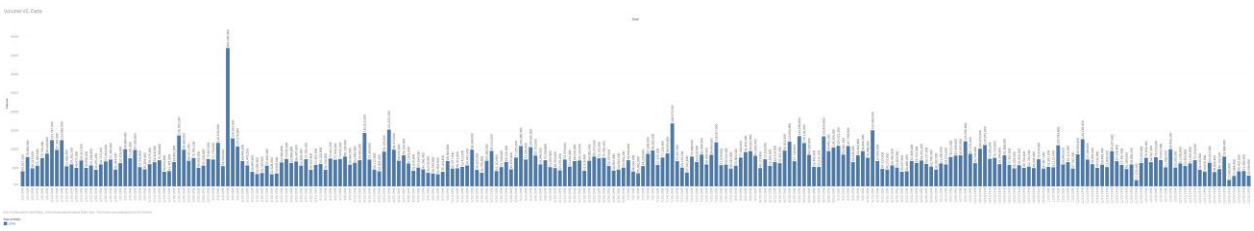
1) (15 pts) For this problem, we'll look at data about Intel stock (Intel-1998 dataset from the website). The data covers stock market trading for the Intel corporation in 1998. Each row is a day, with the following columns: Date, Trading Day (integer day number, including skips), Open (price at market 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).

a. Graph the closing price vs. the date with an ordinary 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".

I have Date as X-axis and Closing price as Y-axis. As market summary in real-life website (like Google or Yahoo Finance) aggregates individual date when one-year summary is chosen, I made X-axis continuous to display months instead of individual days.

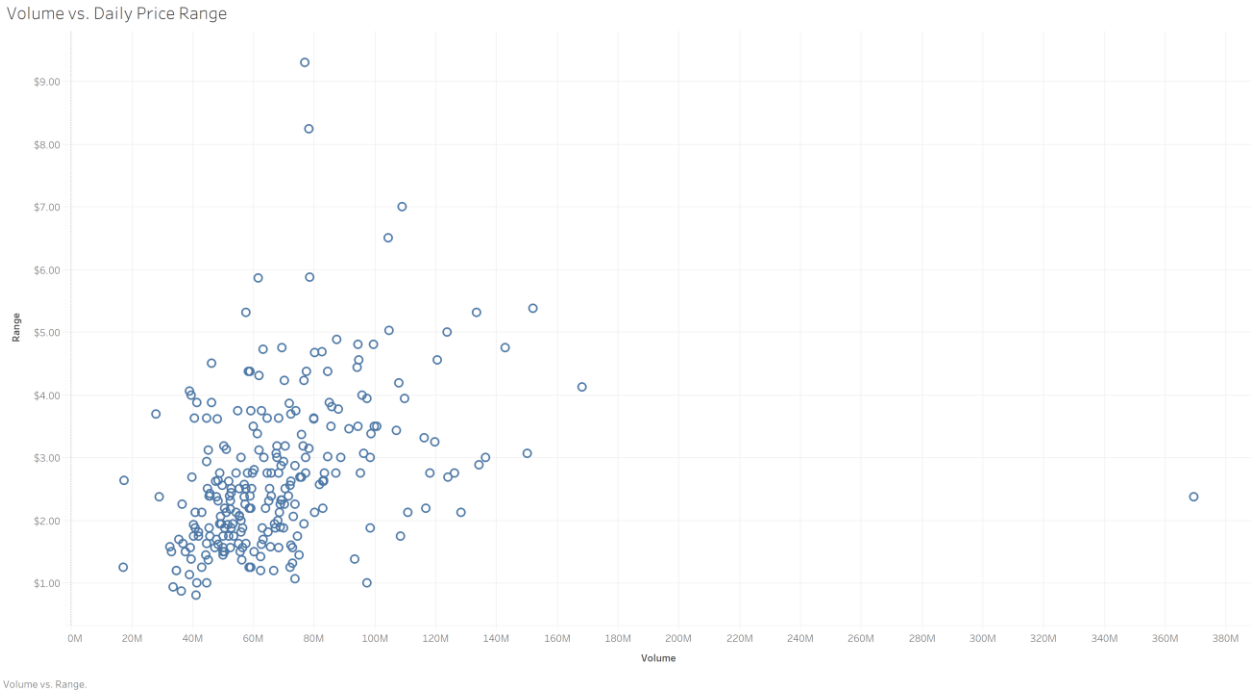
b. Graph the Volume vs. the exact Date as in the last part with a bar graph.

For volume, I simply set date as X-axis (discrete) and volume as Y-axis (continuous). The graph shows volume of every single trading day in 1998.



c. 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. 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".

I followed instruction to create a new column – "range" in Tableau, then set Volume on the x-axis and range on the y-axis. Everything keeps default from Tableau.

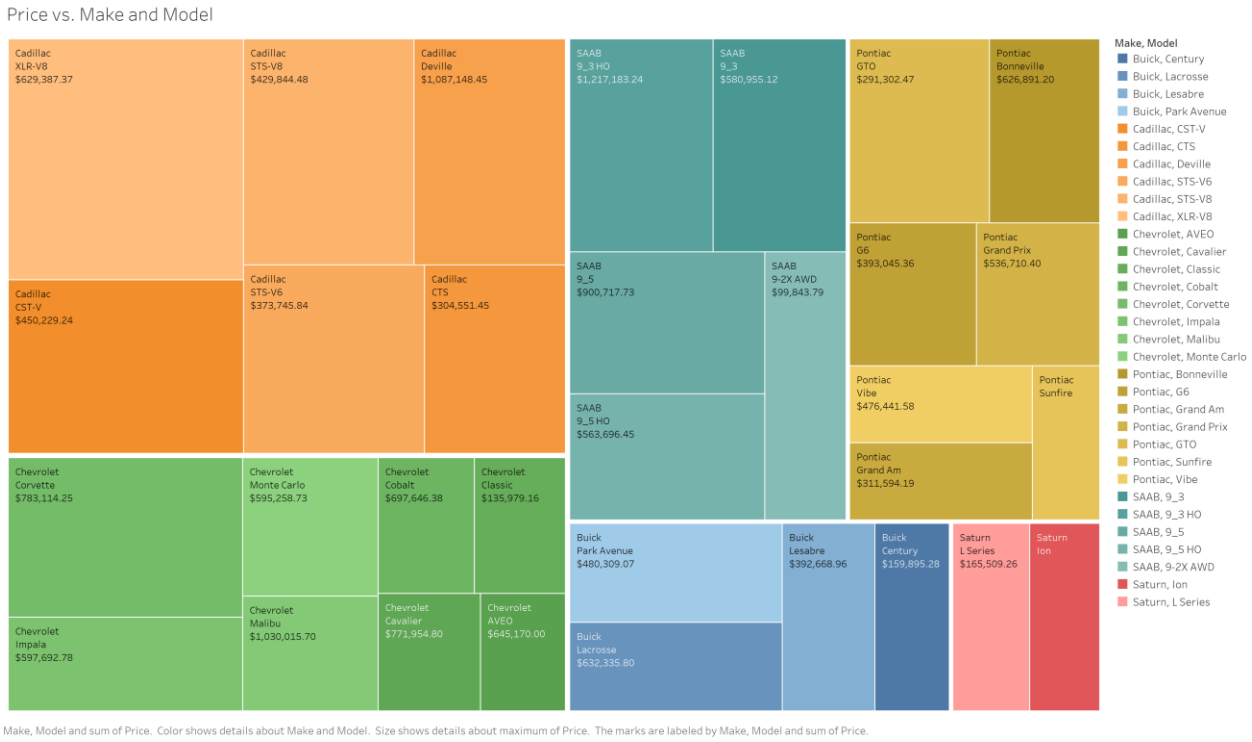


Volume vs. Range.

2) (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. Hint: use the "Show Me" menu.

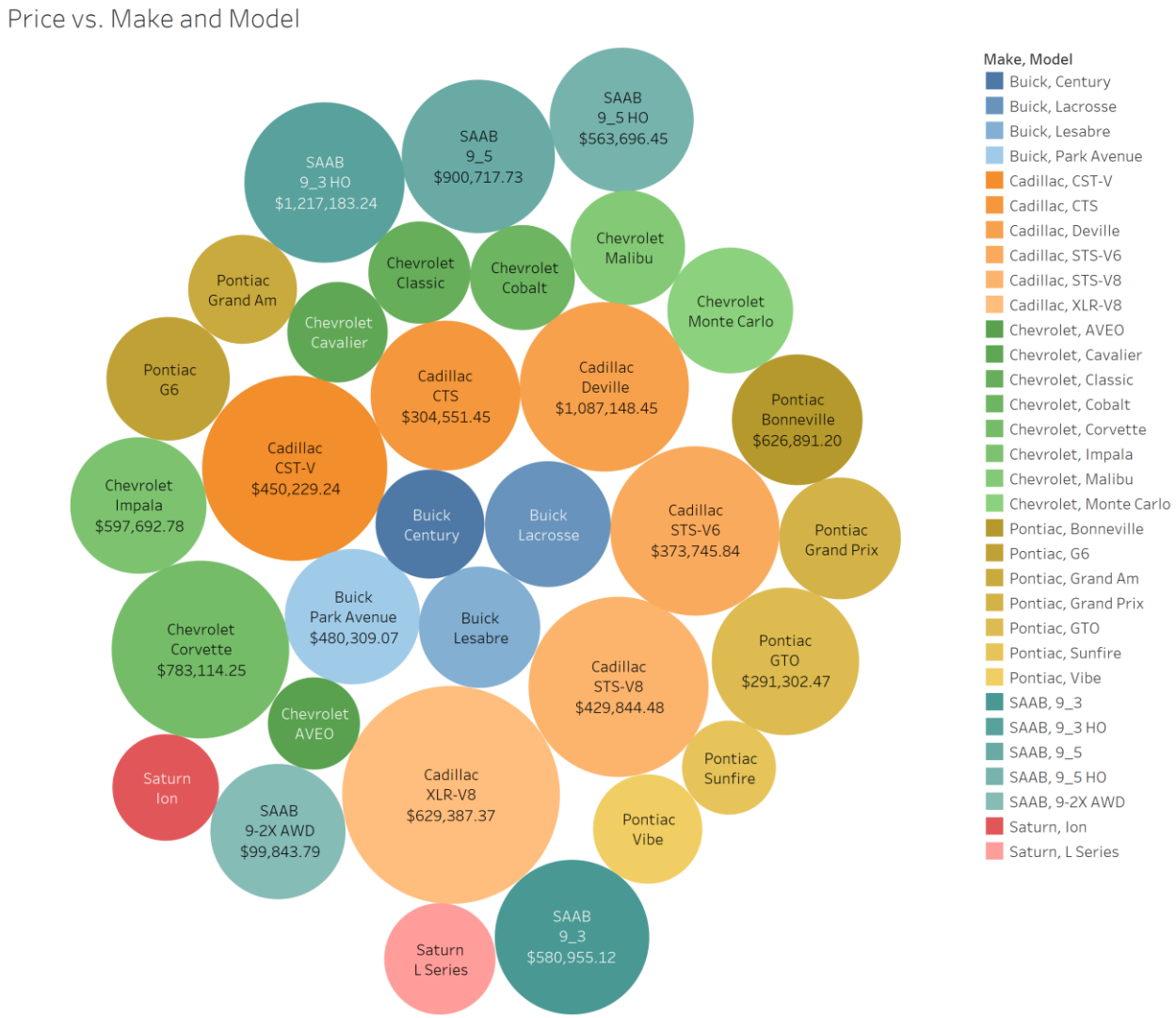
a. A treemap based on Price with a main subdivision for the Make of the car and a minor subdivision based on the Model. Because each row of the data file represents a single car but each box in the treemap represents all the cars with a given make and model, pay very close attention to what kind of aggregation is being used.

I first categorized data based on vehicle's Make, each color represents one Make. Then I used different shades under the color to represent various Model, and the size of the rectangle is determined by price. I also labeled each rectangle/square with Make, Model and Price.



b. A packed bubble chart of the same type.

I did same thing as the first question. Use different color to represent different Make, then display Model of the Make by different shades. The size of circles is determined by price. Also, label there is easier to read.



Make, Model and sum of Price. Color shows details about Make and Model. Size shows maximum of Price. The marks are labeled by Make, Model and sum of Price.

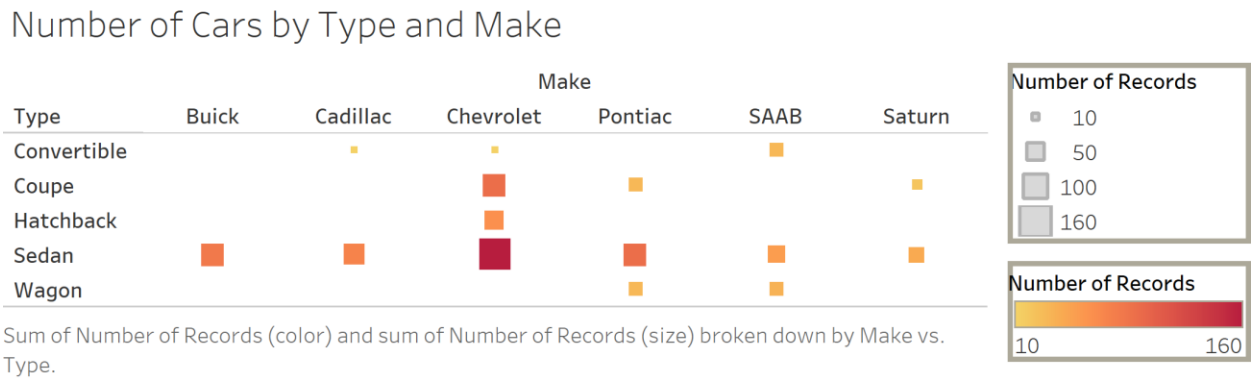
c. Write a short paragraph discussing the differences between the two plots. Describe for each something that displayed more clearly than with the other.

Treemap has better layout than packed bubble does. From the graph, treemap places same color category in one area, while packed bubble organizes graph based on both color and bubble size. As different colors, for both graphs, indicates different Makes of cars, have same color category together will be easier to compare different prices of different models under same Make.

Packed bubble is easier at comparing price because all numbers are represented in circle, and the size of circle is determined by price amount. In contrast, is hard to tell which rectangle is larger in treemap.

d. Create a contingency plot (Tableau calls it a heat map under Show Me) showing with color the number of cars (Number of Records) of each Type sold by each Make. Explain at least one observation about that data that this chart makes it easy to see.

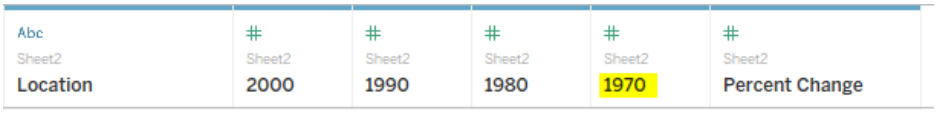
Different shades are used to represent number of records. The darker the shades the larger the number of records. Obviously, from the chart, Chevrolet's sedan has the largest number of records. Also, it's easy to assume that sedan is the most popular vehicle type, as every Make carries sedan, and for Maker has multiple models, Sedan has the highest sale records, no exceptions.



3) (20 pts) This problem works with a dataset containing the population of Montana and of each of the 7 Native American reservations within it (reservation70-00.xlsx). There is a measurement for each decade between 1970 and 2000. Sheet1 has the original data.

We will use Tableau for this question, but Sheet1 has a header that confuses Tableau. If you're interested, check out the "Data Interpreter" feature in Tableau to learn how to deal with this. Otherwise, use Sheet2, where I've removed the header. We need a few transformations to get the data ready to work with:

1. Renaming the 1970* field so it has no * and can be converted to a number



2. "Pivoting" the year fields in a similar manner to how it was demonstrated in the tutorial.

3. Changing the name of the pivot fields to Year and Population and changing the type of the year field to "whole number".

4. You can also hide the "Percent Change" field as it only contains information for change over the entire period, not per decade.

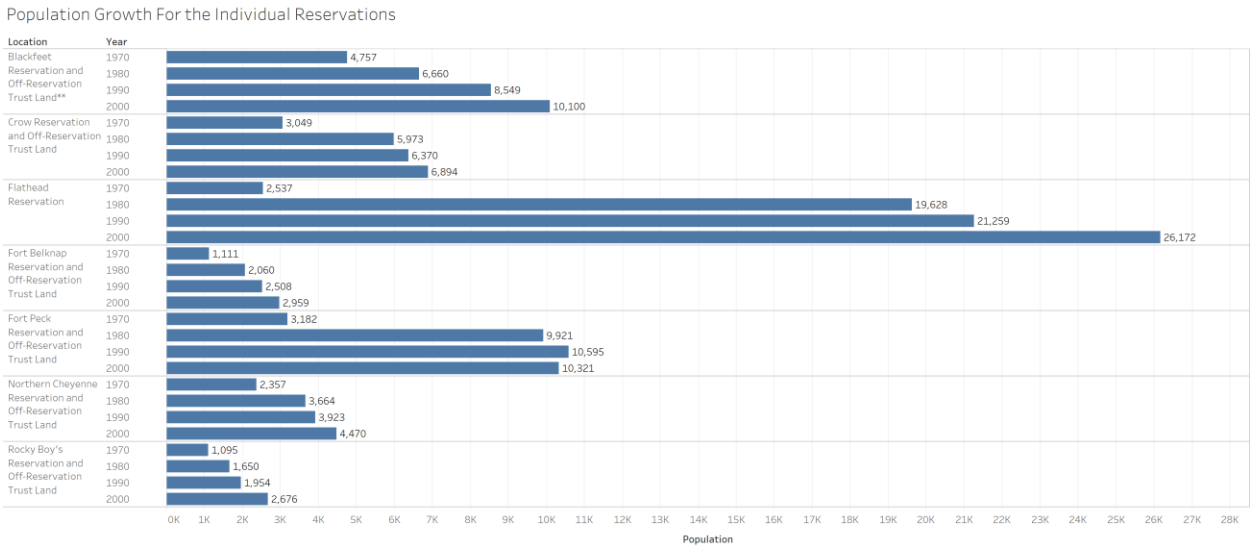
5. If you would like to have an actual Date field for the Year, so that it is treated by Tableau as a time instead of just a number, you need to create a "Calculated Field". It should construct a Date using the Year, i.e. make a Date field that is on January 1 of the specified year: makedate([Year], 1, 1)

6. We are not interested in the Montana population, only the reservation populations. When you have used Location on your graph, you can right mouse click (or click the down arrow within it) to apply filters. You can also use "Exclude" from the right click menu on the legend just below the "Marks" configuration.

Create graphs to show the following information, using appropriate graph types. Make sure that the graphs are properly labeled and that the axis scales properly reflect the type of data represented.

a. One chart that graphs the population growth over the years for the individual reservations.

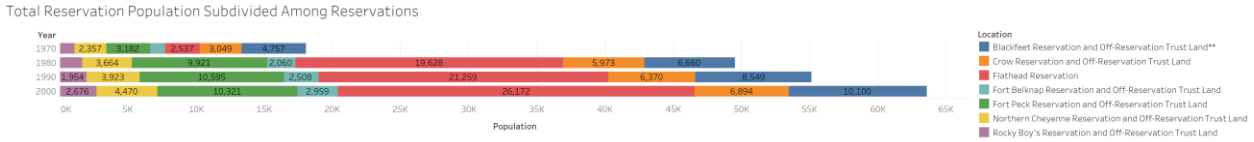
I put population on x-axis, and location and year on y-axis. Rows are grouped by location first, then under each location, population for each year is displaying by bar, so growth trend of population is easy to read.



Sum of Population for each Year broken down by Location. The marks are labeled by sum of Population.

b. One that graphs the total reservation population for each year, subdivided among the different reservations. The difference between this and (a) is that in (b) we are not looking only at each population individually but at the growth of the total population of all of them together, then subdivided by the reservations.

I put population on x-axis, and year on y-axis, then place different location into different color.



Sum of Population for each Year. Color shows details about Location. The marks are labeled by sum of Population.

4) (10 pts) For this question, answer only with text. You may include an illustration if you would like, but you do not need to visualize data for this question. a. Explain what we mean by 'pre-attentive' attributes. Are these as effectively recognized by human perception when they are used in combinations? b. Use Weber's Law to explain why it is important to include 0 in the numerical axis of a bar chart.

a. Pre-attentive means around 200ms, the time for an eye movement. I think they are effectively recognized by human perception when they are used in combinations because it is so easy to see that human won't need to think. If there's an important point of line that you want audience to pat attention, pre-attentive attributes can catch audience's attention easily.

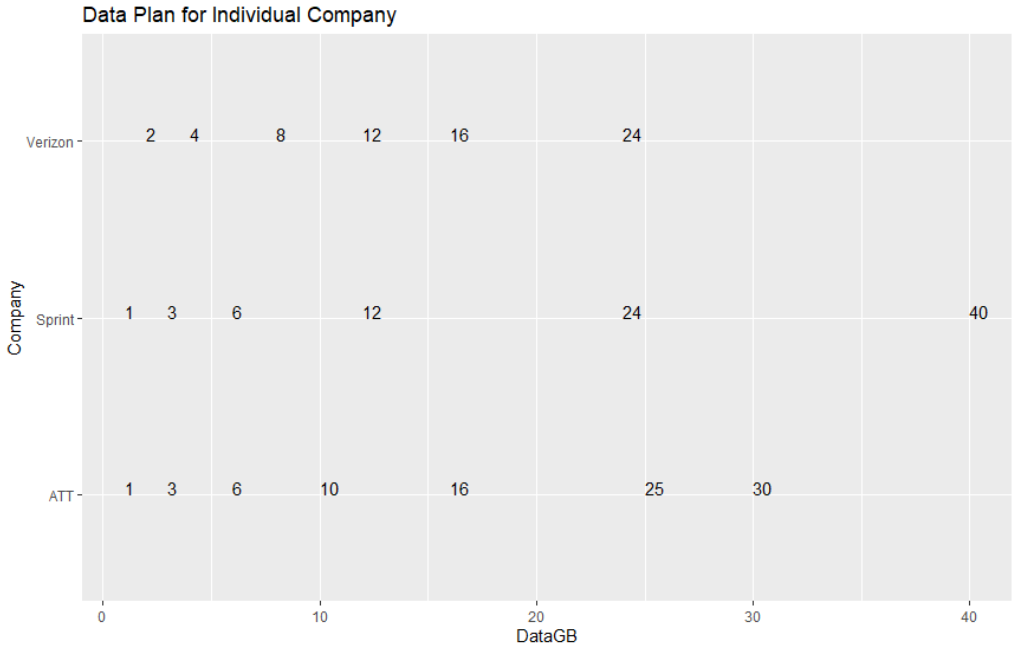
b. Weber's law means just-noticeable difference between two stimuli is proportional to the magnitude of the stimuli. It will be hard for human to compare two stimuli if they are not starting at same point or not aligned. Therefore, include 0 in the numerical axis of a bar chart will make it easier for human to determine the starting point to tell the difference between two stimuli.

5) (25 pts) This graph of cell phone pricing plans is not very easy to use. Use R for this question and recreate this graph in two different ways of your choice. For each one, explain what you are trying to help the user see. For example, one might be to compare the cell phone companies to see what kind of plans they have. Another might be best for examining the trend of the relationship between price and data bandwidth. That relationship may hold overall, or you could look to see if it is different per company. You can decide what to visualize, i.e. what question to answer with your visualization, but make sure to explain what this visualization should be showing. To get full credit, you must produce a graph which makes the answer to your question immediately clear. It must also be well implemented, i.e. following the guidelines at the top for a clean graph.

a. Compare the cell phone companies to see what kind of plans they have

G1 <- ggplot(cellPlans, aes(x=DataGB, y=Company))+geom_text(aes(label=(DataGB)),hjust=0,vjust=0)+ggtitle("Data Plan for Individual Company")

Instead of showing points and label them, I used text feature to show DataGB on x-axis, and have companies on y-axis. So, it's easier to read what kind of plans each company has.



b. Compare price per GB to see which company has the lowest $/GB

```
cellPlans1 <- mutate(cellPlans, PricePerData = Price / DataGB)
head(cellPlans1)

G2 <- ggplot(cellPlans1, aes(x=DataGB,y=PricePerData,group=Company,
colour=Company))+geom_line( size=1)+geom_point(size=4,shape=21, fill="white")+ggtitle("Price Per DataGB for Individual Company")
```

This graph is to show which company has the lowest dollar amount per dataGB. It's hard to compare directly because each company has different data plan with different price. The only way to know which company has lowest cost is to calculate price/GB. From the graph, we can say Sprint has the lowest cost among all three.