Cynthia Cho
DSC 465
Due: May 3, 2019

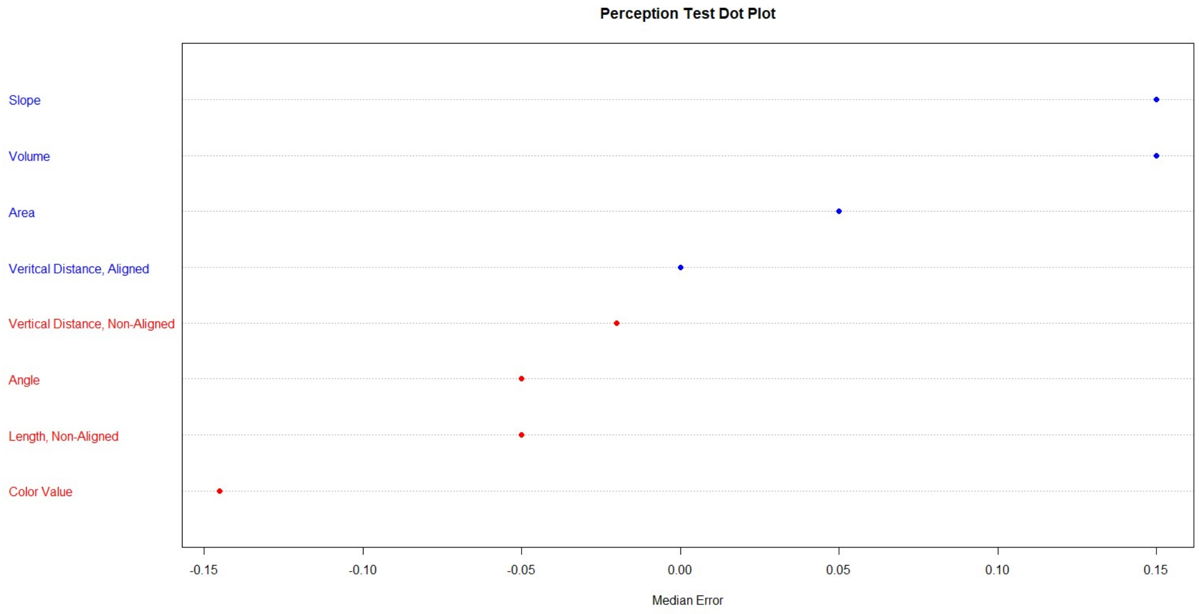
Homework 2

Problem 1
Read Cleveland sections 3.1-3.13 (Not to turn in)

Problem 2
In-class participation for lecture 3. Analyze three graphs for their visual inaccuracies or misrepresentations.

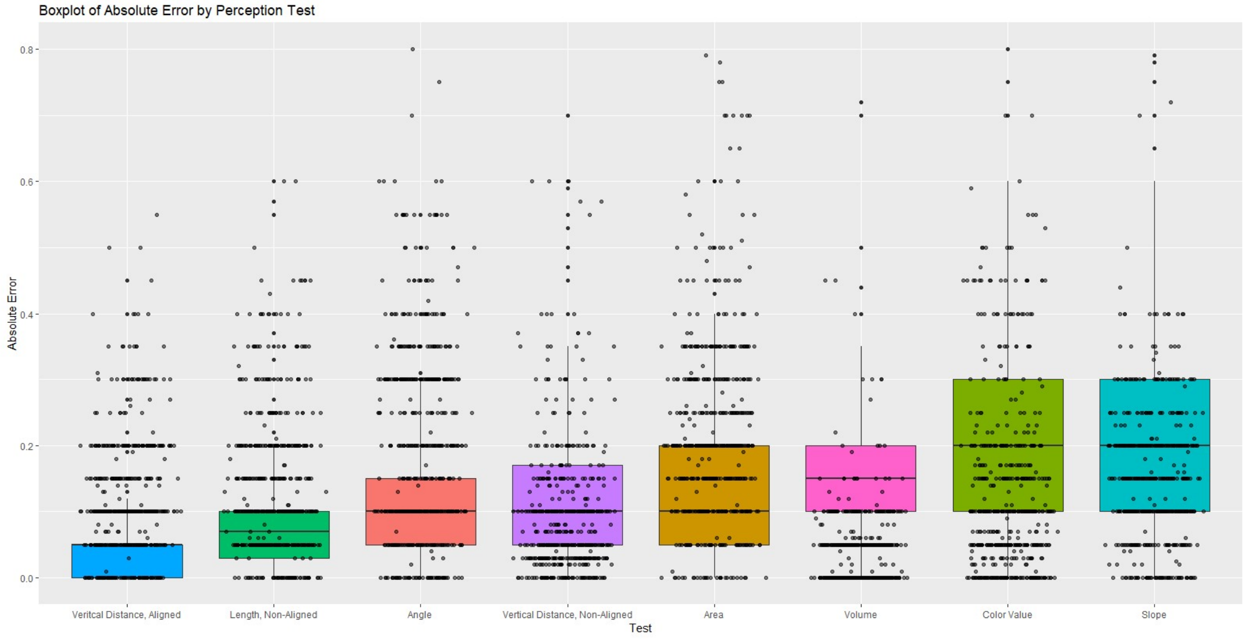
Problem 3: Done in RStudio using perception data except problem 3f. Code used for RStudio is attached separately along with the homework

a) Dot chart/dot plot for Perception Test
Both the dot chart and bar chart allow us to see how the median Errors are distributed for each test. The dot chart provides a clean visual and efficiently uses data ink to deliver the context of the graphics by using a dot to represent each categorical value. And because of this, it is an ideal graph to use when there are many categories. The bar graph on the other hand can be used as a grouped bar chart or stacked bar chart. However, they are not suitable for larger datasets with a lot of categories. For our homework, both charts were able to present the data clearly.



b) Perception Test Absolute Error Box Plot with Jittering: the normal distribution was used for jittering (see code attached to appendix)

The boxplot is from homework 1 problem 3e with a jittered categorical overlay. The Errors histogram from homework 1 problem 3 gave us a visual of the distribution and exhibited a normal distribution. The boxplot provides a visual of the summary statistics for each perception test and allows for easier comparison across the tests. The jittering overlay gives us an idea of the distribution for each test. While the boxplot is good at providing more detail than a histogram, it is unable to provide context on the distributions. From the jittering overlay, it looks like some of the tests may have bimodal/multi-modal distributions, something that we wouldn't have known by just looking at the boxplot. So adding the jitter overlay to the boxplot provided additional context to the graphics (an idea of the distribution with the summary statistics). Also, there appears to be clumping that can be seen from the graph starting from 0.0 that increments every 0.05 (5%).

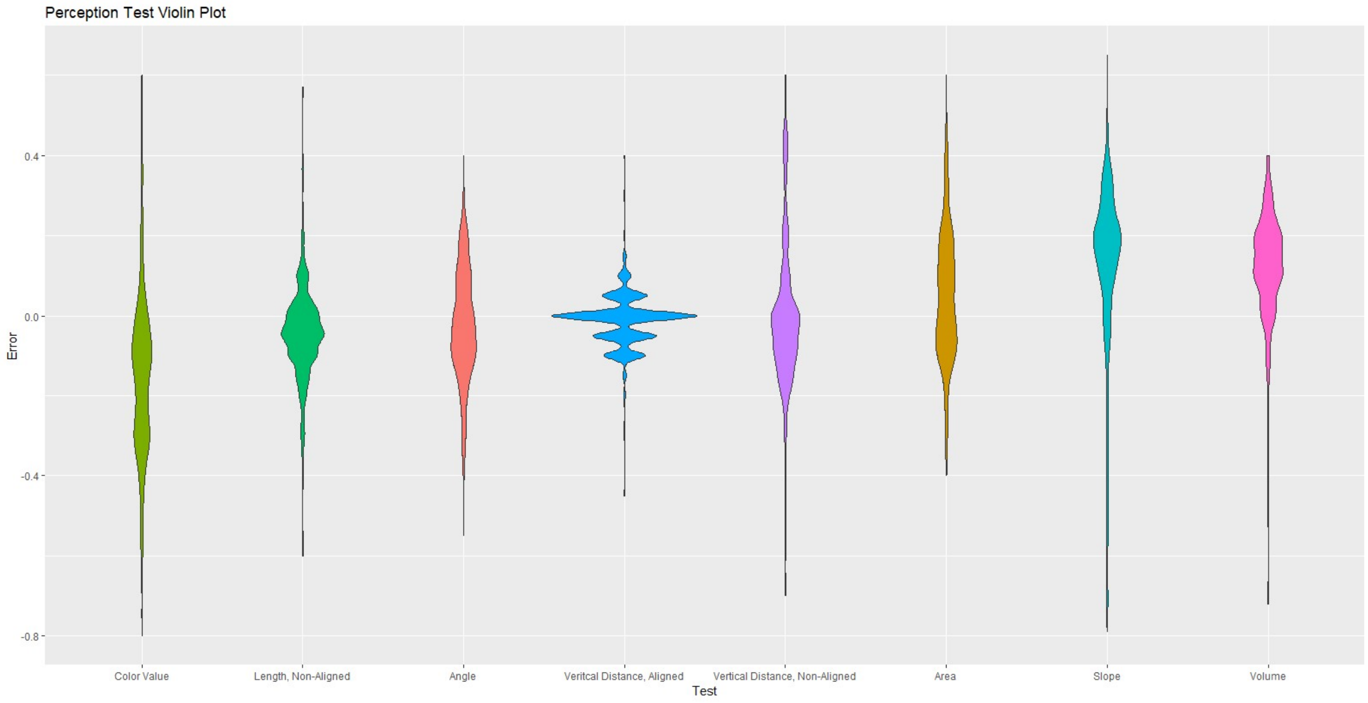


c) Extra Credit: The data shows quantization in the responses on multiples of 5%. The possible harm is the data distortion that can be taking place. With the qualization occuring for every 0.05 increment, it looks as through each increment is a cluster, which can cause a misreading of the distribution by associating each cluster as a distribution peak.

d) The violin plot explores the the error field with respect to each perception test. Because each test is categorically colored, the legend was removed.

The following tests people underestimated the data: Color Value, Length – Not Aligned, and Angle. The following tests people overestimated the data: Slop and Volume. Based on the distribution, the other two tests, Vertical Distance – Aligned and Area appear to have a somwhat even distribution around 0; however it is a little challenging to tell since we don't know what the median is by just looking at the violin plots. What we can see is that the distributions show signs of kutosis and skewness. What is also noticeable from the plots is the Tests are not normally distributed; several of them are bimodal/multimodal.

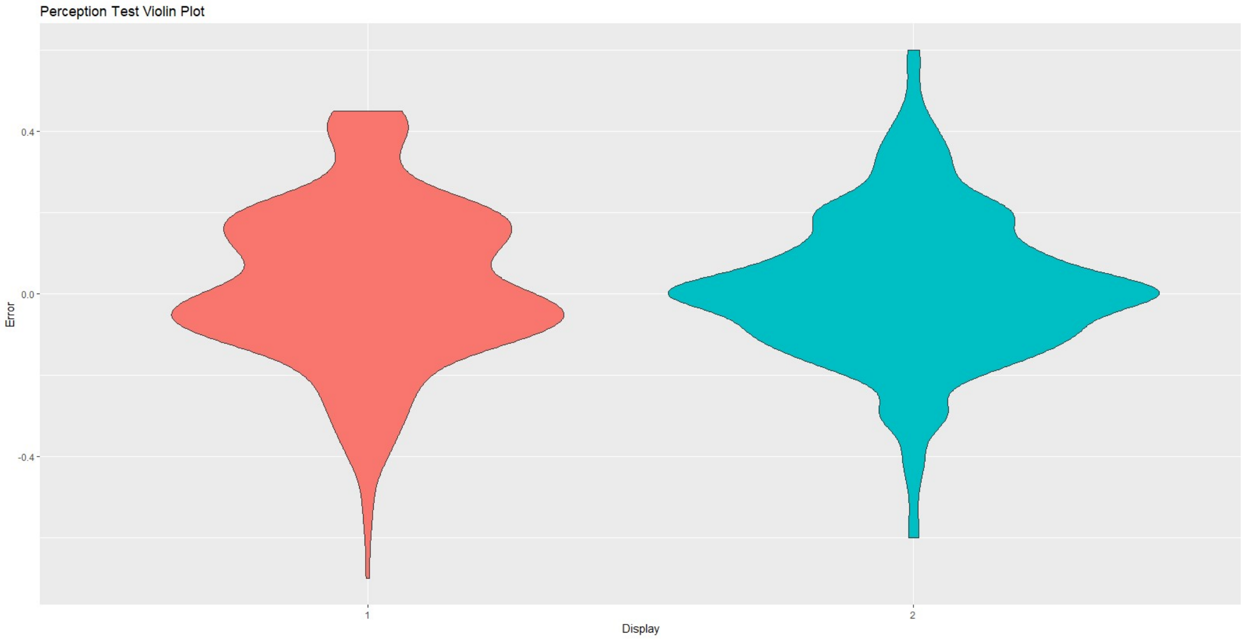
Adding a jitter plot to the violin plot would be redunant since the violin plots already provide us with the distribution. And from what've seen, because of the quantization, it would cause more focus on the quantization rather than the shape of the distribution.



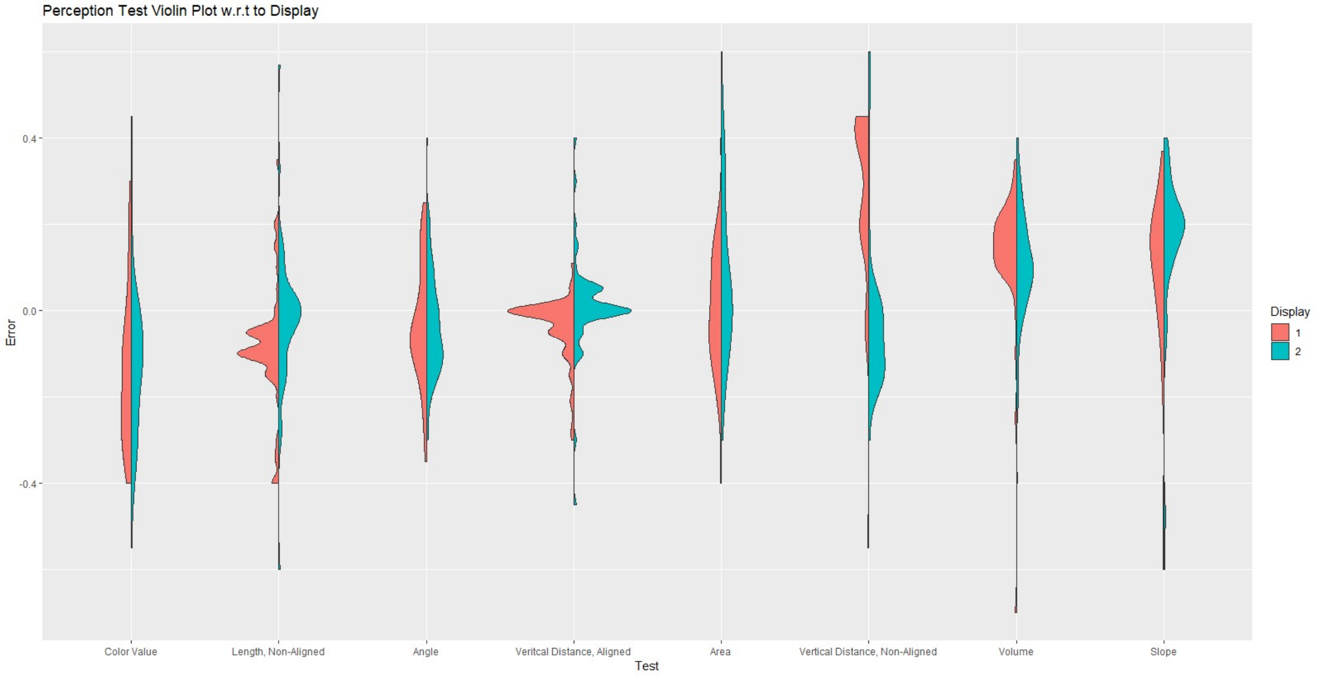
e) In part i, the violin plots are broken out by Display. As the Error approaches 0.4, there is a trunctation that takes place for Display 1. For Display 2, the plot gradually tapers off around 0.45. What is also different between the two Displays is the distribution; Display 1 has a multimodal distribution, and for Display 2, the particpants must have improved their assessment because the peaks are not as pronounced.

In part ii, the Vertical Distance, Non-Aligned shows a stark contrast between Display 1 and Display 2. The split violin reveals that the particpants severely overestimated the Trials (B, C, D) for Display 1, while for the Trials in Display 2, the particupants did not particularly over or underestimate the test. The trunctation that takes place in Display 1 for the violin plot in 3ei) is also present in the split violin. While the difference in the split violin plot could indicate that the participants' assessments are improving from Display 1 and Display 2, it is not consistent with how people are performing on the other Tests/Displays. The other tests do not exhibit a change in their judgement– this should be investigated to see what is causing the change.

i. Violin Plot for Display 1 and Display 2



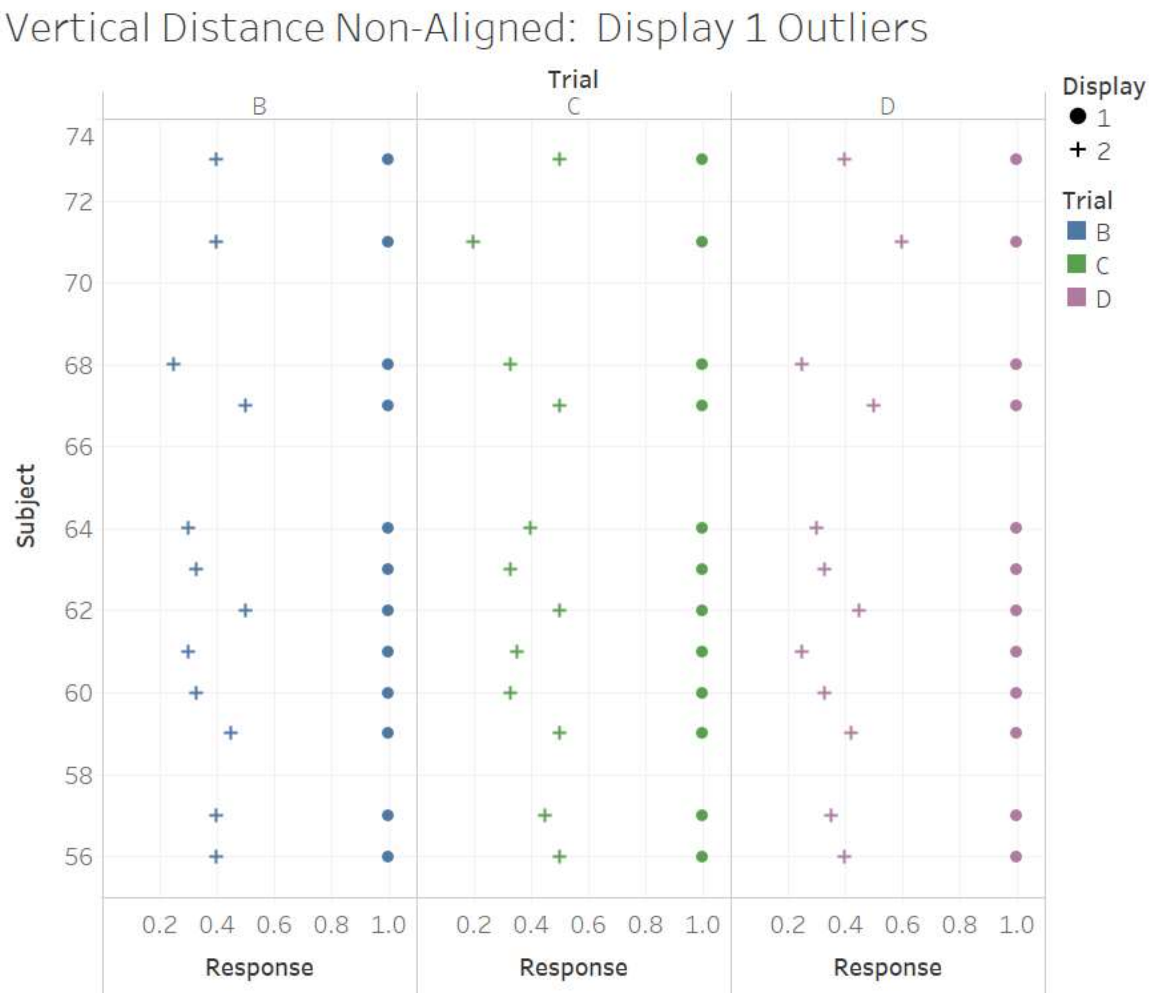
ii. Violin Plot with Violin Splits



f) Two graphs are presented in this problem: The first graph is of both Display 1 and 2 for the Vertical Distance, Non-Aligned Test focusing on the Subjects under observation. The second graph is an overall graphic that shows all responses from all subjects for the Test, limited to Display 1. This is where the pattern is most noticeable.

After looking into the excel file with the raw scores, there were Subjects ("participants") that responded with a value of 1 for all three Trials (B, C, and D) in Display 1; the participants were mostly in consecutive order. While a value of 1 is a valid response since it's based on a participant's assessment, it appeared to be less a perception issue, but rather an issue with the Display because of the exhibited pattern from the graphs below. Because the responses are of an erroneous stimulus, the outliers should be removed from the analysis because it distorts the data and is inaccurate.

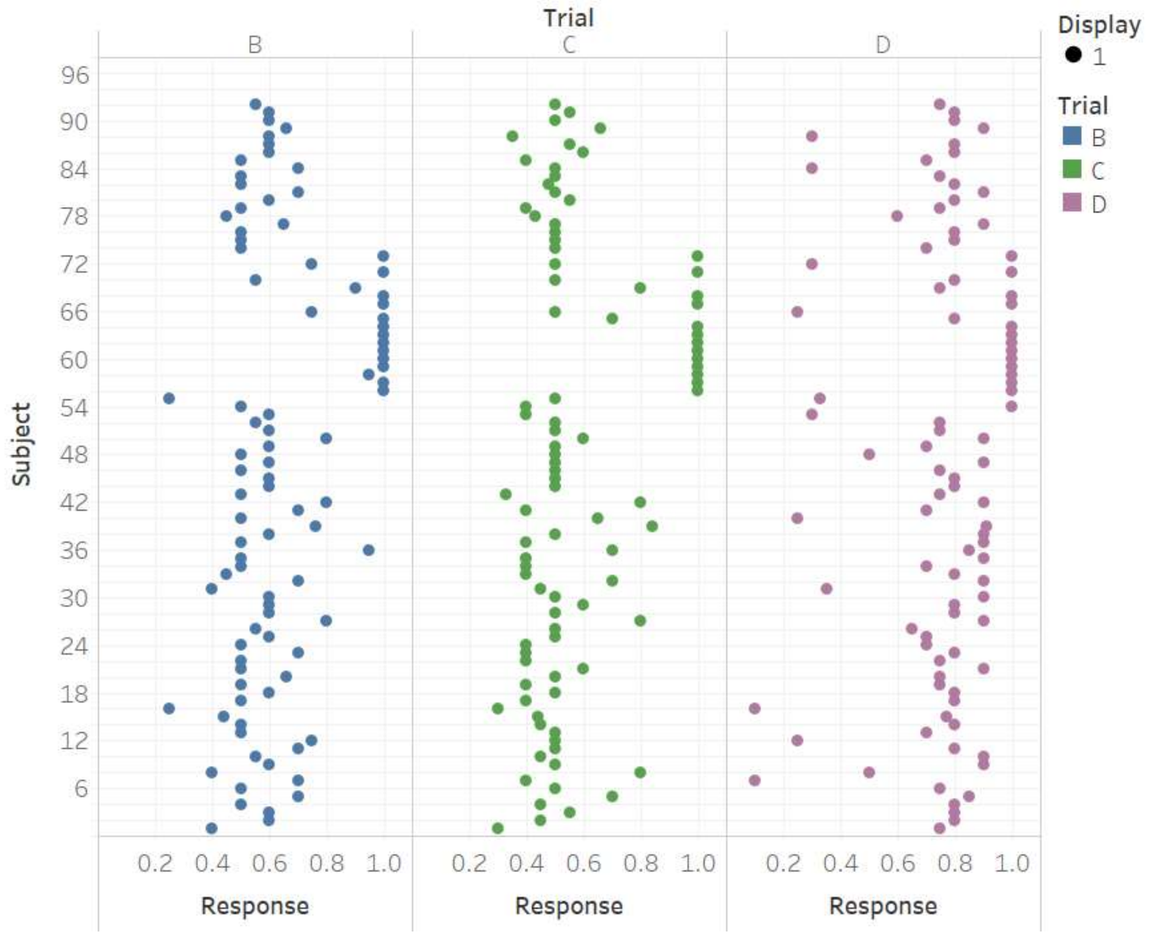
The graph shows a collection of scatter plots narrowing in on the responses of Subjects 56-73 with respect to each Trial for the Vertical Distance, Non-Aligned test. The Display is coded by shape, and the Trials are coded by Color. The responses for Display 2 Trials vary within approximately 0.2 and 0.6, whereas the Trials for Display 1 all have a response of 1. Based on what can be seen here, it appears that a different stimulus must've been used for Display 1 for the subjects where the response is 1. The results are outliers and are inconsistent and distorted and therefore should not be used in the analysis.



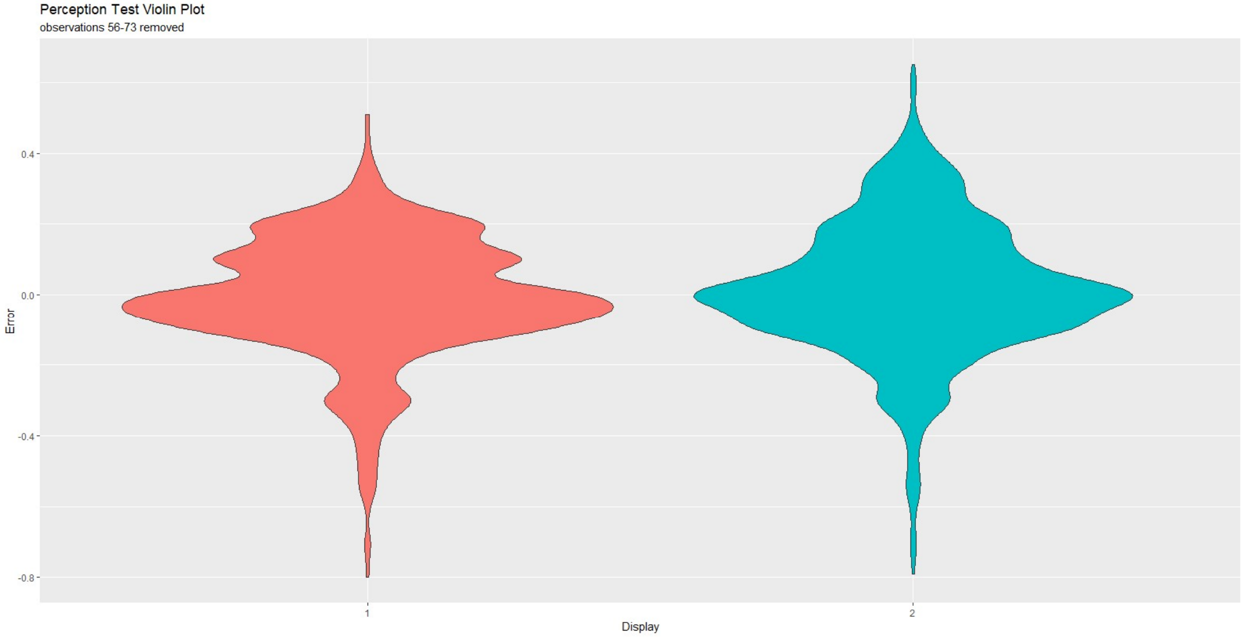
To further pin-point the problem and why the outliers should be removed, a scatter plot collection of all subjects for only Vertical Distance, Non-Aligned, Display 1 is also provided.

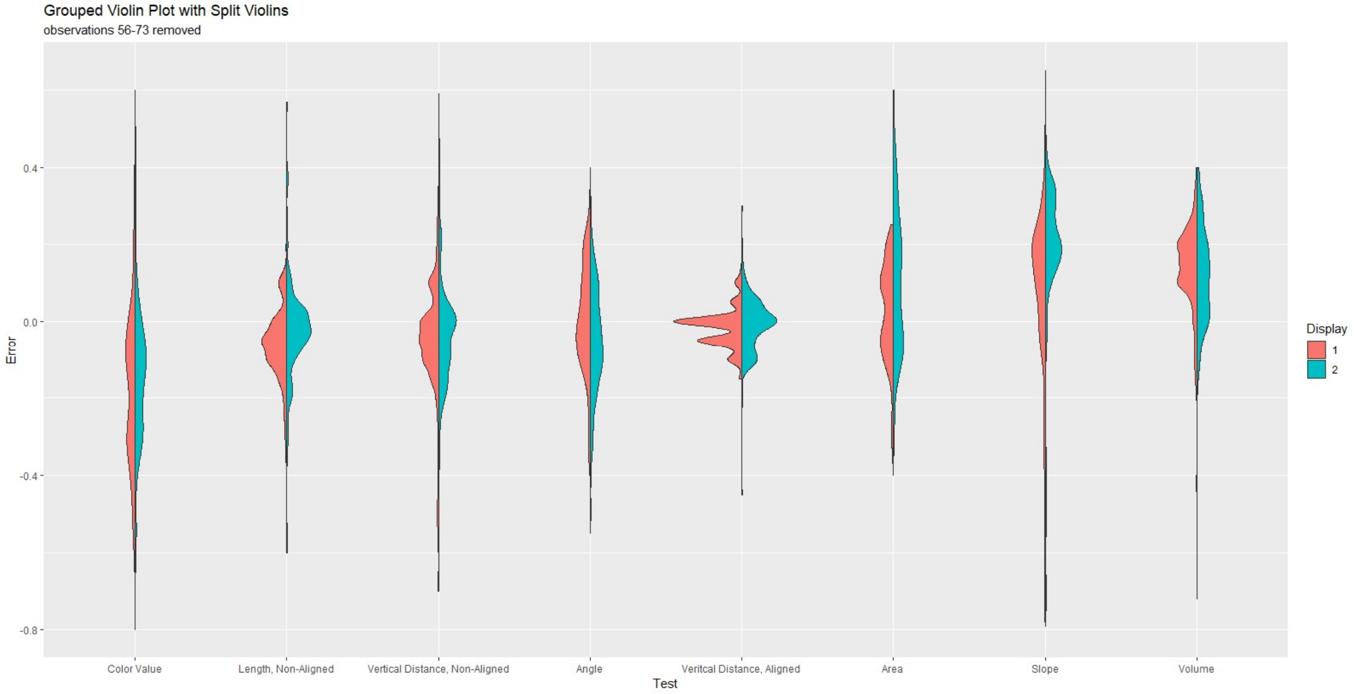
When all the subjects' responses for the Vertical Distance, Non-Alighed Test, Display 1 are graphed, we can see the pattern that exists for the faulty responses. We expect to see variations in responses for all Trials, and while we do see that in most of the subjects, there is a cluster where all the subjects responded with a 1. The responses vary within each Trial and against each other, except for observations 56-57, 59-64, 67-68, 71, and 73. The listed observations all have the exact same response for each Trial for all of Display 1. Reiterating what was said before, it would appear the Display 1 used for the specific set of Subjects were different and should not be a part of the analysis, otherwise it's comparing apples to oranges.

Vertical Distance Non-Aligned: Display 1 Outliers
(with all observations to illustrate pattern)



g) With the outlier overservations removed, the trunctation that happened in Display 1 around 0.4 for part ei) is now gone. The errors are tapering off at the tail end because the errors of the outliers were generating a high difference between the True Value and their response (“1”). And from the split violin plot, we can see that the Errors for Vertical Distance, Non-Aligned do not have a significant swing in estimation from Display 1 to Display 2, which is consistent with what we have been seeing for the other Tests. Overall, removing the Subjects shown in 3f) for the Vertical Distance-Non-Aligned, corrected the distortion.

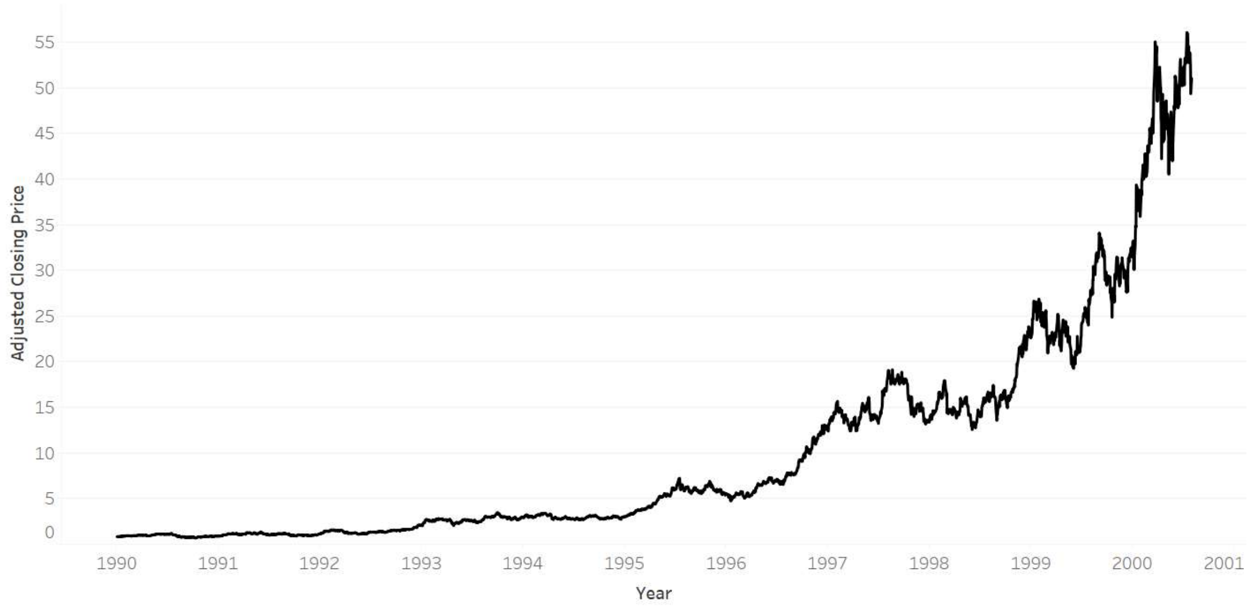




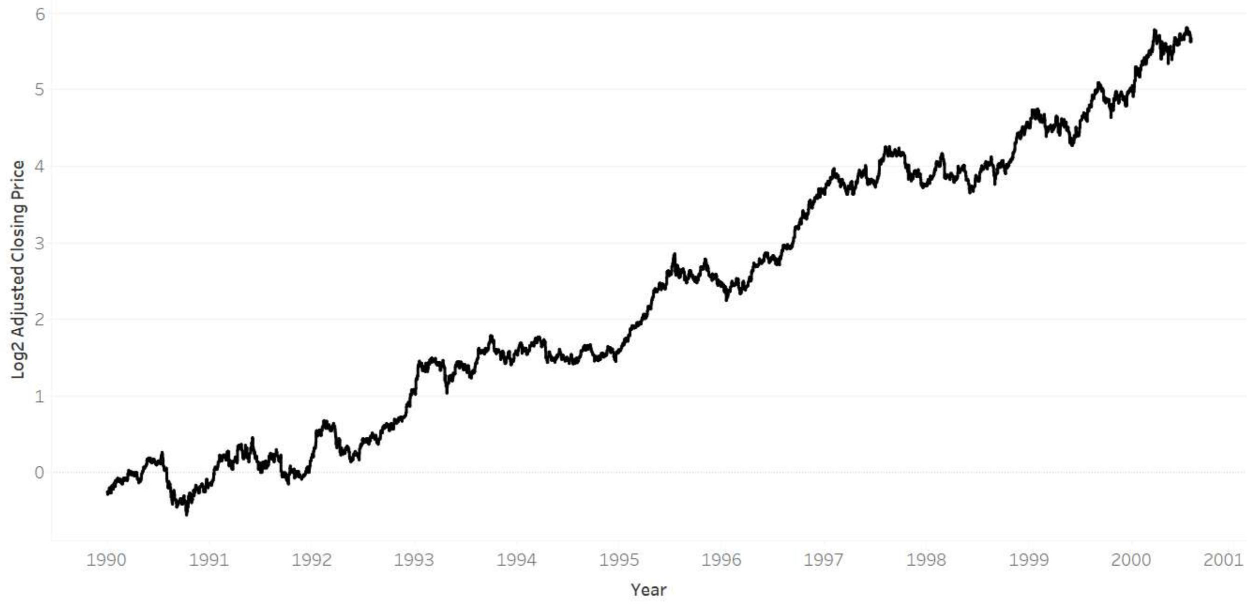
Problem 4: Tableau

a) The logarithm scale improves our ability to read the graph; especially for points laying close to the x-axis. The scaling also reduced extreme ranges of values in the data, shows change in data using percentage, and it reduces the size of the numbers on the axis – it changes the scale, but not the graph itself. By using log2 to adjust the price on the second graph, we can see how the price changed over the course of 10 year indicating a constant rate of growth.

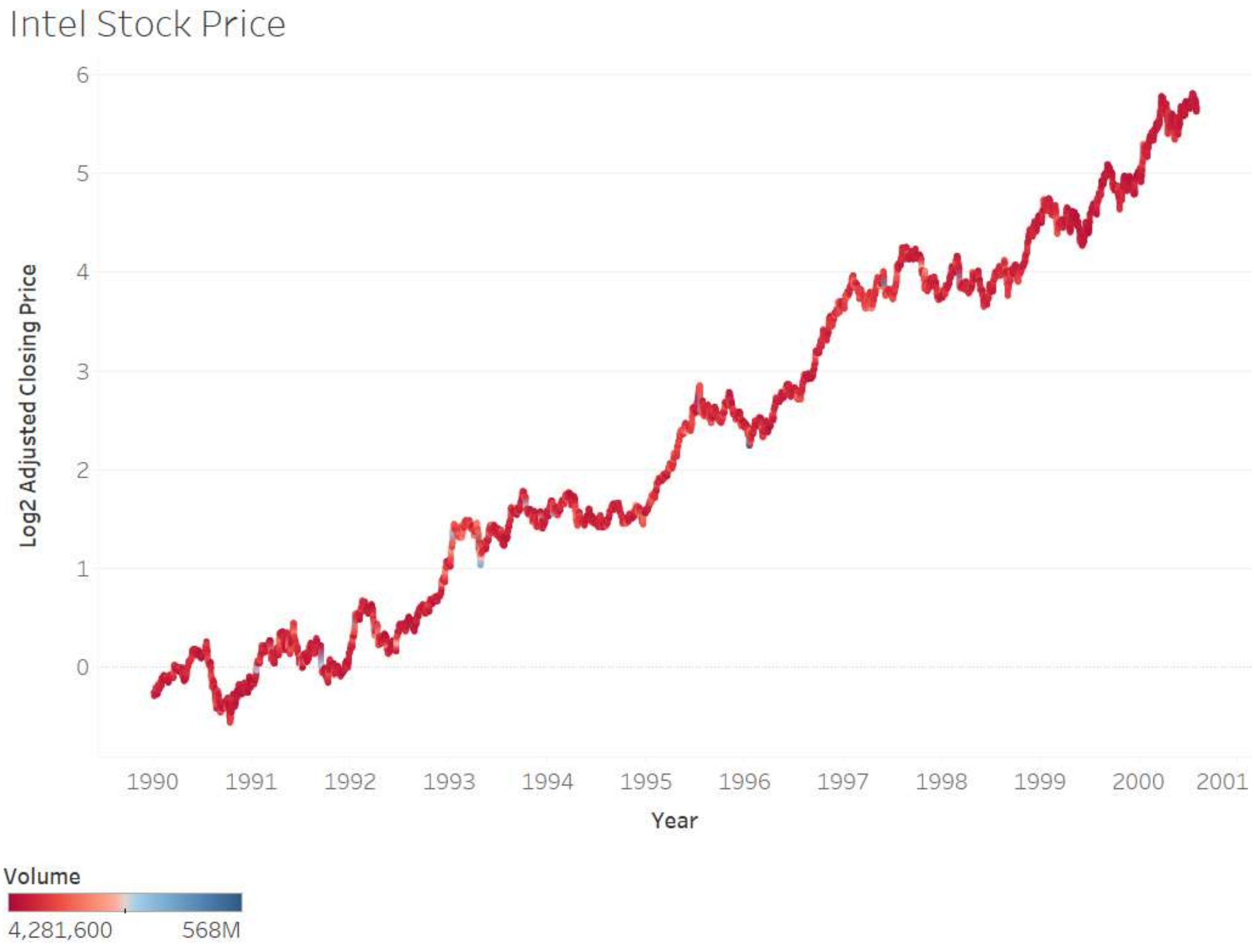
Intel Stock Price



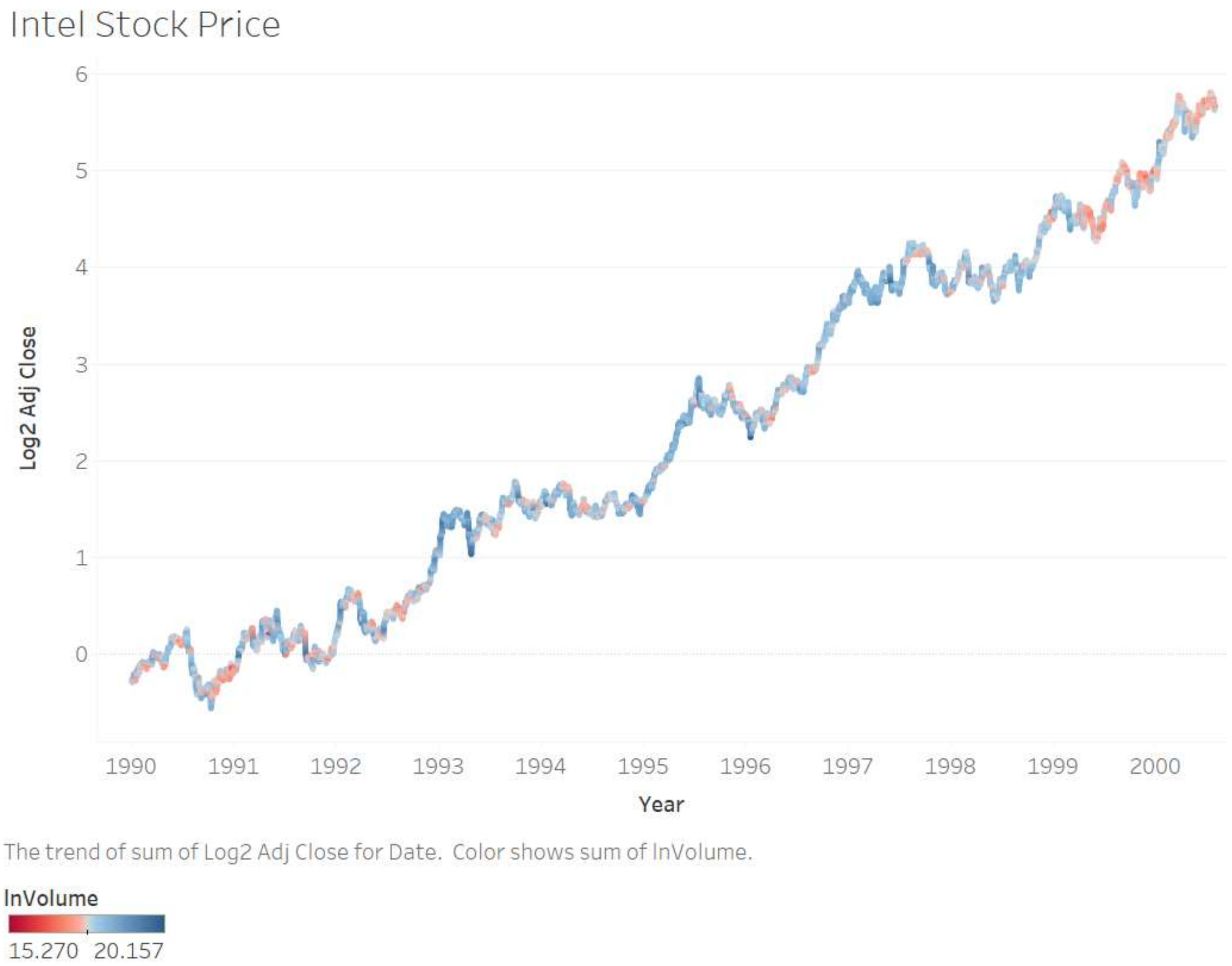
Intel Stock Price (logarithm)



b) Using a standard line graph, to graph Date against the log2 Adjusted Closing Price, Volume is an added parameter. The Red-Blue gradient was chosen for Volume to highlight the occurrences of the large Volume trades (blue). Unfortunately, the wide range in volume overshadows the fewer occurrences of large volume trades.



c) A calculated field for "logVolume" = ln(Volume) was created and used for the same graph from 4b). Because volume was scaled and now has a range of 15.270 - 20.157, we can see both the small and large volume trades. Scaling the Volume has the same effect as scaling the price in problem 4a) by reducing extreme ranges. In comparison to graph 4b), the movement in both price and volume can be seen; there appears to be a positive relationship between stock prices and stock volume. By adding a third parameter to the Intel Stock Price line graph and using it as a color feature, the logVolume provides additional context to the graph.

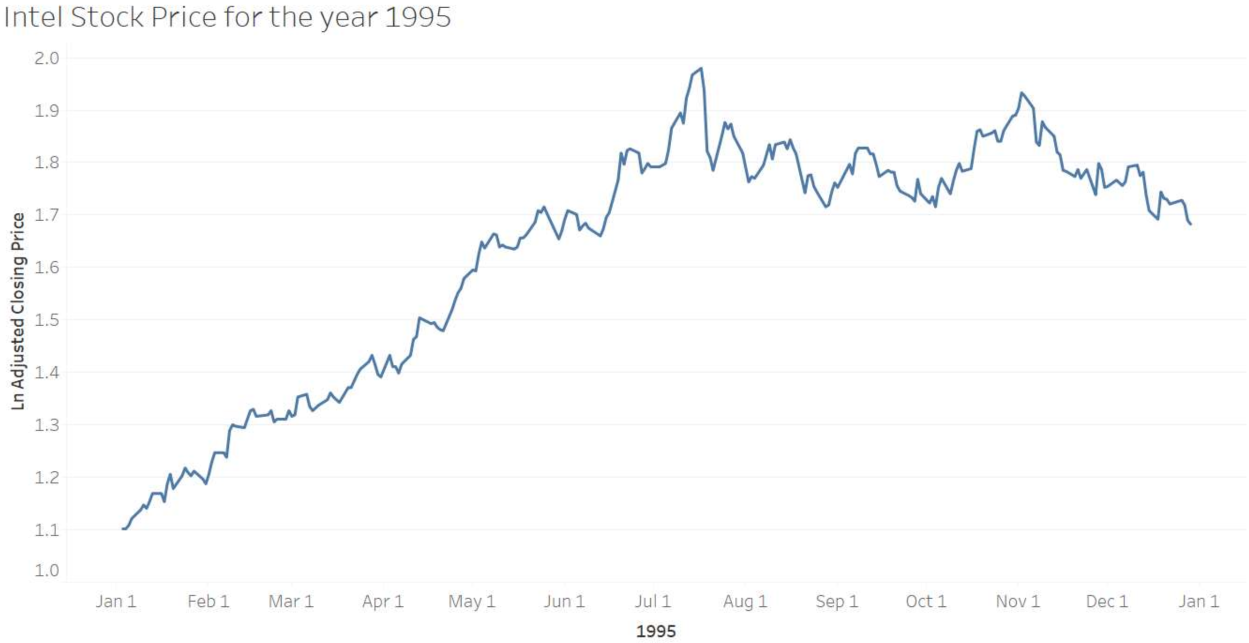


Intel Stock Price

The trend of sum of Log2 Adj Close for Date. Color shows sum of lnVolume.

d) After filtering for the year 1995, the three surges in price between 10% - 20% occur between:

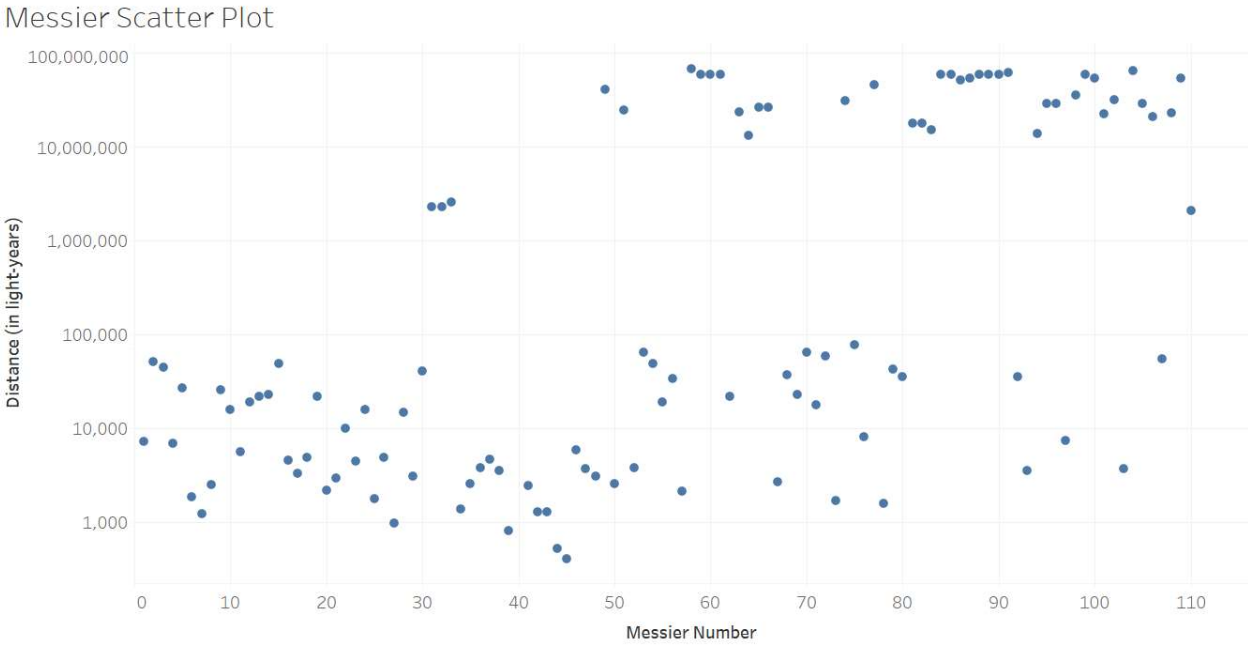
1. End of April to early May (April 21 – May 2), there is a 12% increase in price
2. Mid-June towards the end of June (June 13 – June 20), there was approximately a 15.8% increase in price.
3. Early July to mid-July (July 3 – July 10), there was approximately a 10% increase in price.



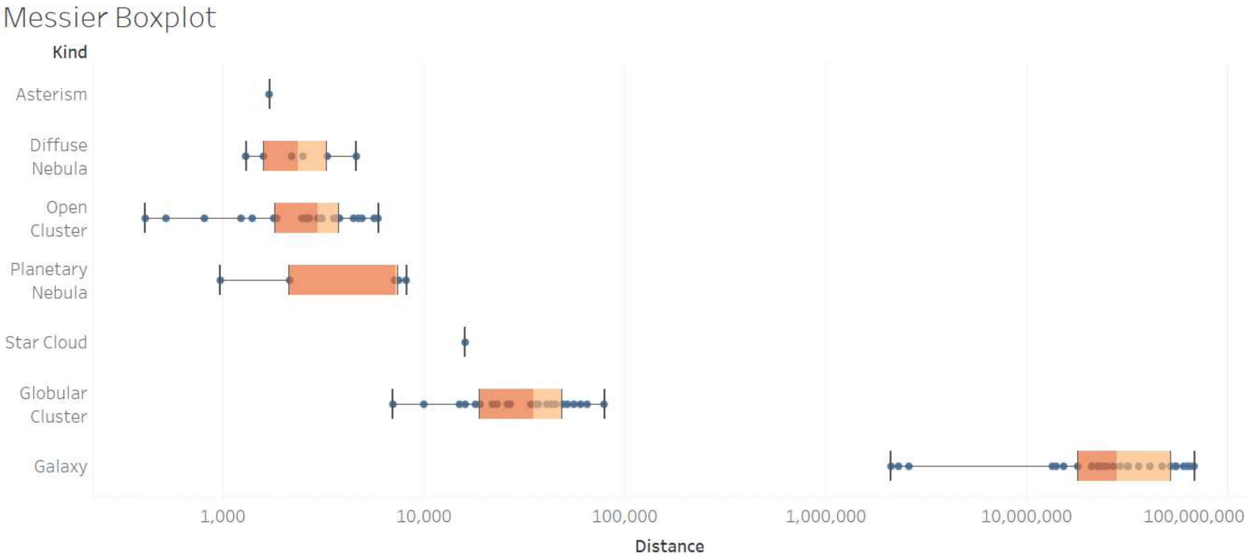
Problem 5: Distance parameter is scaled using log10 for all graphs using distance. But the values are still displayed in light years for readability.

a) In order to adjust for the large number of points laying along the x-axis, the y-axis was scaled using log10 for distance (in light years). **The three plots with the three different variables against Messier number are shown in the Appendix.

The graph that exhibits a pattern with respect to the Messier list is the scatter plot of the Messier Number (Object) plotted against the Distance. Roughly the first half of the objects have a closer distance, but as the Messier numbers increase, more objects are farther away.



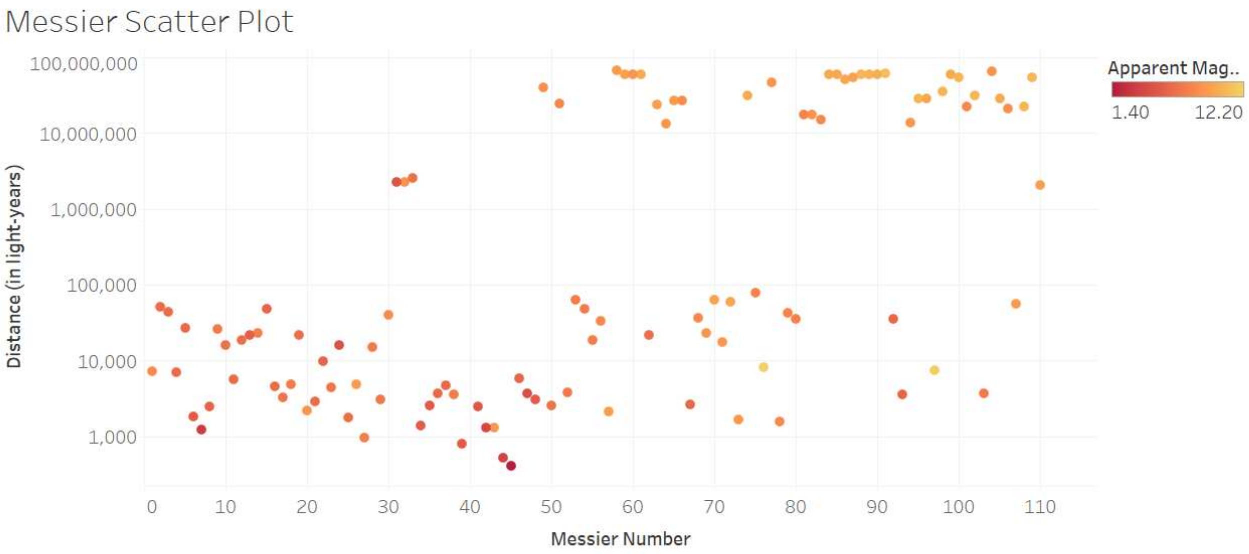
b) Messier Boxplot for Distance with respect to Kind:



The boxplot for Distance is shown with respect to each Kind, sorted based on Median value. The Distance was scaled using log10 to prevent most of the Kinds from being squashed against the axis. Immediately from the boxplot we can see that Asterism and Star Cloud only have a single object. The farthest Kind is Galaxy, where the closest object is approximately 2 million light years away, and the closest object is approximately about 450 light years away.

c) The scatter plot is the Messier number (object) against the Distance. The distance is scaled using log10.

By using a gradient color scheme to identify the measure of how bright the objects are in the sky, it shows the range of brightness, the extreme values, and how far they are in relation to distance (light years). The red plots are the brightest and have a closer distance, while the yellow plots are dim and less clearly visible and are farther away. There appears to be a general relationship between distance and apparent magnitude.



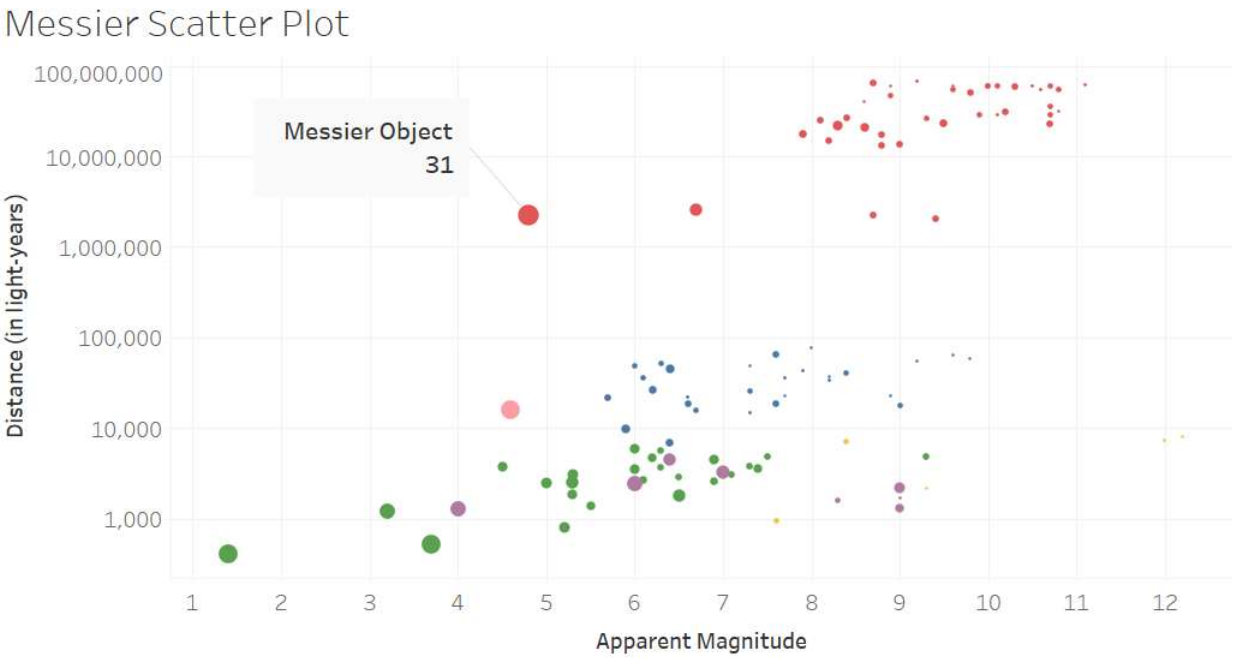
Each object's Apparent Magnitude is reflected by the color legend. A low magnitude (red) indicates how bright the object is in the sky, and a high magnitude (yellow) indicates how dim the object is in the sky.

d) The four parameters used to graph:
   a. Distance: y-axis (scaled using log10)
   b. Apparent Magnitude: x-axis
   c. Kind: Color (colored categorically). To increase opacity, the plots/bubbles were filled
   d. Size: Size of each plot/bubble

The four parameters were used in the order listed above to provide context and identify any patterns that exist. (The legend was moved to the bottom of the graph to provide a larger view). Because the Size parameter (numerically) is not easily understood because it deals with the angular size of an object in radians, a brief caption was provided to explain what the size relates to.

From the graph, you can see where the objects are, how they group together, how far and bright/dim they are, and the size of the objects. The farthest Kind is the Galaxy and their apparent magnitude is the dimmest, except for Messier object 31 where its apparent magnitude is approximately 4.7. The closest are the Open Clusters and happen to be the brightest in comparison to the other Kinds.

By using the four parameters, the graph is effectively telling a story through data visualization. Specifically using Kind and Size as two parameters in addition to the ones used for the axis, additional context was added which allowed for an interesting story about the Messier objects that would not be seem with just two variables.



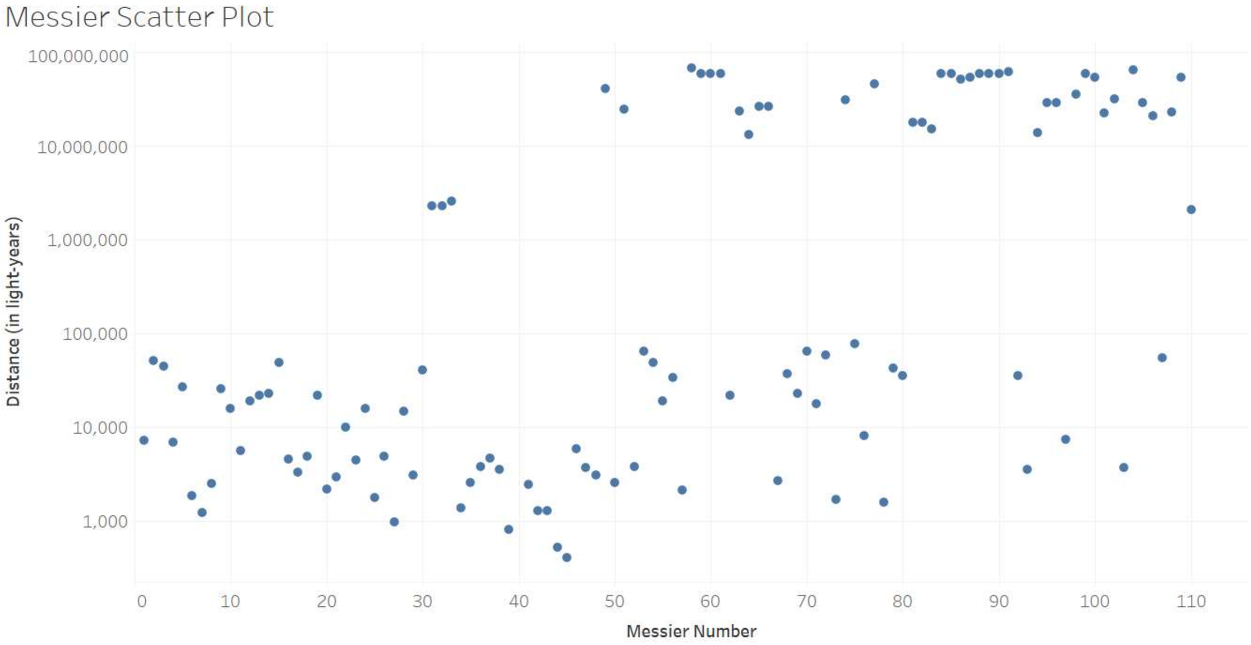
The size of the plots are the angular size of each object. Angular size is the amount of space that an object takes up in your field of view (in radians). The larger the size, the more it takes up your field of vision and vice versa.

Kind
🟫 Asterism
🟪 Diffuse Nebula
🟥 Galaxy
🟦 Globular Cluster
🟩 Open Cluster
🟨 Planetary Nebula
■ Star Cloud

Appendix:

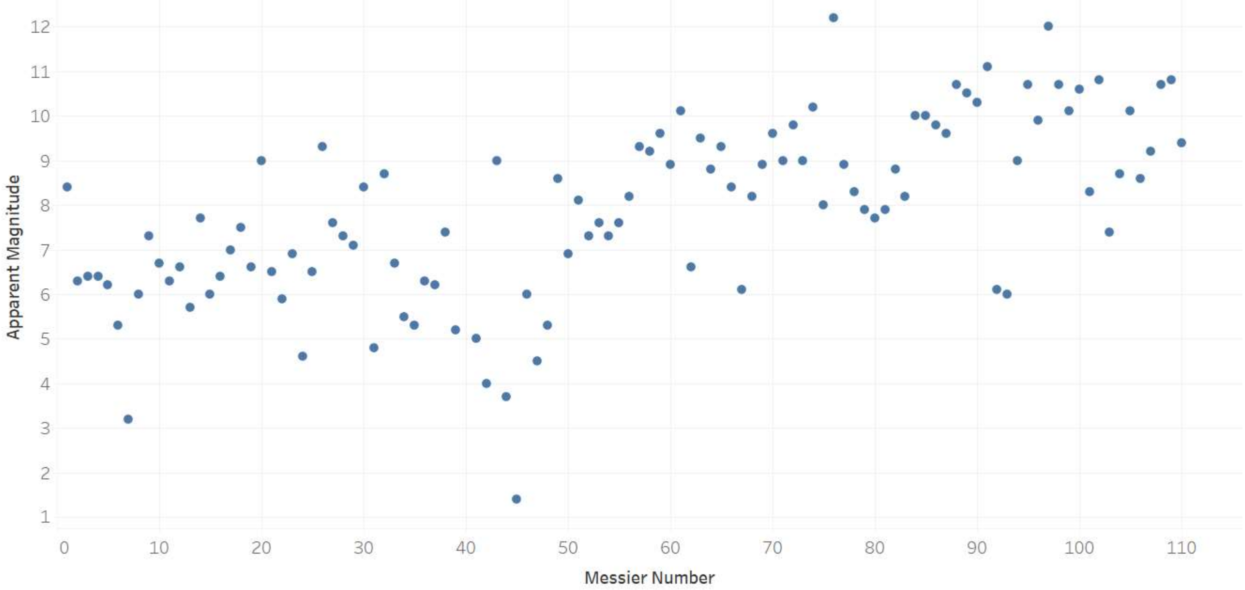
Problem 4a: Three graphs for Messier data. The chosen graph is shown on 4a.

The first plot is of the Messier Number against the Distance variable. The y-axis was scaled using log10 to show the order of magnitude of the distance (in light years).



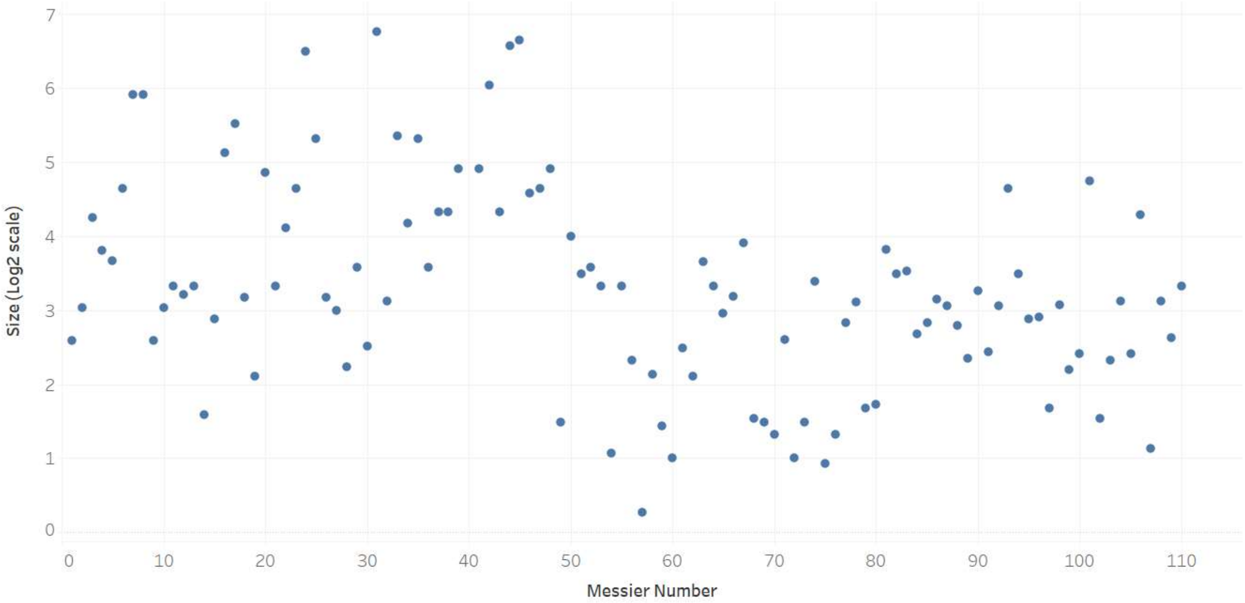
The second plot is of the Messier Number against the Apparent Magnitude variable. The y-axis was not scaled because plots did not cluster towards the x-axis.

Messier Scatter Plot



The third plot is of the Messier Number against the Size variable. The y-axis was scaled using log2 because without scaling there were many points laying along the x-axis.

Messier Scatter Plot


Size (Log2 scale)
Messier Number