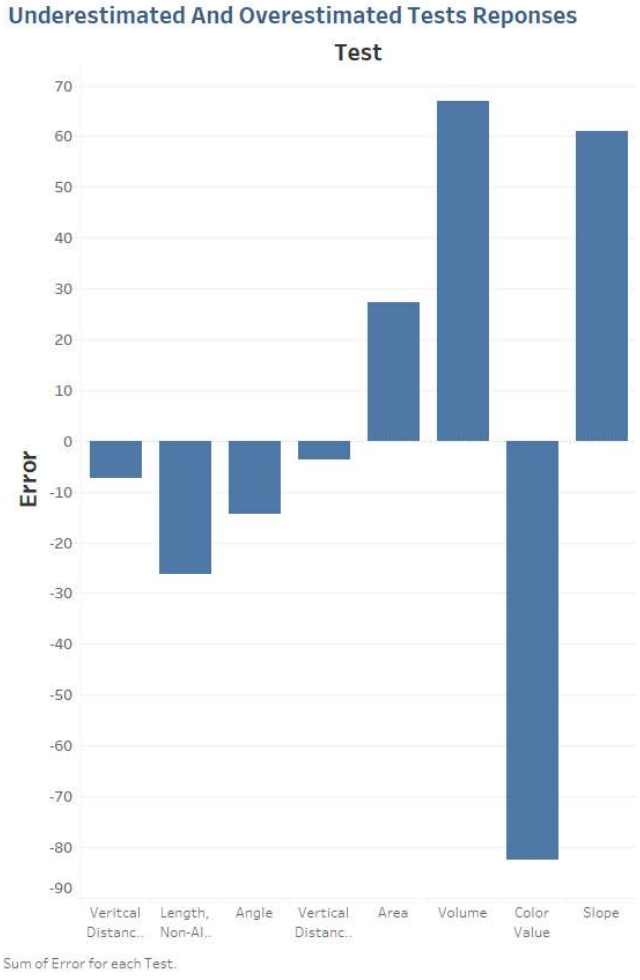
Umair Chaanda
DSC 465 Data Visualization
Homework-3

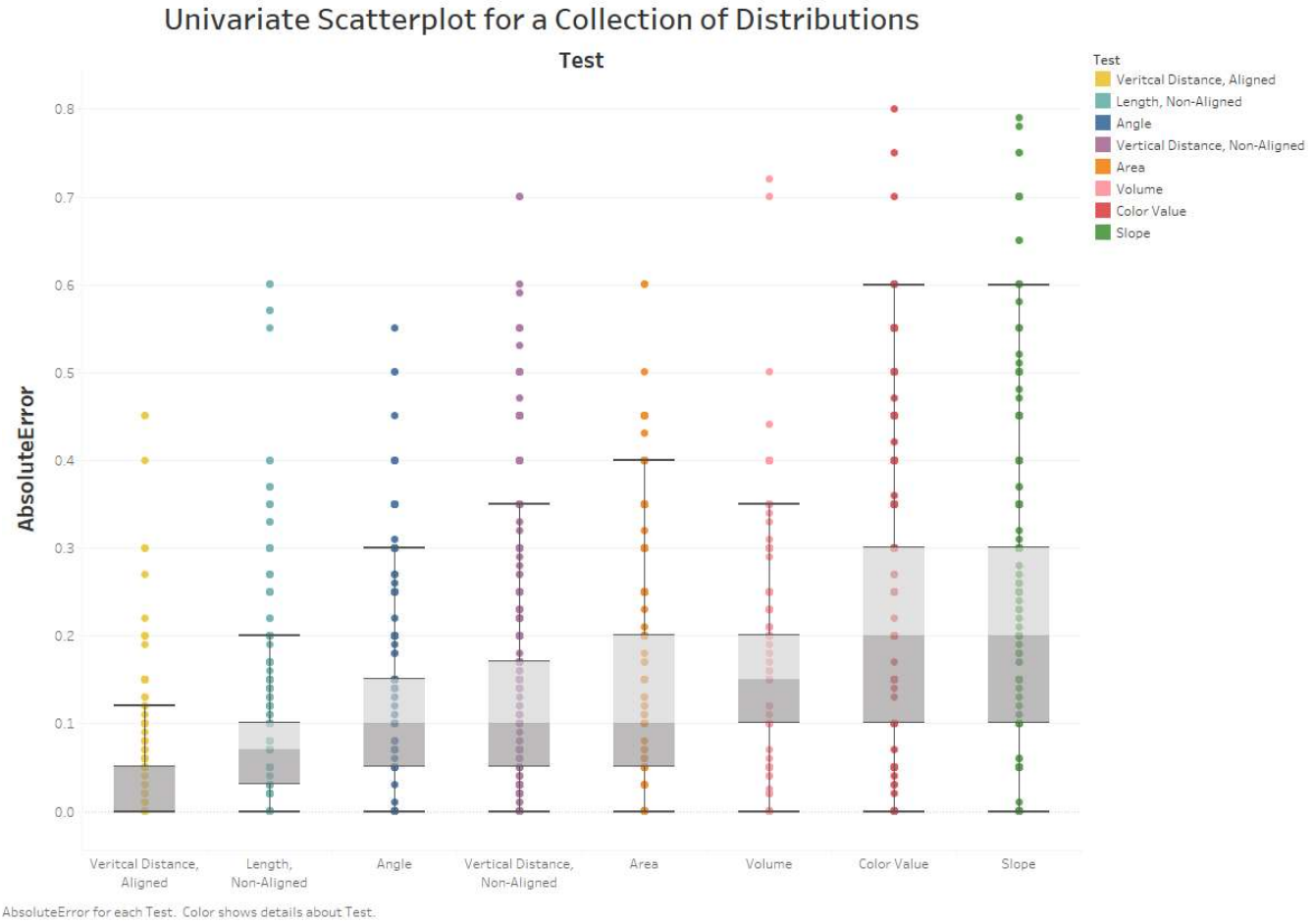
Problem 1: Perception Experiment:
Here is how the data are laid out in columns: each type of encoding is a Test, and each one got displayed with two separate slides. The individual PowerPoint slides are called Displays. Each individual Display of each Test has a unique Test Number. Each sample that you estimated a value for was labeled B, C or D as its Trial. The Subjects are the students and the estimates they made are the Responses. Each row has a copy of the True Value, i.e. the correct value that the student should have entered (if the whole point weren't how hard it is).

a) Were there any tests where people generally underestimated or overestimated the data? Explain what field you can graph to test this, what graphical method reveals this clearly. Analyze the results and explain in a short paragraph.



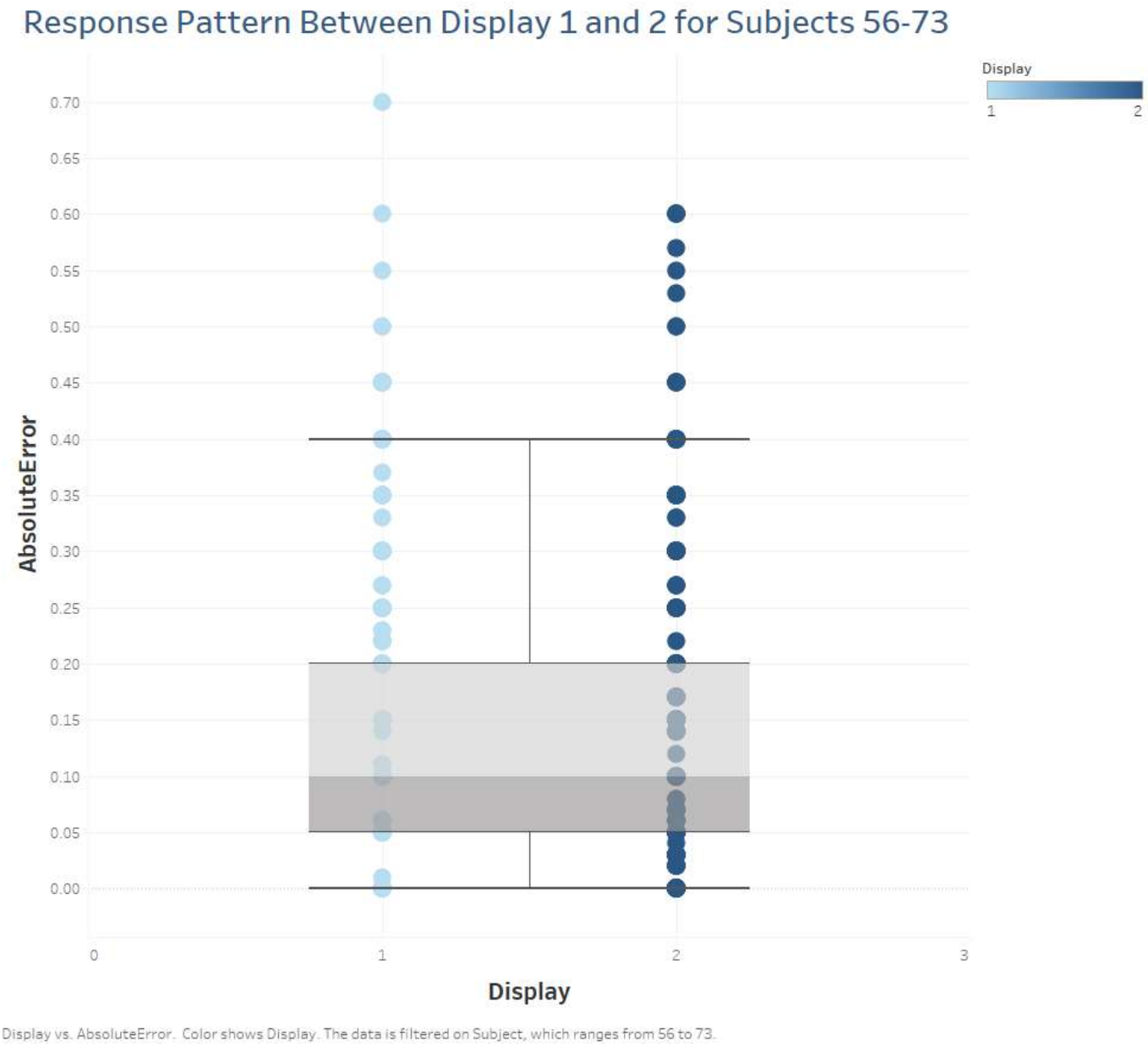
- We can examine the underestimated and overestimated test responses by plotting Error vs. Test. Anything below 0 or negative values on the graph are underestimated data and anything above 0 or positive values are overestimated data.
- The most underestimated test results are for Test type "**Color Value**" and the most overestimated test results are related to "**Volume**" and "**Slope**".
- The results for "**Vertical Distance**" came out to be almost perfect with very minimum error.

b) Use a Univariate Scatterplot or another technique that shows fine detail for a collection of distributions. For each Test (don't divide between Display 1 & 2 or Trial B, C and D) plot the AbsoluteError (absolute value of Error). 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?



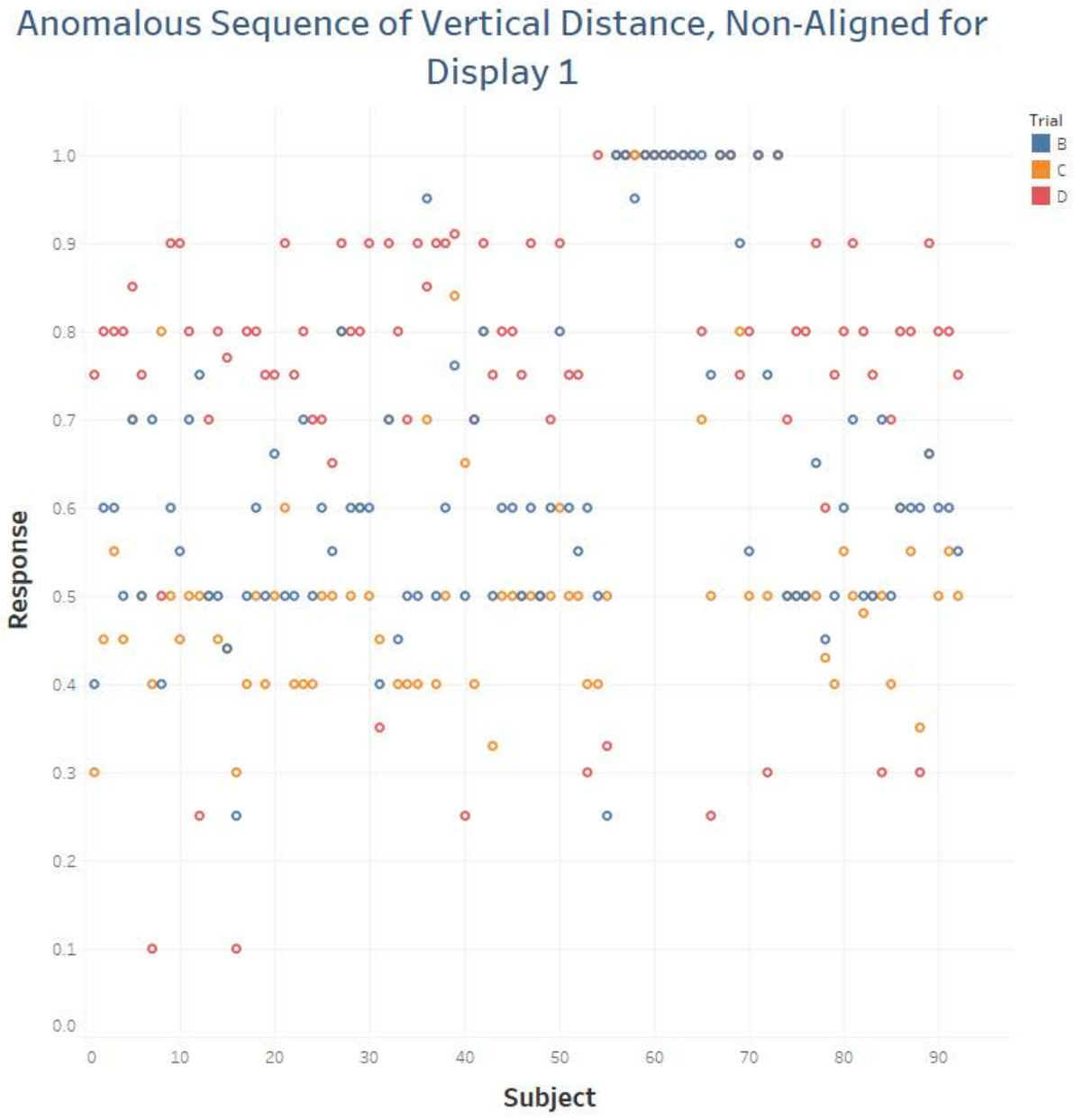
- As we can see in the above graph that "**Vertical Distance, Aligned**" encoding test has the lowest absolute error and the smallest distribution with very few outliers.
- The distribution of "**Vertical Distance Aligned**" is mostly at the lower range and it is concentrated between 0.0 to 0.1 error ranges.
- Encoding tests which have largest distributions are "**Color Value**" and "**Slope**".
- Both "**Color Value**" and "**Slope**" have most of the values in the upper error range and their IQR = Q3 – Q1 = 0.3 – 0.1 = 0.2.
- The 2nd and 3rd visual encoding test which has low absolute error are "**Length, Non-Aligned**" and "**Vertical Distance, Non-Aligned**".

c) Compare the data for Displays 1 and 2 for subjects 56-73 (you will need to filter the data in Tableau or R). Create a visualization that shows any differences in the response patterns between the two. 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? Did the participants get better at judging after having done it once?



- Display 1 and Display 2 were plotted against the "**AbsoluteError"** with data filtered on variable "**Subject"** from 56 to 73.
- The dark blue dots represent Display 2 data and Light blue dots represent Display 1 data.
- We can clearly see the difference in the values for Display 2 as the points are more clustered around 0.01 to 0.08 Absolute Error whereas for Display 1, most points are between 0.20 to 0.40.
- It shows that the participants got better results at judging with lower error.

d) An erroneous stimulus was used for the first Display of "vertical distance, non-aligned" for a small subset of the subjects. They manifest themselves as an anomalous sequence of "1" Responses across Trial B, C and D. Look closely at the original raw scores and identify the sequence of subjects (hint: they are contiguous). Visualize the raw scores in a way that highlights these values and makes their anomalous nature clear. It should make it clear not only that they are outliers but should show any features that distinguish them from ordinary outliers. Some features that you might think about exploiting they are identical values across all three Trials, regardless of what the true values for the Trial are; they are only for a small subset of subjects.

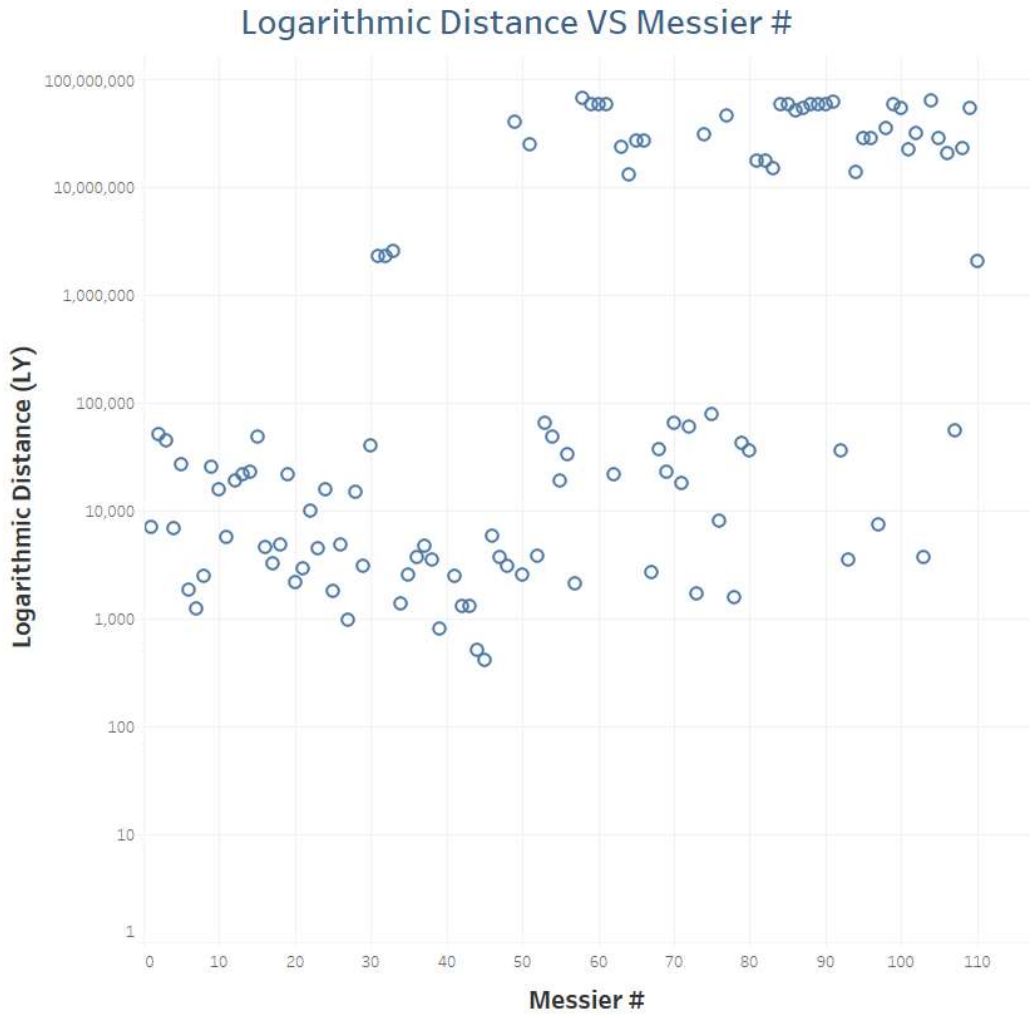


- The data is filtered on variable "**Test**" with "**Vertical Distance, Non-Aligned**". The data is also filtered with **Display 1**only. The color shows the different Trails B, C, and D.
- We can clearly see the anomalous sequence of Responses "**1.0**" on top of the graph. They are identical across all three Trials B, C, and D for a small subset of subjects.

Problem 2: 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 the18th 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.

For this dataset, you will have to pick suitable scales to make the data readable in your graphs. You should not wind up with most 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

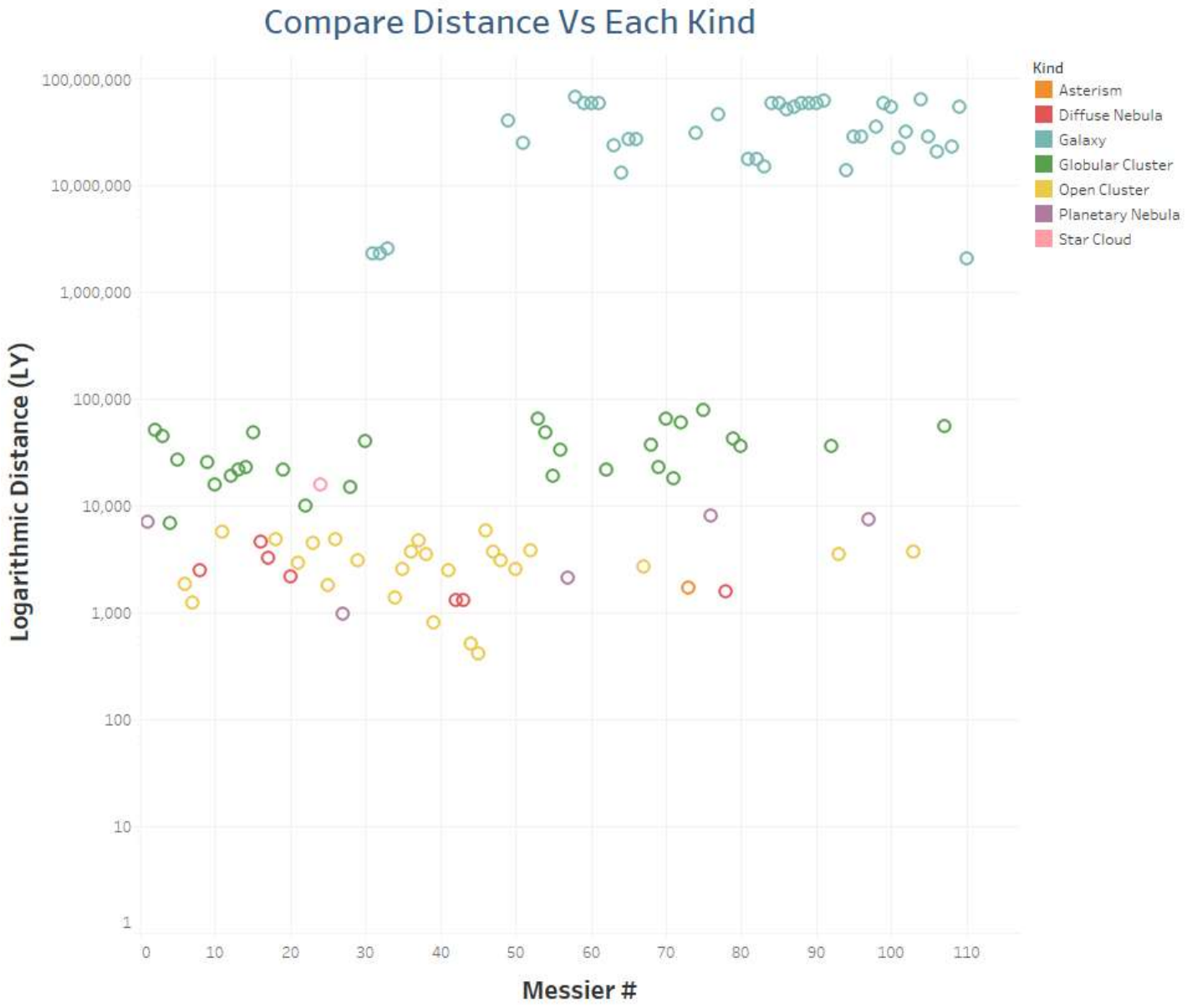
a) Start by trying to graph one or more properties of the objects against the Messier Number. Remember, there is nothing 'intrinsic' about this number; it is just the order of Messier's list. Is there any property that exhibits a pattern with respect to the ordering in his list?



Messier # vs. Distance (LY).

- For variable "Distance (LY)", the scale of the y-axis was changed to Positive Log10 in Tableau.
- After converting the scale to log10, it is clearly visible that there are two distinct clusters of points.
- The cluster of values on top right corner of the graph starts Messier # 50.
- Although there is nothing 'intrinsic' about the Messier Number, but it will be interesting to find out why there are two distinct clusters of points.

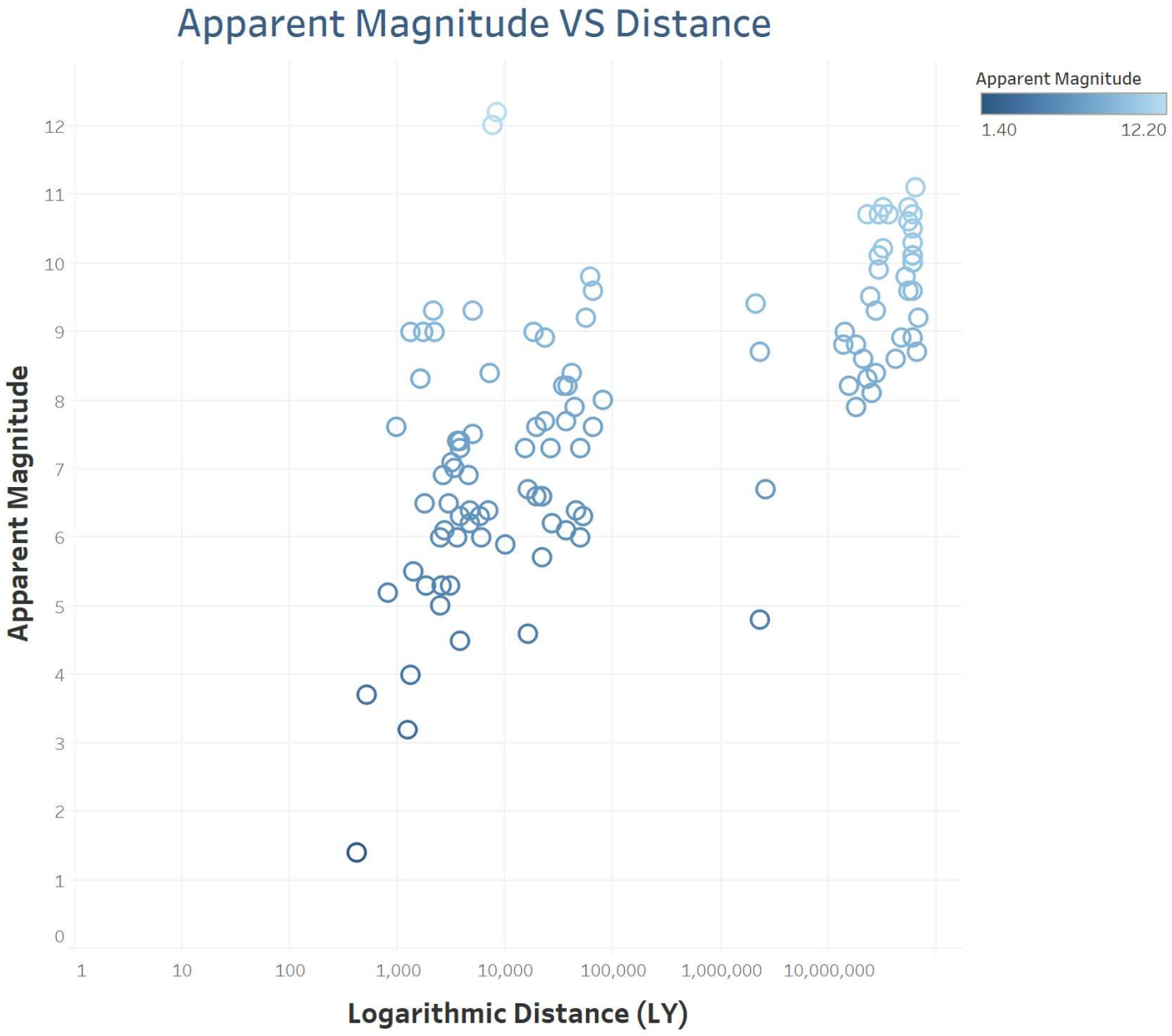
b) Create a visualization that compares the distributions of the distances to the objects in each Kind. Note that the Type variable is a very different category and is really a subcategory of Kind. Do not use that here. Sort the distribution displays in a way that makes the relationship clear.



Messier # vs. Distance (LY). Color shows details about Kind. The view is filtered on Kind, which excludes Null.

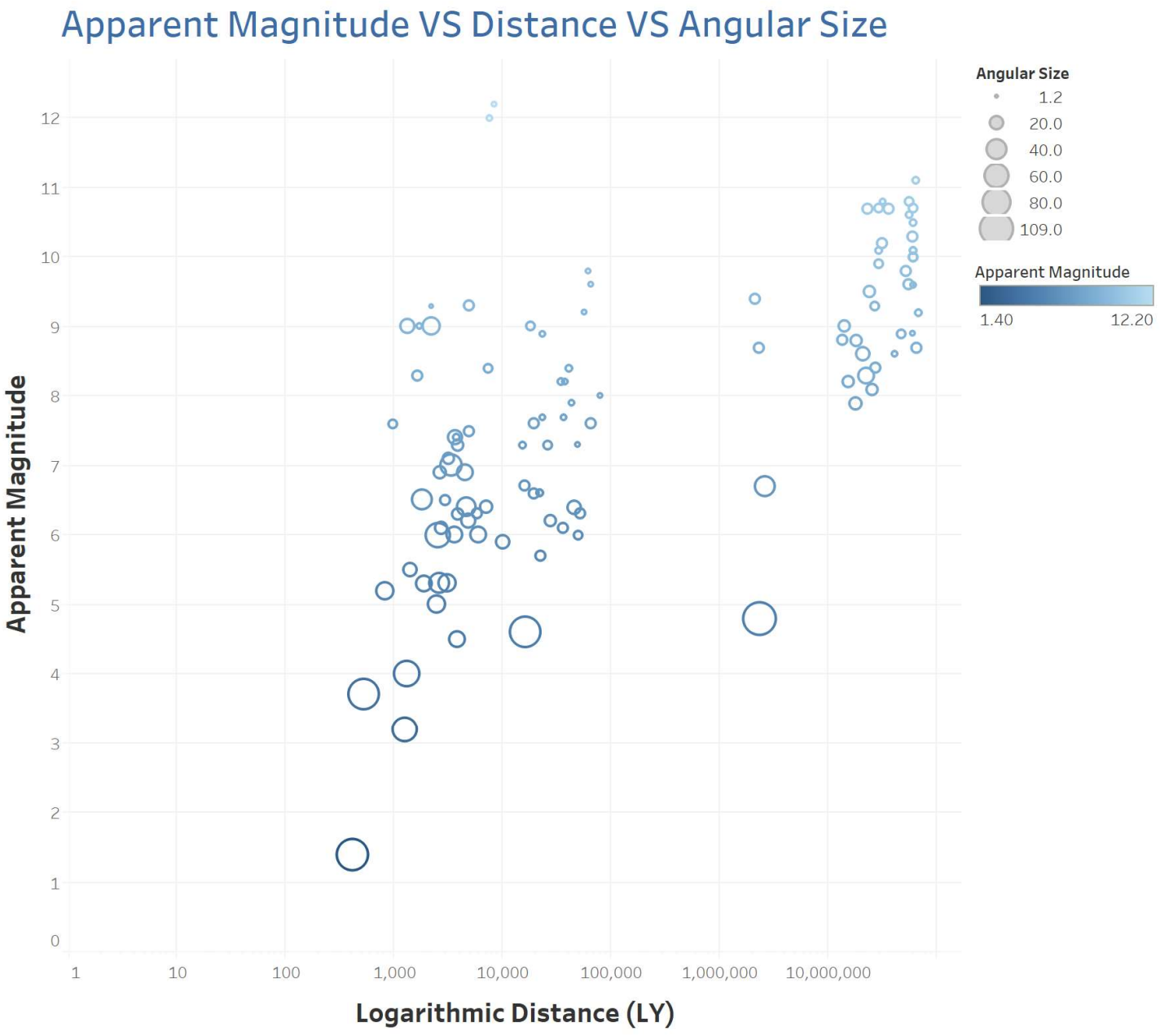
- The view is filtered on Kind which excludes the Null/Blank values. The different colors show the details about different kinds of objects.
- There are clearly three clusters of points identified by different colors Galaxy (sky blue), Globular Cluster (green), and Open Cluster (yellow).
- The points for Galaxy are more clustered around the Messier# between 50 - 110.
- There are further two groups of values for Globular Cluster. The first group is between 0 to 30 Messier # and the second group is between 51 to 80 Messier #.

c) Create a scatter plot with the distance to the Messier objects plotted against their Apparent Magnitude (it's their visual magnitude, a measure of how bright they are in the sky). Note that these values may be… backwards from what you would think. The **higher** the number the **fainter** the object is in the sky. Try to **incorporate that into your visualization to make the relationship clear**.



Distance (LY) vs. Apparent Magnitude. Color shows Apparent Magnitude. The data is filtered on Kind, which has multiple members selected.

d) Augment the visualization in (c) by adjusting the size of the points in the scatter plot based on the angular Size of the objects in the sky. Evaluate how easy it is to analyze all encoded aspects of the data from this graph and give a suggestion on how you might modify the graph to display all this information more readably.

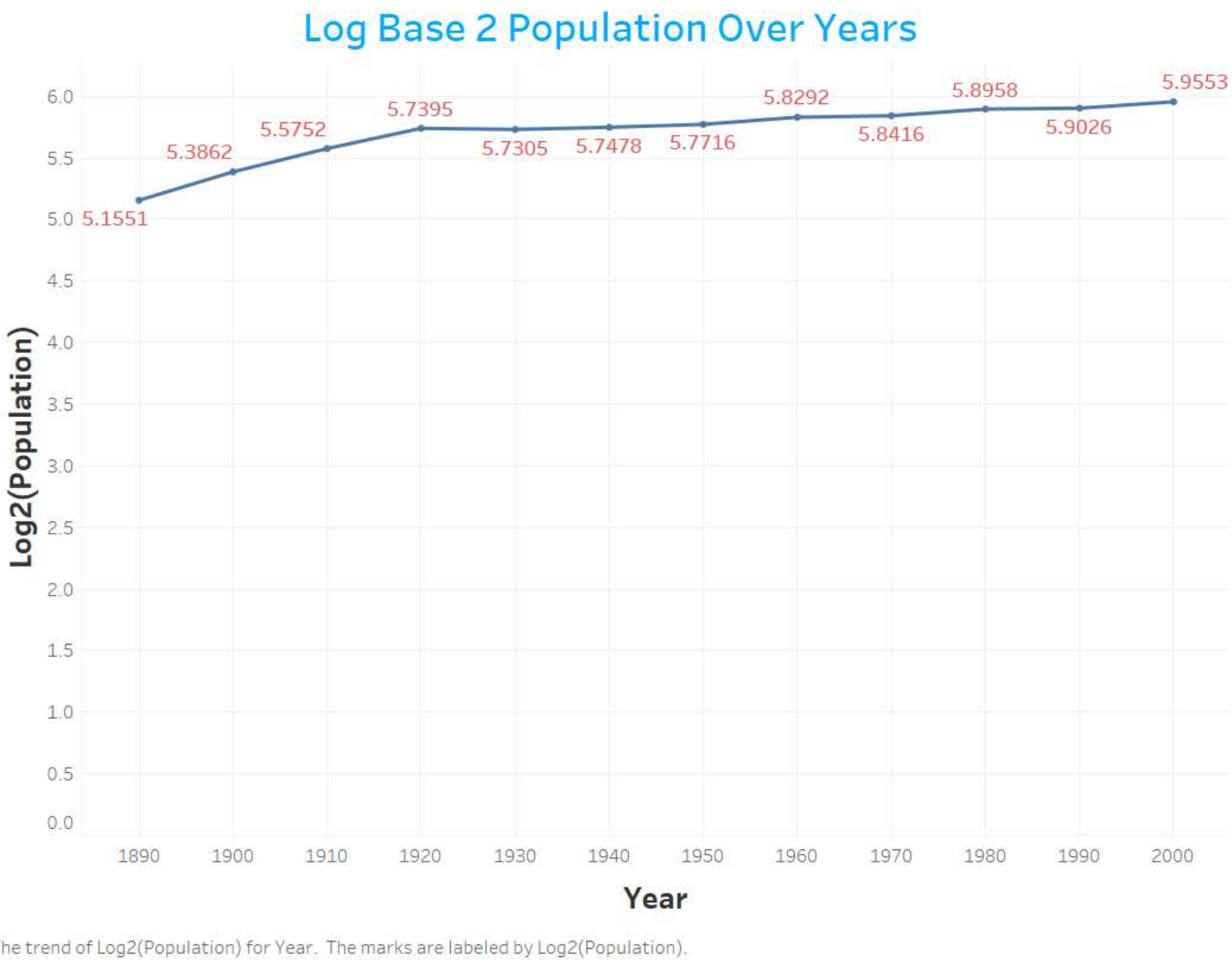


Distance (LY) vs. Apparent Magnitude. Color shows Apparent Magnitude. Size shows Size('). The data is filtered on Kind, which has multiple members selected.

- After adjusting the size of the points in the above scatter plot based on the variable '**Size**' (Angular size of the objects in the sky), it is very easy to see that:
  - Further away the objects in the sky, they become fainter and smaller in size.
  - The closer the objects are, they are bigger in size and brighter in color.

Problem 3: Download and graph the Montana Population data set (different from the one we used previously). Create visualizations using logarithmic scales, and intended for a technical audience, that **clearly** demonstrate **visually** the answers to the following questions. Viewers should be able to read the answers to these directly off the graph scales. Different logarithmic scale techniques may be appropriate for each part. If you use a single graph to answer multiple parts, make it clear that you are doing so.

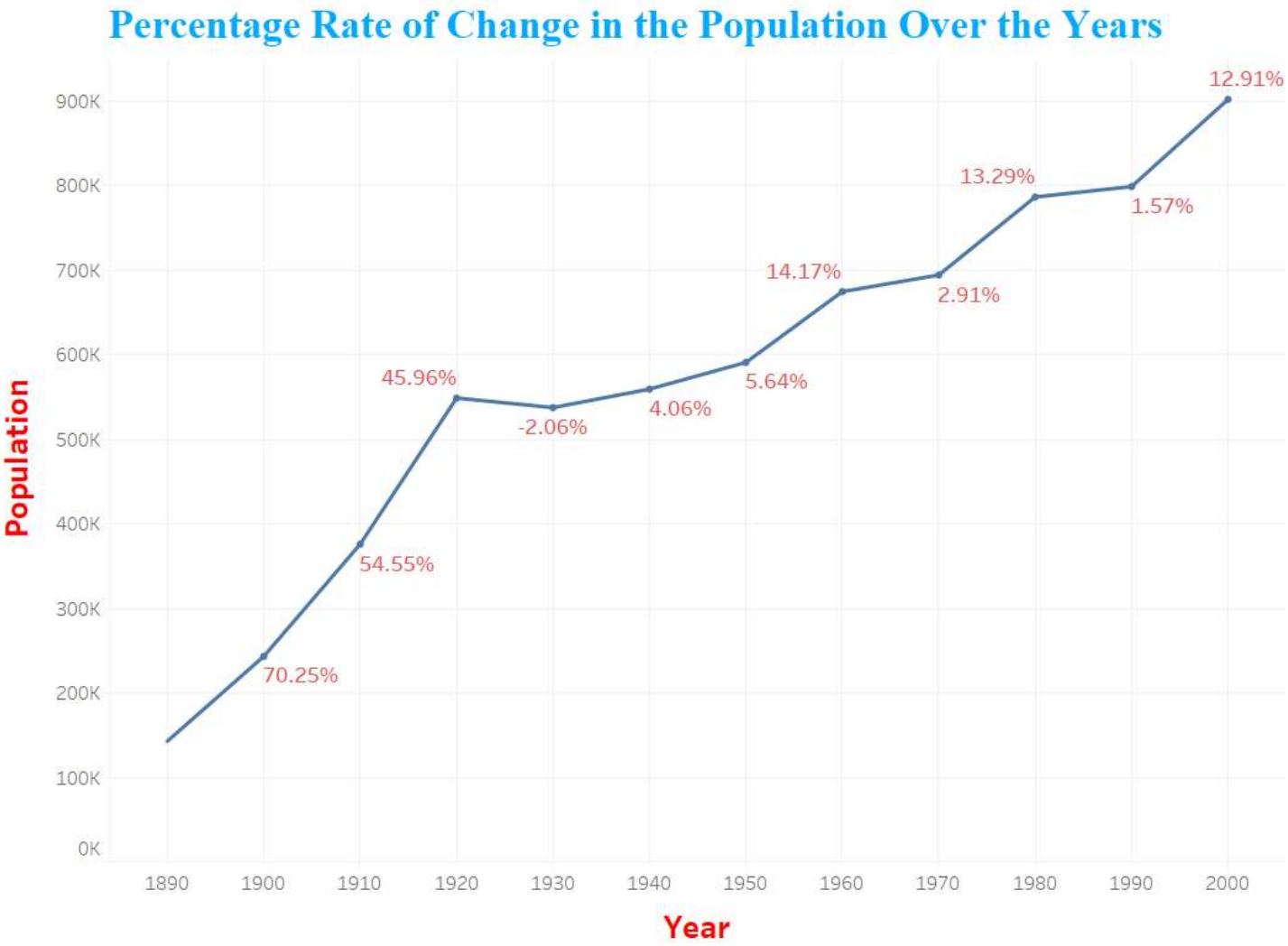
a) How many times has the population doubled since 1890?



The trend of Log2(Population) for Year. The marks are labeled by Log2(Population).

- The population doubled from 1890 to 1920.

b) Has the percentage rate of change in the population increased or decreased over the years? What years had the greatest increase in population %-wise?



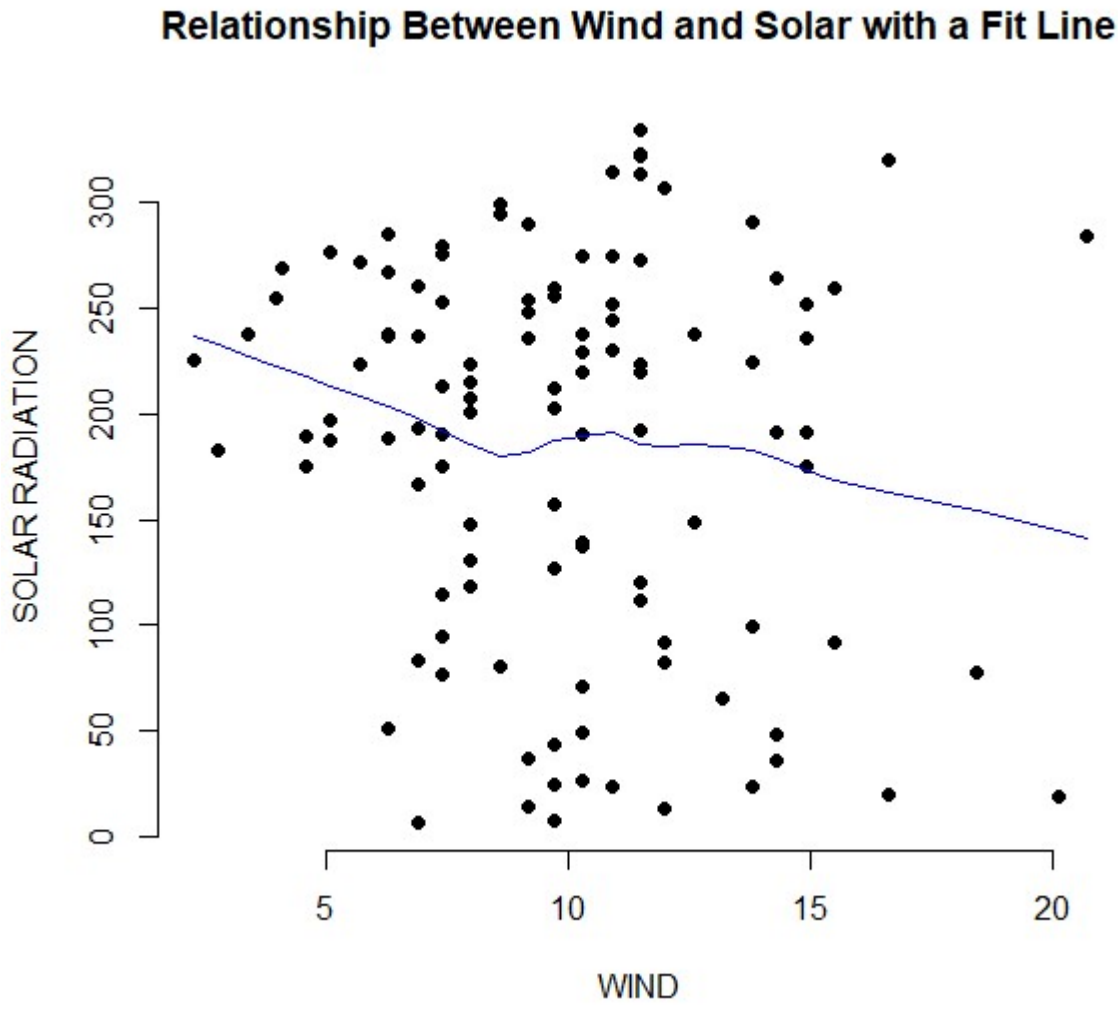
- The trend of population shows that the there has been percentage rate of increase in population since 1890 except the decrease of percent population from 1920 to 1930.
- The biggest increases in percentage wise population were from 1890 to 1920.

c) What years was the population percentage increase greater than 15%?

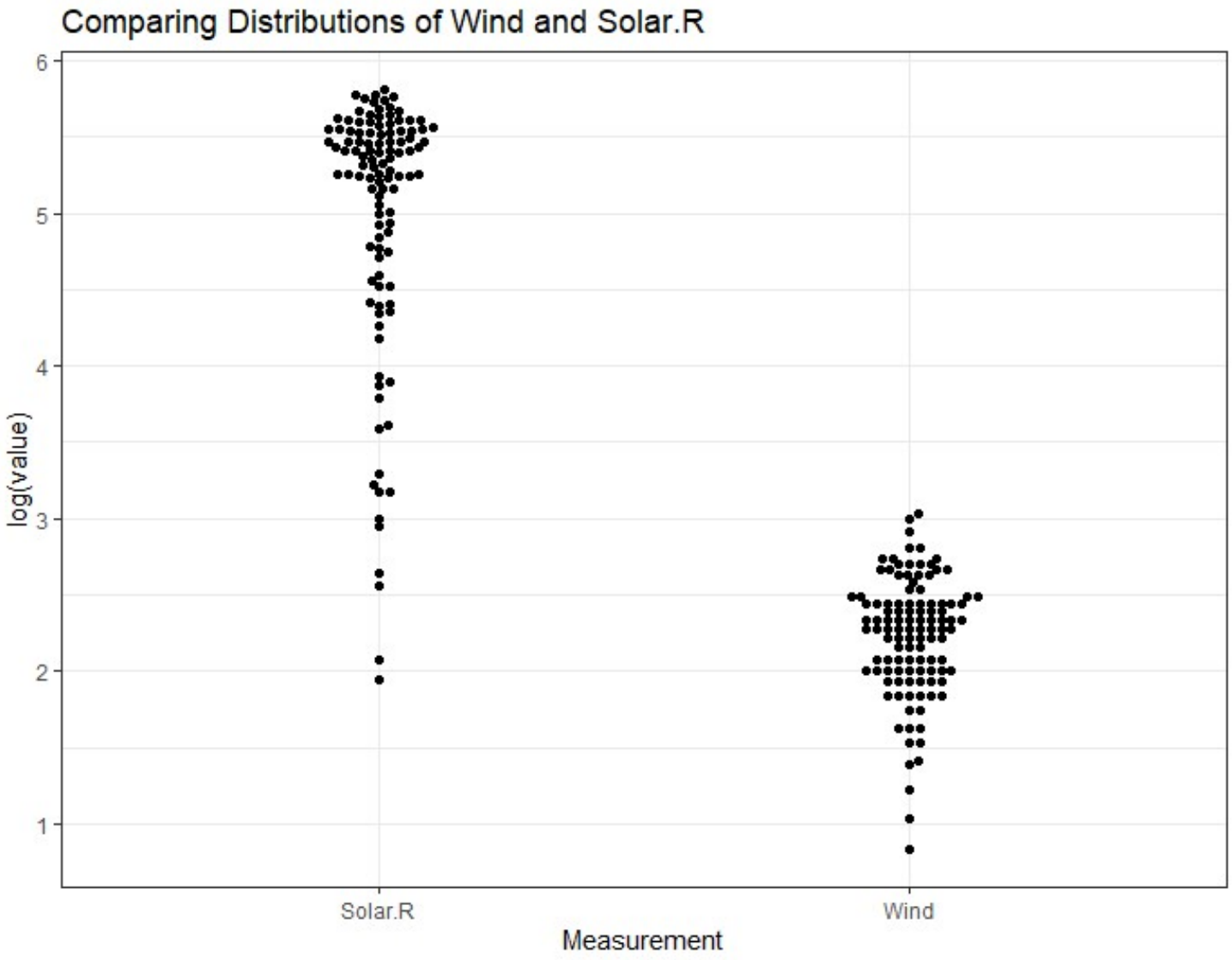
- As we can see in the previous graph (b) that there were big population percentage increases in 1900, 1910, 1920. Those were the greater than 15% increase.

Problem 4: We will look at data on air quality, captured from May to September in New York. This is built into R, but not as a data frame. There is a copy on the D2L site.

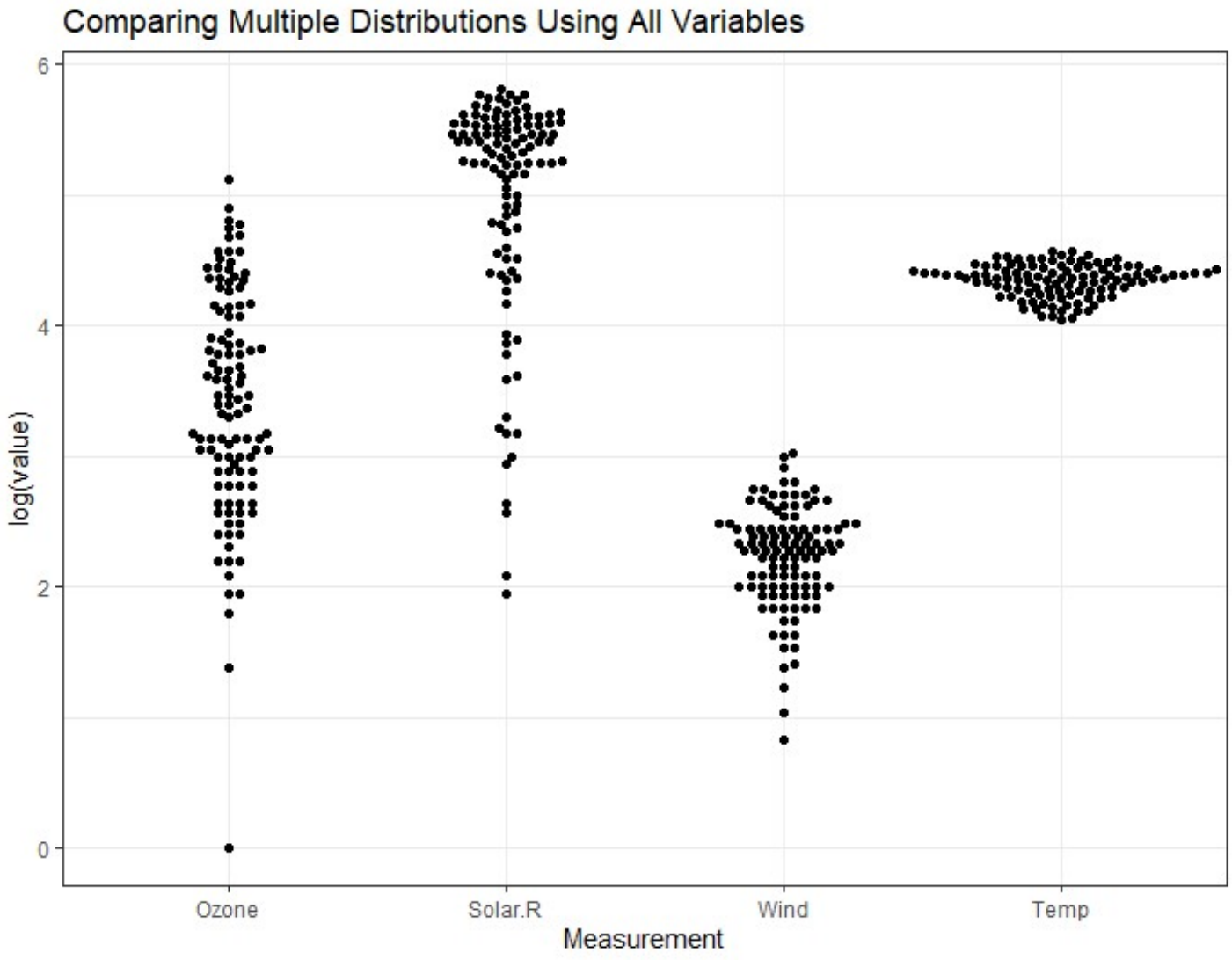
a) Use a scatter plot to look at the relationship between Wind and Solar.R (solar radiation). Show a fit line. Make sure to produce a clean visualization with emphasis on the trend. This provides one view of the relationship.



b) Use a plot that will show the distributions of Wind and Solar.R and allow you to compare with fine detail.



c) Finally, show these distributions in context of the rest of the variables by using a technique for comparing multiple distributions.



d) For extra credit, compare Wind and Solar.R again with a QQ plot. What does this tell you?