Alex Teboul
DSC 465
Homework 2
Submit a PDF file with your answers.

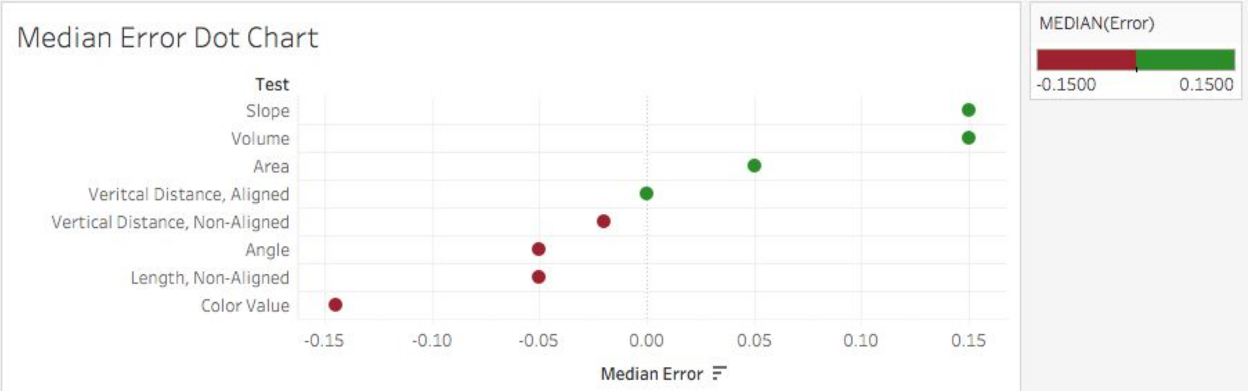
Clearly label which answer and visualization goes with which question. If it is not easy to find your answers, you may lose credit. Include written responses to questions and images of your visualizations (from screenshots or copying and pasting right from Tableau or RStudio into your document), you may use either, but some of these require techniques very difficult to achieve in Tableau or that are far simpler in RStudio.

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

2) (5 pts, due Sunday April 21) Submit the in-class participation for lecture 3 online. This participation consisted of analyzing three graphs for their visual inaccuracies or misrepresentations.

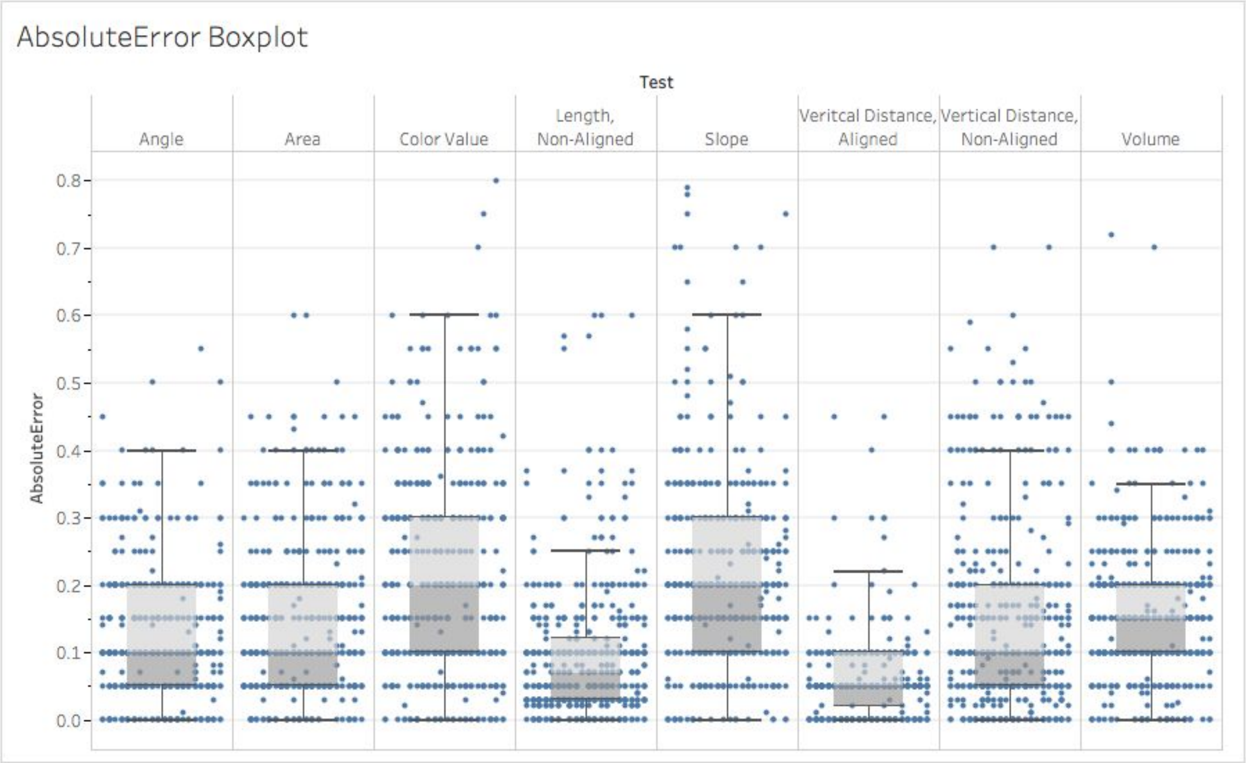
3) (30 pts) This problem continues analyzing the data from the perception test that we started in Homework 1. In this problem, we will dig deeper into the distributions of each perception test and look for patterns that reveal any strengths or weaknesses. First, I recommend re-reading the description of the data from HW1 as this data has some subtleties to it. For the problem, explore the data for the following features and display them as clearly as possible using any techniques that we have covered for displaying and comparing univariate distributions. You may do this either in R or Tableau, but be aware that R will give you more options for your visualization. In either case, be thorough in looking at what methods are appropriate. Focus on the clarity of the display, keeping in mind the criteria from the lectures on clarity and accuracy. You will re-use your calculated fields for error and absolute error from HW1.

a. Create a dot chart of the median error by test. This is the same graph as the last homework part c), but instead of using a bar graph, you are using a dot-chart. Rather than coloring the dot chart categorically, color it so that the color emphasized positive and negative values with different colors. Describe in two or three sentences how the two compare in their communication. The plot should be clean and uncluttered.



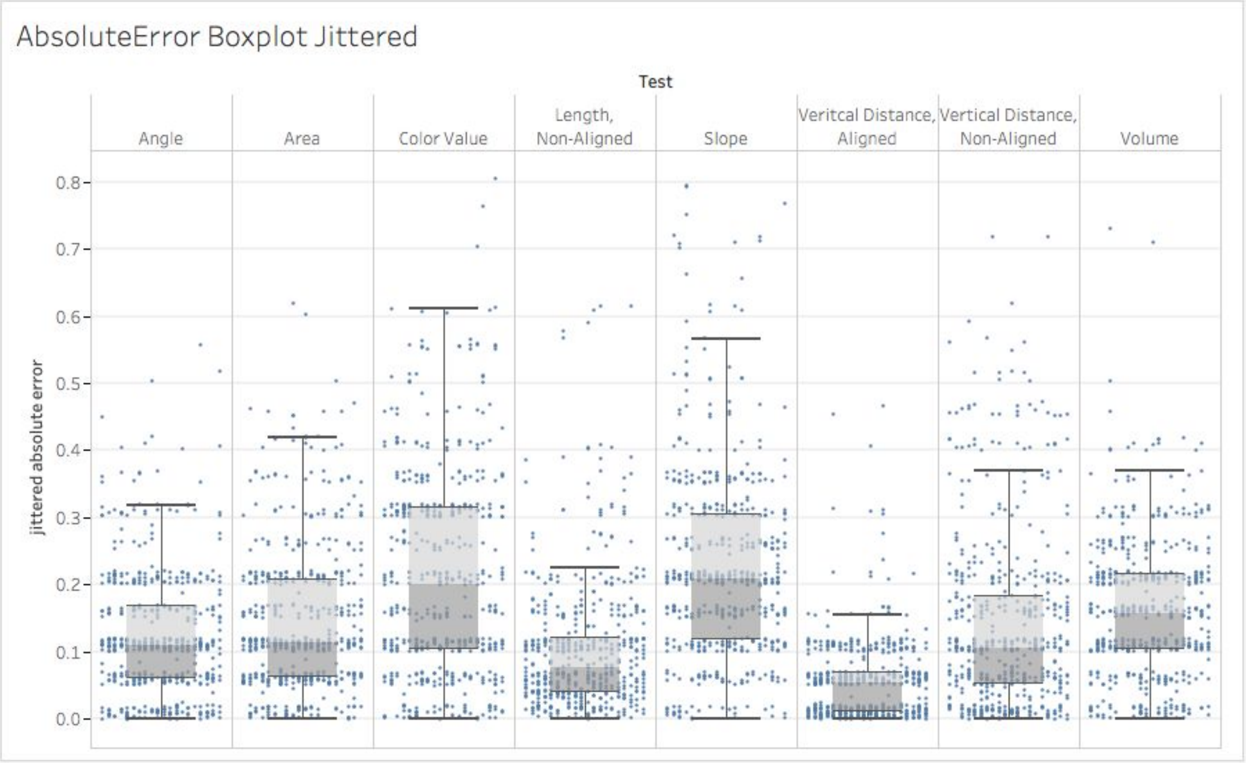
- I know in lecture, the benefits of dot-charts were discussed, but in this case I still prefer the bar graph. That said, the dot provides a more concise plot, especially given the long test names and number of tests.
- In Tableau the gridlines are slightly off, but still, it takes the average viewer longer to discern the values of the different categories in the dot as well as their implications when compared to a bar graph. While dot charts save on pixels and can simplify, in this case the dots appear to indicate a relationship between the x and y axis, but the y-axis is categorical. The positive-negative coloring is helpful on the dot chart but on a vertical bar graph the positive-negative relationship is already evident based on the direction the bars face (up-down).

b. Build on your graph from the last homework, part e) by adding a jittered categorical scatterplot overlayed with the box plot to display, for each Test (don't distinguish between Display 1 & 2 or Trial B, C and D), the absoluteError of the responses. Then write a short paragraph of analysis. How do the distributions of the data compare across the different methods our perception test studied for encoding numerical data visually? Is there any noticeable clumping of responses for any of the methods? (2 points of e.c. will be given if you use a normal distribution for the jitter. Make it clear if you have done this).



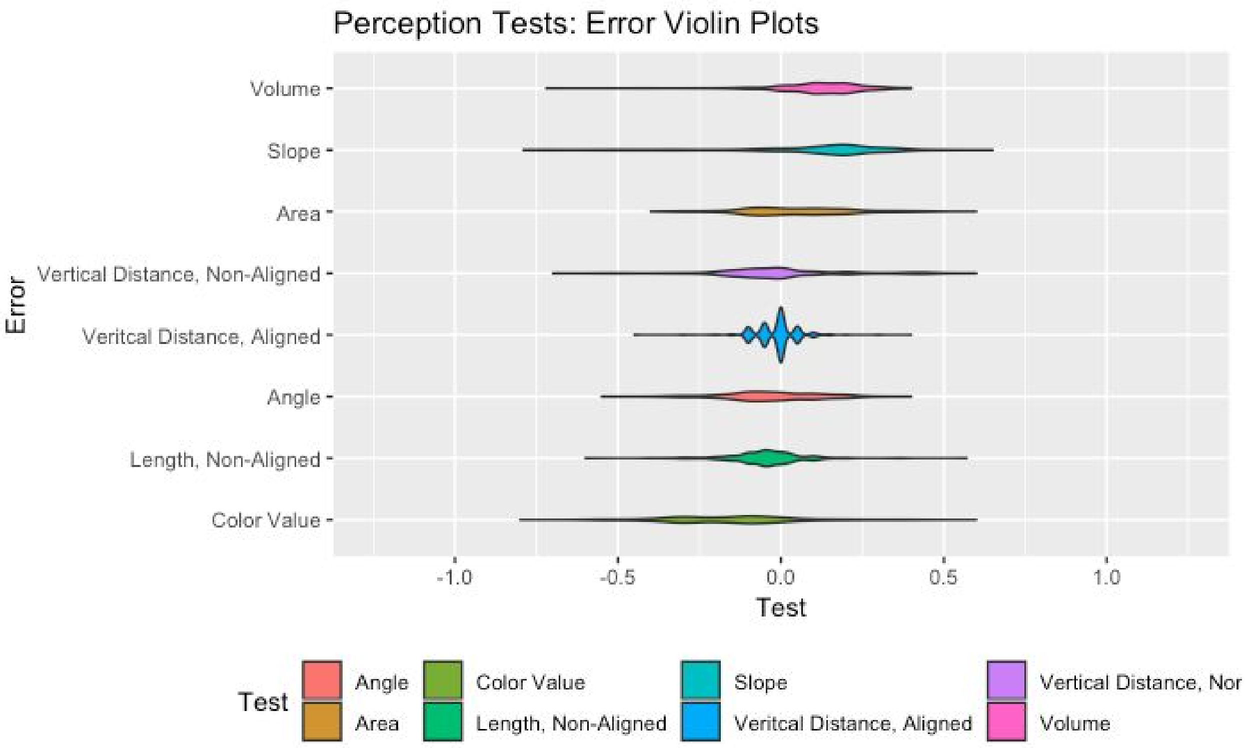
- The first method (without using jittering) shows significant clumping which fails to illustrate the distribution of the data. There is still some clumping even with the jittering I performed given the great number of overlapping results for all the subjects (participants in the survey). It doesn't help that the the vast majority of participants responded a with guesses in increments of 0.05, increasing overlap (data is quantized). Jittering is a useful technique for visualizing the distribution over the boxplots.
- With respect to the test methods studied, Length Non-Aligned, and Vertical Distance Aligned had the smallest Absolute Errors and distributions. This makes sense and confirms that perceptually we just lengths and distances reasonably well when compared to other perceptual tasks like Color Value. Interestingly, Angle and Area had nearly identical Absolute Error Distributions. Color Value and Slope had the widest range of values as well as the highest median Absolute Errors. Again, because participants estimated on 5% increments, there is clumping/quantization present in the plots.

c. (3 points of e.c.) The data shows quantization in the responses on multiple of 5% experiment with jittering along the direction of the response values (absoluteError) by a small amount … i.e. small enough that you aren't changing the quantized range of the responses. Write a short paragraph of analysis on what visual benefit/harm this provides as well as whether this is distorting the data or not.



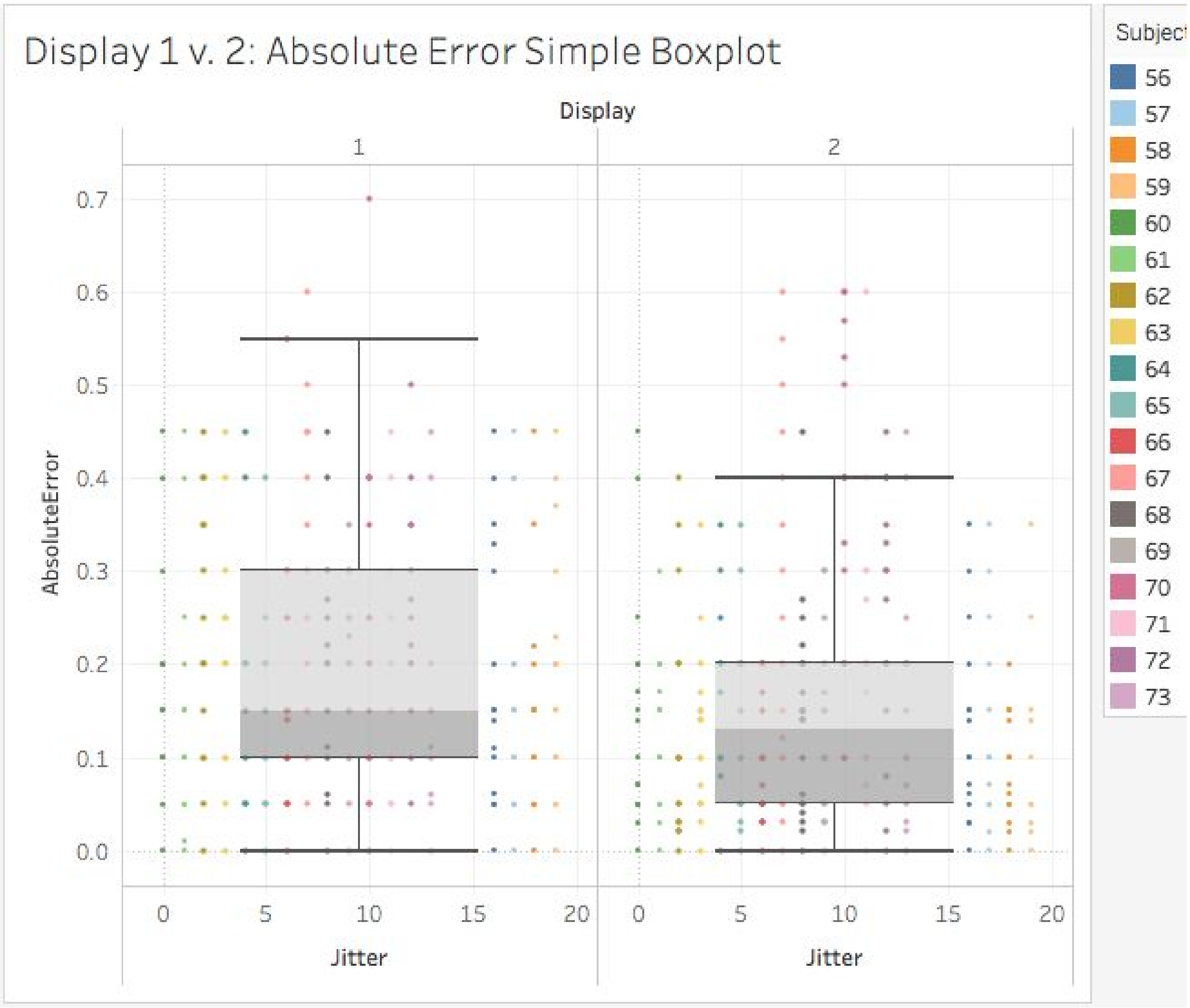
- If I understood the question correctly we were asked to add a slight jitter effect to the absolute error as well. The visual benefit of this is that overlap at the 5% participant response increments is reduced, and we can see the distribution more clearly - at least in the sense that more of the data points are present on the graph. The harm in this is that it is not completely accurate, but for the purposes of the visualization it is useful. In truth, those jittered values are off slightly from the y-axis but by equally random spacing above and below the y-axis values. Overall, the technique is helpful, as the quantization remains mostly intact, and we are still able to visualize the distribution much more effectively. This makes it more clear why the boxplots have formed the way they have while retaining the integrity of the quantized data.

d. Use a collection of violin plots to explore the Error field you calculated from last week. For which perception tests did people generally underestimated or overestimated the data? Would a jittered scatter plot overlay be helpful here in understanding the distributions? Analyze the results and explain in a short paragraph.



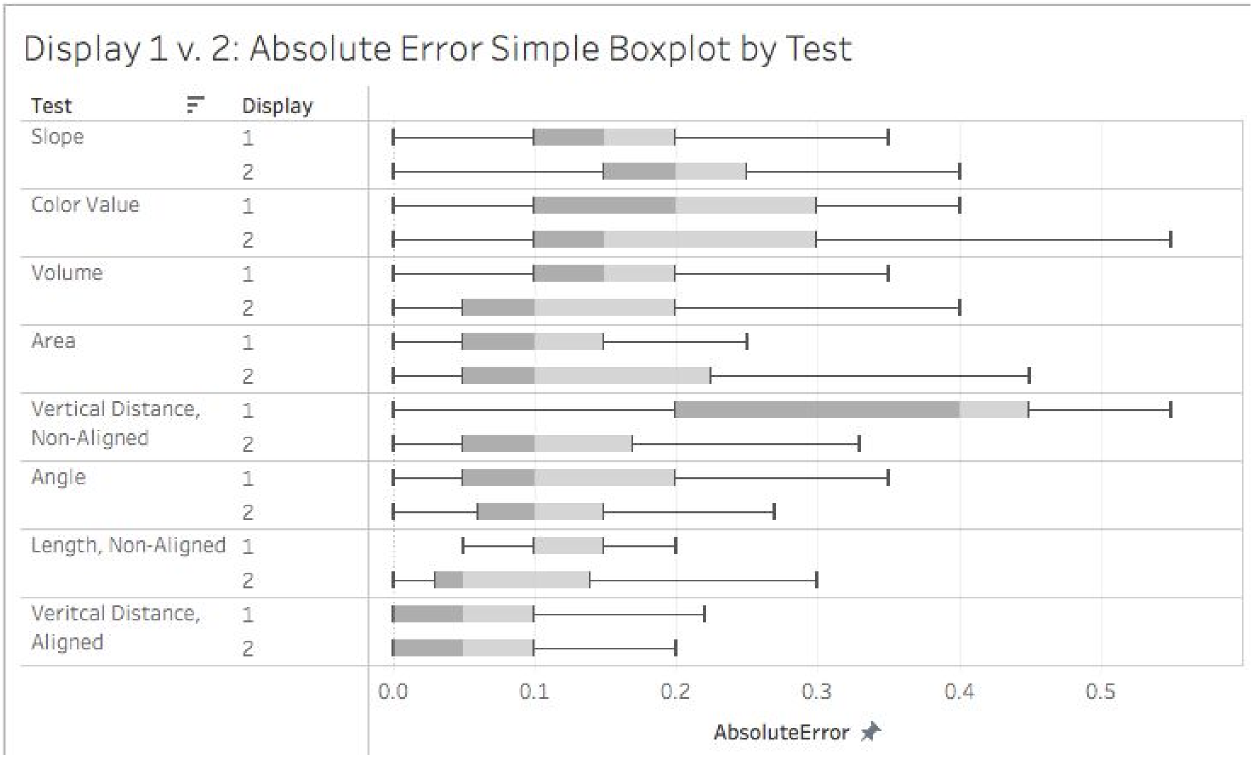
- The jittered scatterplot would help better show the quantized nature of the Errors. But for so many tests, the violin plot is already a bit cluttered and the violins are quite thin to be really seeing the distributions anyways. The combination of this plot with a separate plot to hone in on distribution might be more effective, though less concise. So, no I don't think a jittered scatterplot overlay should be used in this case.
- Most clearly overestimated (positive error): Volume, Slope
- Most clearly underestimated (negative error): Color Value, Length - Non-Aligned
- Note - switched to R for this. Trying to learn Tableau but couldn't get it to work.

e. Create a visualization that compares the data for Displays 1 and 2 for subjects 56-73 (in Tableau, you will need to filter the data here, and in R you will need to subset). The visualization should have two graphs.
i. One that compares overall results, not broken out by Test,



- From this simple boxplot we can observe that the Display 2 responses had generally lower Absolute Errors - meaning subjects were better at perceiving the tests.

ii. Another graph that shows each test divided into Display 1 and 2.



- From this graph we can tell that:
  - Slope: Performance on the perceptual test was actually worse in the second display in terms of median Absolute Error, but the wide range of error values suggests this variation has more to do with the general challenge of estimating slope than the influence of a prior Display on perception.
  - Color Value: Median Absolute Error on this test decreased with display 2 suggesting an improvement, but the range of error values was much higher, suggesting again that the impact of prior display could be marginal.
  - Volume: The same case for volume as Color Value with lower Median Absolute Error in Display 2.
  - Area: For area the median Absolute Error stayed the same.
  - Vertical Distance Non-Aligned: Display 2 has a significantly smaller Median Error as well as range of errors. Here there appears to be clear improvement
  - Angle: Angle stayed about the same.
  - Length Non-Aligned: There also appears to be an improvement here but the range of errors is also greater.
  - Vertical Distance Aligned: Stayed about the same, good perception in both Display 1 and Display 2.

Then in your analysis, answer the following: these subjects all saw the first set of Displays before the second set.
-Is there any difference in the values for Displays 1 and 2?

- There is not a significant difference that can be attributed to the actual Display effect with the exception of Vertical Distance Non-Aligned in which nearly all the participants improved significantly on Display 2 perception in terms of Absolute Error.

-Did the participants get better at judging after having done it once?

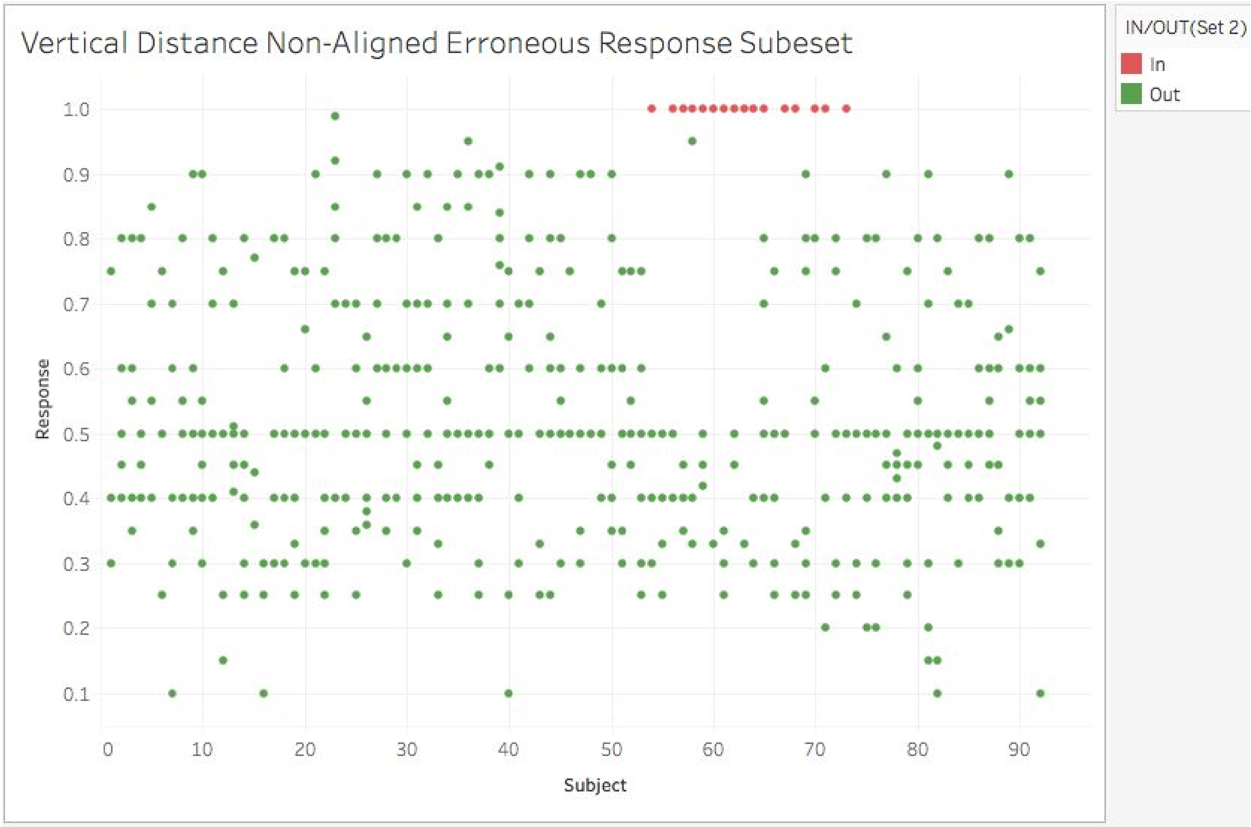
- In general the participants got better or stayed the same at judging based on Median Absolute Error in their perception for Display 2. But with this small sample size, it's hard to draw any serious conclusions. That said, Vertical Distance Non-Aligned should be explored more as it had a striking difference (improvement) in Display 2 perception.

f. Visualizing erroneous data: An erroneous stimulus was used for the first Display of "vertical distance, non-aligned" for a small subset of the subjects. Imagine that you are trying to explain to your team that these responses are compromised and the responses from these subjects need to be removed from their analysis.

You will have to find the erroneous responses by looking through the data (Excel?). They are an anomalous sequence of "1" responses across Trials B, C and D for specific respondents (i.e. Subject ID numbers).

1. Your first task is to look closely at the original raw scores and identify the sequence of subjects (hint: they are contiguous) and later in the dataset.
    - Subjects:
        - 56-73
        - Is this why we graphed these subjects in Part e of the question?

2. Next, visualize the raw scores (not the errors) as a scatterplot or collection of scatterplots, with the subject ID as one of the axes and the response in the other axis. Filtering the data will be important. Use color and other visual features to clearly show that these values are different from the other responses. Your graph should make it clear not only they are outliers of with a very specific pattern but are most likely due to a bad stimulus.



3. Some features that you might think about exploiting in the visualization:

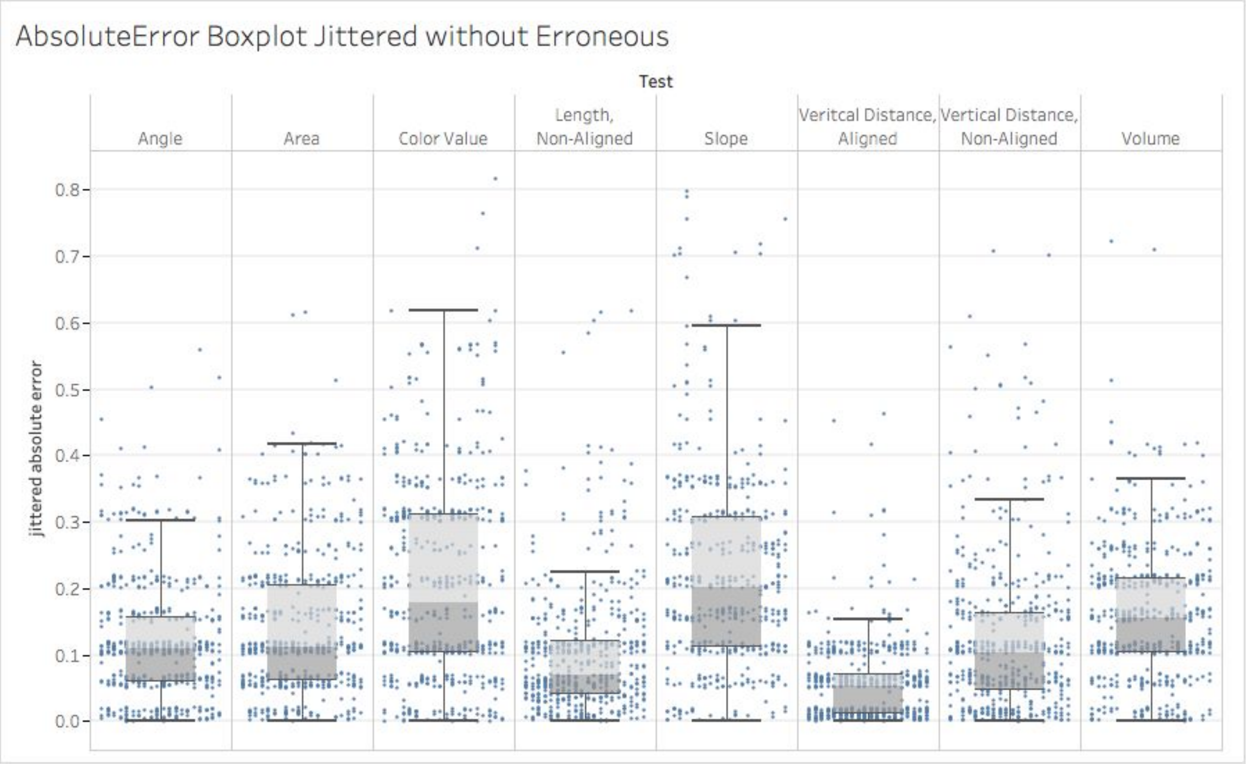
    • they are identical values across all three Trials, B, C and D, regardless of what the true values for the Trial is. This is why response is the proper field to plot

    • they are only for a small subset of subjects

    • they are only for display

    Because of this, filtering (i.e. subsetting in R) will be key in building this visualization.

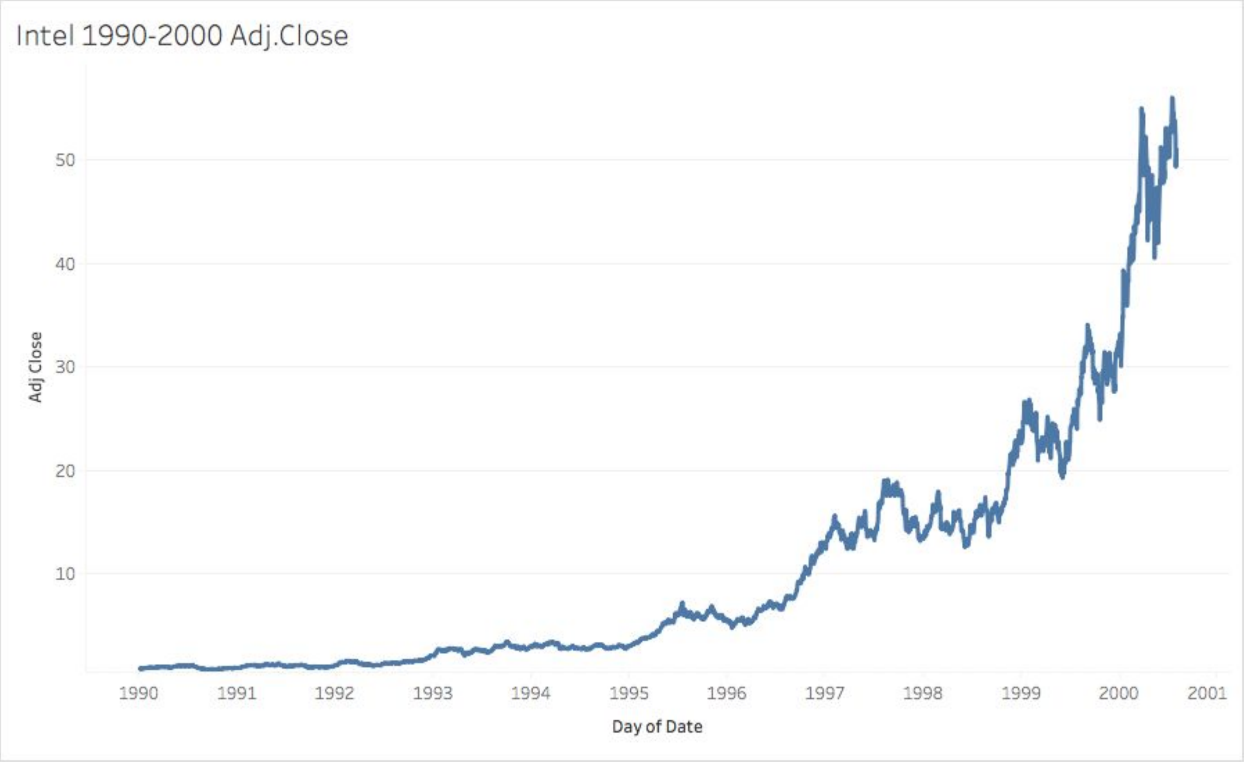
g. For the graph in b), recreate your visualization with the subjects from part e) removed. Explain whether and how the exclusion of these subjects changes the results. The following problems requires the use and understanding of logarithmic scales from lecture. Notice that in R, the natural log is the function “log”, and it has two other functions, log2 and log10 for the other bases. In Tableau, the natural log is “ln(value)”, and it has another function called “log(value, base)”.

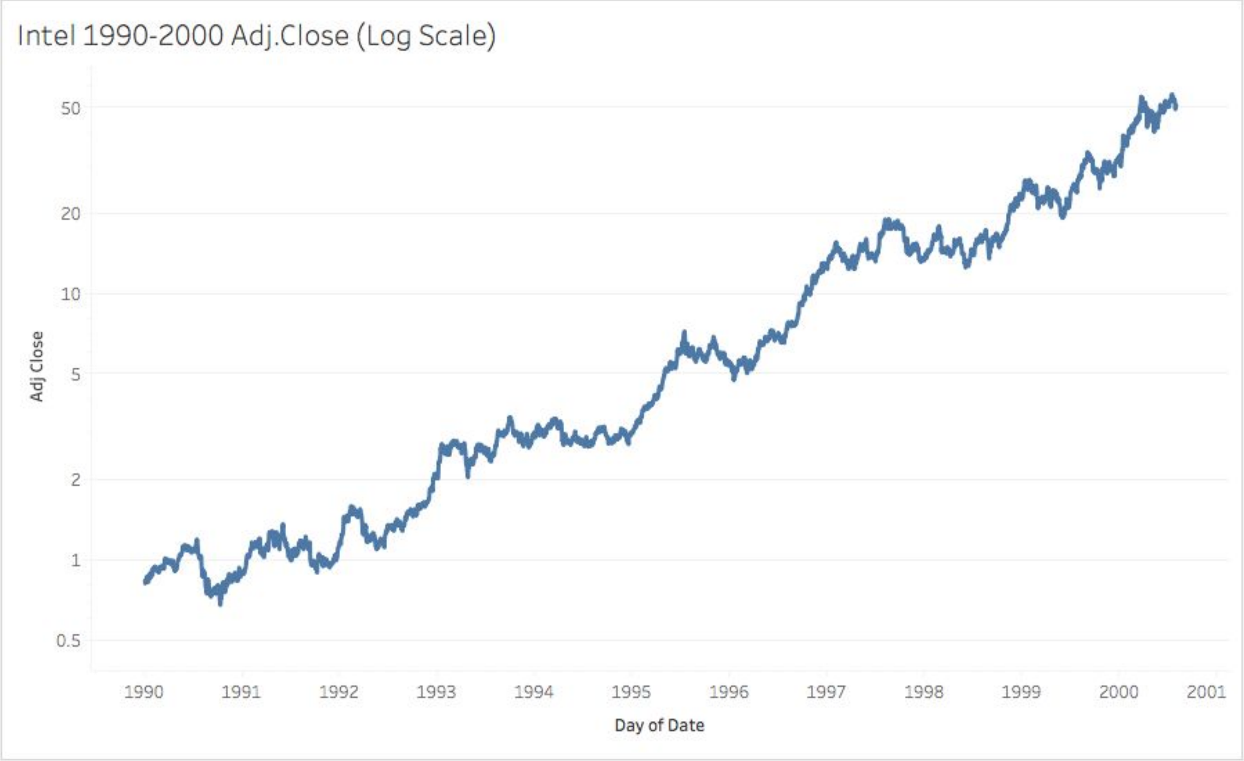


- So this didn’t seem to impact the results really, given the number of data points. Those outliers didn’t have such a huge impact overall.

4) (20pts) Download the stock data for Intel. This time the file contains data over a longer period, just up to the 2001 .com bust. Graph the data in the following ways with one graph on each page of the workbook

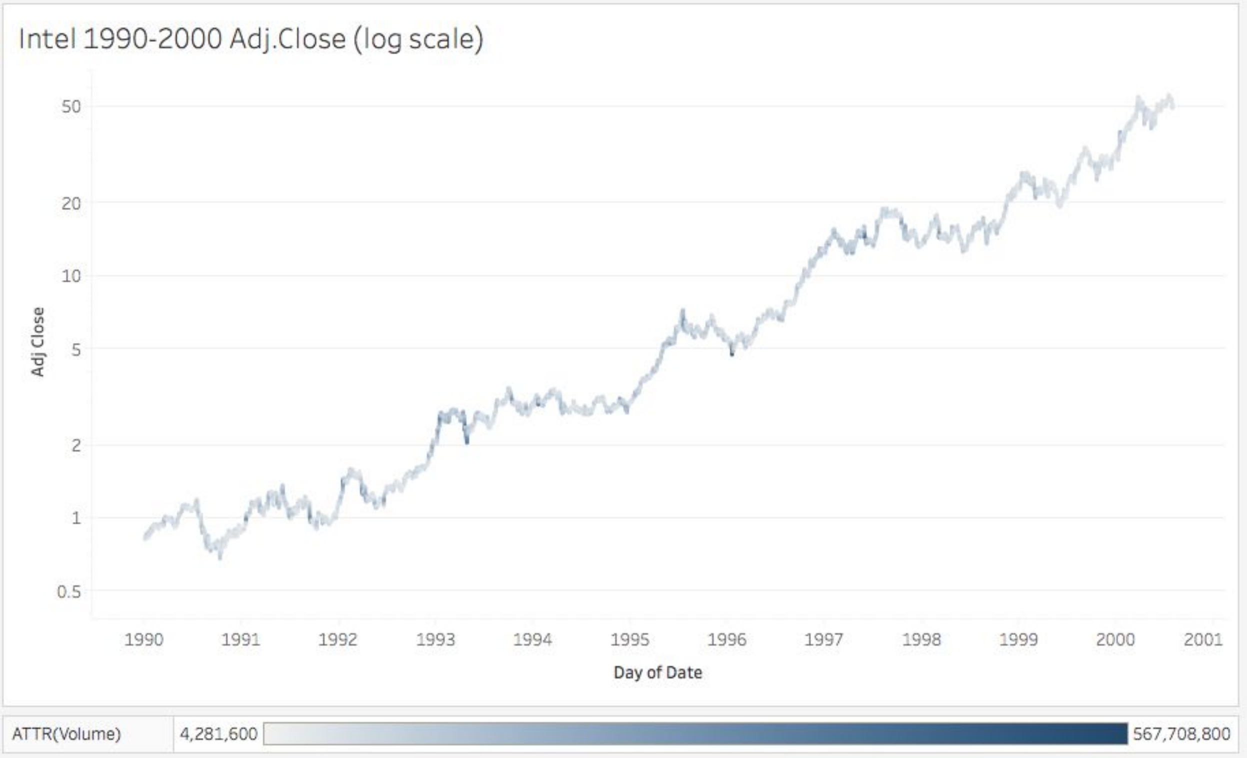
a. Create a standard line graph of the Adj.Close, both with and without a logarithmic scale. You may use the automatic log scale in Tableau or the scale_x_log ggplot scale in R. How does the logarithmic scale alter the visualization? Does it allow you to see any aspects of the data more clearly?



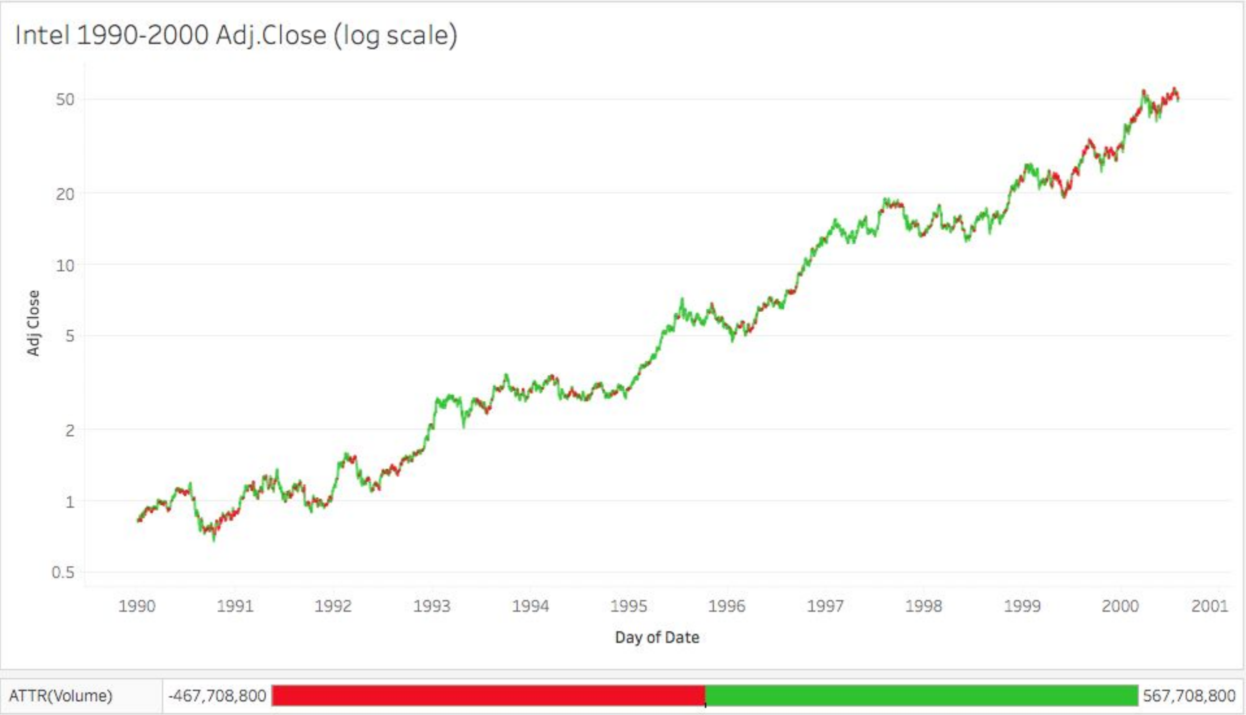


- The log scale allows for a much clearer interpretation and visualization of the slope or in this case percentage rate at which the stock's price is changing annually. The first graph shows the Adj.Close growing exponentially approaching the year 2001. When the log scale is applied in the second graph, it produces the much from interpretable graph with a near 45 degree slope. What this really shows is that on average between 1990 and 2001, Intel’s stock price was growing percentage wise at a near-constant rate.

b. Again, use a standard line graph and use the Volume field to alter the color of the line at each point. You may have to make the line thicker to see the result, and transparency (alpha in R, or the “Opacity” slider in the color properties in Tableau). Strike a balance between the visibility of the color and the definition of the line. Use a logarithmic scale for price.

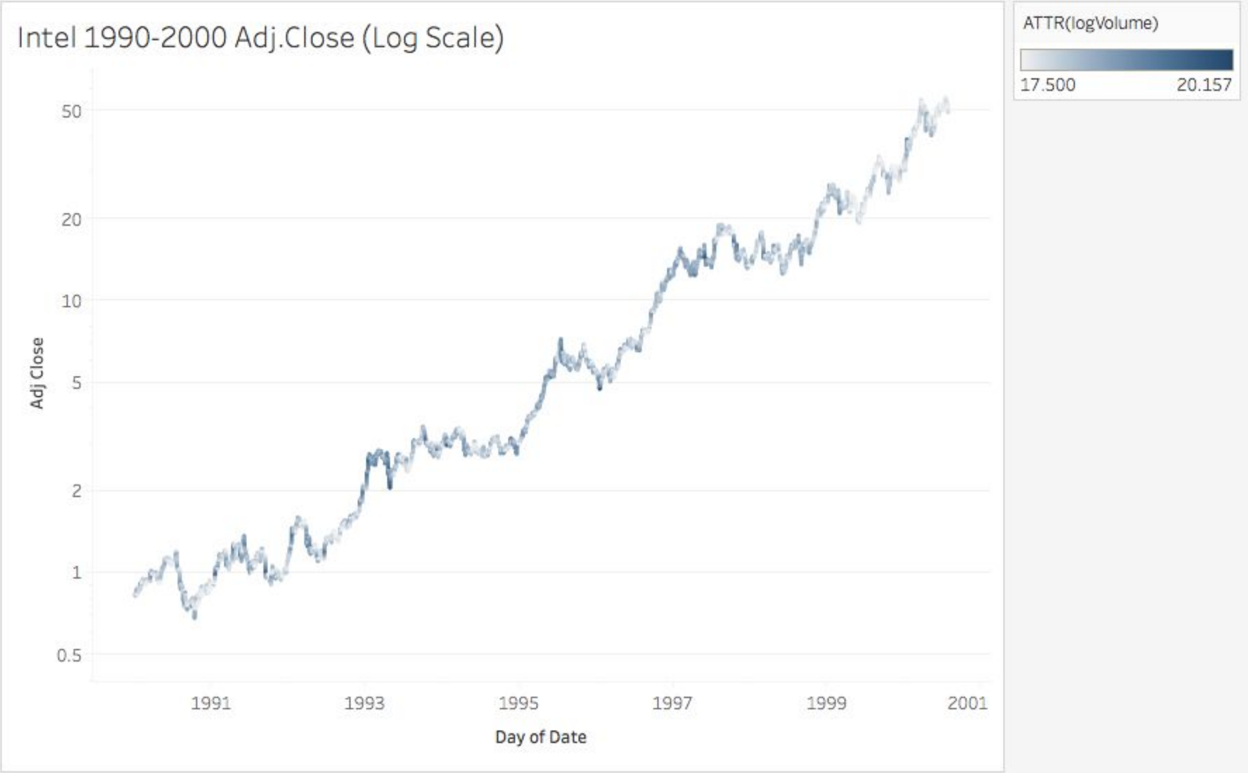


- Couldn’t determine a particularly strong correlation between the volume and price movements using this approach.

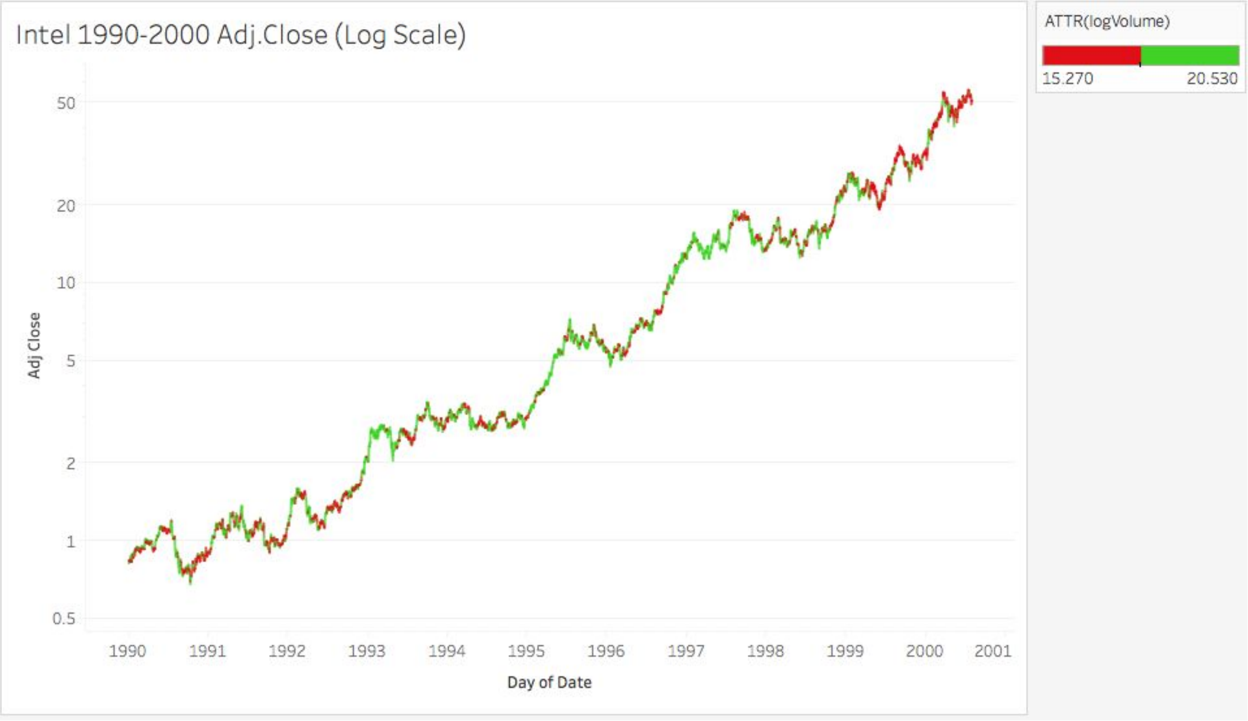


- Works alright when a different center is chosen at around 50,000,000 and a red-green diverging color scheme is used. A slight correlation begins to present between lower volumes and price corrections.

c. Create a calculated field for “logVolume” = ln(volume). Use the log of the volume for the color property in your line graph from b. Does this help or hinder the efforts to visualize the trading volume along the curve?

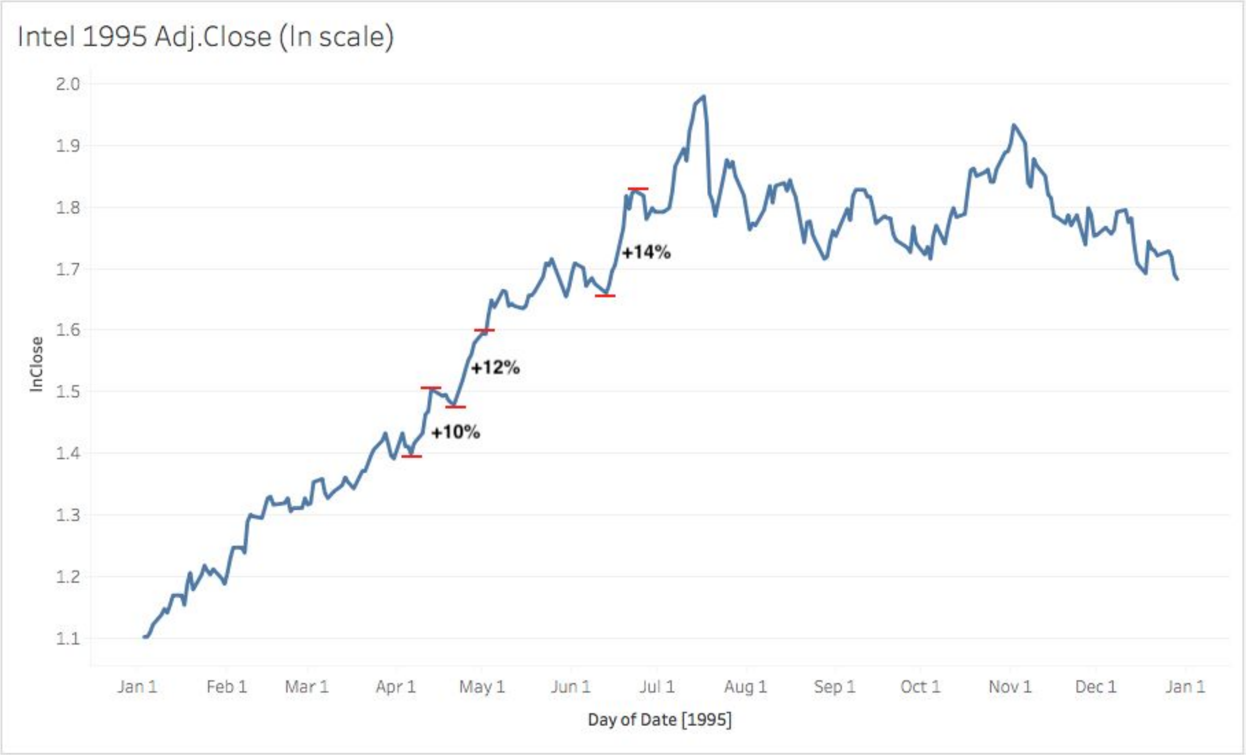


- Yes, with the same parameters it does a better job of distinguishing volume changes, but still does not produce a particularly informative graphic. At least in the sense that we cannot deduce how volume is impacting the Adj Close price.



- Using the divergent red-green color scheme provides a much more useful and informative visual that shows a correlation between price downturns and decreased volume. But this is only exploratory.

d. Limit your graph to the single year of 1995 (filter on the date). Change the log scale to a natural log, and instead of having the adj.close numbers on the scale, display the natural log of the adj.close. To do this, you will have to create a lnClose = ln(adj.close) field for this. Then graph the ln of the adj.close for the year 1995. In your description for this part, identify three surges in price (from a local minimum to a local maximum) that are between 10 and 20% increases. For each, estimate as precisely as you can, just from the graph, the %-wise increase during the price surge.



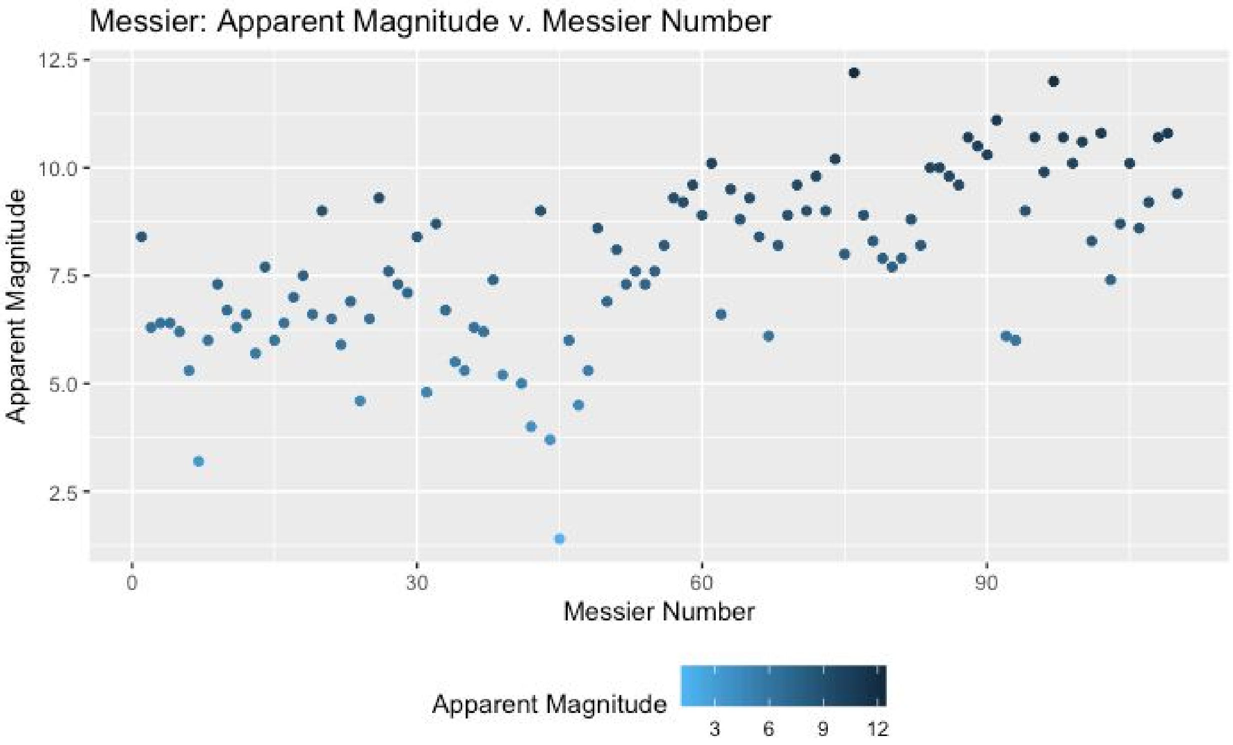
- Based on the graph, I estimated 3 different price surges at 10%, 12%, and 14% gains as indicated by the increase in ln(adj.close). These surges are indicated on the line graph above.

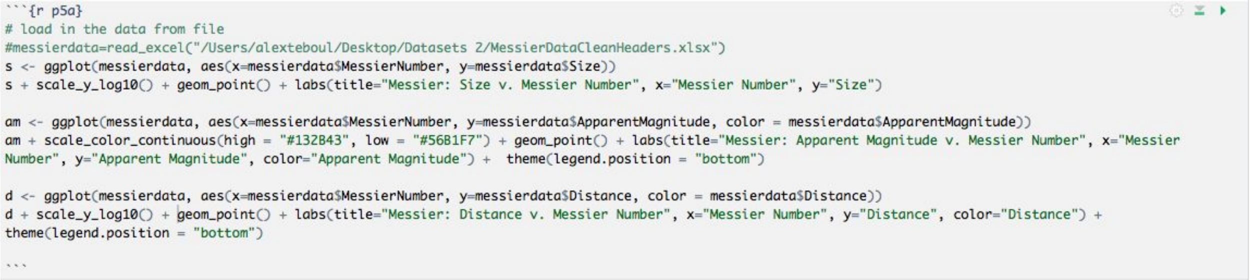
5) (20pts) Download the astronomical data for the Messier objects. These are objects that can be seen in a dark sky with binoculars or a telescope that Charles Messier cataloged in France in the 18th century so that they wouldn't be confused with comets. Some of these are clusters of stars or great clouds of gas in our galaxy, some are galaxies that are much farther away. The dataset contains a list of 100 deep sky objects along with their distances from the earth in light-years. Graph this data in the following ways to explore the information provided about these interesting objects.

Important note: For this dataset, you will have to pick suitable scales to make the data readable in your graphs. You should not wind up with a majority of the points squashed down along the one axis. For distances, the scale should show the "order-of-magnitude" of the distance in light years (10, 100, 1000, etc.) clearly.

a. Pick three of the variables in the data other than Messier Number. For each, plot the value for each of the objects against the Messier Number on the x-axis, one-by-one. Remember, there is nothing 'intrinsic' about this number, it is just the order of Messier's list. Is there any property that exhibits a pattern with respect to the ordering in his list? Submit the graph that exhibits a pattern. Remember, you should not have a large number of points lying along the axis, so if you do, how can you adjust for this?

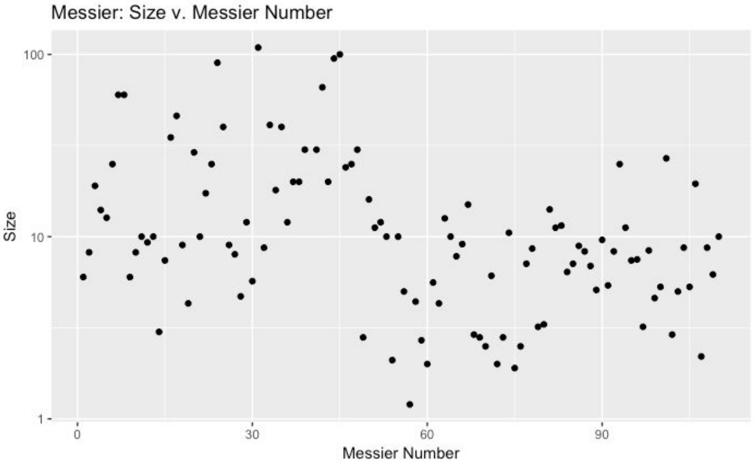
- I chose variables Size, Apparent Magnitude, and Distance. Of these three variables Apparent Magnitude seems to have the clearest pattern. Specifically, higher Apparent Magnitude is linked to higher Messier Numbers. This makes sense because the brighter objects in the night sky seem to generally be discovered and catalogued by Messier before the dimmer (more distant objects).





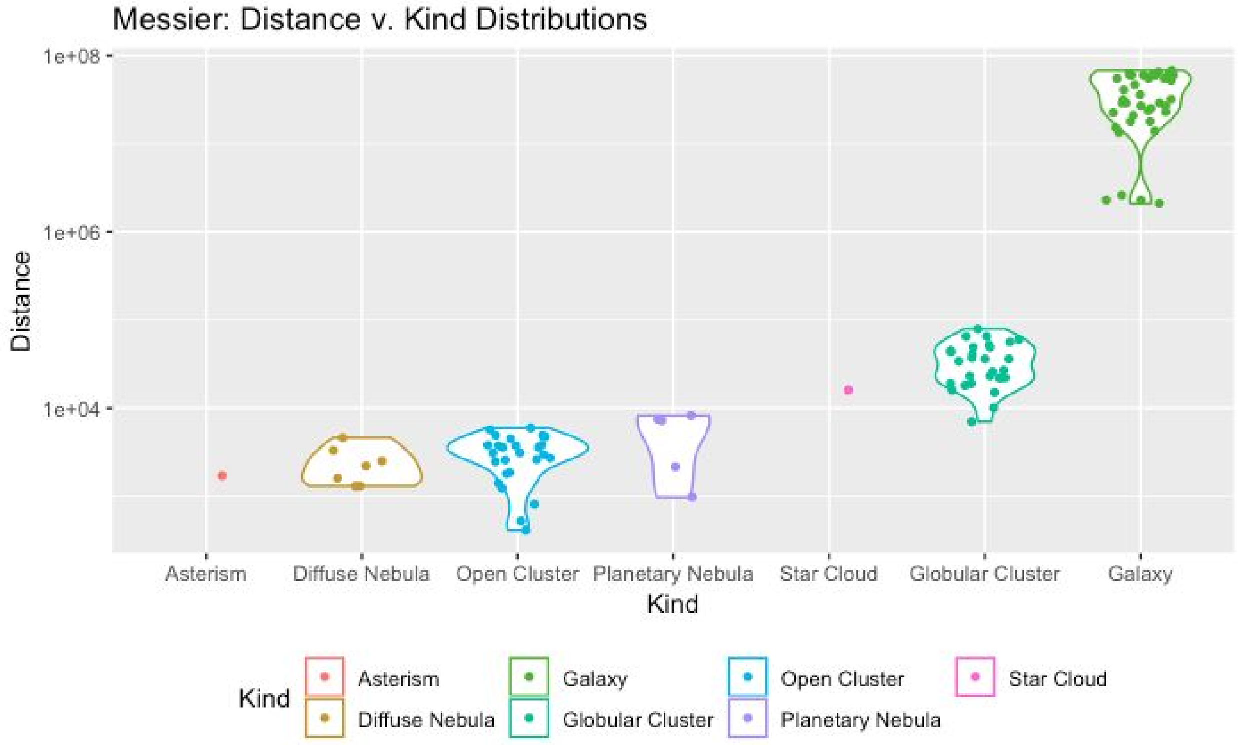
~~~
```{r p5a}
# load in the data from file
#messierdata=read_excel("/Users/alexteboul/Desktop/Datasets 2/MessierDataCleanHeaders.xlsx")
s <- ggplot(messierdata, aes(x=messierdata$MessierNumber, y=messierdata$Size))
s + scale_y_log10() + geom_point() + labs(title="Messier: Size v. Messier Number", x="Messier Number", y="Size")

am <- ggplot(messierdata, aes(x=messierdata$MessierNumber, y=messierdata$ApparentMagnitude, color = messierdata$ApparentMagnitude))
am + scale_color_continuous(high = "#132B43", low = "#56B1F7") + geom_point() + labs(title="Messier: Apparent Magnitude v. Messier Number", x="Messier
Number", y="Apparent Magnitude", color="Apparent Magnitude") +  theme(legend.position = "bottom")

d <- ggplot(messierdata, aes(x=messierdata$MessierNumber, y=messierdata$Distance, color = messierdata$Distance))
d + scale_y_log10() + geom_point() + labs(title="Messier: Distance v. Messier Number", x="Messier Number", y="Distance", color="Distance") +
theme(legend.position = "bottom")

```
~~~

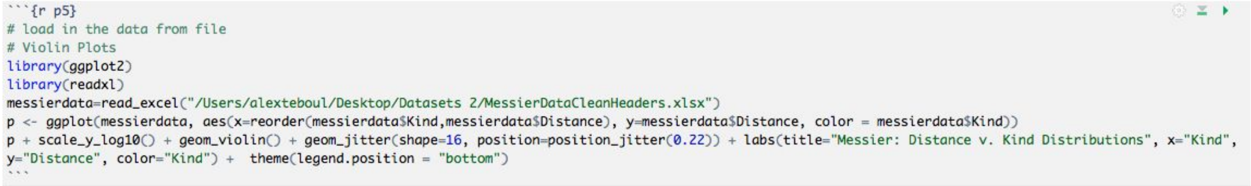
- Size also had a somewhat noteable pattern. Specifically, when using the log scale, Size exhibited a pattern with respect to the Messier Number. The pattern observed below is that the higher the Messier Number the smaller the Size. This could make sense given the first 45 Messier objects were published in the original catalog, so their larger sizes could indicate they were easier to find. In subsequent additions to the catalog, perhaps better telescopes were used in order to detect smaller objects in the sky. *This is the example of what to do when a lot of points lie along the axis - adjust the scale to better visualize.



b. Create a visualization that compares the distributions of the distances to the objects in each Kind. Note that the Type variable is a very different category and is really a subcategory of Kind. Do not use type here, rather use kind. Sort the distribution displays in a way that makes the relationship clear. Make sure your distance axis is transformed in such a way as to not have most of these "Kinds" squashed down to the bottom of the graph along the axis.

- Below is the visualization comparing the distributions of distances to objects for each Kind. Note that *Asterism* and *Star Cloud* do not have distributions as there is only a single example of each in Messier's List, I also had to drop Messier Number 40 because it had a *null* value for Distance. Clearly, galaxies are the furthest objects present in the list and globular clusters are next. Diffuse nebula, open clusters, and planetary nebula are all under 10^4 kilo-light-years away with open clusters having the widest range of values.

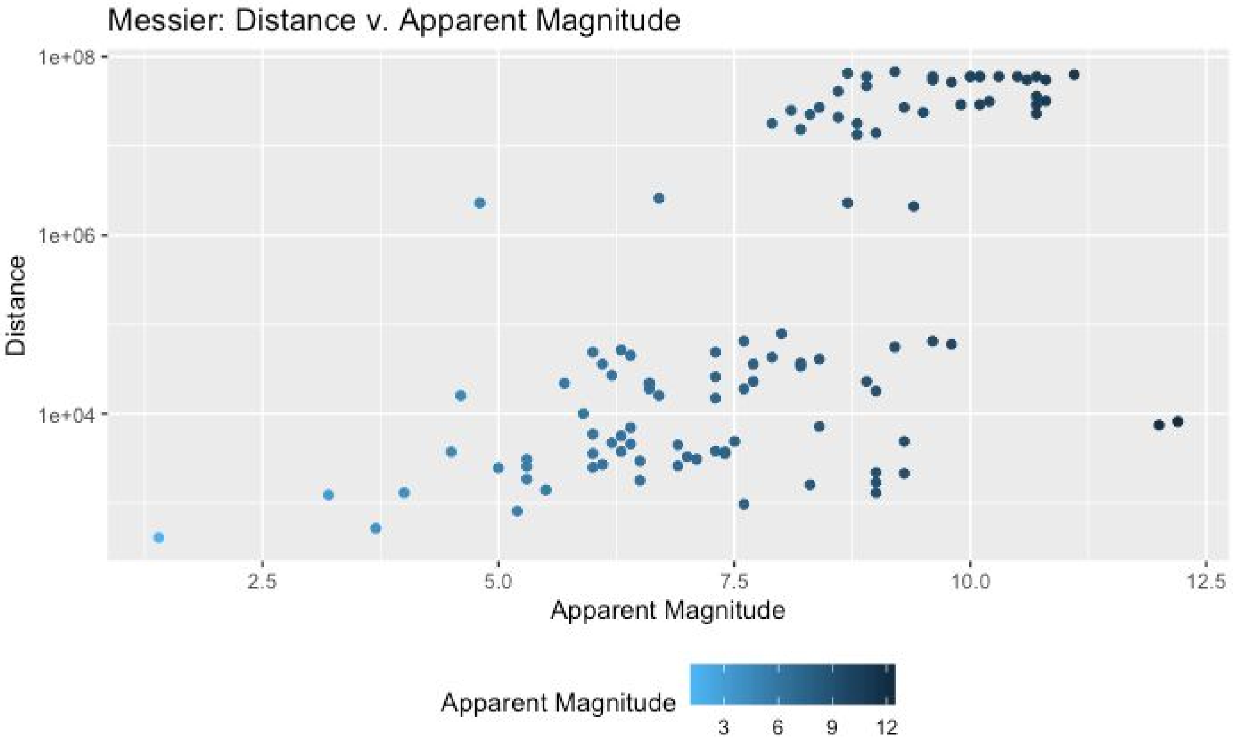


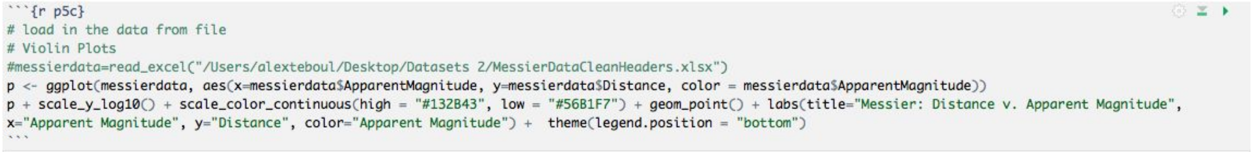


~~~
```{r p5}
# load in the data from file
# Violin Plots
library(ggplot2)
library(readxl)
messierdata=read_excel("/Users/alexteboul/Desktop/Datasets 2/MessierDataCleanHeaders.xlsx")
p <- ggplot(messierdata, aes(x=reorder(messierdata$Kind,messierdata$Distance), y=messierdata$Distance, color = messierdata$Kind))
p + scale_y_log10() + geom_violin() + geom_jitter(shape=16, position=position_jitter(0.22)) + labs(title="Messier: Distance v. Kind Distributions", x="Kind",
y="Distance", color="Kind") +  theme(legend.position = "bottom")
```
~~~

c. Create a scatter plot with the distance to the Messier objects plotted against their Apparent Magnitude (this is a measure of how bright they are in the sky). Note that these values are backwards from what you might think. The higher the number, the fainter the object is in the sky, magnitude 1 = very bright, magnitude 9 = very dim. Incorporate that into your visualization to make the meaning of the Apparent Magnitude clear. Again, pay attention to how you handle distance so that the full range is clearly displayed and the distance to all objects are clearly readable.

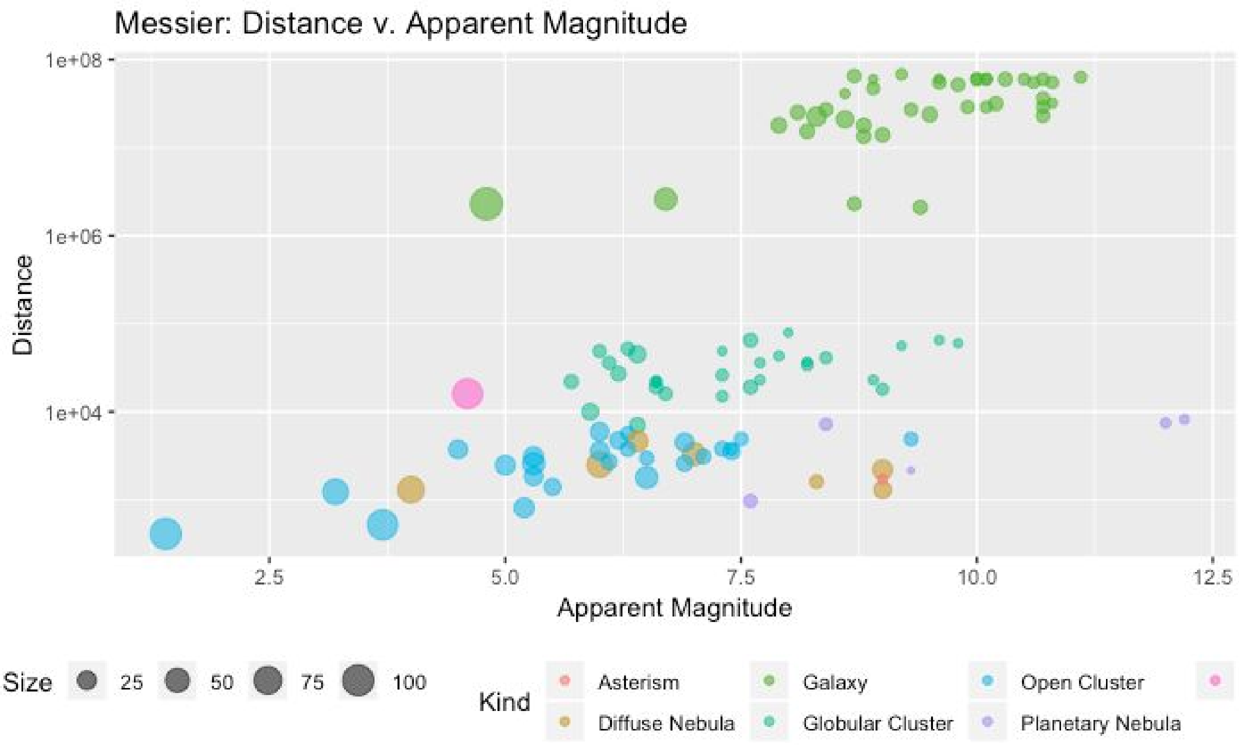
- Below is the scatter plot showing the Distance v. Apparent Magnitude of Messier Objects. It makes sense that generally objects that are further away are dimmer (have a higher Apparent Magnitude).

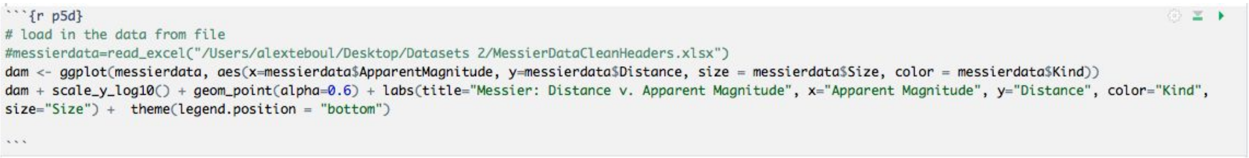



~~~
```{r p5c}
# load in the data from file
# Violin Plots
#messierdata=read_excel("/Users/alexteboul/Desktop/Datasets 2/MessierDataCleanHeaders.xlsx")
p <- ggplot(messierdata, aes(x=messierdata$ApparentMagnitude, y=messierdata$Distance, color = messierdata$ApparentMagnitude))
p + scale_y_log10() + scale_color_continuous(high = "#132B43", low = "#56B1F7") + geom_point() + labs(title="Messier: Distance v. Apparent Magnitude",
x="Apparent Magnitude", y="Distance", color="Apparent Magnitude") +  theme(legend.position = "bottom")
```
~~~

d. Finally, create a scatter plot or another type of visualization that that displays, for all four of the parameters: Distance, Kind, Apparent Magnitude and the angular Size of the objects readably in one scatter plot. Think closely about what the two axes should be and what would be better presented by a color, size, shape, etc. (One interesting idea: you might look up how these object "kinds" are presented on astronomical maps). Evaluate how easy it is to analyze all four aspects of the data from this graph.

- I created a bubble scatter plot that displays the Distance v. Apparent Magnitude, with the size of Bubbles indicating the Size of the objects and color indicating the Kind of Messier object. I know there that celestial maps offer another scatter plot like way to visualize this data, but I opted for the bubble scatter as is easier to analyze.
- Specifically analyzing the data in this scatter plot, we can tell that in general, for each Kind of object the further away that object is (distance increase) the smaller (size decrease) and dimmer (apparent magnitude increase) it is. This makes sense that for a kind of object that the bigger and closer they are, the the brighter they would appear. So I would say it is a good plot for analyzing the four aspects of the data.





~~~
```{r p5d}
# load in the data from file
#messierdata=read_excel("/Users/alexteboul/Desktop/Datasets 2/MessierDataCleanHeaders.xlsx")
dam <- ggplot(messierdata, aes(x=messierdata$ApparentMagnitude, y=messierdata$Distance, size = messierdata$Size, color = messierdata$Kind))
dam + scale_y_log10() + geom_point(alpha=0.6) + labs(title="Messier: Distance v. Apparent Magnitude", x="Apparent Magnitude", y="Distance", color="Kind",
size="Size") +  theme(legend.position = "bottom")

```
~~~