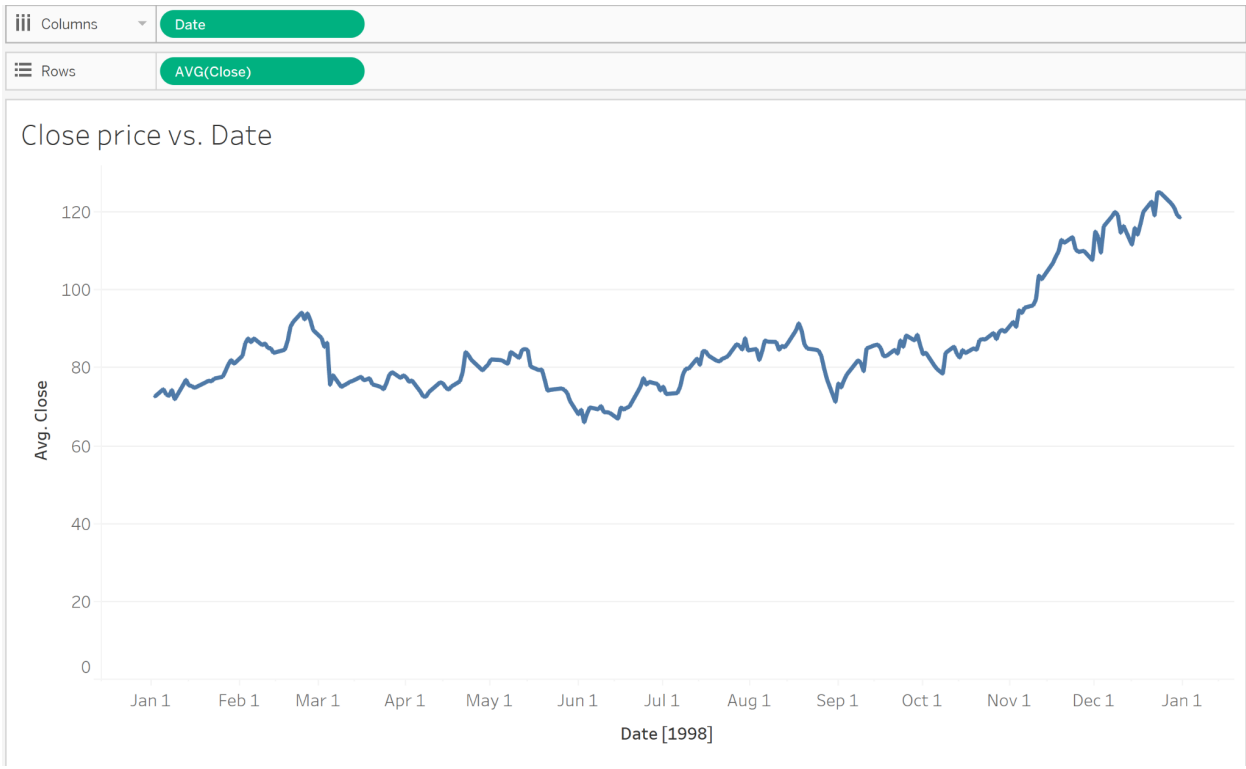
Elijah Caluya
DSC 465
4/10/2021

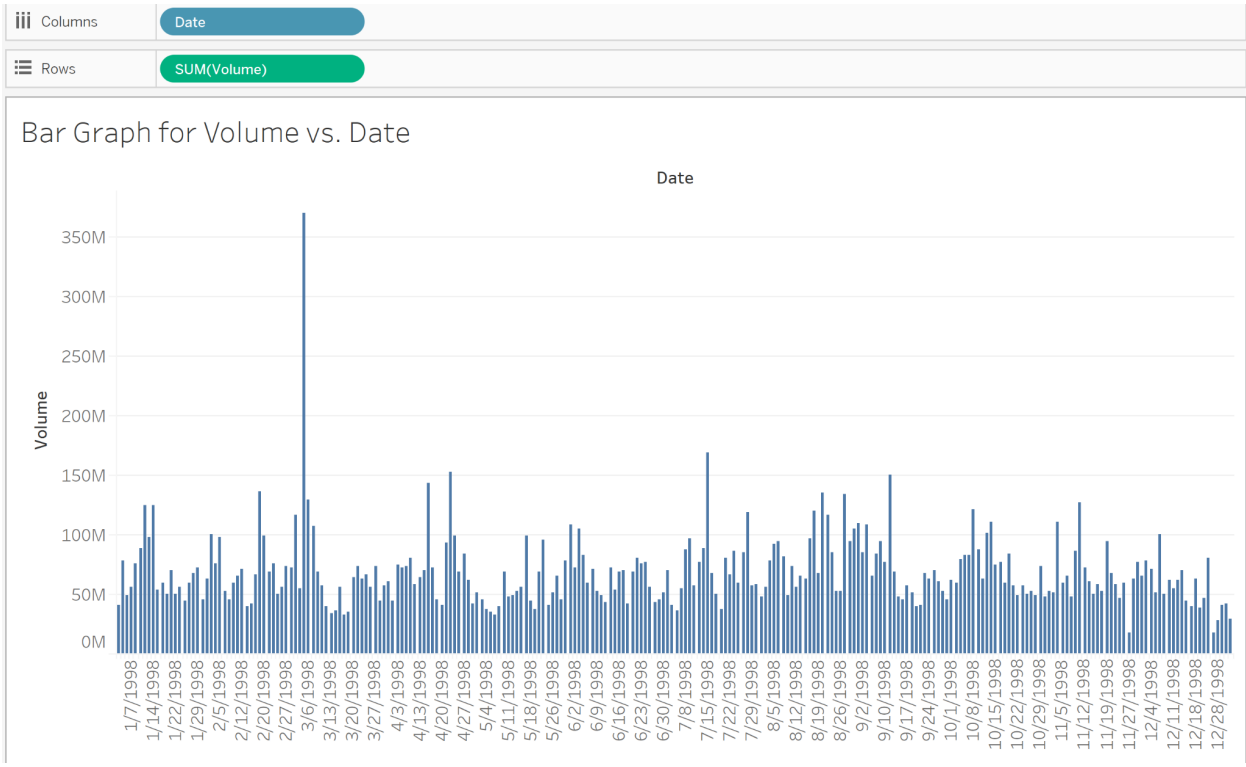
Homework 1

1) Questions using the Intel-1998.csv data (Only C is in R, the other questions are in Tableau)

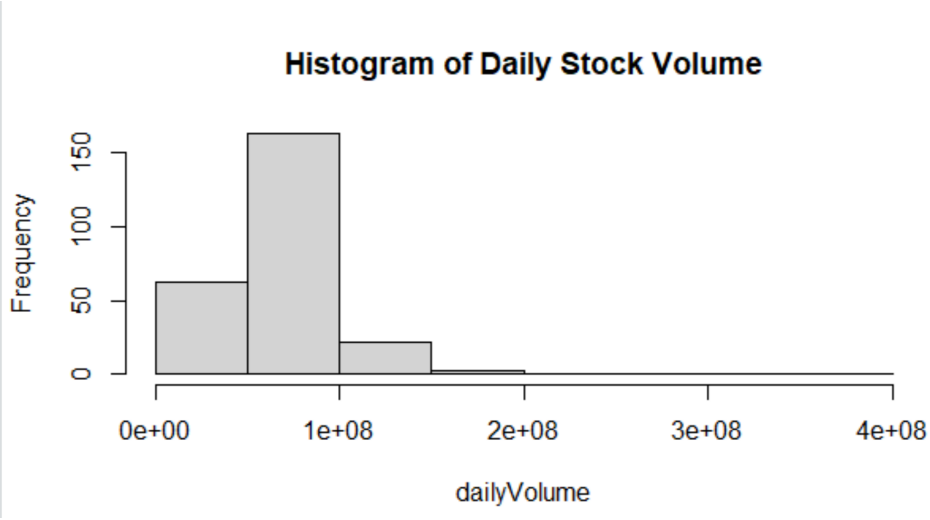
   a) Line graph for Close price vs Date using Tableau



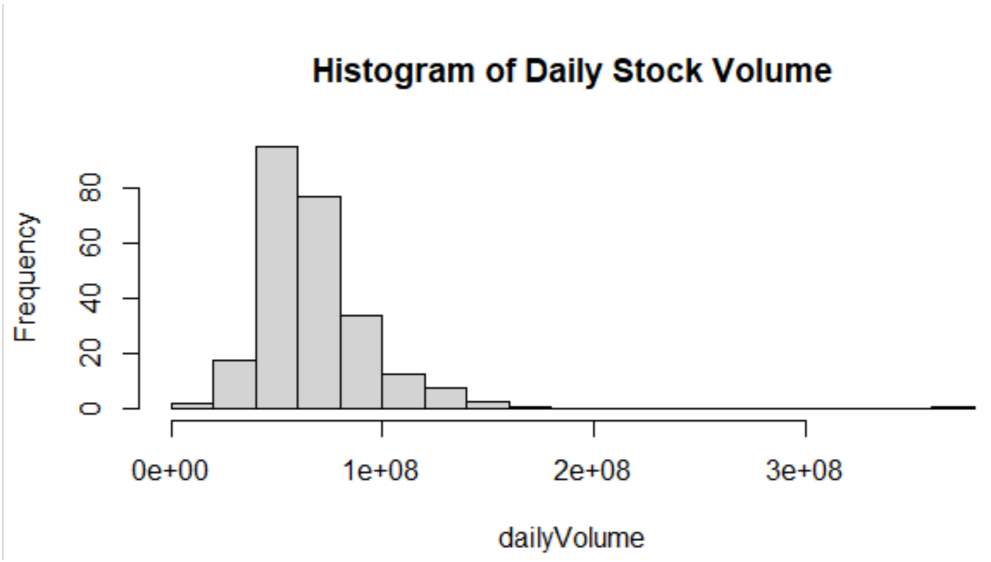
b) Bar graph for Volume vs. Date using Tableau



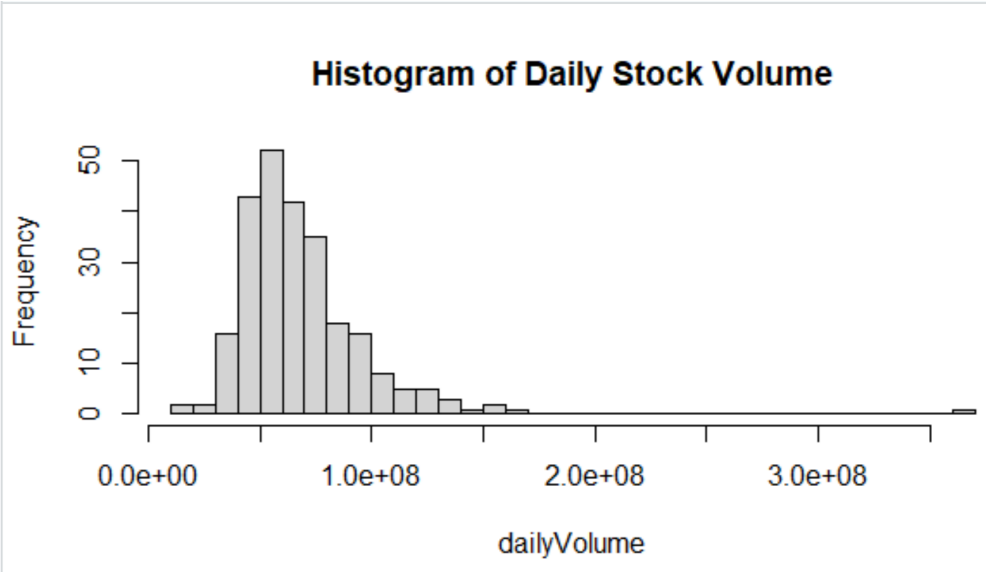
c) Histogram of the daily stock Volume (Used R for this problem)
First Histogram using the following R command:
dailyVolume <- Intel.1998$Volume
hist(dailyVolume, main="Histogram of Daily Stock Volume)



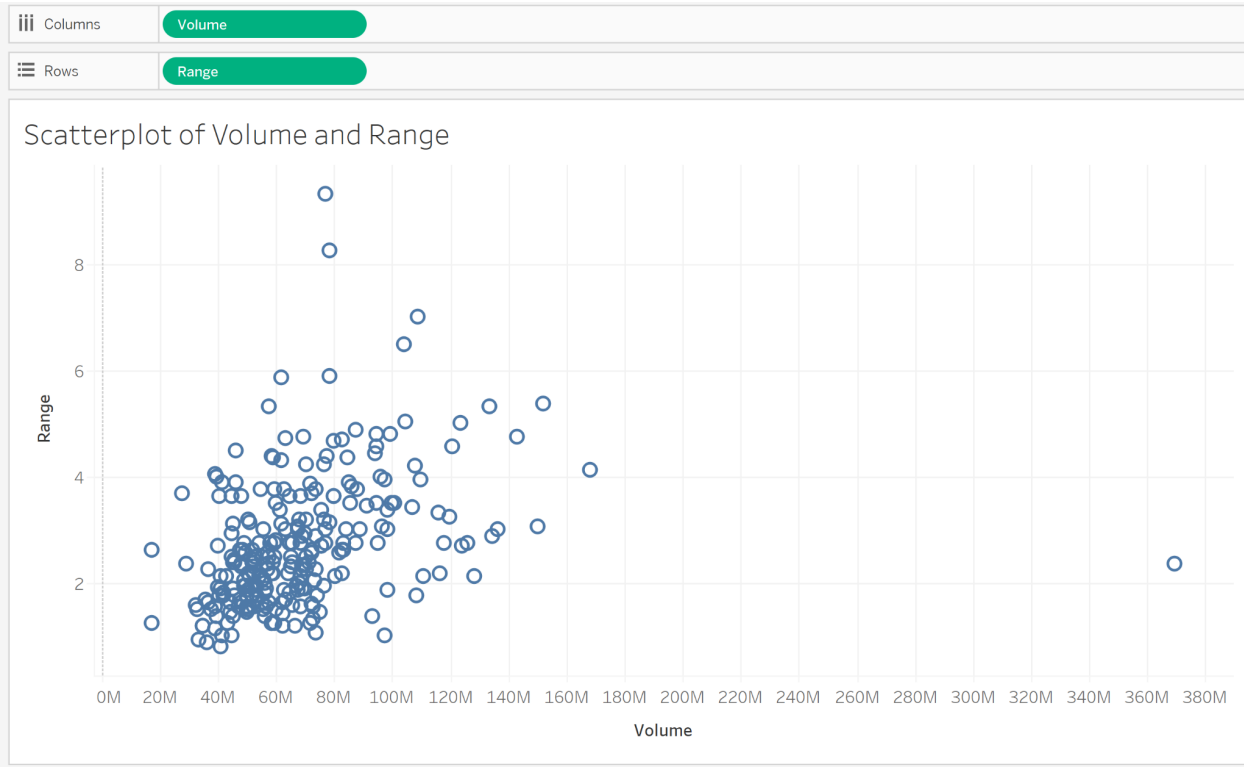
Changing bin size to 20 with the following R command:
hist(dailyVolume, breaks=20, main="Histogram of Daily Stock Volume")



Changing bin size to 40 with the following R command:
hist(dailyVolume, breaks=40, main="Histogram of Daily Stock Volume")

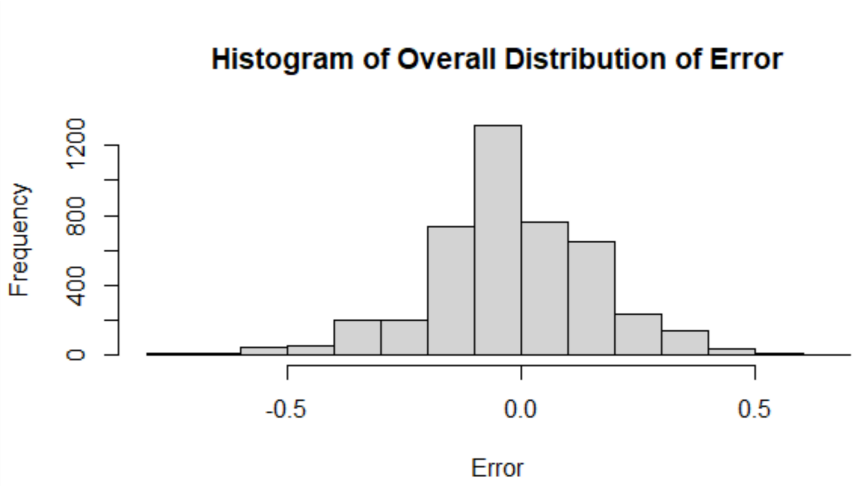


d) Scatterplot with Volume and daily price range

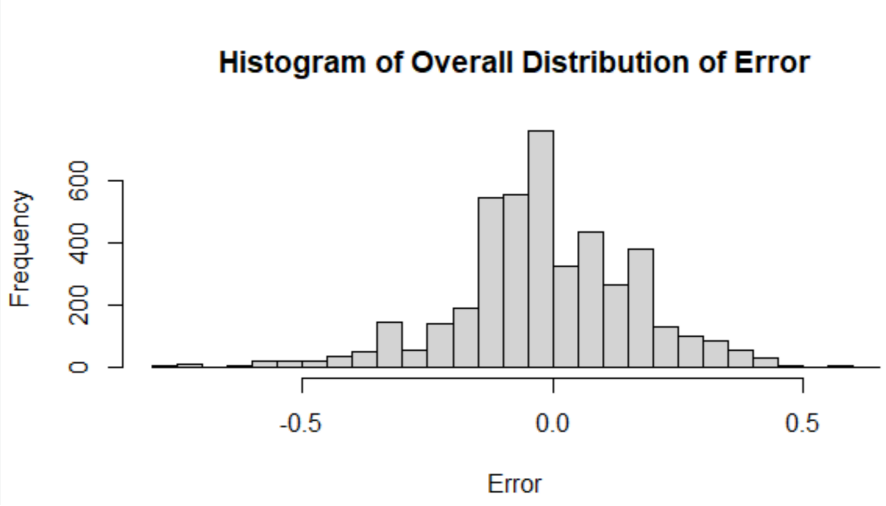


2) Questions using the PerceptionExperiment.csv data (Only question A uses R, the other questions are through Tableau)

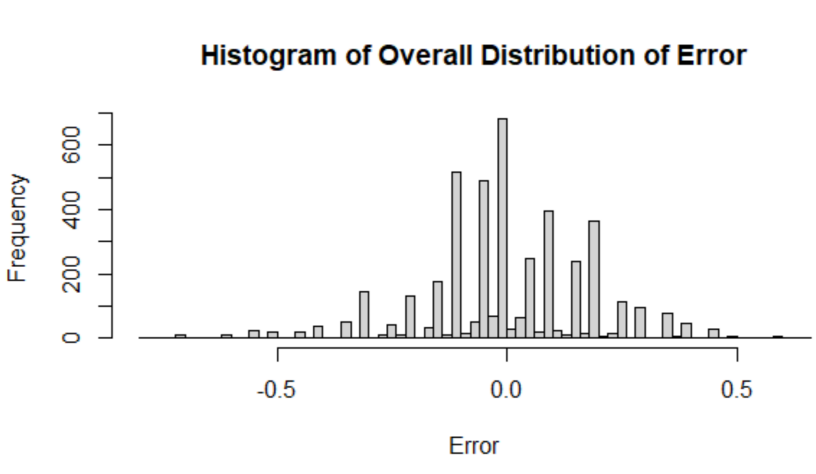
   a) Histogram of Overall Distribution of Error (Using R commands)
   First Histogram is created through the following R commands:
   Error <- PerceptionExperiment1$Response - PerceptionExperiment1$TrueValue
   hist(Error, main="Histogram of Overall Distribution of Error")



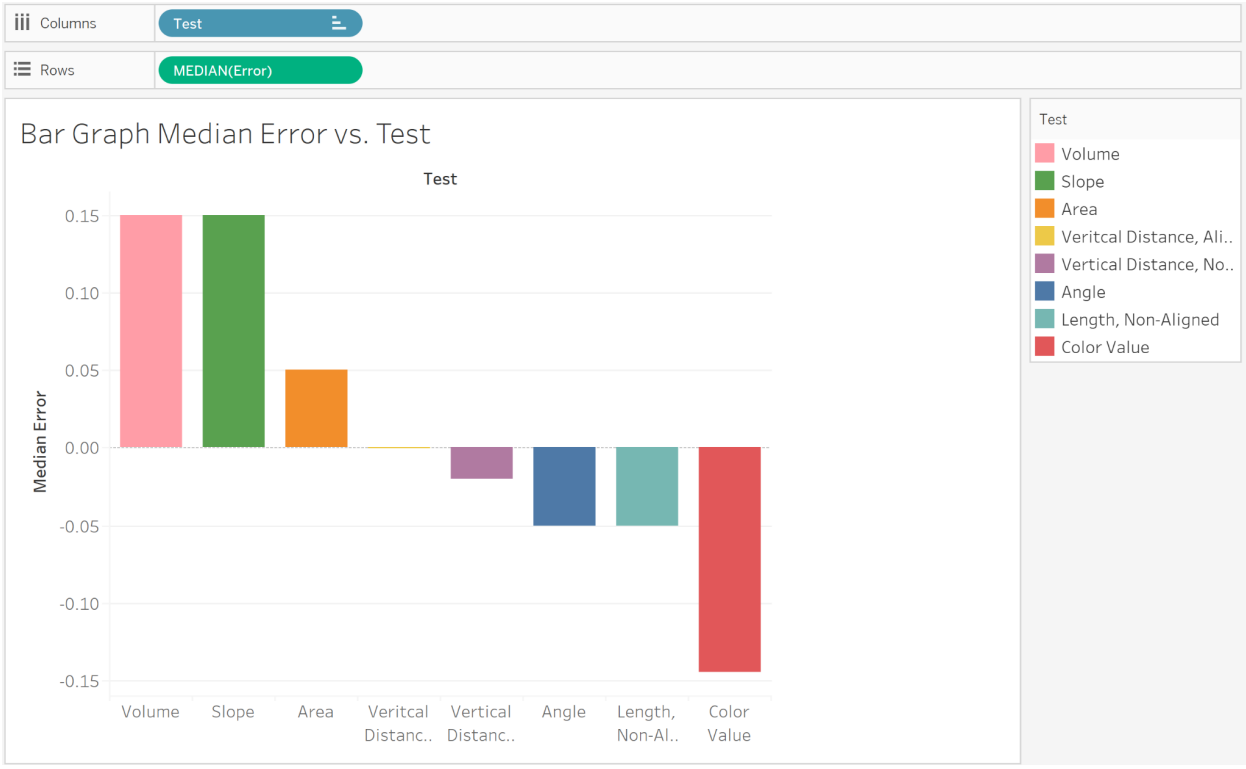
Changing bin size to 40 with the following R commands:
hist(Error, breaks=40, main="Histogram of Overall Distribution of Error")



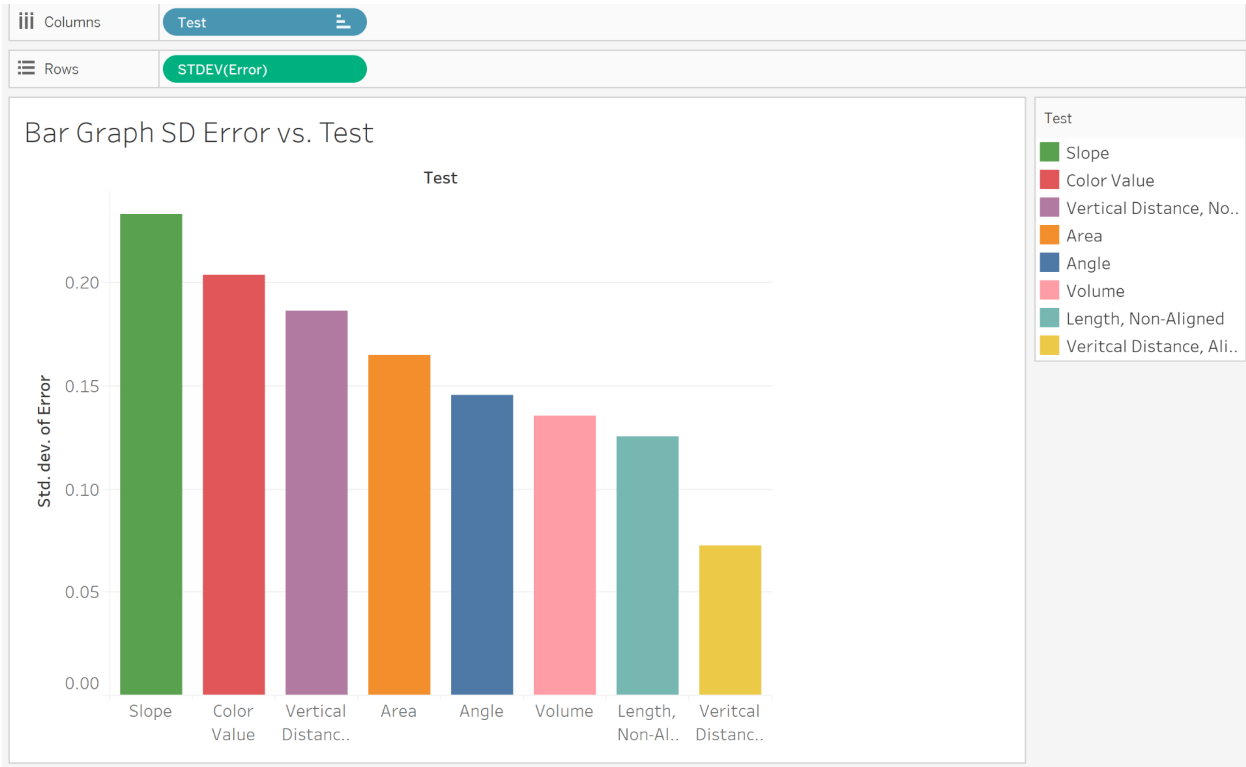
Changing bin size to 60 with the following R commands:
hist(Error, breaks=60, main="Histogram of Overall Distribution of Error")



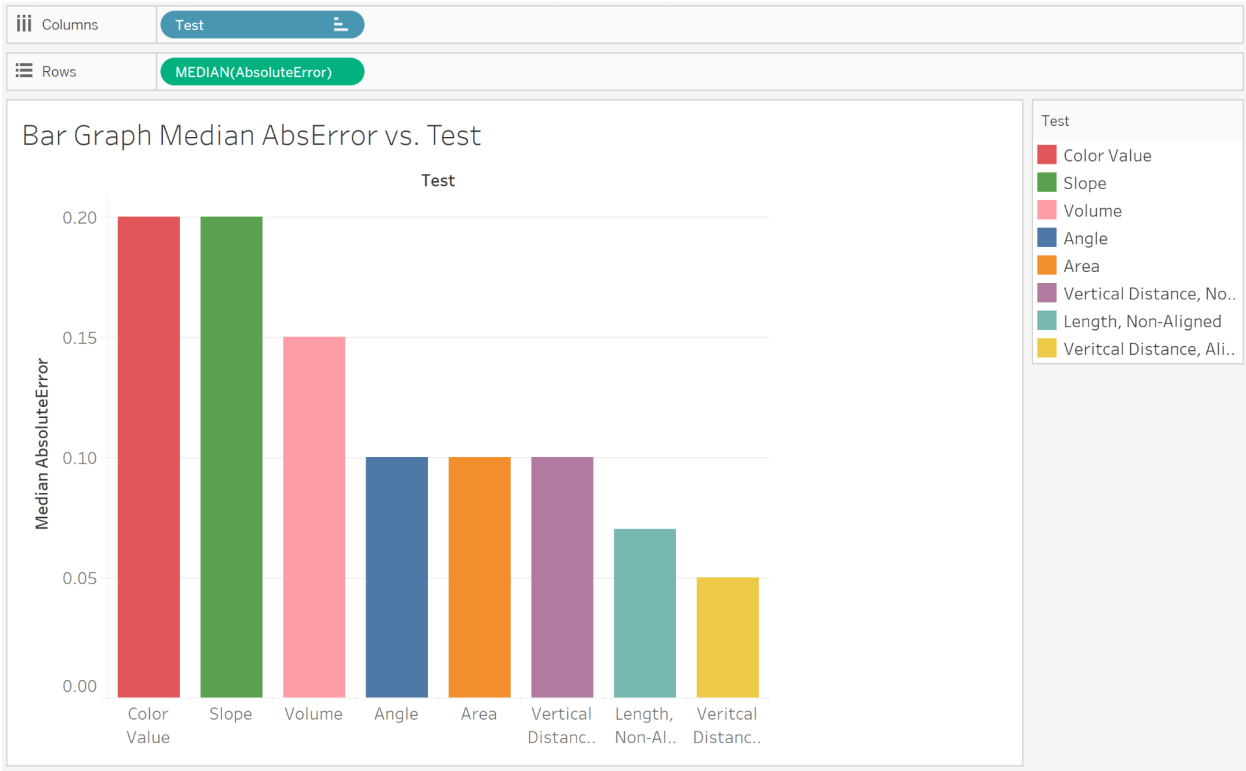
b) Bar Graph of Median Error by Test



c) Bar Graph for Standard Deviation of the Error by Test



d) Bar Graph for Median Absolute Error by Test



e) The first graph in part A describes a histogram of the overall distribution of error and we can see that it has a fairly normal distribution. We can also say that the distribution is left skewed as there are slightly more to the right side of the distribution than the left.

The second graph for part B shows the medians for the tests Volume, Slope, and Area being positive while Vertical Distance Non-aligned, angle, Length Non-Aligned, and Color value have negative medians. The only test that has a median of 0 is Vertical Distance Aligned. This tells us that most of the tests have a negative error as 4 tests were negative. However, the values of the positive errors are greater than the negative ones, so in the data, there are more positive error values.

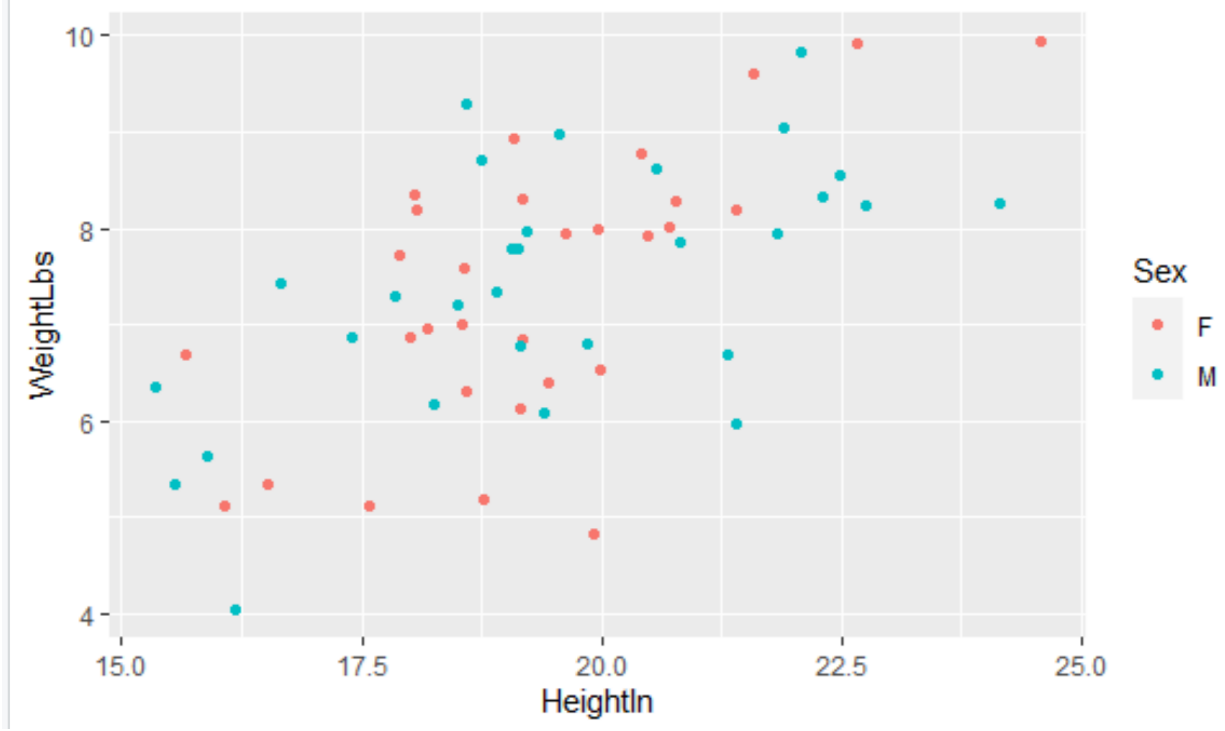
The third graph for part C is a bar graph that is shown in descending order for the Standard Deviation between Error and Test. From the graph, we can see that Slope has the highest standard deviation while Vertical Distance Aligned has the smallest. This means that most of the values for Vertical Distance Aligned are not very far away from the mean.

The fourth graph for part D describes the medians of Absolute Error vs. Test. In this graph, there are no negative values as it is the ABsolute Error. We can see from the graph now Color Value and Slope have the same median where in the

regular error graph color value was negative. It is also interesting to see Angle, Area, and Vertical Distance Non-Aligned have the same value where in the regular error graph they were different.

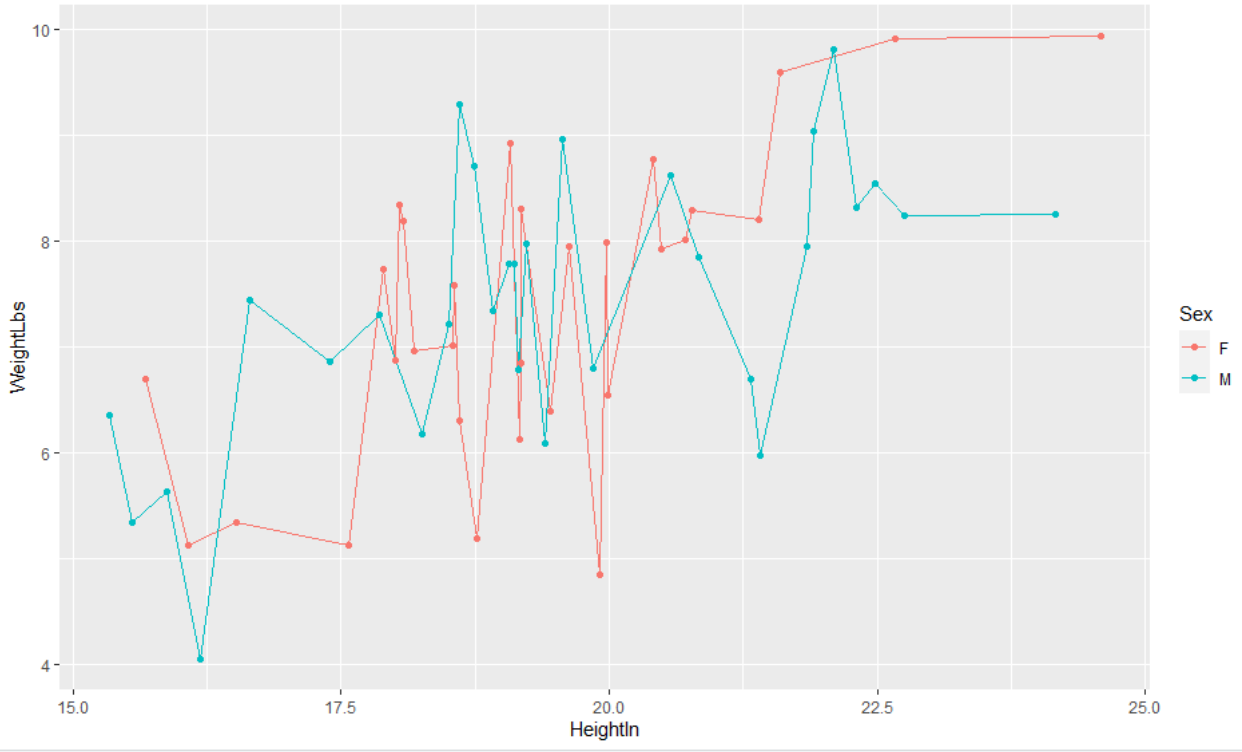
3) Questions using InfantData.csv. In order to load in the data into RStudio, I used the following command:
InfantData1 <-
read.csv("C:/Users/elica/Desktop/depaul_grad/DSC_465/InfantData1.csv")

   a) Scatterplot for HeightIn and WeightLbs with Sex differentiated.
   Used the following R command:
   ggplot(data=InfantData1, aes(x=HeightIn, y=WeightLbs, color=Sex)) +
   geom_point()



   b) Graph with adjusted weights for separate trend lines
   Used the following R command:
   ggplot(data=InfantData1, aes(x=HeightIn, y=WeightLbs, color=Sex)) +

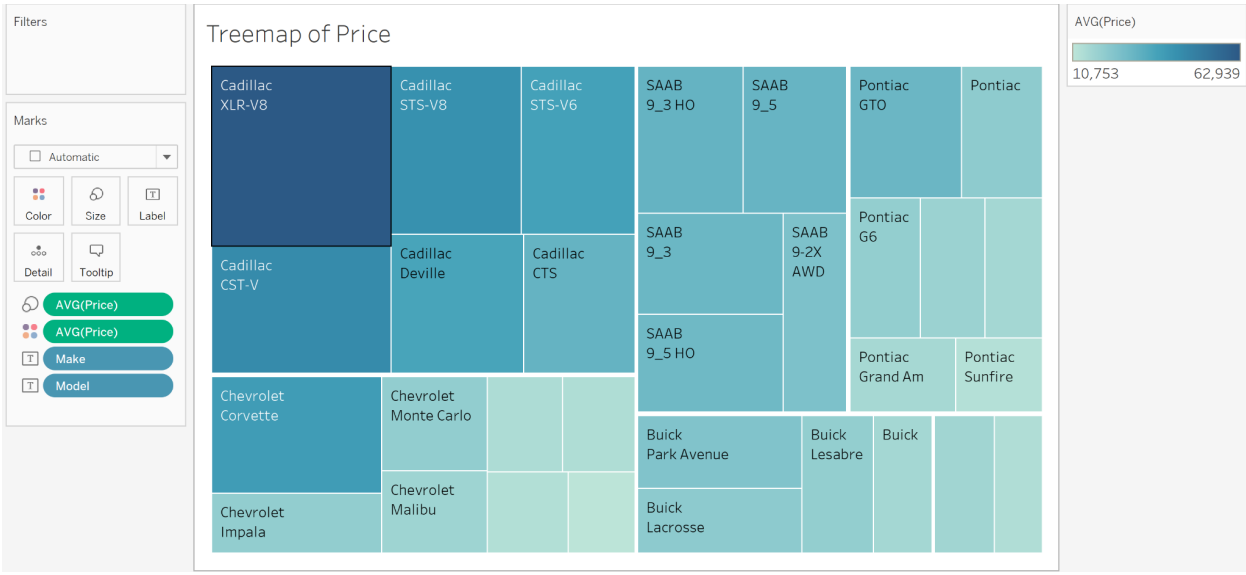
geom_line() + geom_point()



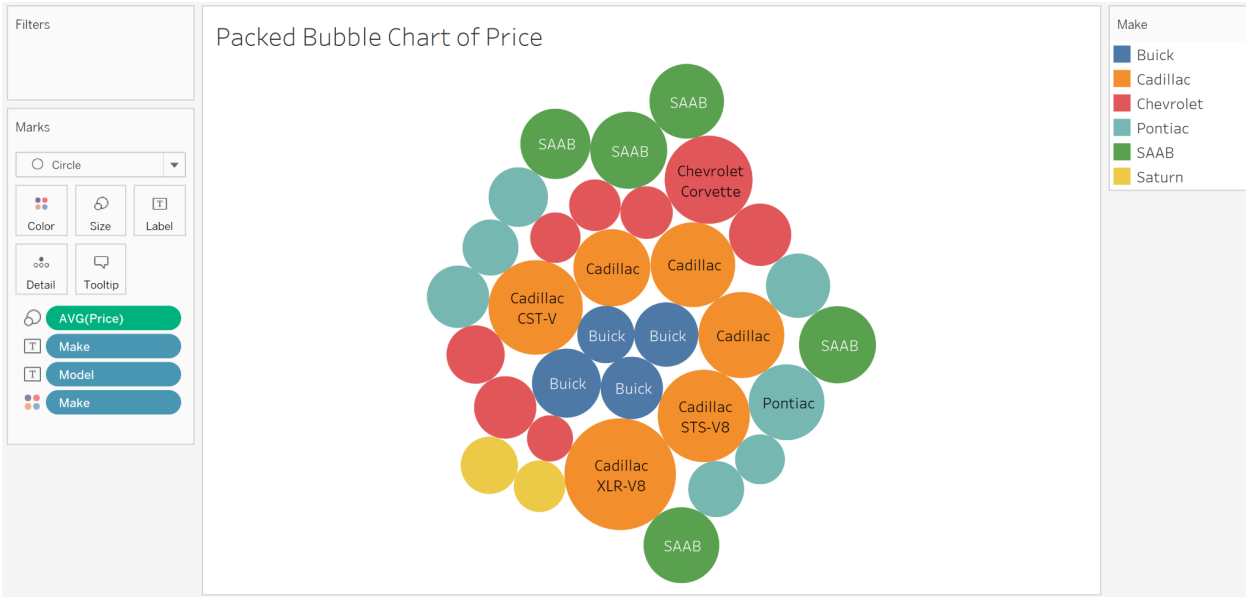
c) For the graphs I wanted to make sure that it was easy to differentiate between F and M by making the points and lines different colors. By making the lines different colors, it is much easier to determine which tren belongs to a specific sex. Also, by having the trend lines, it is easier to follow the patterns of the points and see that the height increases as the weight increases.

4) Questions using gmcar_price.txt

a) Treemap based on Price and main subdivision of Make and minor subdivision of Model.



b) Packed bubble chart based on Price and main subdivision of Make and minor subdivision of Model.

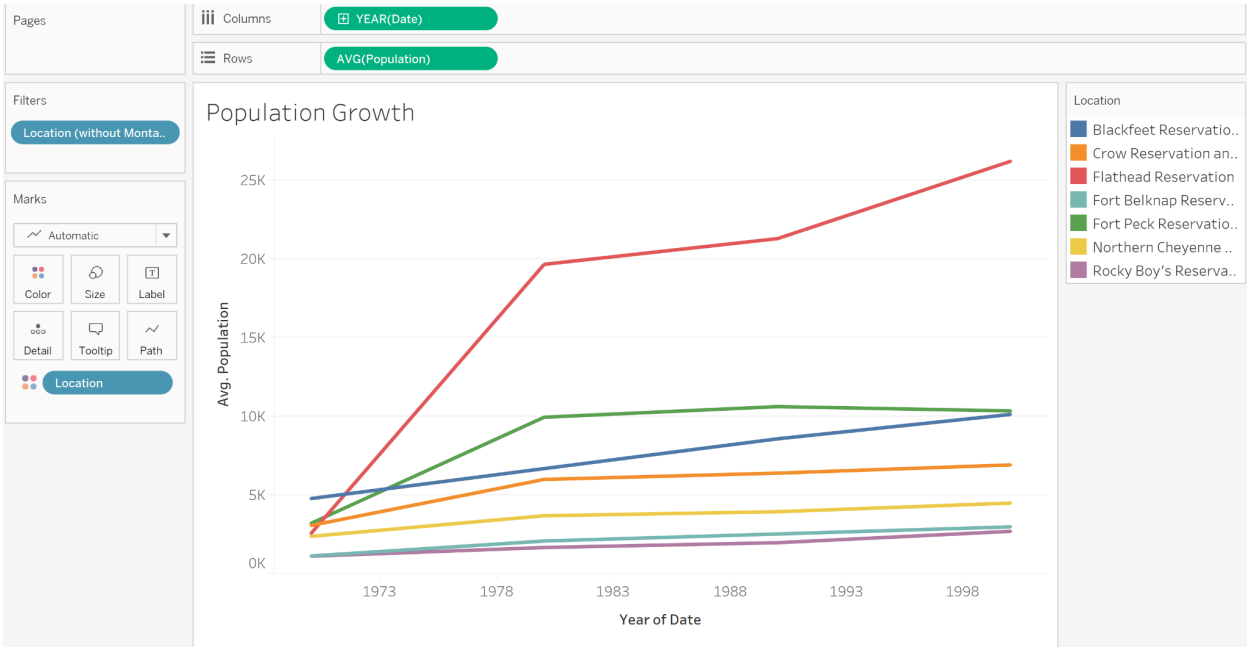


c) Between the charts made above, I found that the Treemap was a little easier to read. This is because the difference in color represents the price where the darker the color, the more expensive the price. You can also see that the more expensive makes and models are closer to the top left of the chart as well as have a larger square. On the bubble chart, the price is determined by the size of the bubble and I found that it was harder to distinguish the sizes between some circles so I could not tell which ones were more expensive than the others. The bubble chart on the other hand, is very good at distinguishing between the different makes because they are color coordinated. The treemap separates the

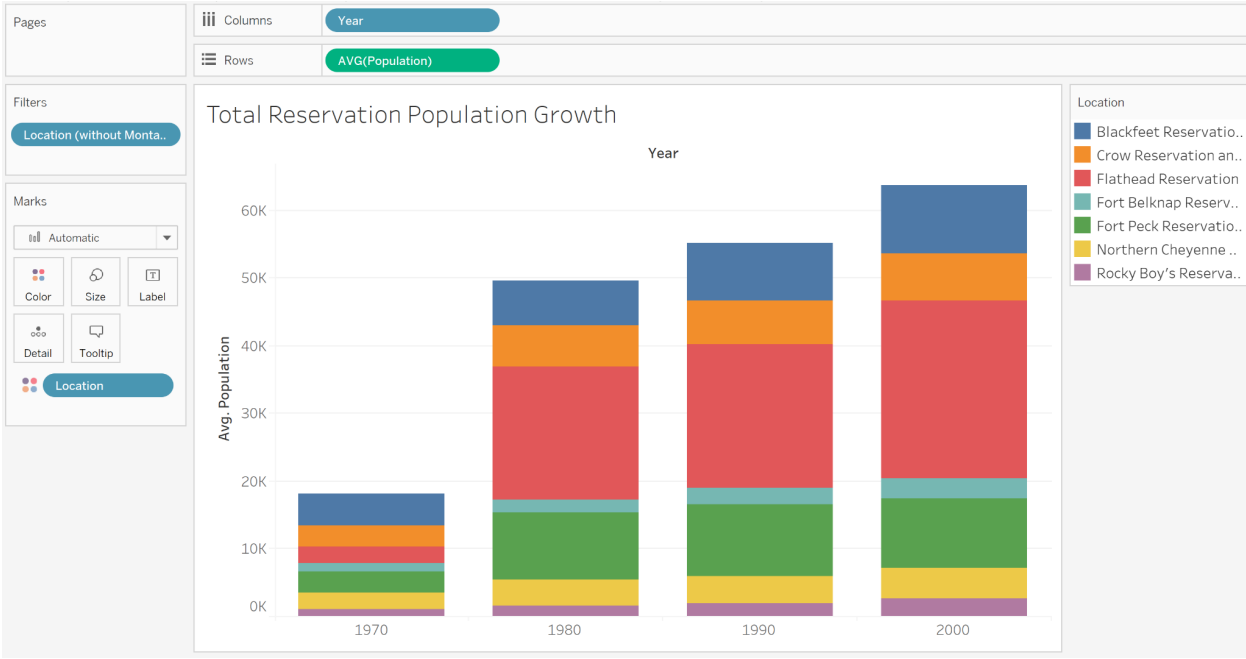
brands visually by having thicker lines between groups but it is harder to distinguish them as all of the squares are some shade of blue and tightly packed. In both of the charts we can see that Cadillac has the highest average price and Saturn has the lowest average price but on the Treemap it is more apparent than the bubble chart.

5) Questions using reservation70-00.xlsx data. In order to exclude the row with 'Montana', I had to create a calculated field that had 'Exclude' if the location was Montana 'Include' if it wasn't. I then used this field as the filter and checked the 'Include' box so that only locations that were not Montana would be present in the graphs.

a) Chart to show Population growth over the years.



b) Total Reservation Population Growth subdivided among the different reservations.



c) Population Distribution for each Reservation in Box-and-Whiskers plot.