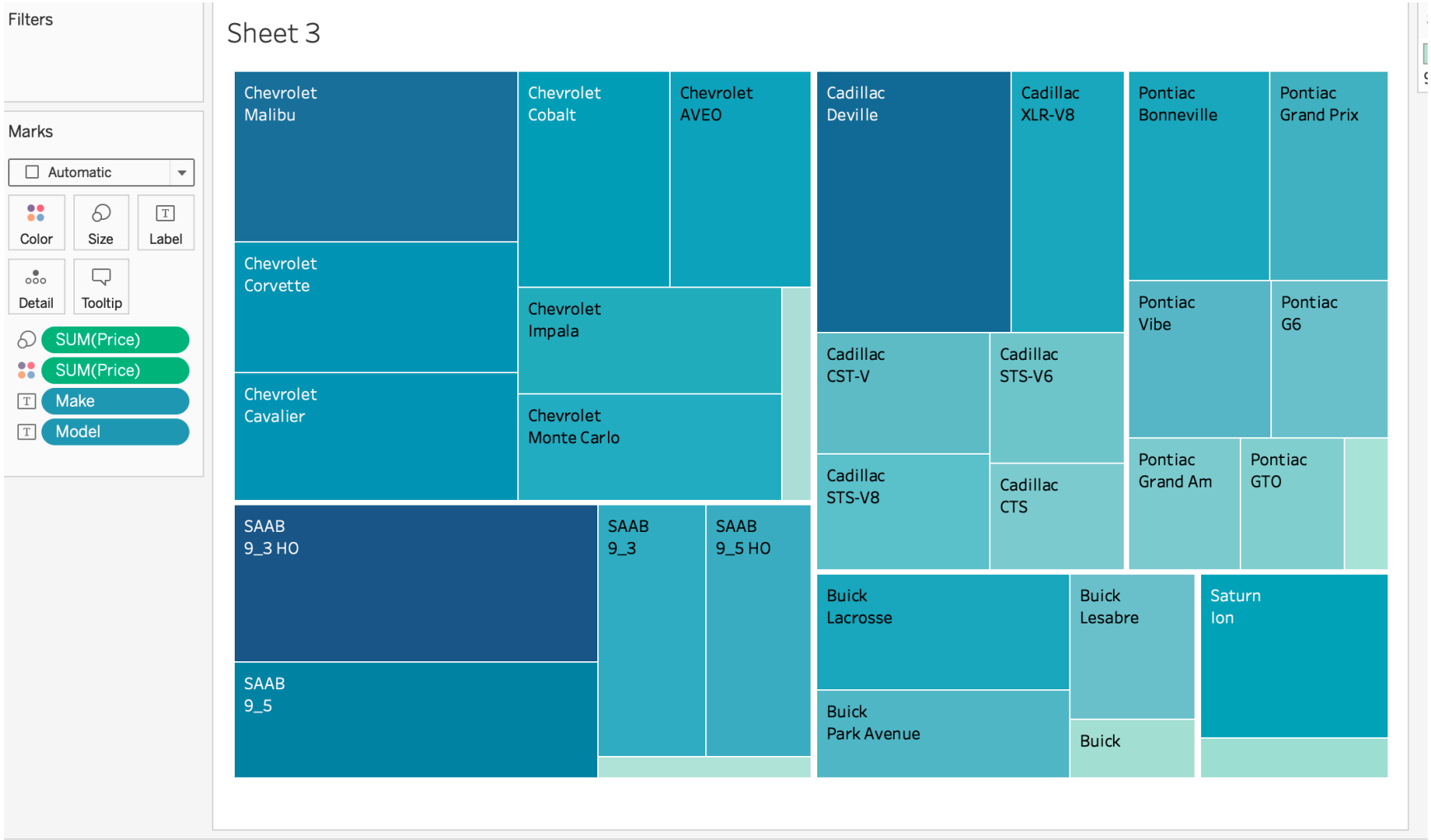
```
ggplot(data=ds, aes(x=Height.in, y=Weight.lbs, color=Sex)) + geom_point() + geom_smooth()
```

3. 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.

In this part – I assigned colors to differentiate between M and F values while I was plotting Height vs Weight. Due to this – we can see better results and identify the heights and weights from both Male and Female on one graph.

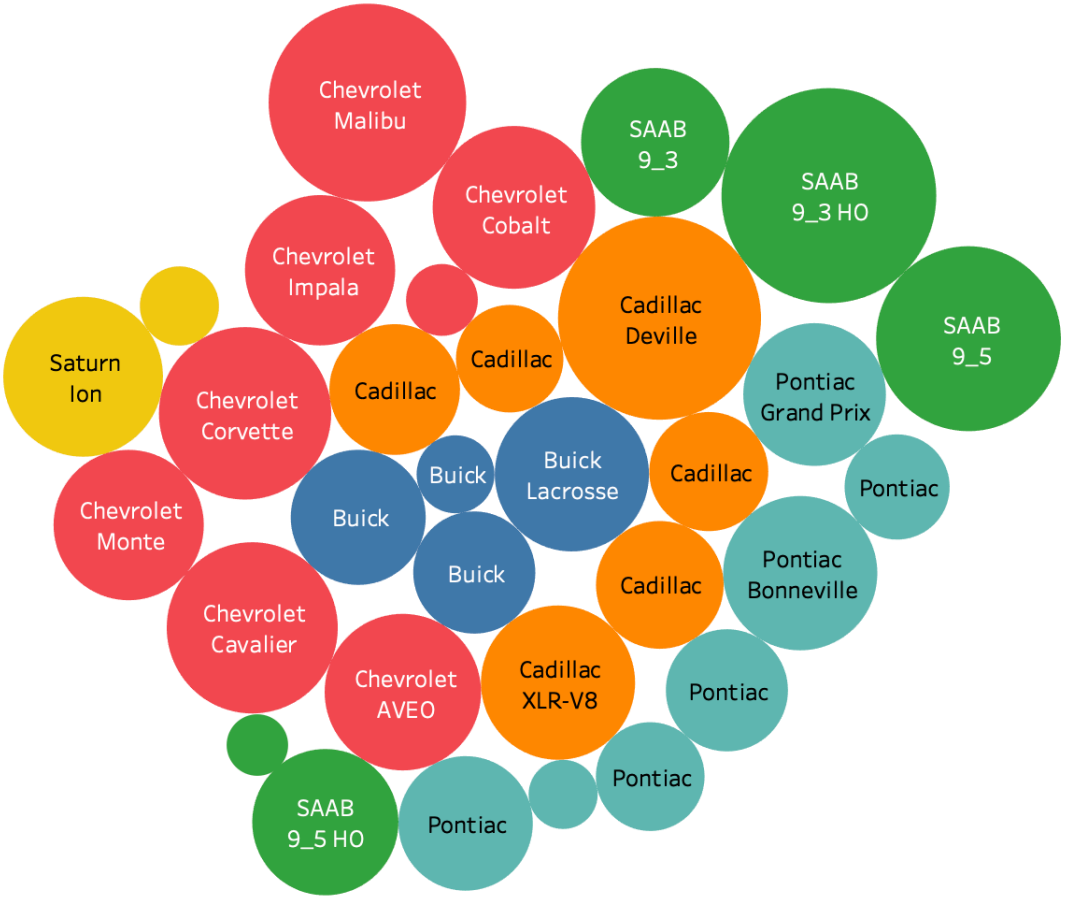
Question 5 (20 pts) Use Tableau for this question. Open the GM cars dataset included with this assignment (gmcar_price.txt). Each row represents a different car that was sold and includes information about features like the mileage and the price of sale. Create the following plots (we will look more 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.

4. 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.



5. A packed bubble chart of the same type.

Sheet 3


Chevrolet Malibu
SAAB 9_3
SAAB 9_3 HO
Chevrolet Cobalt
Chevrolet Impala
Cadillac Deville
SAAB 9_5
Saturn Ion
Chevrolet Corvette
Cadillac
Cadillac
Pontiac Grand Prix
Buick
Buick Lacrosse
Cadillac
Pontiac
Chevrolet Monte
Buick
Buick
Cadillac
Pontiac Bonneville
Chevrolet Cavalier
Chevrolet AVEO
Cadillac XLR-V8
Pontiac
Pontiac
SAAB 9_5 HO
Pontiac

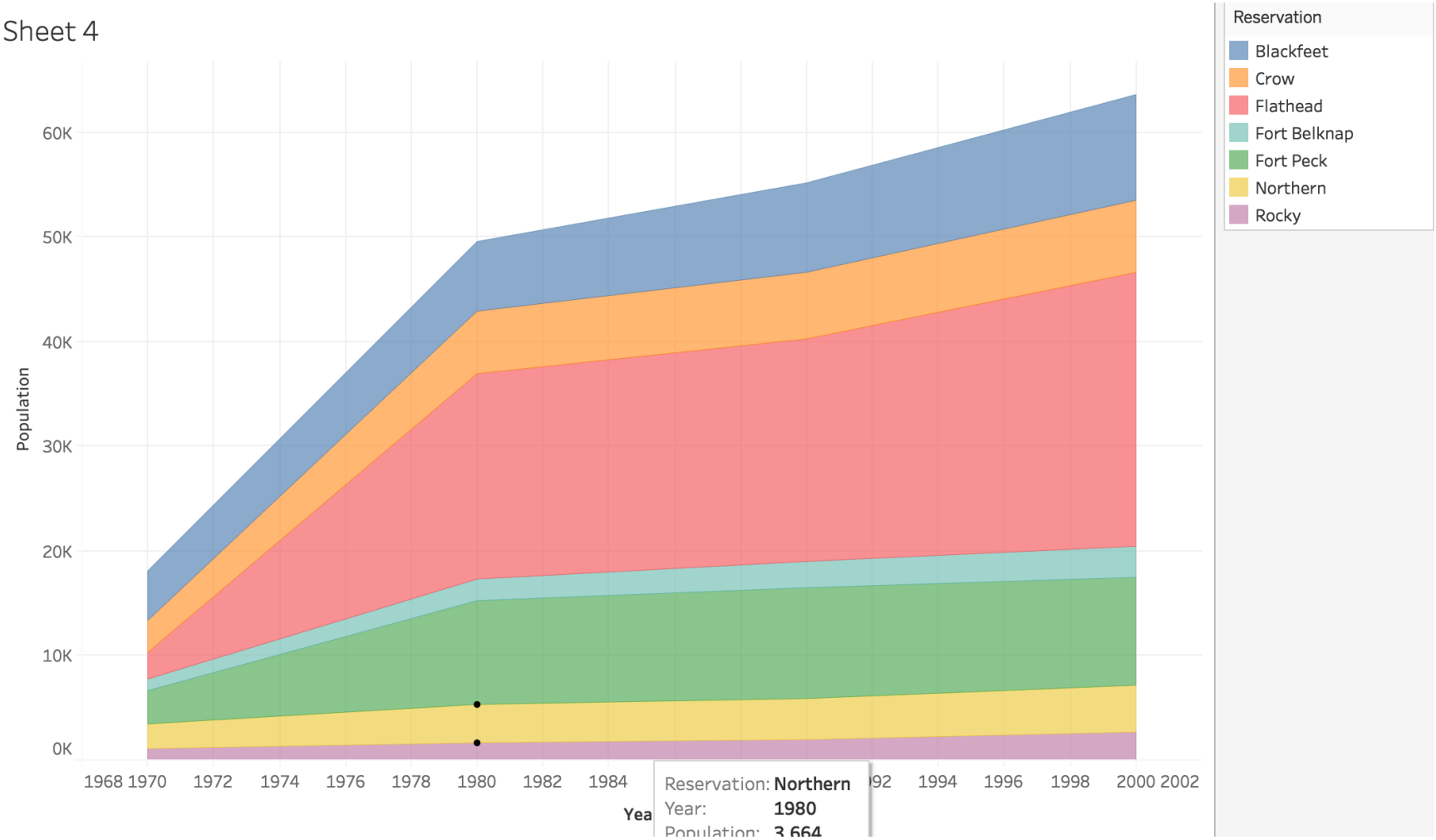
6. Write a short paragraph discussing what each plot reveals. Describe the

   differences between the two plots. Describe for each something that displayed more clearly than with the other.

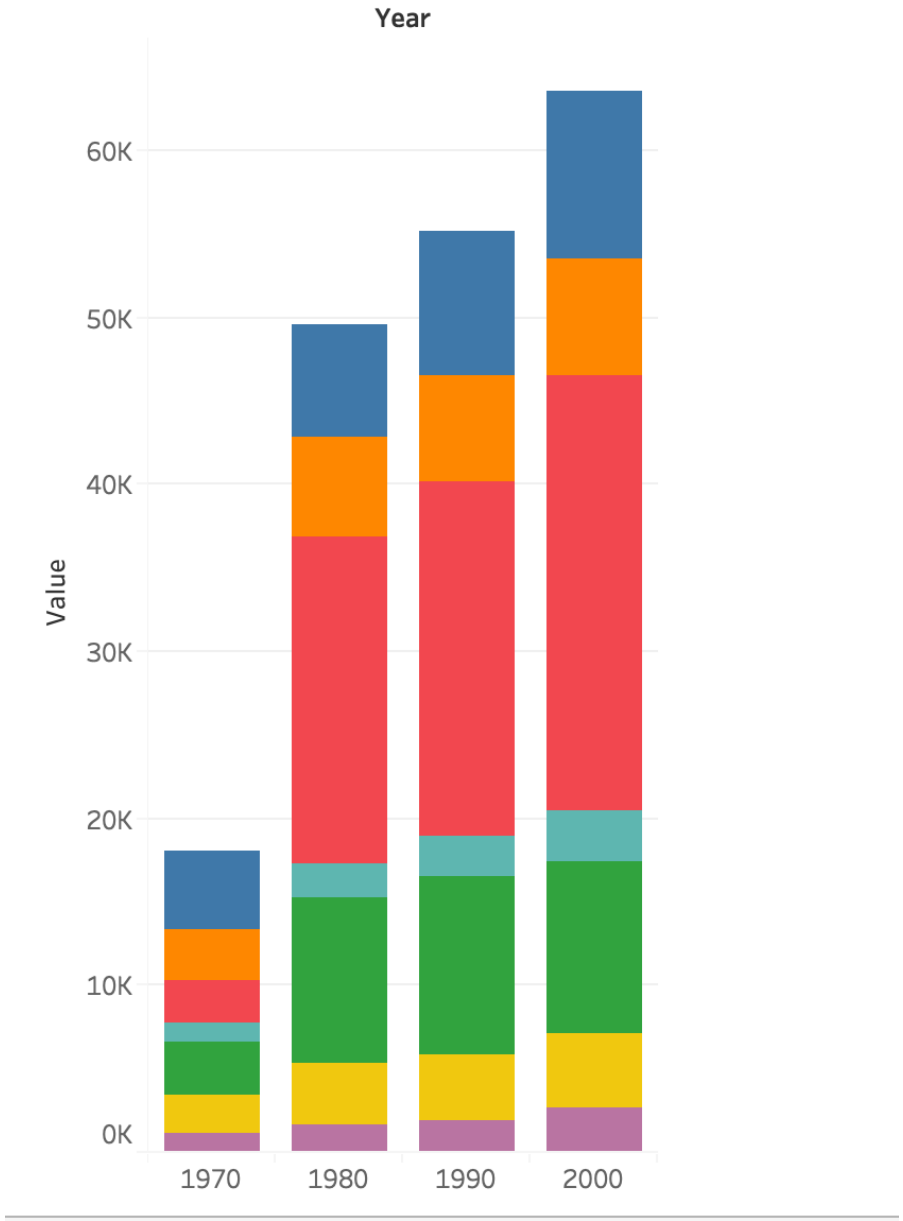
   Tree maps displayed in part 1 display hierarchical data as a set of rectangles close to each other. Each branch of the tree is given a rectangle which helps in looking at a clear picture which is further on titled with smaller rectangles that are representing subbranches. On the other hand a bubble chart displays data in a cluster of circles – in which dimensions define the individual bubbles and measure define the size and the color of the circles individually. The tree map also helps to narrow down and see the make and model of the car with the highest price – which means its easy to look at the data : this is represented in the chart as the largest rectangle with color proportions. The bubble graph doesn't gives a clear picture to narrow down the make and model of the car with the highest price – in this it is represented by the largest circle.

Question 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.**

1. 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.



2. 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.


Year
Value
60K
50K
40K
30K
20K
10K
0K
1970
1980
1990
2000

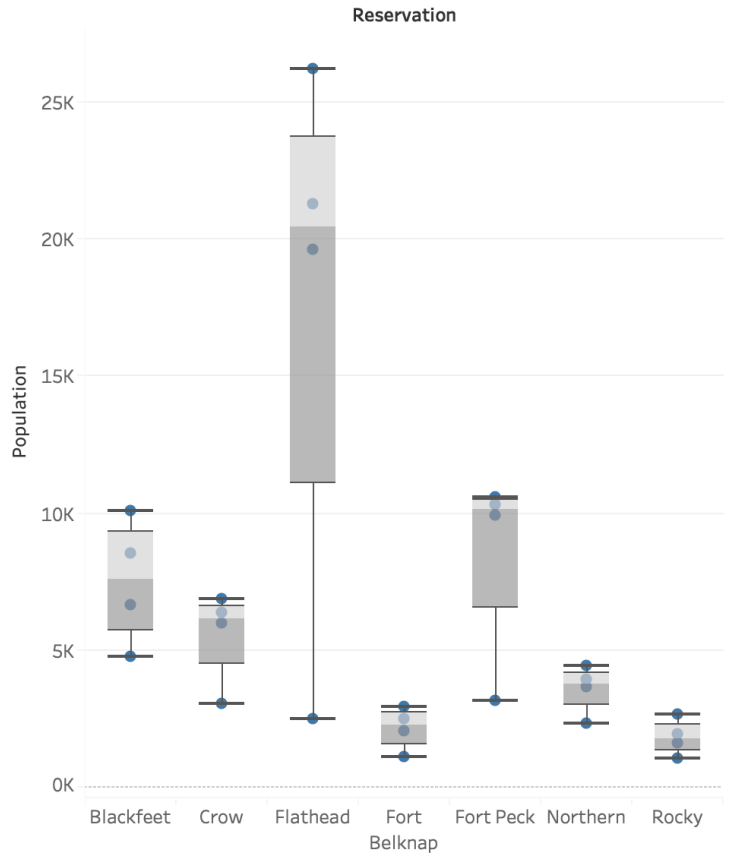
3. One that graphs the population distribution vs. years for each reservation with a box-and-whisker plot. The x-axis should be the reservations, and the y-axis should be the reservation populations. Each reservation will have four values which will be summarized by the box-and-whisker plot.

   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.

Columns | Reservation

Rows | Population

Sheet 5



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

Question 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?

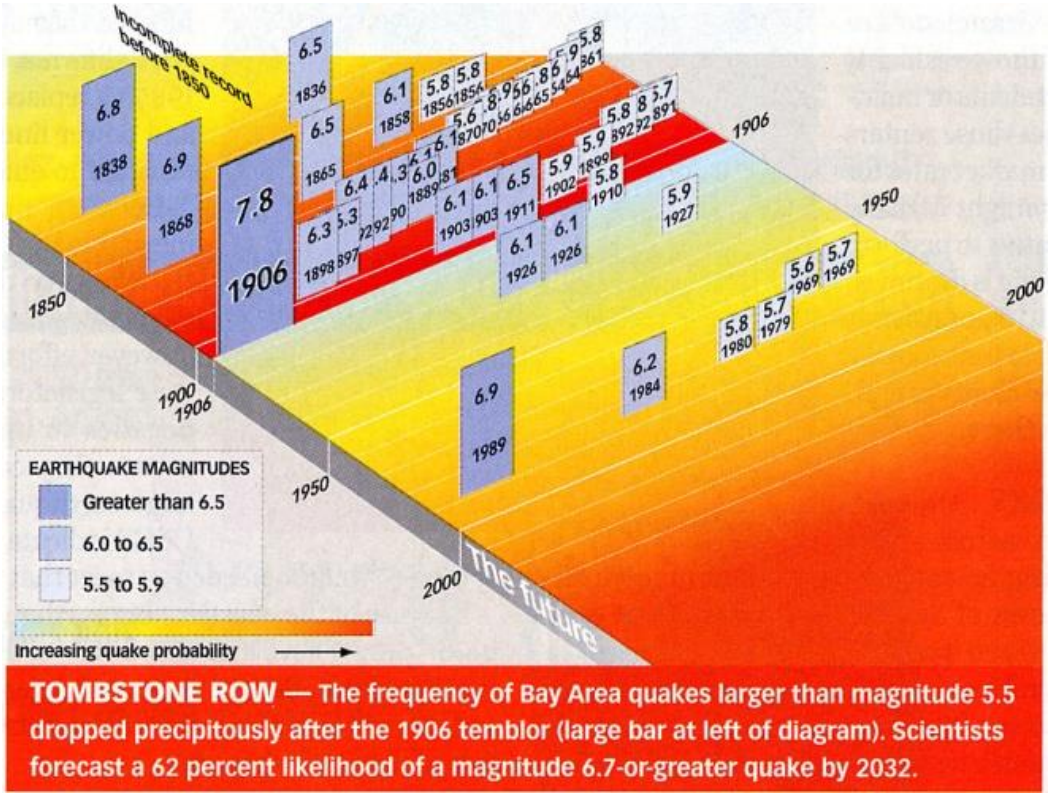
The few things that I see wrong with the image below are – firstly that the scale does not seem very accurate. It is hard to compare 7.8 Earthquake magnitude with 6.9 Earthquake magnitude. The heat map doesn't explain well and is not clear of what exactly are we looking at.
Furthermore, I also think that there is a visibility issue – there are some values on the heat map that are not ridable. It is hard to understand the map like this if you are trying to understand it yourself or explain it to somebody.
Lastly, I think the scale at the bottom of the Heath Map which shows the incoming quake probability is not very clear to explain what is happening. It has to be more clear, with more visual – which also relates to having the visibility clear and the scale.

The scale is wrong

7.8 – 6.9 compare

Visibility issue

The heat map – doesn't explains well



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.