Cynthia Cho
DSC 465
Due April 17, 2019

Homework 1

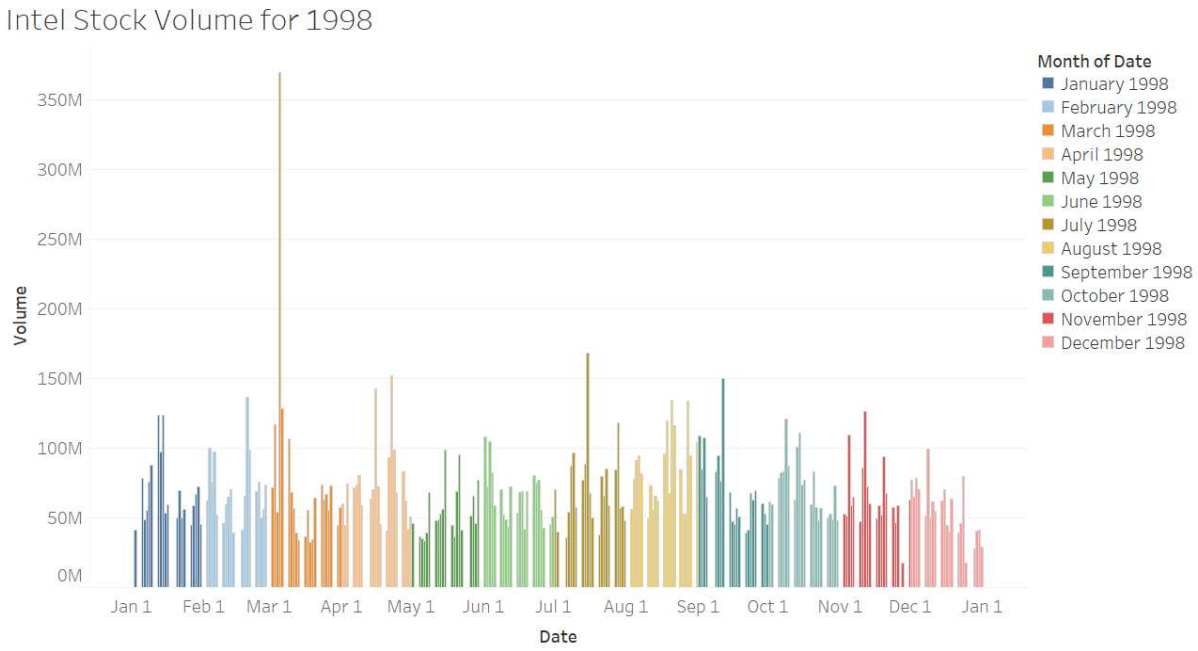
Problem 1 – reading of textbook

Problem 2: Done in Tableau using Intel-1998.csv

a. Line graph: To create the graph, the Columns are the exact dates, and the Rows are the Adj. Close. The title and axes were relabeled for readability.
What can be seen from the graph is that throughout the year, the price oscillates between $15 and $22. However, around the 4th quarter, the price shows an upward trend.

b. Bar graph for Volume versus Date: The Columns use the exact dates, and the Rows are the Volumes for each date. Each month is a different color, and the bars were updated to be thin enough to not create overlap. The legend for each color-coded month is to the right of the graph. The title and axes were relabeled for readability.

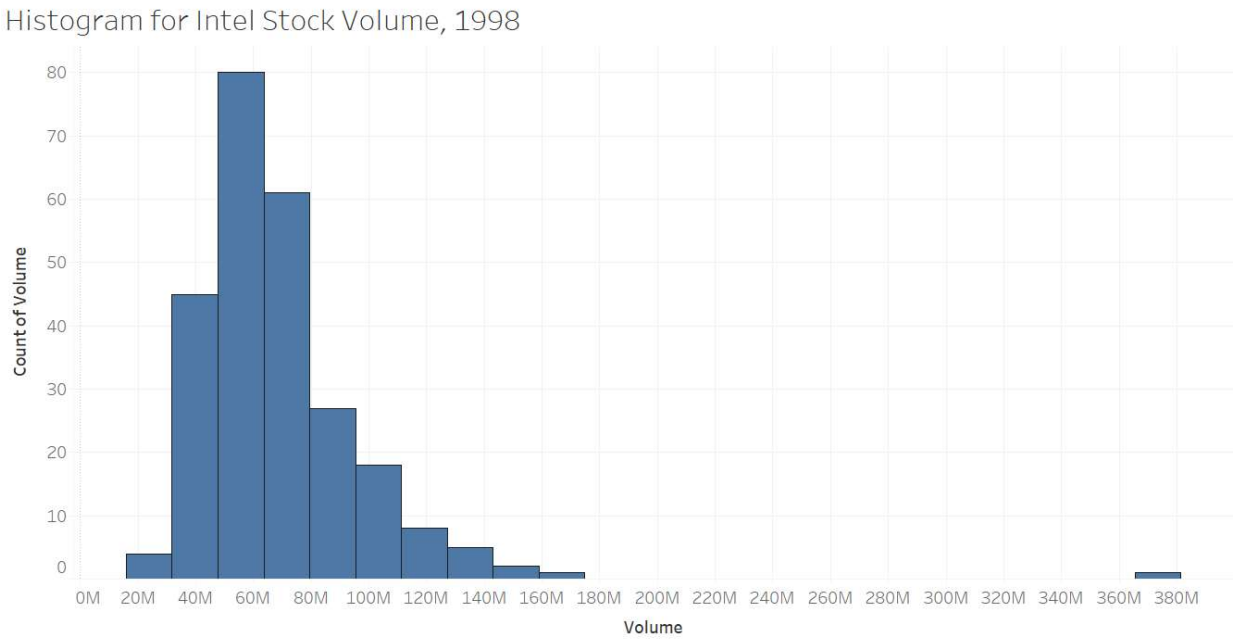
From the bar graph, what is immediately noticeable is the spike in stock volume in the first couple of weeks in March 1998, which is approximately 370M reflecting the largest volume traded that year. The smallest volumes are reflected in the last couple of weeks in November and December with a value of approximately 17M.



The plot of Volume as an attribute for Date. Color shows details about Date Month.

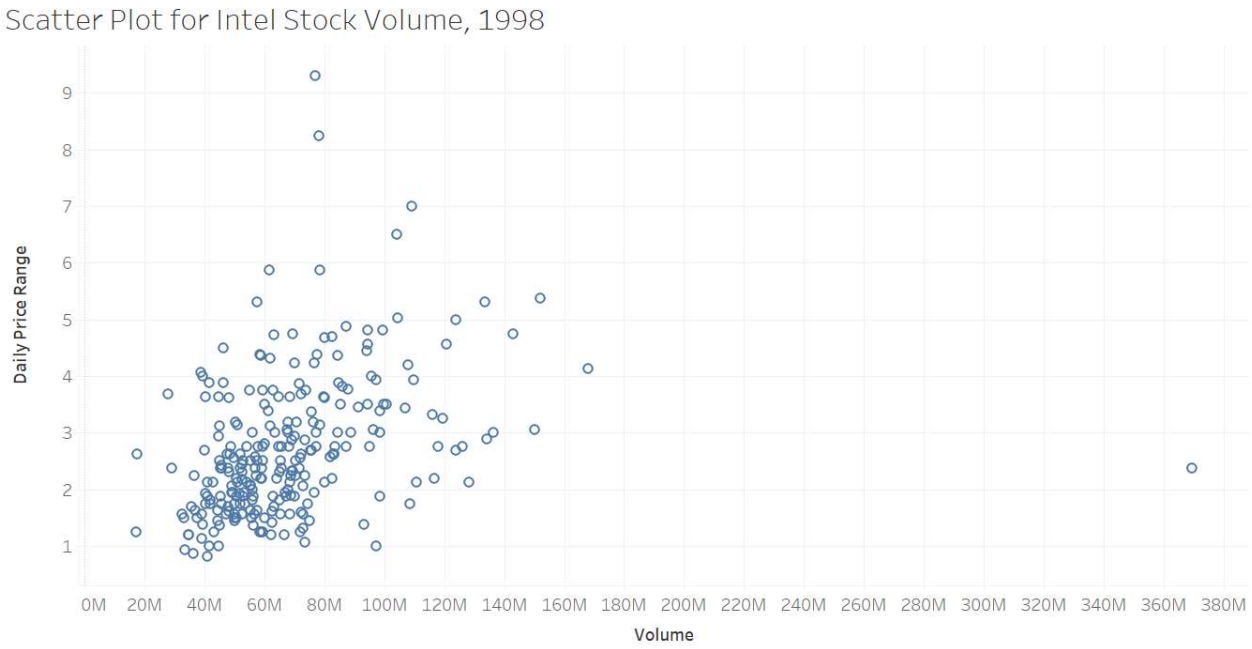
c. Volume histogram: The Columns use Volume as a dimension, and the Rows use count of Volumes within each bin. The default bin size was used because after testing out different sizes, the default option provided a cleaner and concise graph. The title and axes were relabeled for readability.

The histogram shows that due to outliers in the Volume variable, the distribution is slightly skewed to the right, which is shown as the spike in Volume from question 2b. The skewness is reasonable since there is a cutoff on the lower bound, and no real cutoff for the upper bound.



The trend of count of Volume for Volume .

d. Scatter Plot for Volume versus the Daily Price range. The Columns are the Volume, and a separate variable – "Daily Price Range" for the Rows was created using the Calculated Field in Tableau. The Daily Price Range is the difference between the High and Low (Price) for each day. The scatter plots were not filled to show the overlap between the various clustered points. The title and axes were relabeled for readability.



Volume vs. Daily Price Range.

The scatter plot above shows an upward trend between the volume of intel stocks traded and its price range on a given day. The higher the price range, the higher the volume of stocks traded, which could indicate a positive relationship between volatility and the daily price range.

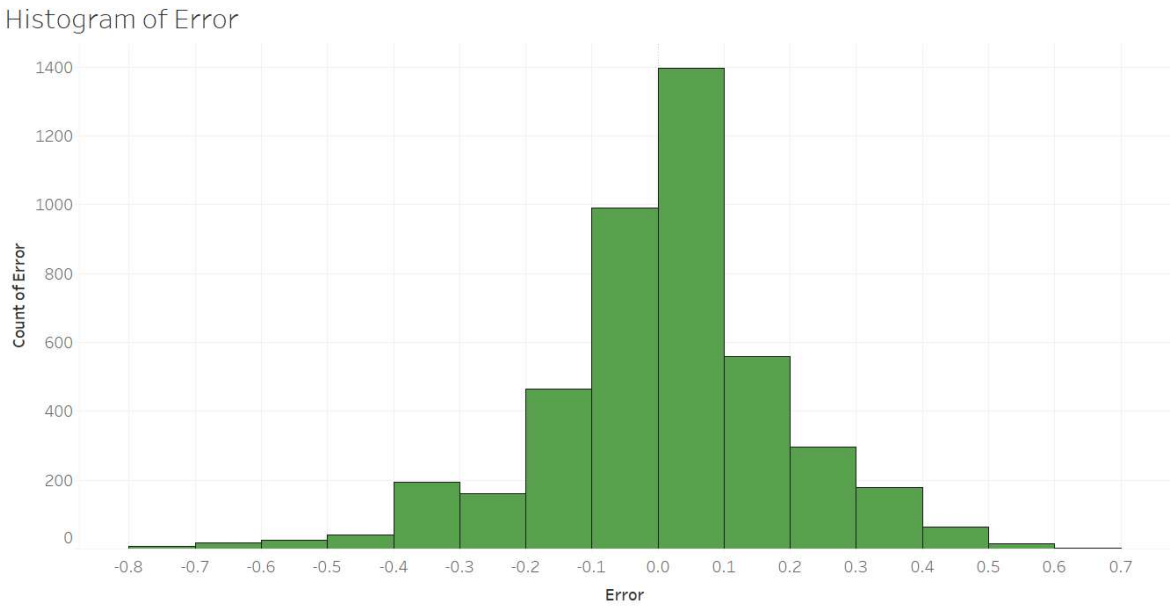
There is one outlier on the far right of the graph where the price range is moderate, but with a large volume traded on March 5, 1998. Without knowing the events of this day, it is difficult to state what caused this large increase in volume. However, such large volume trade doesn't seem to occur again in 1998.

Problem 3: Done in Tableau using PerceptionExperiment.csv

a. Error = Response – TrueValue
A new column was created for Error that took the difference between the Response values and the TrueValues. The Response values are the answers to the perception test provided by the students from the previous year. The TrueValue are the correct values that the Responses are judged against.

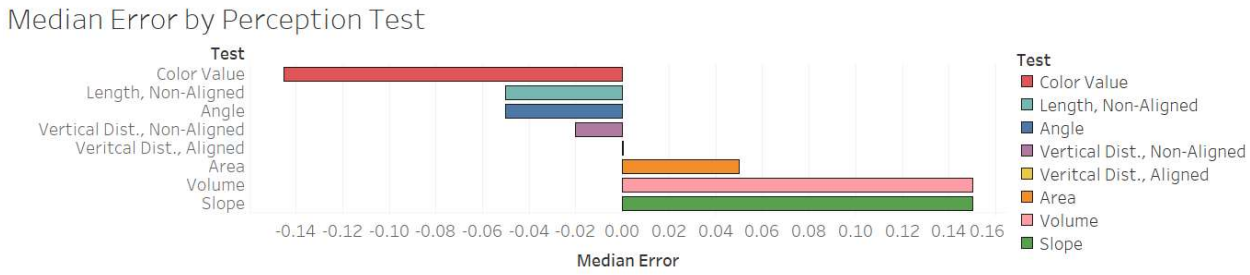
b. Histogram of Errors: The Columns use Error as a dimension, and the Rows use count of Errors within each bin. The default bin size was changed to 0.1 because the new bin size provided a cleaner and concise graph. The title and axes were relabeled for readability.

The Perception Experiment dataset has 4,416 observations and after creating a histogram of the 4,416 Error values, the histogram looks approximately normally distributed. The histogram shows the presence of potential outliers with the existence of minor tails to the left and to the right of 0.0 mark (center).



c. Bar graph of the median Error vs. Test: The Columns are the sorted median error value for each test in ascending order and the rows are the perception tests. The title and axes were relabeled for readability. Each color represents a different test.

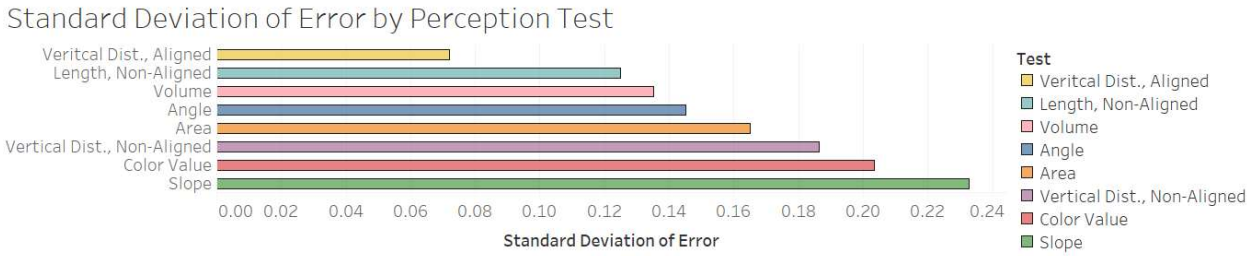
From the bar graph, we can see that of the perception tests, 4 tests are underestimated by students: Color Value, Length Non-Aligned, Angle, and Vertical Distance Non-Aligned. The largest underestimation comes from the Color Value, when a value of approximately -0.15. The three tests: Area, Volume, and Slope are the ones that students tend to overestimate. Volume and Slop have the highest median error value with an approximate value of 0.15. Vertical Distance Aligned reflects the smallest median error value – close to 0.



Median of Error for each Test. Color shows details about Test.

d. Bar graph of the standard deviation of the Error by Test: The Columns are the standard deviations of each tests error, and the rows are the perception tests. The bar graph was flipped horizontally to make it easier to see the tests on the categorical axis. Each color represents a different test.

From the bar graph, the highest standard deviation comes from the Slope test (the higher the standard deviation, the greater the spread of how widely subjects varied in their response against the TrueValue). The lowest standard deviation value comes from the Vertical Distance, Aligned with a standard deviation of approximately 0.075.



Standard deviation of Error for each Test. Color shows details about Test.

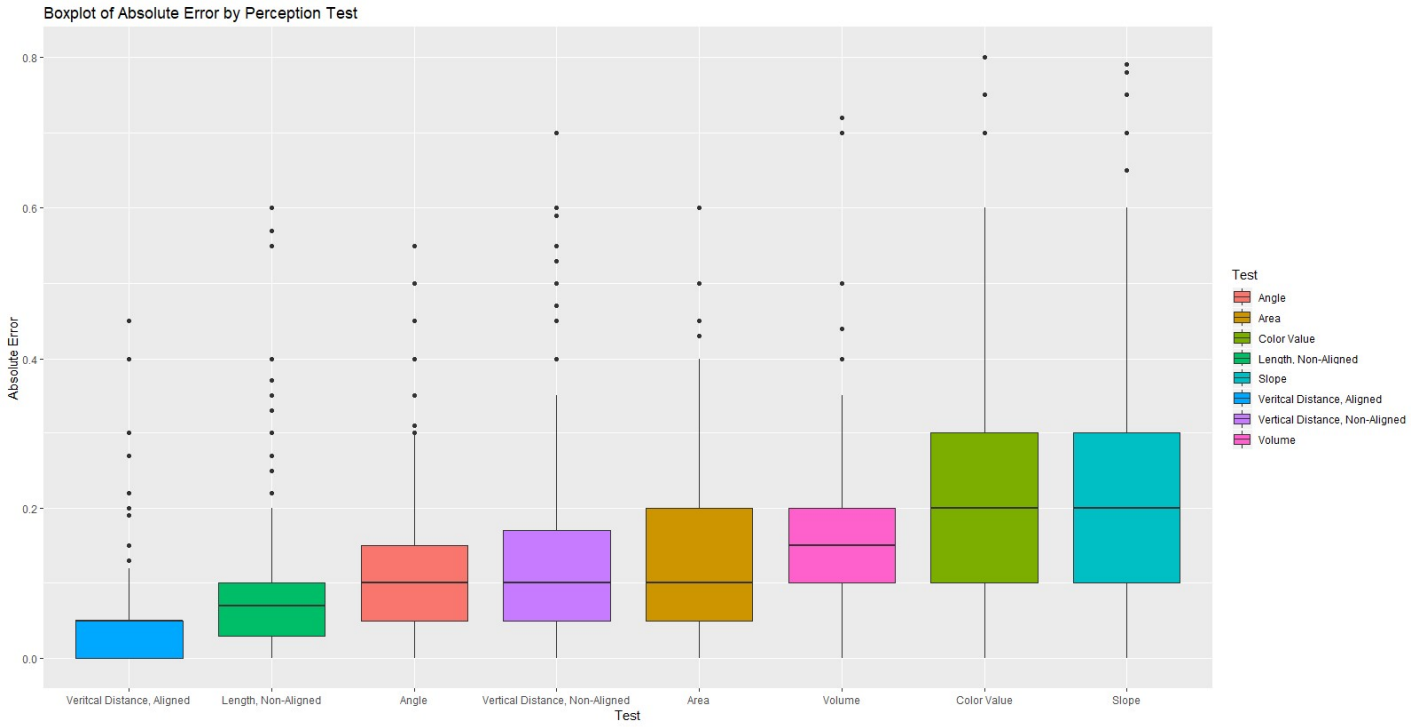
e. Boxplot of Absolute Error using RStudio

Code:

```
ggplot(perception, aes(reorder(Test,abs(Error)), y=abs(Error), fill=Test)) +
geom_boxplot() +
labs(title="Boxplot of Absolute Error by Perception Test", x="Test", y =
"Absolute Error") + coord_flip()
```

The boxplots are updated with different colors to reflect the perception tests and are put in ascending order of the absolute error values. From the boxplot, we can see that all the boxplots have extreme outliers, which are shown as points that extend past the whiskers. What's interesting is that the boxplots appear to be mirroring the order of the tests from standard deviation bar graph. From the boxplot, we can also see that the range for the Color Value test and Slope test are the largest amongst the perception test. Lastly, another point to mention is

that while the Errors histogram did not reflect skewness, the individual boxplots illustrate that some tests do exhibit skewness. This can especially be seen in the Vertical Distance Aligned test and Area test.

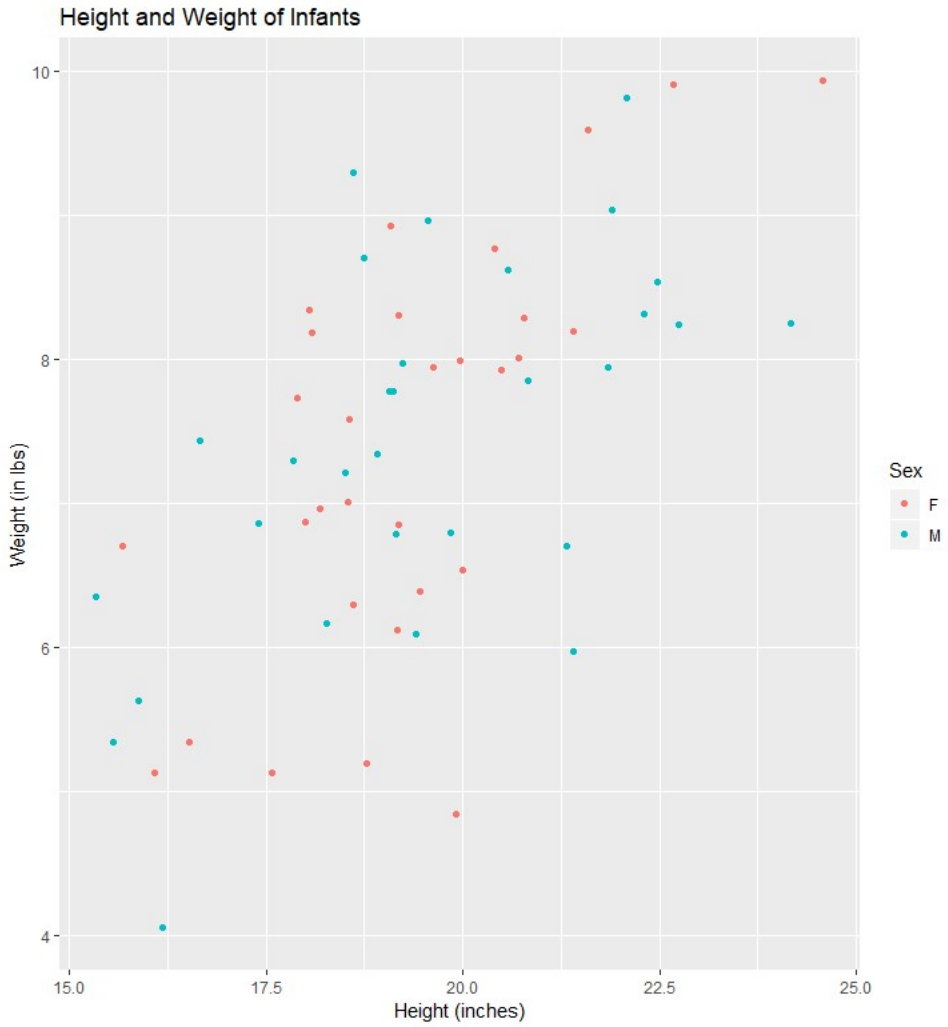


Problem 4: Done in RStudio using InfantData.csv

a. Scatter plot of Height (inches) and Weight (lbs): (Discussion about the graph is presented in 4c)

Code:

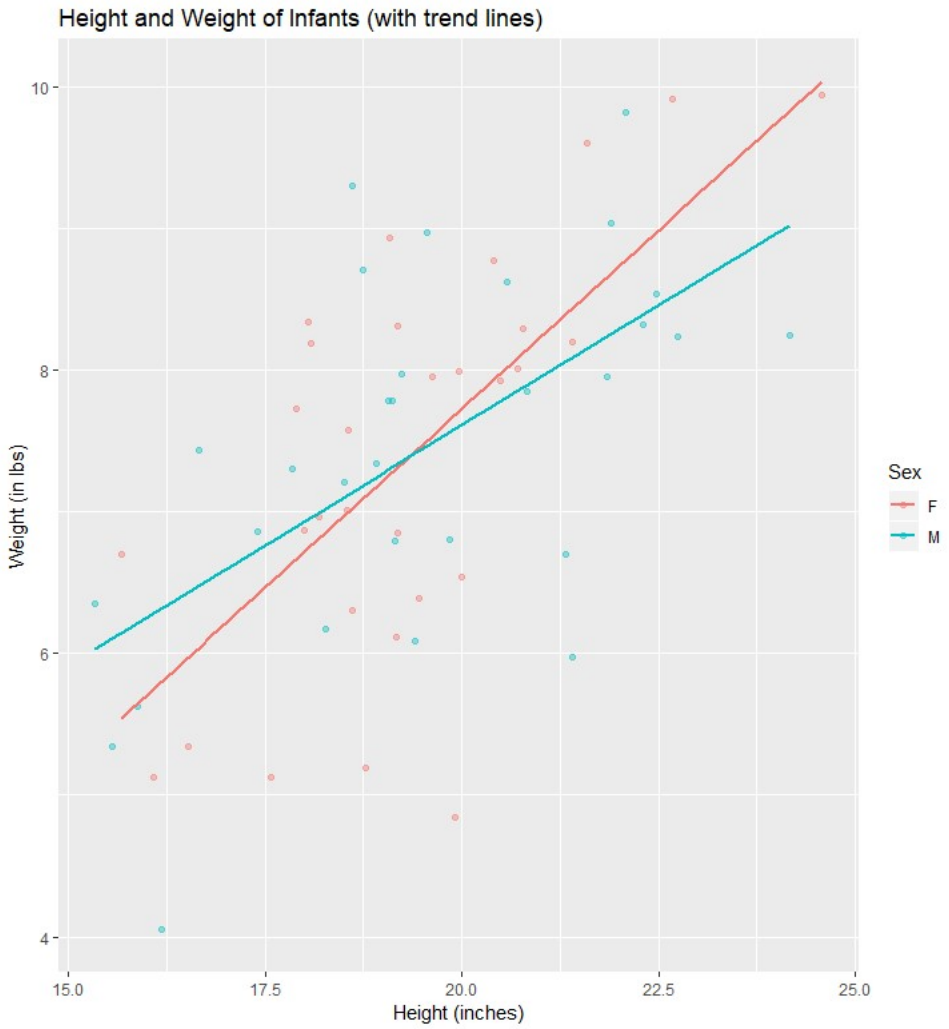
```
ggplot(InfantData, aes(x = Height.in, y = Weight.lbs, color = Sex)) +
  geom_point() + labs(title="Height and Weight of Babies", x="Height
  (inches)", y = "Weight (in lbs)")
```



b. Scatter plot of Height (inches) and Weight (lbs) with trend line: (Discussion about the graph is presented in 4c)

Code:

```
ggplot(InfantData, aes(x = Height.in, y = Weight.lbs, color = Sex)) +
  geom_point(alpha=.4)+ geom_smooth(method = "lm", se = F) +
  labs(title="Height and Weight of Babies (with trend lines)",
     x="Height (inches)", y = "Weight (in lbs)")
```

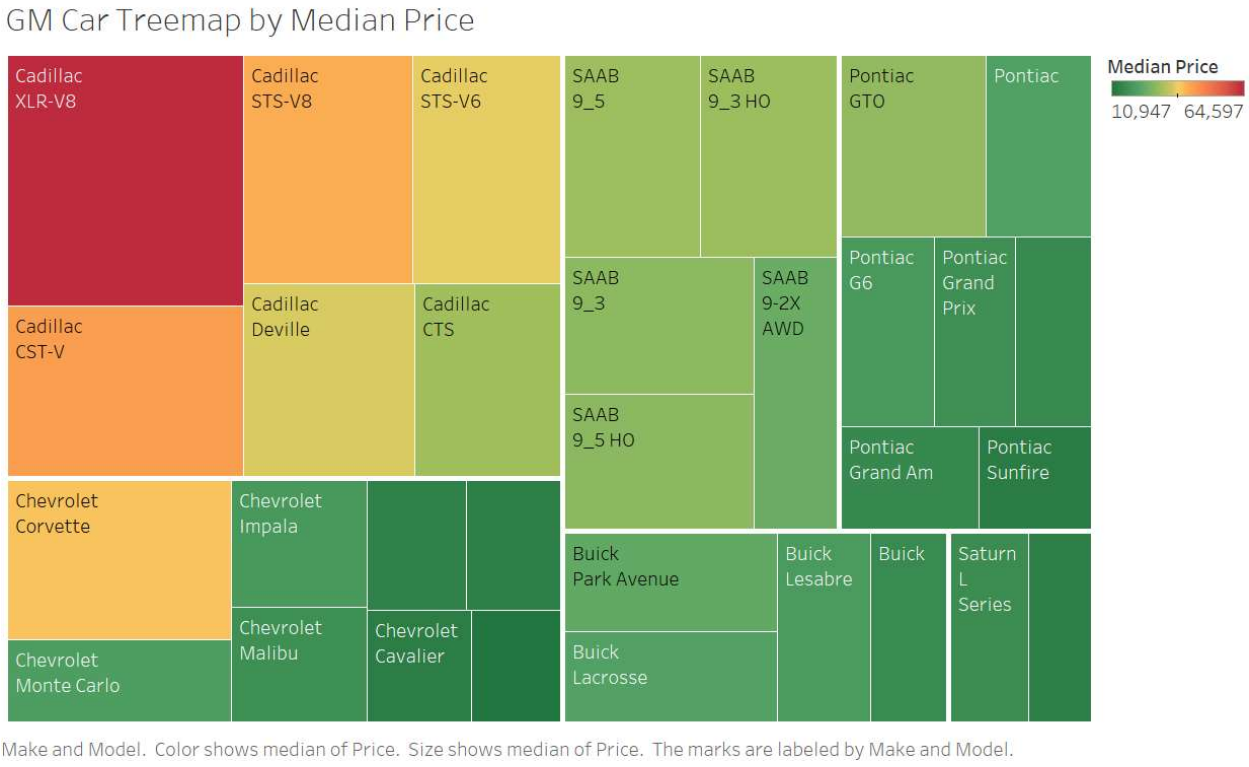


c. For the Height and Weight Graph of Infants, prior to adding the trend lines, while we can see that there is a positive relationship between Height and Weight for both populations, however, it isn't clear what kind of impact sex has on the relationship between the two variables. After adding the trend lines for the two populations allowed for a better understanding of the relationship between Height and Weight considering the sex of the infants. What becomes clear is that the slope of the Female trend line is steeper than that of the Male trend line. Until the equilibrium point where Height is approximately 19.5 inches, at the various height measurements, Male infants weigh more than the Female infants. However, past the equilibrium point, Female infants begin to outweigh the Male infants for the same height.

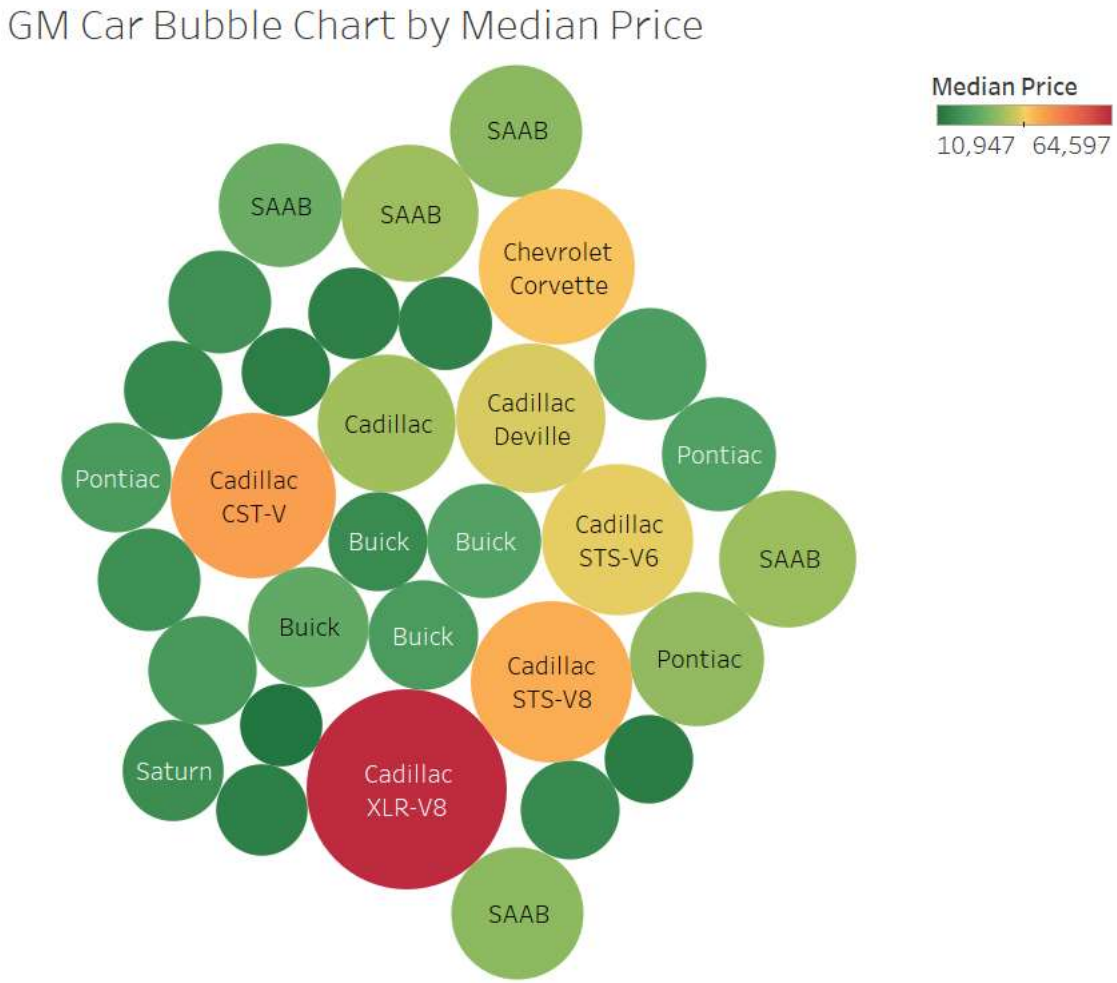
Additionally, the graph does not start at zero because the data doesn't have values close to the zero-lower bound. And unlike bar graphs, the zero is not necessary as it doesn't contribute to the graph and its representation of the data.

Problem 5: Done in Tableau using gmcar_price.txt

a. Treemap: The GM car treemap is based on the median price of the different makes and models. The main subdivision is the Make of the car and the minor subdivision based on the Model. The color legend on the right side displays the median price from lowest to highest and their respective color. The color choice was purposefully designated as red-gold-green diverging to show that low prices are green, and as prices increase, the boxes change to gold and then to red. The objective of choosing this color theme was to put emphasis on any critical points.



b. Bubble Chart: The GM car bubble chart essentially has the same criteria as the treemap, except for the type of graph that is used to convey the data.



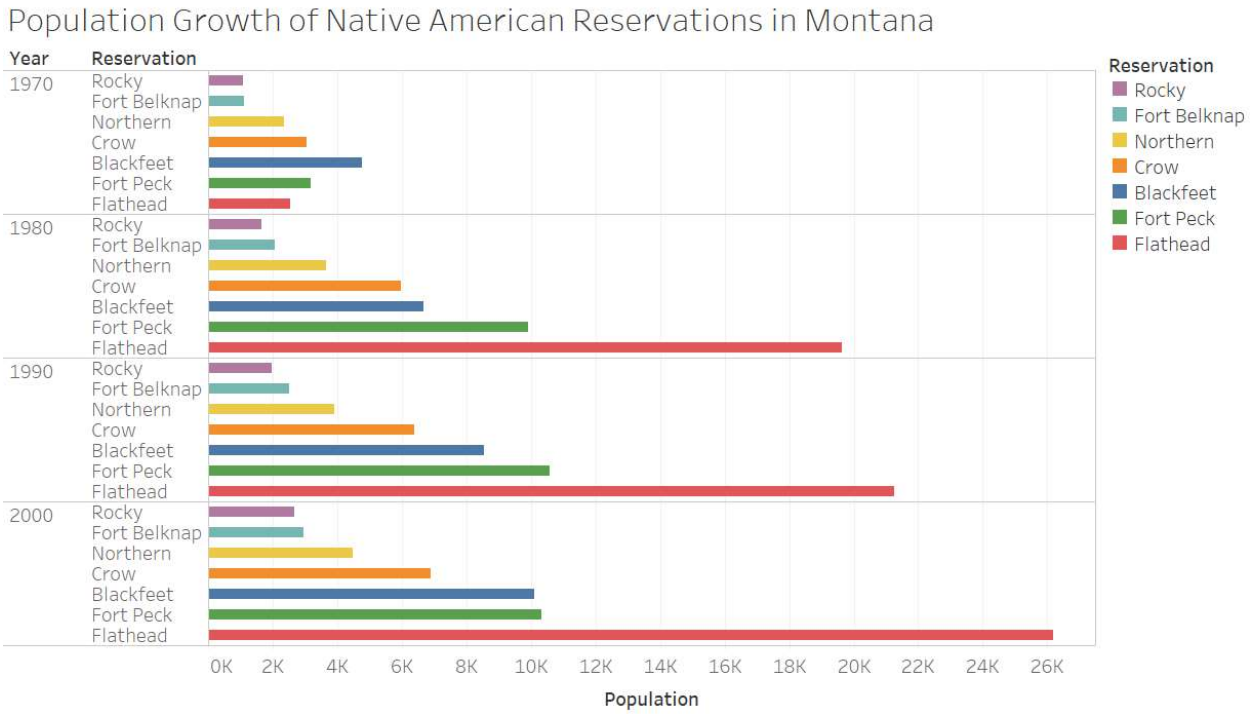
Make and Model. Color shows median of Price. Size shows median of Price. The marks are labeled by Make and Model. The view is filtered on Make, which keeps 6 of 6 members.

c. Describe the differences between the two plots

The treemap displays a larger number of the make and models of the car than the bubble chart through optimizing space. There is less information provided in the bubble chart due to the size limitation of the bubbles, so some bubbles only display the make of the car or no information at all. Both the treemap and bubble chart use color effectively to "illustrate" pricing for each model and make of GM cars. Just from the color range, immediately we see that the Cadillac XLR-V8 is the most expensive of the group of cars in the dataset, followed by the Cadillac CST-V and STS-V8. What the bubble chart is slightly more effective in doing that the treemap had a harder time conveying is the color scale for the Cadillac CST-V and the Cadillac STS-V8. In the bubble chart, the CST-V is slightly more orange than the STS-V8. In the treemap it is slightly harder to tell. This may very well be the case because the color progression in more subtle since the colors are not staggered like the bubble chart. However, overall, the treemap was more effective in summarizing the data than the bubble chart.

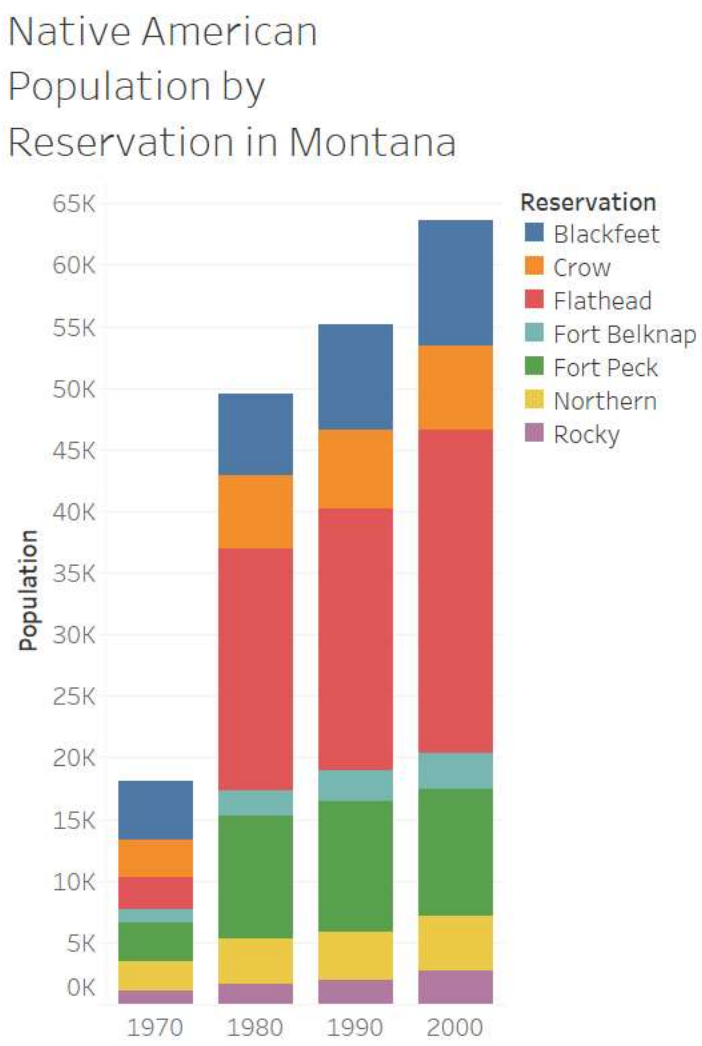
Problem 6: Done in Tableau using reservation70-00.csv

a. Population Growth



The bar graph displays the continuous population growth for 7 Native American reservations starting from 1970 for the next three decades. The bar graph is shown horizontally for the purpose of displaying the names of each reservation and avoid clutter. The x-axis displays the population, and the y-axis displays the years.

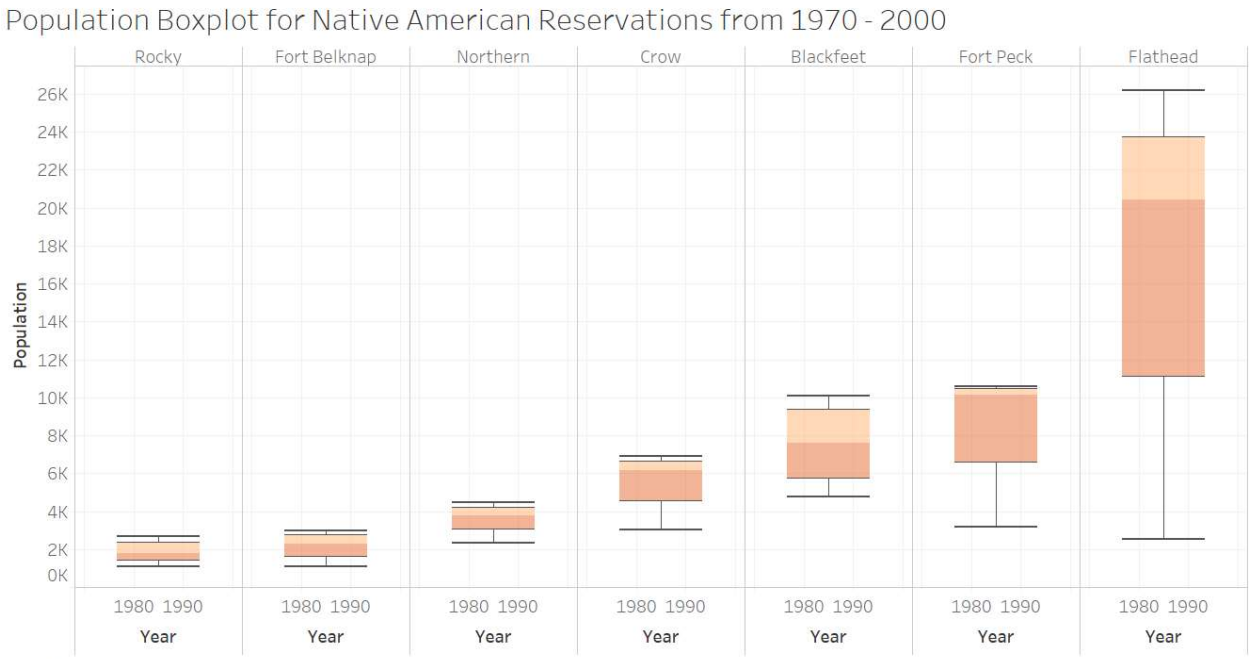
b. Total Reservation



The stacked bar graph displays the total reservation population for each year subdivided into each reservation's population with the sort order based on ascending population total. The colors allow for quick distinction between the different reservations.

An interesting thing that I have noticed between the bar graph in a) and the stacked bar graph in b) is what the different graphs convey. With graph a), what you first notice is how the individual reservations change in population over time. Immediately you notice that the Flathead reservation had the largest growth in comparison to the other six reservations. With the stacked bar graph, you not only see the breakdown of each reservation and how they make up the entire reservation, you noticed the large population change from 1970 to 1980. While the same information can be gathered from graph a), it isn't the first thing that comes to mind. It's interesting to see how graphs may use and present the same data, but different presentation styles can change the way that the reader interprets the data (granularity versus big picture).

c. Boxplot



The boxplot displays the population distribution vs. years for each reservation. The x-axis is are the reservations, and the y-axis is the population. The four values for each reservation are summarized by the boxplot. The box-and-whiskers for each reservation shows the distribution of population values at that location during this overall period.

From the graph, what we can see is that except for Fort Belknap and Blackfeet, the five other reservations are showing skewness as their median is not aligned with the mean. The largest population spread is from the Flathead reservations, which is supported by graph a).

Problem 7

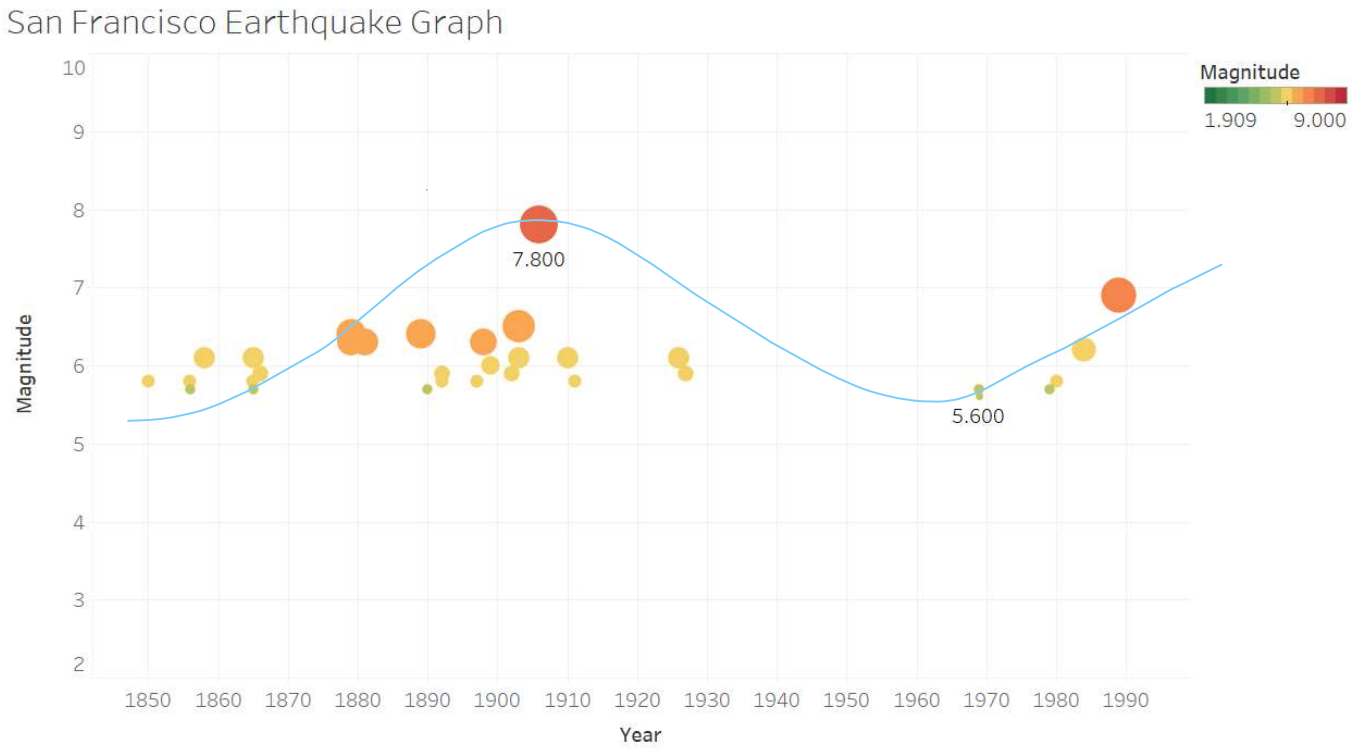
Three issues with the graph:

1. The gratuitous 3D aesthetic used on the graph is unnecessary. It doesn't add to the story that the graph is telling. Because of the 3D aesthetic, the graph appears cluttered.

2. The tick-marks that have the year and earthquake magnitude written inside each box are redundant because the axis already displays the year (time line) and the magnitude is both presented as the size of the box and again on a color scale. Squeezing print onto the tick marks and then the overlap causes the reader to attempt to read each box and lose sight of the story.

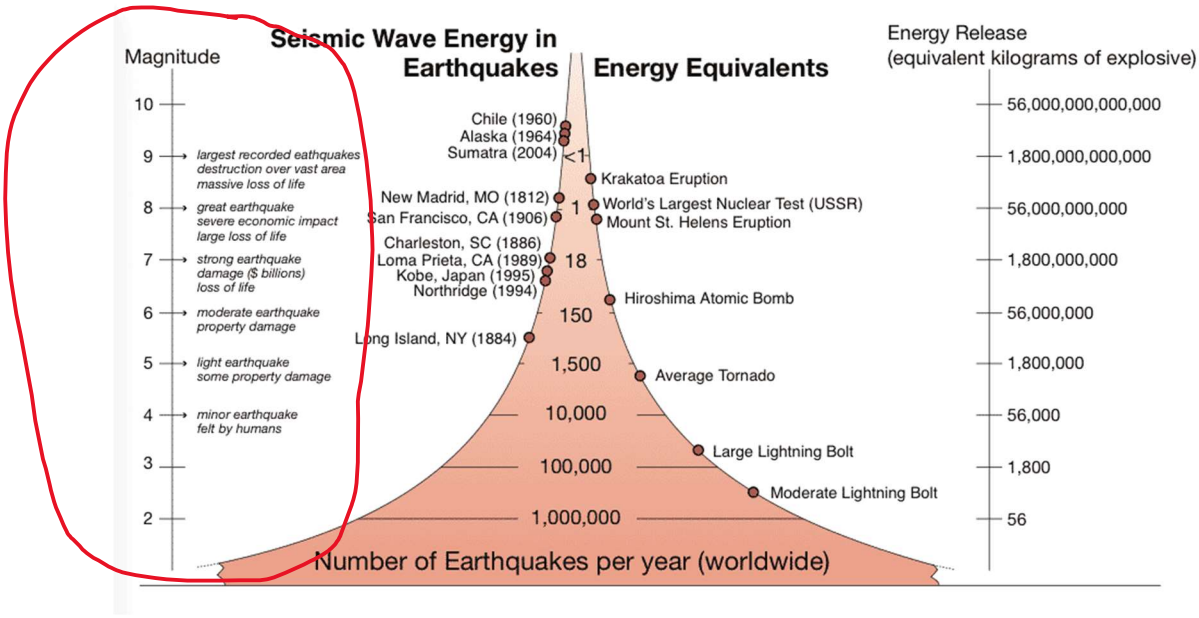
3. The third issue with the graph is lack of a scale that that explains what changes in magnitude (based on the Richter scale) mean because going from a magnitude of 6.0 to 7.0 is not proportionate.

Proposal: Use scatterplot where the plots vary in size of magnitude; the larger the magnitude, the larger the bubble and coordinate that with a color scheme as well. Using a red-green-gold diverging seems appropriate to illustrate the dangers and concerns of higher magnitudes. Instead of labeling every single scatter plot with the respective year and magnitude, reference the highest and the lowest as reference points without all the clutter. The addition of the frequency line essentially summarizes the caption on the bottom of the 3D graph – there are predictions for a large magnitude earthquake in the near future. Lastly, on the last page, there are two examples of what could be used as a legend to help explain the effects of a magnitude change.



Year vs. Magnitude. Color shows details about Magnitude. Size shows details about Magnitude. The marks are labeled by Magnitude and Year.

Source of picture: https://earthquake.usgs.gov/learn/topics/mag-intensity/



Source of picture: Google: earthquake magnitude scale