CSC 465

Homework 1

Name: Kavana Manvi Krishnamurthy
Student ID: 2158984
Email: kmanvikr@depaul.edu

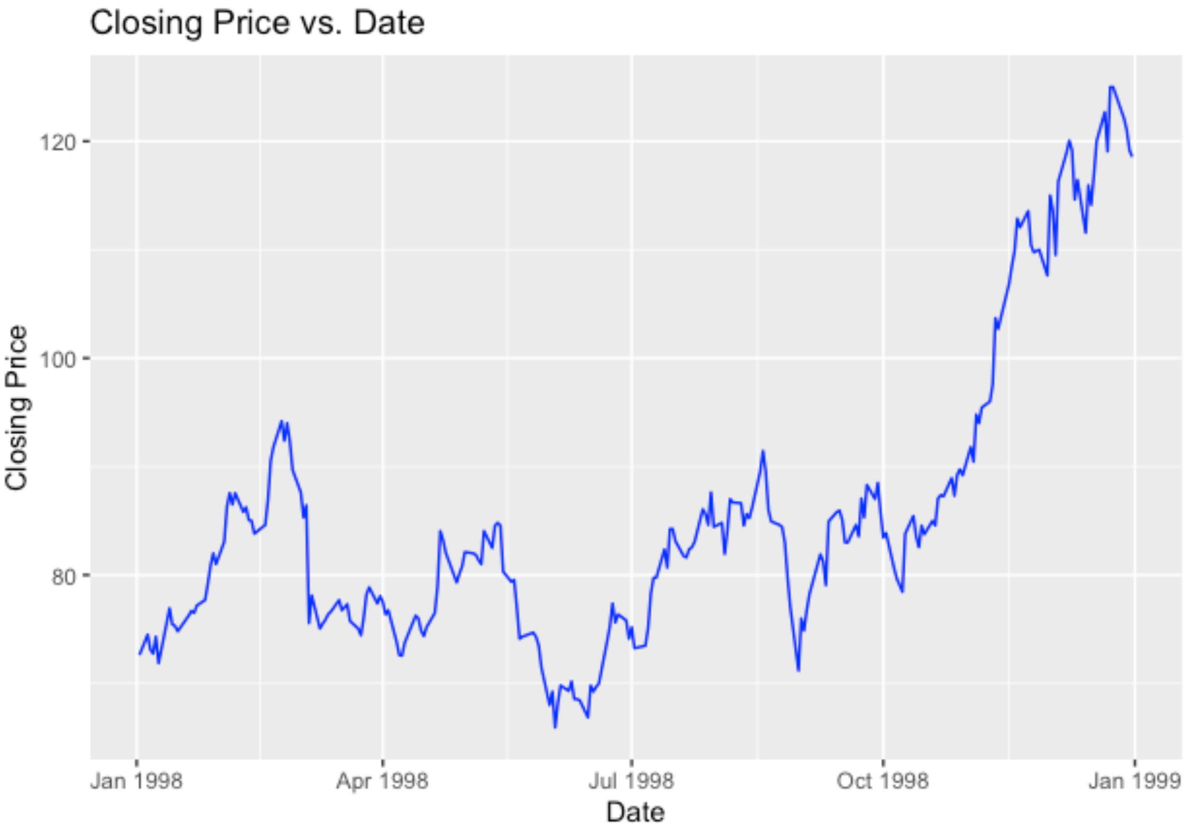
QUESTION 1

1) (20 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).

Make the specified graphs in either R or Tableau:

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".

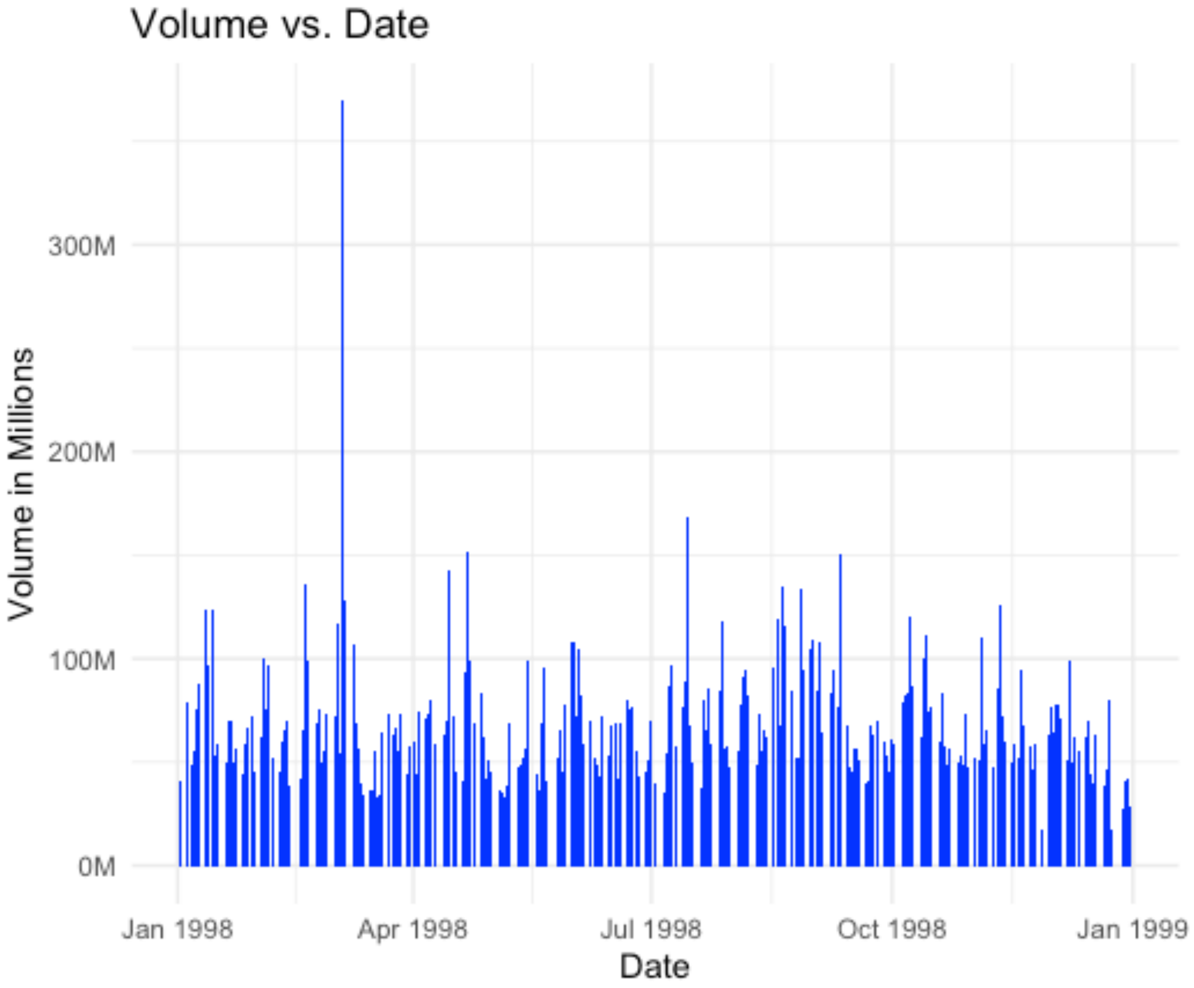
```
Intel.1998 <- read.csv("~/Desktop/DV/HW1/Intel-1998.csv")

library("dplyr")

library("ggplot2")

library("mosaic")

library("lubridate")



Intel <- Intel.1998 %>%

  mutate(Date = mdy(Date))



ggplot(data=Intel, aes(x=as.Date(Date), y=Close)) +

   geom_line(color="blue") +
  labs(title="Closing Price vs. Date", x="Date", y="Closing Price")
```



This is a line graph which displays the closing price of Intel stocks in 1998. The x-axis represents the Date plotted against the y-axis which shows Closing Price values. The graph shows increase and decrease in the stock's closing price value over the span of a year. It is simple and easy to interpret, provide an intuitive way to track changes over time.

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

```
library("scales")

ggplot(data=Intel, aes(x=as.Date(Date), y=Volume)) +

 geom_col(fill="blue") +  # Use geom_col() for bar heights

 labs(title="Volume vs. Date", x="Date", y="Volume in Millions") +

 scale_y_continuous(labels = label_number(scale = 1e-6, suffix = "M"))
```



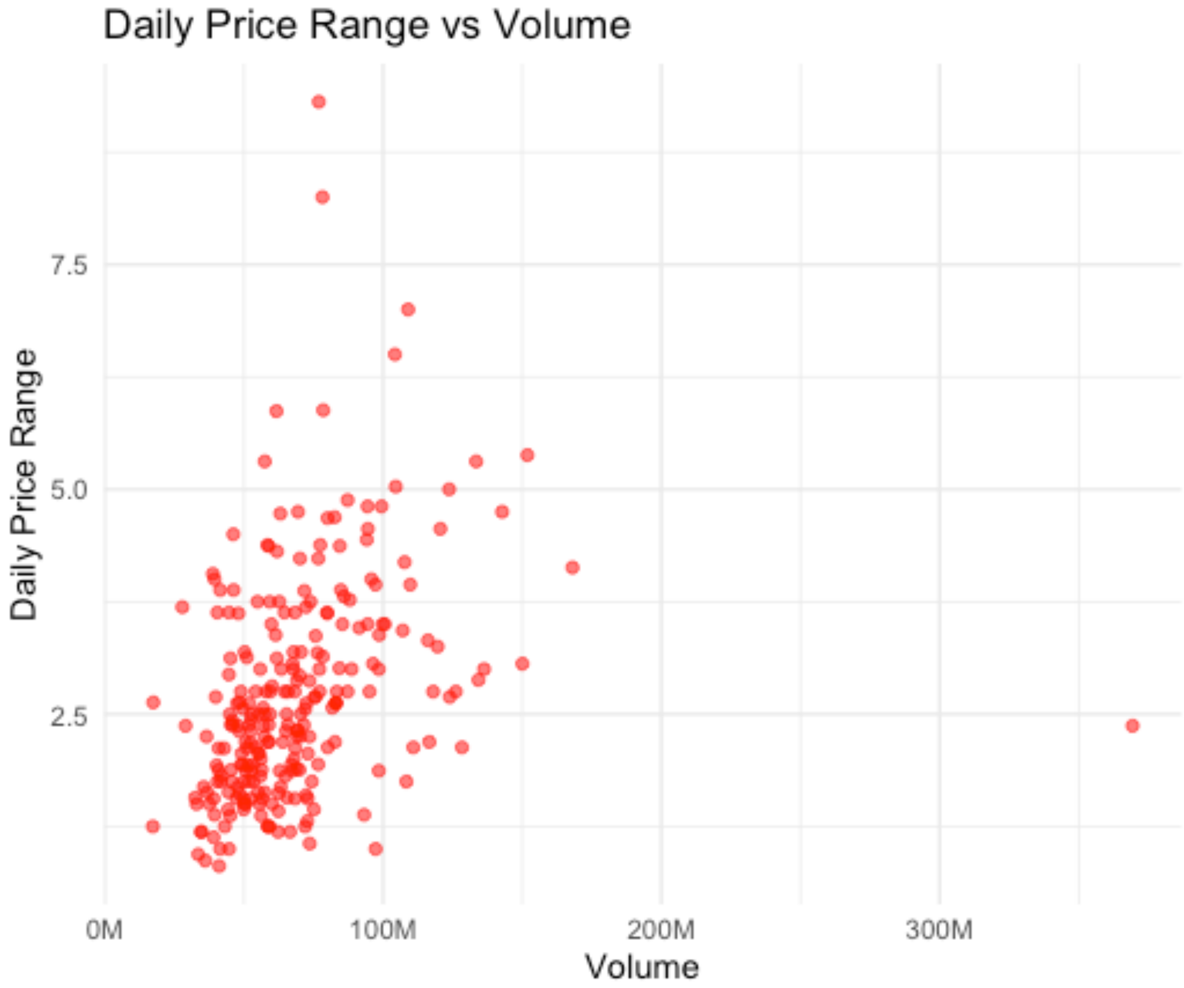
The bar graph above illustrates Intel's trading volume over the course of one year in 1998. The x-axis represents the date, while the y-axis displays the trading volume. The graph highlights the volume of stocks traded throughout the year, with noticeable gaps in the middle indicating periods of no activity, possibly due to holidays. In contrast, a line graph would represent these gaps as sudden drops, which could mislead the viewer into thinking there was a significant drop in volume, rather than a lack of trading activity.

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

```
Intel$Range <- Intel$High - Intel$Low

head(Intel)

ggplot(data=Intel, aes(x=Volume, y=Range)) +

  geom_point(color="red", alpha=0.6) +  # Scatter plot with blue points
```

```
 labs(title="Daily Price Range vs Volume", x="Volume", y="Daily Price Range") +

scale_x_continuous(labels = label_number(scale = 1e-6, suffix = "M"))
```



The above graph shows a scatter plot of Volume vs Daily Price Range. The Daily Price Range is calculated as the difference between the high and low column values, representing the fluctuation in stock price within a single trading day. Additionally, the dots in the scatter plot have an alpha value of 0.6, making overlapping points more visible. The density of the color indicates areas with a higher concentration of points, providing insight into the frequency of certain price ranges and volumes. We can also identify outliers pretty easily.

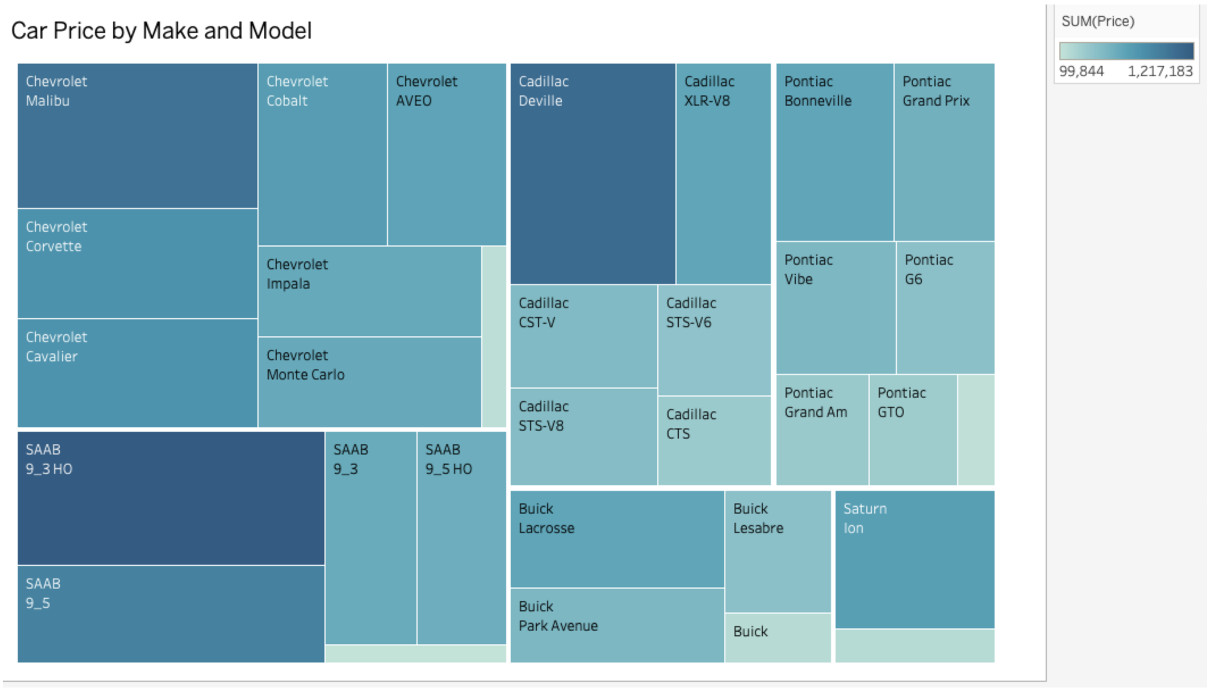
Question 2

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.

Steps used in Tableau:

- Drag the Price to columns.
- Drag Make and Model to rows.
- Select Tree Map under Show me.

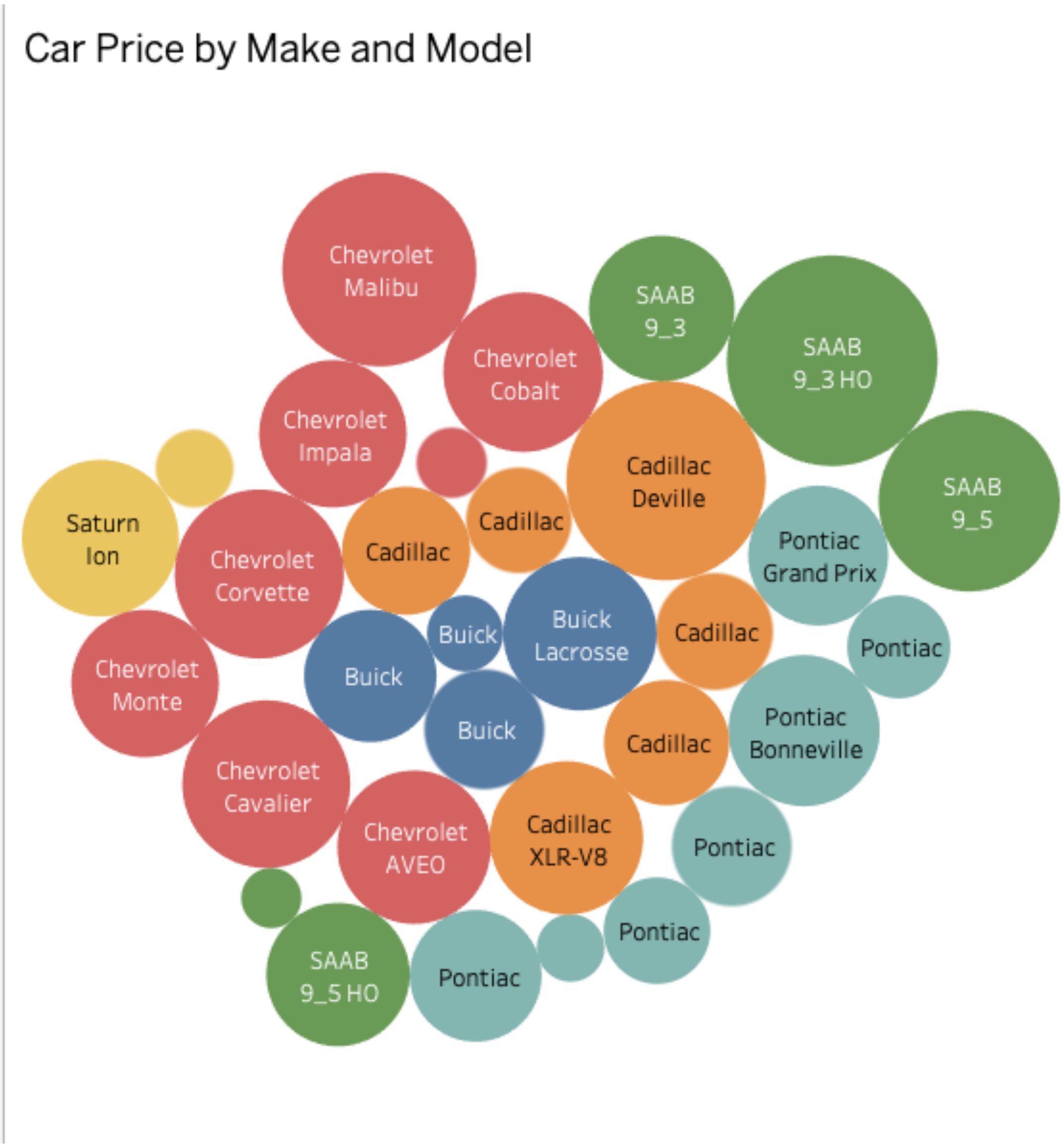


A **tree map** in data visualization is a graph which displays hierarchical data. It basically uses nested rectangles. Each rectangle represents a category of Make and minor subcategory of Model, and its size represents sum of price value. It you to see the distribution at a glance.

C. A packed bubble chart of the same type.

Tableau steps:

- Drag Make to Rows.
- Drag Model to rows.
- Drag Price to Columns.
- Select Packed bubbles under show me.



The **packed bubble chart** is used to display GM Cars data in a cluster of circles. Each bubble represents Make data, and the size of the bubble is nothing but the sum of price value. Packed bubble charts are useful for showing relationships between categories, especially when comparing proportions or seeing how items contribute to a larger whole. The position of bubbles is not meaningful, but the size represents sum of price and the colour represents the Make of each car. If you hover over each bubble, it will show you the Make, Model and price.

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

The **tree map** and the **packed bubble chart** both provide different ways to visualize hierarchical and categorical data, but each has unique strengths.

In the **tree map**, hierarchical relationships between car **Makes** and **Models** are displayed more clearly due to the nested rectangles, with each rectangle's size representing the sum of the price values. This makes it easy to see how different models contribute to the total price within a make, and gives a clear view of proportions at a glance.

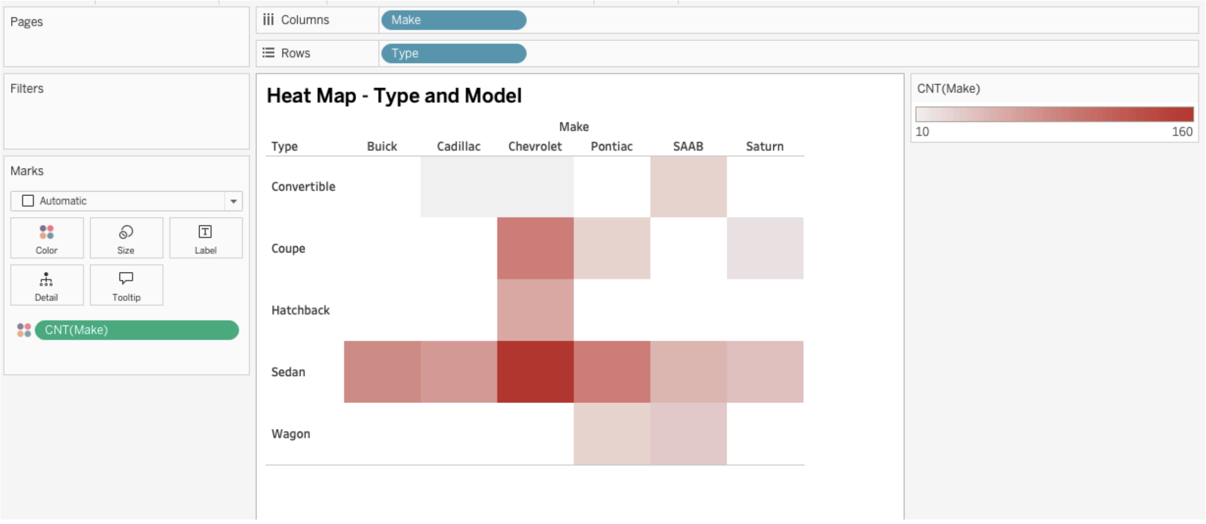
On the other hand, the **packed bubble chart** focuses more on the overall relationship between different **Makes**, with the size of the bubbles representing the total price value for each make. This allows for quick visual comparison between the categories (Makes) in terms of their price contribution. The use of colors and hovering tooltips helps identify individual Makes and Models, which is less structured but more visually appealing for seeing relationships between categories.

While the tree map provides a clear, structured representation of hierarchy and distribution, the packed bubble chart emphasizes category proportions with more flexibility in exploring the data interactively.

E. Create a contingency plot (Tableau calls it a heat map under Show Me) showing with colour 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.

Tableau steps:

- Drag Make to Columns**.**
- Drag Type to Rows**.**
- Drag Number of Records to Color (use the COUNT function).
- Select Heat Map from the Show Me panel.



Columns: Make

Rows: Type

Marks: Automatic

CNT(Make)

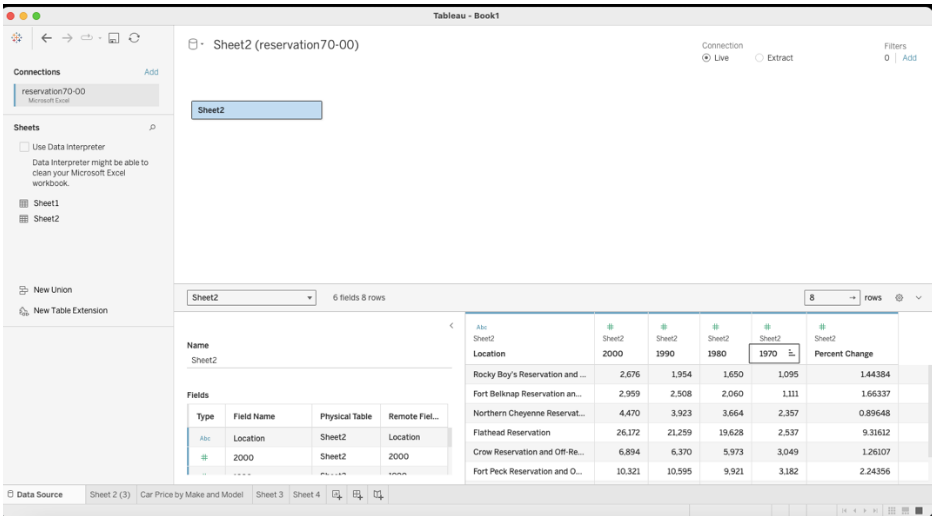
Observation: The heat map shows which Type and Make has a greater number of records. For example, Type– Sedan and Make – Chevrolet has the greatest number of records which is 160. So, it has the highest gradient of colour.

Question 3

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



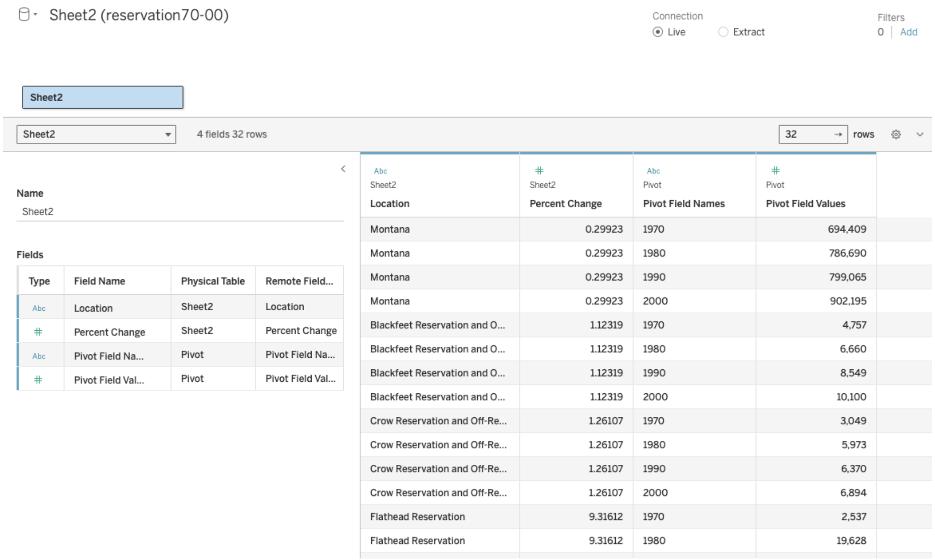
Name
Sheet2

Fields

Type	Field Name	Physical Table	Remote Fiel...
Abc	Location	Sheet2	Location
#	2000	Sheet2	2000

Sheet2 Location	Sheet2 2000	Sheet2 1990	Sheet2 1980	Sheet2 1970	Sheet2 Percent Change
Rocky Boy's Reservation and ...	2,676	1,954	1,650	1,095	1.44384
Fort Belknap Reservation an...	2,959	2,508	2,060	1,111	1.66337
Northern Cheyenne Reservat...	4,470	3,923	3,664	2,357	0.89648
Flathead Reservation	26,172	21,259	19,628	2,537	9.31612
Crow Reservation and Off-Re...	6,894	6,370	5,973	3,049	1.26107
Fort Peck Reservation and O...	10,321	10,595	9,921	3,182	2.24356

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



Sheet2 (reservation70-00)

Connection ● Live ○ Extract

Filters 0 | Add

Sheet2

Sheet2 ▾ 4 fields 32 rows 32 → rows

Name

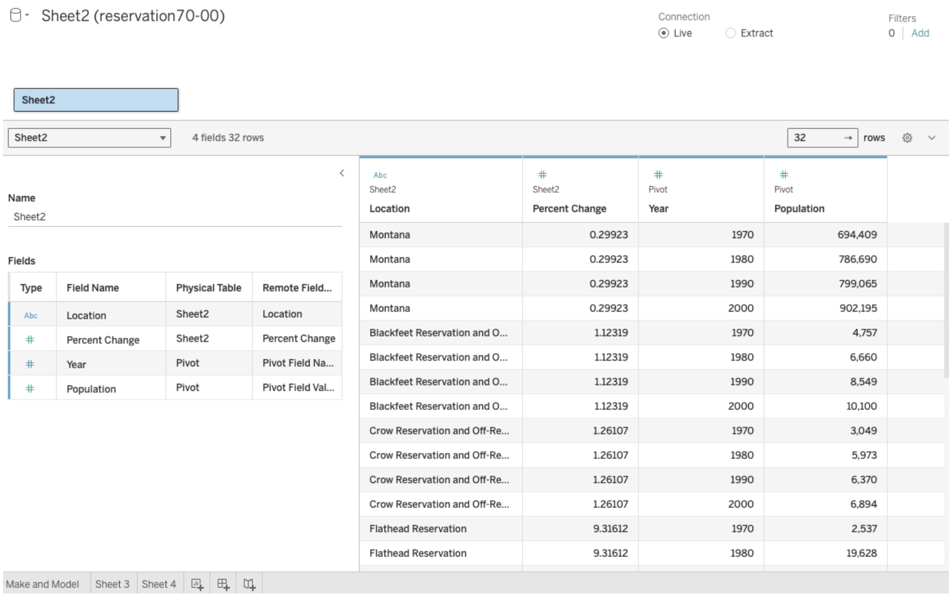
Sheet2

Fields

Type	Field Name	Physical Table	Remote Field...
Abc	Location	Sheet2	Location
#	Percent Change	Sheet2	Percent Change
Abc	Pivot Field Na...	Pivot	Pivot Field Na...
#	Pivot Field Val...	Pivot	Pivot Field Val...

Abc Sheet2 Location	# Sheet2 Percent Change	Abc Pivot Pivot Field Names	# Pivot Pivot Field Values
Montana	0.29923	1970	694,409
Montana	0.29923	1980	786,690
Montana	0.29923	1990	799,065
Montana	0.29923	2000	902,195
Blackfeet Reservation and O...	1.12319	1970	4,757
Blackfeet Reservation and O...	1.12319	1980	6,660
Blackfeet Reservation and O...	1.12319	1990	8,549
Blackfeet Reservation and O...	1.12319	2000	10,100
Crow Reservation and Off-Re...	1.26107	1970	3,049
Crow Reservation and Off-Re...	1.26107	1980	5,973
Crow Reservation and Off-Re...	1.26107	1990	6,370
Crow Reservation and Off-Re...	1.26107	2000	6,894
Flathead Reservation	9.31612	1970	2,537
Flathead Reservation	9.31612	1980	19,628

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



Sheet2 (reservation70-00)

Connection ● Live ○ Extract

Filters 0 | Add

Sheet2

Sheet2 ▾ 4 fields 32 rows 32 → rows

Name

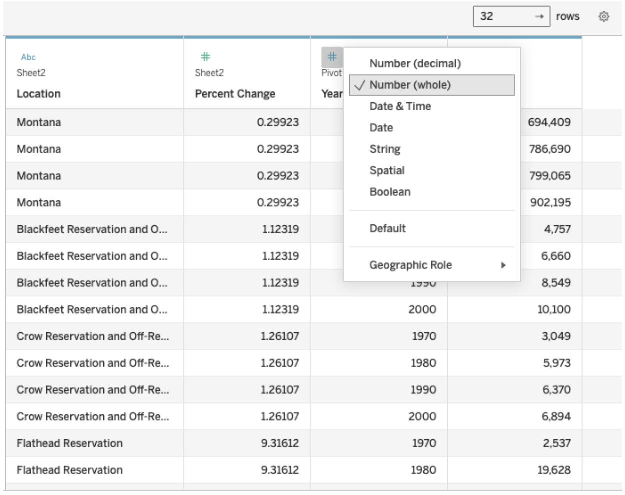
Sheet2

Fields

Type	Field Name	Physical Table	Remote Field...
Abc	Location	Sheet2	Location
#	Percent Change	Sheet2	Percent Change
#	Year	Pivot	Pivot Field Na...
#	Population	Pivot	Pivot Field Val...

Abc Sheet2 Location	# Sheet2 Percent Change	# Pivot Year	# Pivot Population
Montana	0.29923	1970	694,409
Montana	0.29923	1980	786,690
Montana	0.29923	1990	799,065
Montana	0.29923	2000	902,195
Blackfeet Reservation and O...	1.12319	1970	4,757
Blackfeet Reservation and O...	1.12319	1980	6,660
Blackfeet Reservation and O...	1.12319	1990	8,549
Blackfeet Reservation and O...	1.12319	2000	10,100
Crow Reservation and Off-Re...	1.26107	1970	3,049
Crow Reservation and Off-Re...	1.26107	1980	5,973
Crow Reservation and Off-Re...	1.26107	1990	6,370
Crow Reservation and Off-Re...	1.26107	2000	6,894
Flathead Reservation	9.31612	1970	2,537
Flathead Reservation	9.31612	1980	19,628

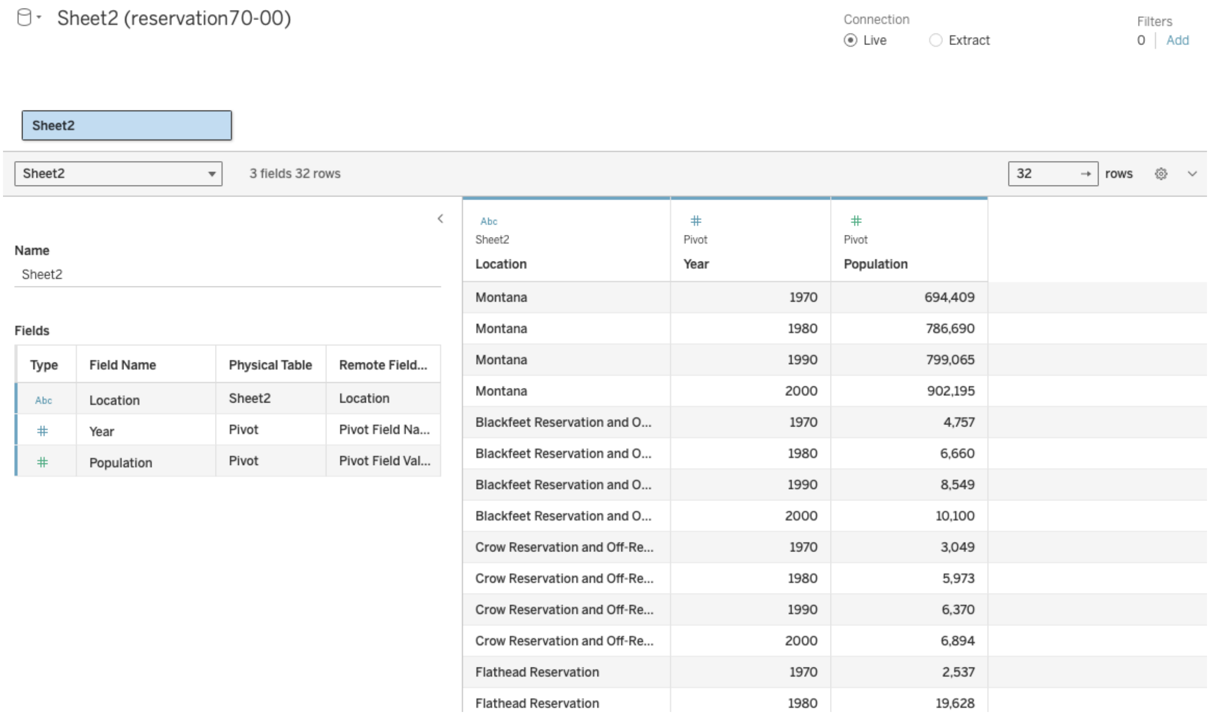
Make and Model | Sheet 3 | Sheet 4



32 → rows

Abc Sheet2 Location	# Sheet2 Percent Change	# Pivot Year	
Montana	0.29923		694,409
Montana	0.29923		786,690
Montana	0.29923		799,065
Montana	0.29923		902,195
Blackfeet Reservation and O...	1.12319		4,757
Blackfeet Reservation and O...	1.12319		6,660
Blackfeet Reservation and O...	1.12319	1990	8,549
Blackfeet Reservation and O...	1.12319	2000	10,100
Crow Reservation and Off-Re...	1.26107	1970	3,049
Crow Reservation and Off-Re...	1.26107	1980	5,973
Crow Reservation and Off-Re...	1.26107	1990	6,370
Crow Reservation and Off-Re...	1.26107	2000	6,894
Flathead Reservation	9.31612	1970	2,537
Flathead Reservation	9.31612	1980	19,628

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



Sheet2 (reservation70-00)

Connection
● Live ○ Extract

Filters
0 | Add

Sheet2

Sheet2 ▾ 3 fields 32 rows 32 → rows

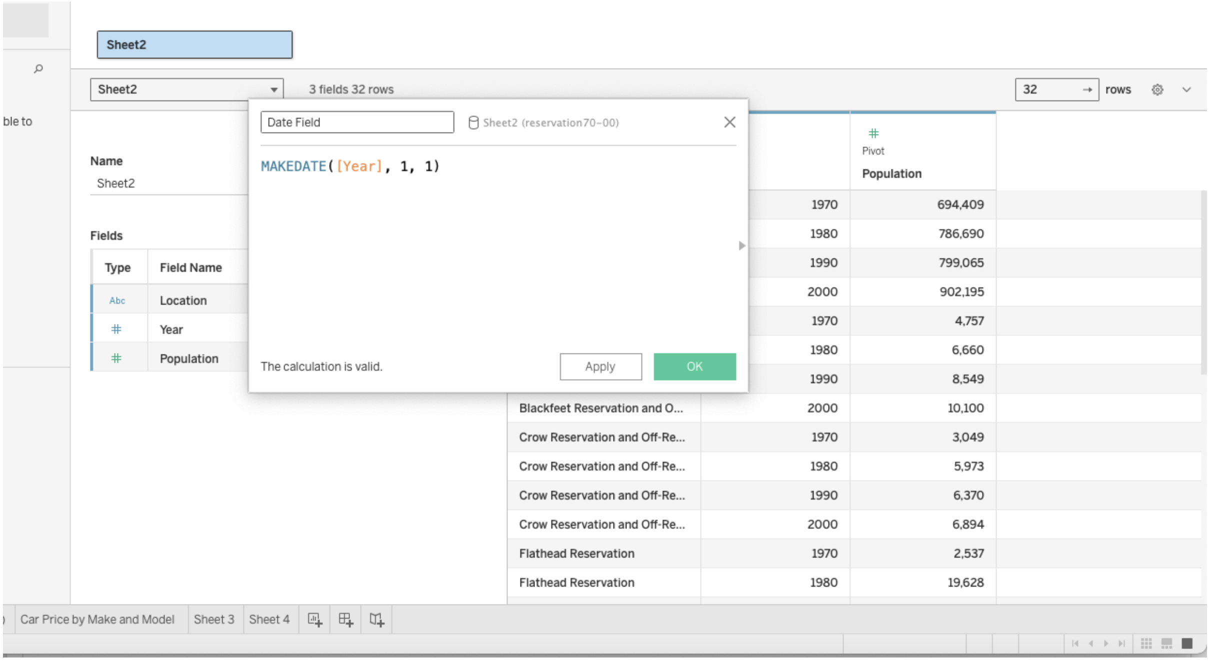
Name
Sheet2

Fields

Type	Field Name	Physical Table	Remote Field...
Abc	Location	Sheet2	Location
#	Year	Pivot	Pivot Field Na...
#	Population	Pivot	Pivot Field Val...

Abc Sheet2 Location	# Pivot Year	# Pivot Population
Montana	1970	694,409
Montana	1980	786,690
Montana	1990	799,065
Montana	2000	902,195
Blackfeet Reservation and O...	1970	4,757
Blackfeet Reservation and O...	1980	6,660
Blackfeet Reservation and O...	1990	8,549
Blackfeet Reservation and O...	2000	10,100
Crow Reservation and Off-Re...	1970	3,049
Crow Reservation and Off-Re...	1980	5,973
Crow Reservation and Off-Re...	1990	6,370
Crow Reservation and Off-Re...	2000	6,894
Flathead Reservation	1970	2,537
Flathead Reservation	1980	19,628

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)



Sheet2

Sheet2 ▾ 3 fields 32 rows 32 → rows

ble to

Name
Sheet2

Fields

Type	Field Name
Abc	Location
#	Year
#	Population

Date Field — Sheet2 (reservation70-00)

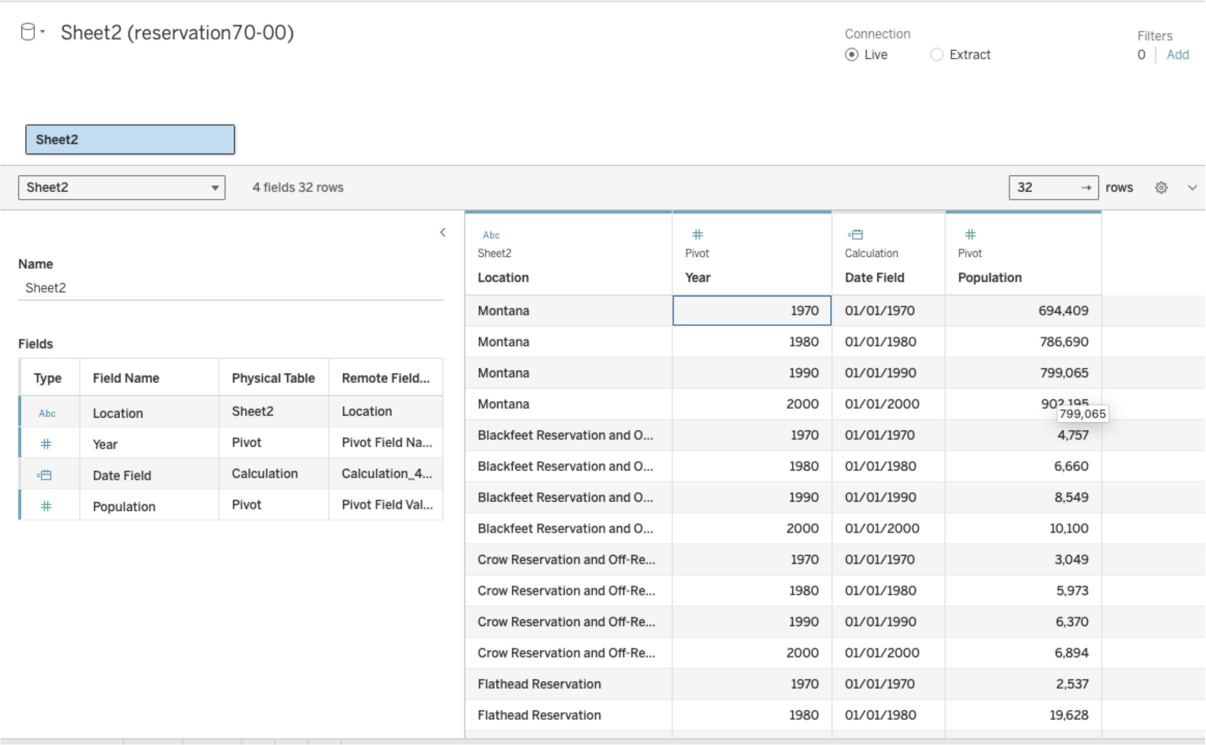
```
MAKEDATE([Year], 1, 1)
```

The calculation is valid.

Apply OK

		# Pivot Population
	1970	694,409
	1980	786,690
	1990	799,065
	2000	902,195
	1970	4,757
	1980	6,660
	1990	8,549
Blackfeet Reservation and O...	2000	10,100
Crow Reservation and Off-Re...	1970	3,049
Crow Reservation and Off-Re...	1980	5,973
Crow Reservation and Off-Re...	1990	6,370
Crow Reservation and Off-Re...	2000	6,894
Flathead Reservation	1970	2,537
Flathead Reservation	1980	19,628

Car Price by Make and Model | Sheet 3 | Sheet 4



Sheet2 (reservation70-00)

Connection
● Live ○ Extract

Filters
0 | Add

Sheet2

Sheet2 ▾ 4 fields 32 rows

32 → rows

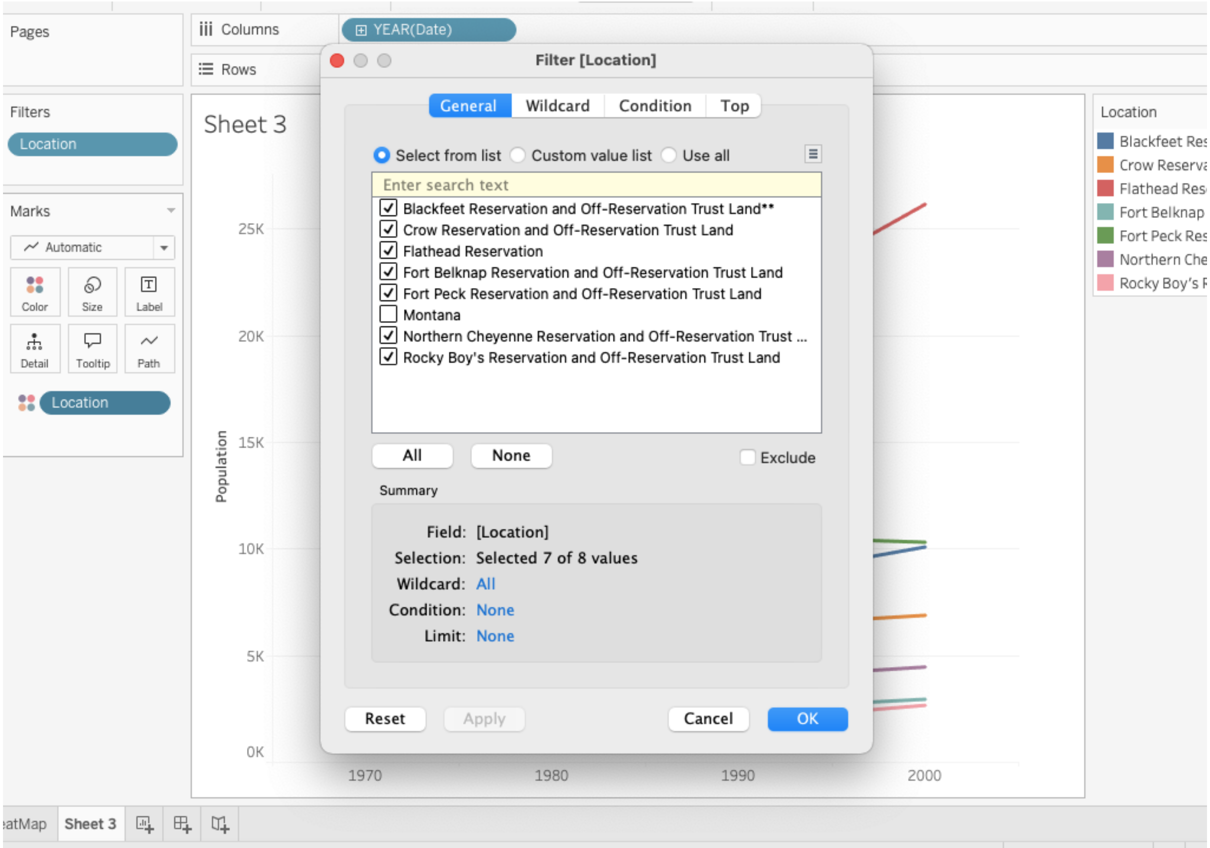
Name
Sheet2

Fields

Type	Field Name	Physical Table	Remote Field...
Abc	Location	Sheet2	Location
#	Year	Pivot	Pivot Field Na...
📅	Date Field	Calculation	Calculation_4...
#	Population	Pivot	Pivot Field Val...

Abc Sheet2 Location	# Pivot Year	📅 Calculation Date Field	# Pivot Population
Montana	1970	01/01/1970	694,409
Montana	1980	01/01/1980	786,690
Montana	1990	01/01/1990	799,065
Montana	2000	01/01/2000	902,195
Blackfeet Reservation and O...	1970	01/01/1970	4,757
Blackfeet Reservation and O...	1980	01/01/1980	6,660
Blackfeet Reservation and O...	1990	01/01/1990	8,549
Blackfeet Reservation and O...	2000	01/01/2000	10,100
Crow Reservation and Off-Re...	1970	01/01/1970	3,049
Crow Reservation and Off-Re...	1980	01/01/1980	5,973
Crow Reservation and Off-Re...	1990	01/01/1990	6,370
Crow Reservation and Off-Re...	2000	01/01/2000	6,894
Flathead Reservation	1970	01/01/1970	2,537
Flathead Reservation	1980	01/01/1980	19,628

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.**



Filter [Location]

General | Wildcard | Condition | Top

● Select from list ○ Custom value list ○ Use all

Enter search text

☑ Blackfeet Reservation and Off-Reservation Trust Land**
☑ Crow Reservation and Off-Reservation Trust Land
☑ Flathead Reservation
☑ Fort Belknap Reservation and Off-Reservation Trust Land
☑ Fort Peck Reservation and Off-Reservation Trust Land
□ Montana
☑ Northern Cheyenne Reservation and Off-Reservation Trust ...
☑ Rocky Boy's Reservation and Off-Reservation Trust Land

All | None | □ Exclude

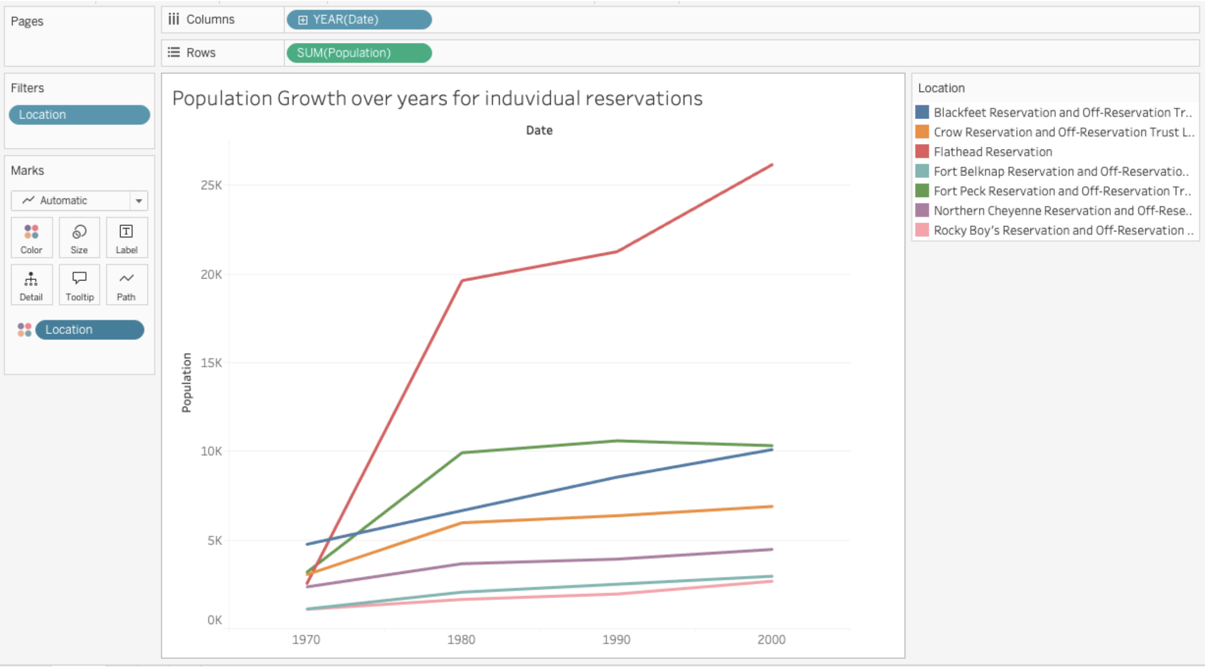
Summary

Field: [Location]
Selection: Selected 7 of 8 values
Wildcard: All
Condition: None
Limit: None

Reset | Apply | Cancel | OK

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.

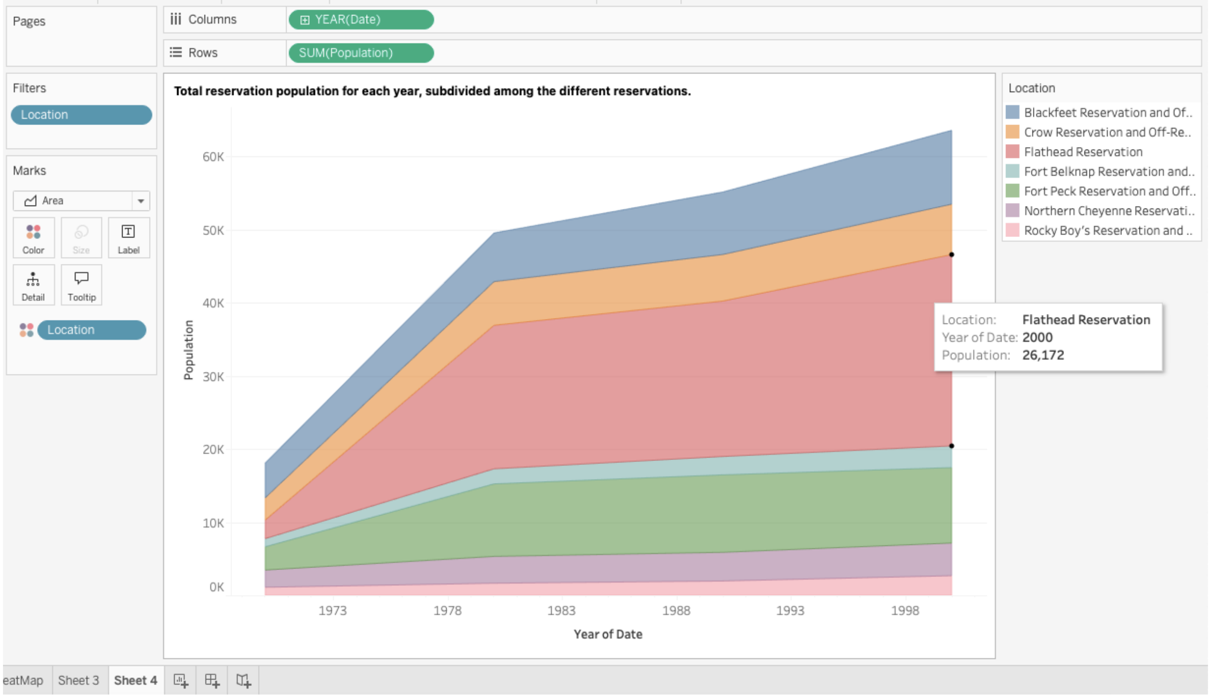
A. One chart that graphs the population growth over the years for the individual reservations.



Pages
Columns
YEAR(Date)
Rows
SUM(Population)
Filters
Location
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Path
Location
Population Growth over years for induvidual reservations
Date
Population
25K
20K
15K
10K
5K
0K
1970
1980
1990
2000
Location
Blackfeet Reservation and Off-Reservation Tr..
Crow Reservation and Off-Reservation Trust L..
Flathead Reservation
Fort Belknap Reservation and Off-Reservatio..
Fort Peck Reservation and Off-Reservation Tr..
Northern Cheyenne Reservation and Off-Rese..
Rocky Boy's Reservation and Off-Reservation ..

This line graph enables viewers to easily follow the Reservation population changes for each individual reservation over a span of four decades. By using different colours for each reservation, the chart effectively highlights individual trends for each population, making it easy to observe patterns of growth or decline in some cases, throughout the years.

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.



Total reservation population for each year, subdivided among the different reservations.

Columns: YEAR(Date)
Rows: SUM(Population)
Filters: Location
Marks: Area
Color: Location

Location: Flathead Reservation
Year of Date: 2000
Population: 26,172

Pages
Filters
Marks
Color
Size
Label
Detail
Tooltip
60K
50K
40K
30K
20K
10K
0K
Population
1973
1978
1983
1988
1993
1998
Year of Date
Location
Blackfeet Reservation and Of..
Crow Reservation and Off-Re..
Flathead Reservation
Fort Belknap Reservation and..
Fort Peck Reservation and Off..
Northern Cheyenne Reservati..
Rocky Boy's Reservation and ..
eatMap
Sheet 3
Sheet 4

The above graph displays a stacked area chart. It offers a comprehensive perspective on total population growth. This type of graph is especially helpful for analyzing overall growth/decline trends and understanding how the population of each reservation has impacted the total over time. It is visually separating the contributions of each reservation.

Question 4

4) 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?

Pre-attentive cues are visual elements that our brains instinctively notice without much conscious effort. These cues, such as colour, size, shape, tilt, proximity, shadow direction and orientation, can quickly capture attention. They stand out. While using a single pre-attentive cue can effectively highlight important information to make it stand out(for example a single red dot in a scatter plot of blue dots), combining multiple cues can enhance visual impact.

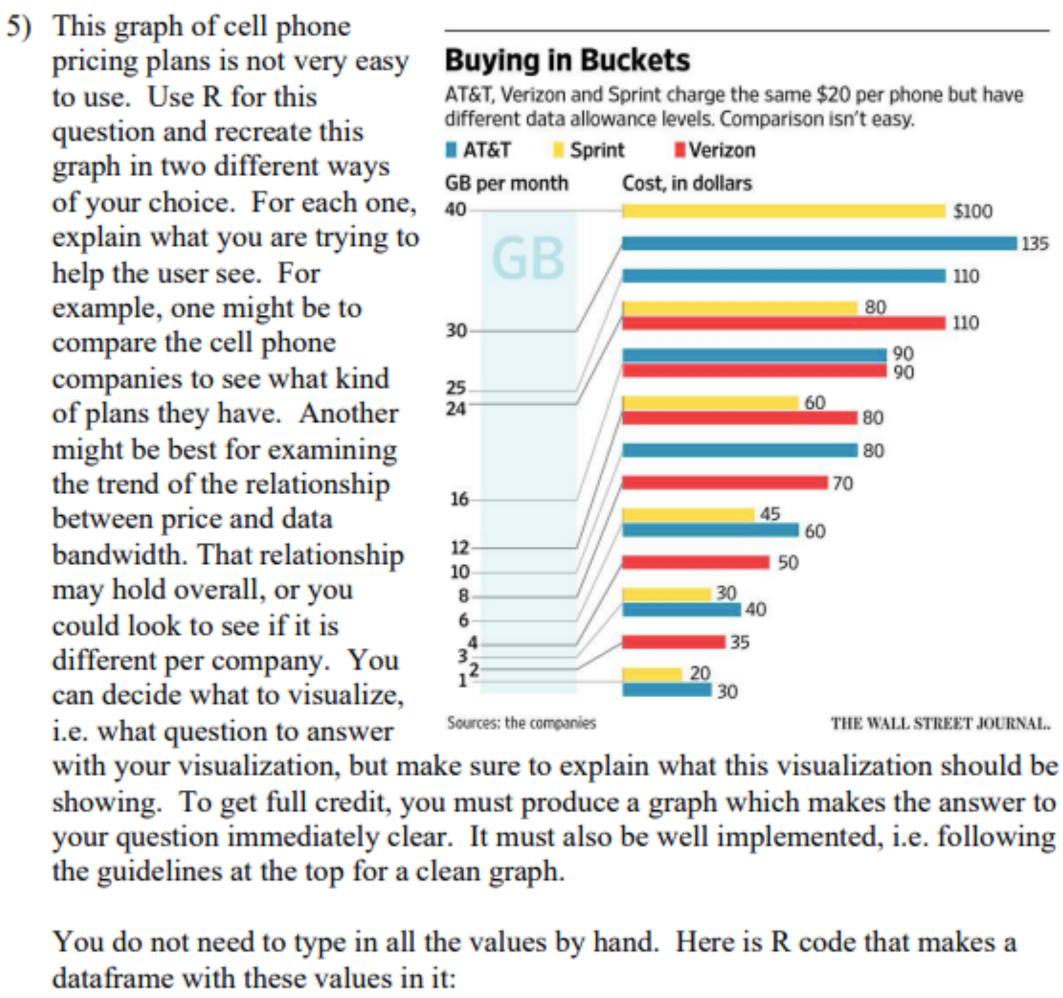
However, it's crucial to exercise caution when combining these cues. Too many competing elements can overwhelm the viewer's cognitive abilities, leading to confusion rather than clarity. For example, a conjunction of attributes like- checking the redness and circle-ness of an object in a scatter plot can be confusing. Therefore, to ensure a visualization remains effective and readable, it's essential to use pre-attentive cues judiciously and avoid creating visual clutter.

B. Use Weber's Law to explain why it is important to include 0 in the numerical axis of a bar chart.

Weber's Law posits that our perception of differences between stimuli is influenced by their initial intensity. In bar charts, this means starting the y-axis at zero is crucial for accurate visual comparisons. When the axis begins at zero, the length of each bar proportionally reflects the data value, ensuring the visual representation aligns with the actual numerical differences. If the y-axis doesn't include zero, the perceived differences can become distorted. Smaller differences might appear exaggerated, while larger ones might seem minimized, potentially misleading viewers.

Starting the y-axis at zero is essential because bar charts rely on the visual length of each bar to convey magnitude. Without a consistent baseline like zero, the proportions become unreliable, violating Weber's Law and distorting how differences are perceived. Additionally, including zero ensures the chart's scale is intuitive, allowing viewers to make straightforward comparisons between quantities. This approach prevents overemphasizing insignificant variations and maintains the integrity of the data representation.

Question 5



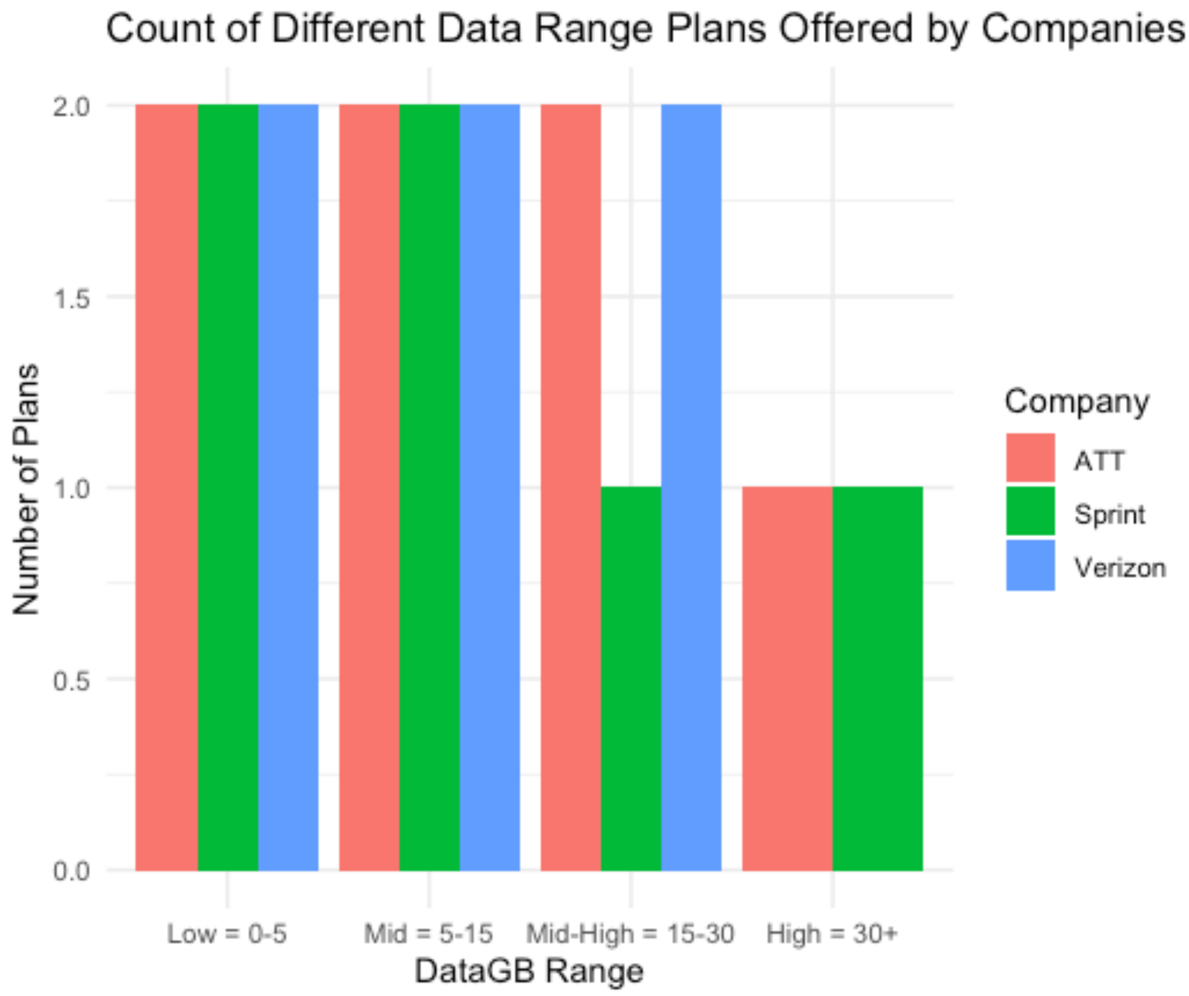
5) This graph of cell phone pricing plans is not very easy to use. Use R for this question and recreate this graph in two different ways of your choice. For each one, explain what you are trying to help the user see. For example, one might be to compare the cell phone companies to see what kind of plans 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.

You do not need to type in all the values by hand. Here is R code that makes a dataframe with these values in it:

```
cellPlans = data.frame(
        c("ATT", "Sprint", "Verizon", "ATT", "Sprint",
          "Verizon", "ATT", "Sprint", "Verizon", "ATT",
          "Verizon", "Sprint", "Verizon", "ATT",
          "Verizon", "Sprint", "ATT", "ATT", "Sprint"),
        c(1, 1, 2, 3, 3, 4, 6, 6, 8, 10, 12, 12, 16, 16,
          24, 24, 25, 30, 40),
        c(30, 20, 35, 40, 30, 50, 60, 45, 70, 80, 80, 60,
          90, 90, 110, 80, 110, 135, 100))
   names(cellPlans) = c("Company", "DataGB", "Price")
```

Visualization 1: How do the different companies compare in terms of the number of plans they offer across various data usage categories (Low, Mid, High, Very High)?

```
cellPlans <- cellPlans %>%
  mutate(DataGB_Bin = cut(DataGB, breaks = c(0, 5, 15, 30, Inf),
        labels = c("Low", "Mid", "High", "Very High"), right = FALSE))

num_plans_df <- cellPlans %>%
  group_by(Company, DataGB_Bin) %>%
  summarise(Plan_Count = n(), .groups = 'drop')

labels <- c("Low = 0-5", "Mid = 5-15", "Mid-High = 15-30", "High = 30+")

ggplot(num_plans_df, aes(x = DataGB_Bin, y = Plan_Count, fill = Company)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Count of Different Data Range Plans Offered by Companies",
       x = "DataGB Range",
       y = "Number of Plans") +
  scale_x_discrete(labels = labels) +  # Use the custom labels
  theme_minimal()
```

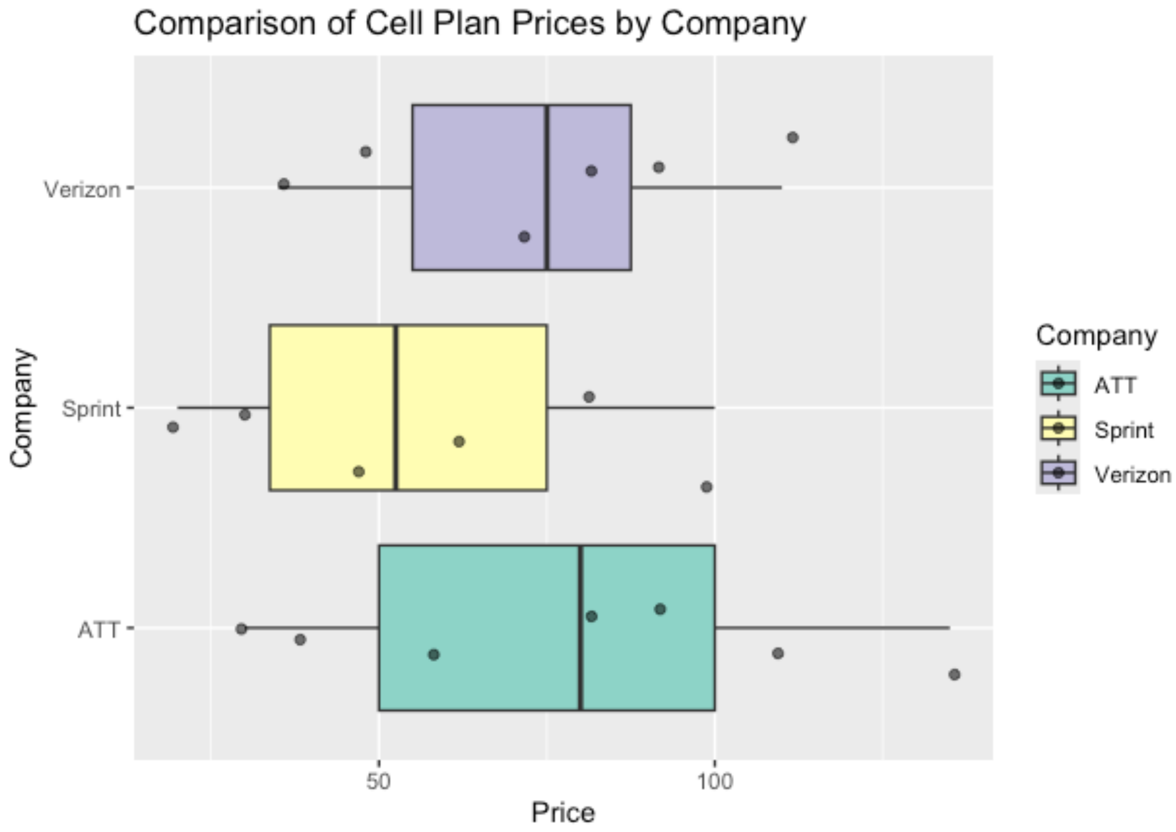


The plot allows for immediate comparison of how many plans each company offers in each data category. For instance, it's easy to see which company has the most offerings in the "Low" range versus the "High" range.

The x-axis categorizes the plans into four distinct ranges: Low (0-5 GB), Mid (5-15 GB), High (15-30 GB), and Very High (30+ GB). This clear categorization helps viewers quickly understand the focus areas of each company. The y-axis indicates the number of plans available for each company within those categories, allowing viewers to identify which companies are more diversified in their offerings. Different colors represent different companies, making it visually straightforward to differentiate between them at a glance. This enhances the readability and effectiveness of the graph.
This visualization clearly communicates the distribution of plans offered by each company across various data usage categories. It allows stakeholders to assess the competitive landscape and identify which companies provide a wider range of options based on data needs, thereby making informed decisions whether for consumers or business strategy.

Visualization 2: How do the prices of cell plans vary across different companies (Verizon, Sprint, and ATT)?

```
ggplot(data=cellPlans, aes(x=Company, y = Price,
fill=Company))+geom_boxplot(outlier.size=0)+geom_jitter(alpha=0.6)+coord_flip()+
scale_fill_brewer(palette = "Set3")+labs(title='Comparison of Cell Plan Prices by Company')
```



The provided box plot effectively addresses the question by visually comparing the distribution of prices for each company.

The horizontal line within each box represents the median price for that company. ATT has the lowest median price, followed by Sprint and then Verizon. Verizon has the largest IQR, indicating a wider range of prices among its plans. Sprint has the smallest IQR, suggesting more consistent pricing. ATT's IQR falls between the two. Individual data points that fall outside the whiskers are considered outliers and are plotted separately. Sprint has the most outliers, suggesting that some of its plans have significantly higher or lower prices compared to the majority.