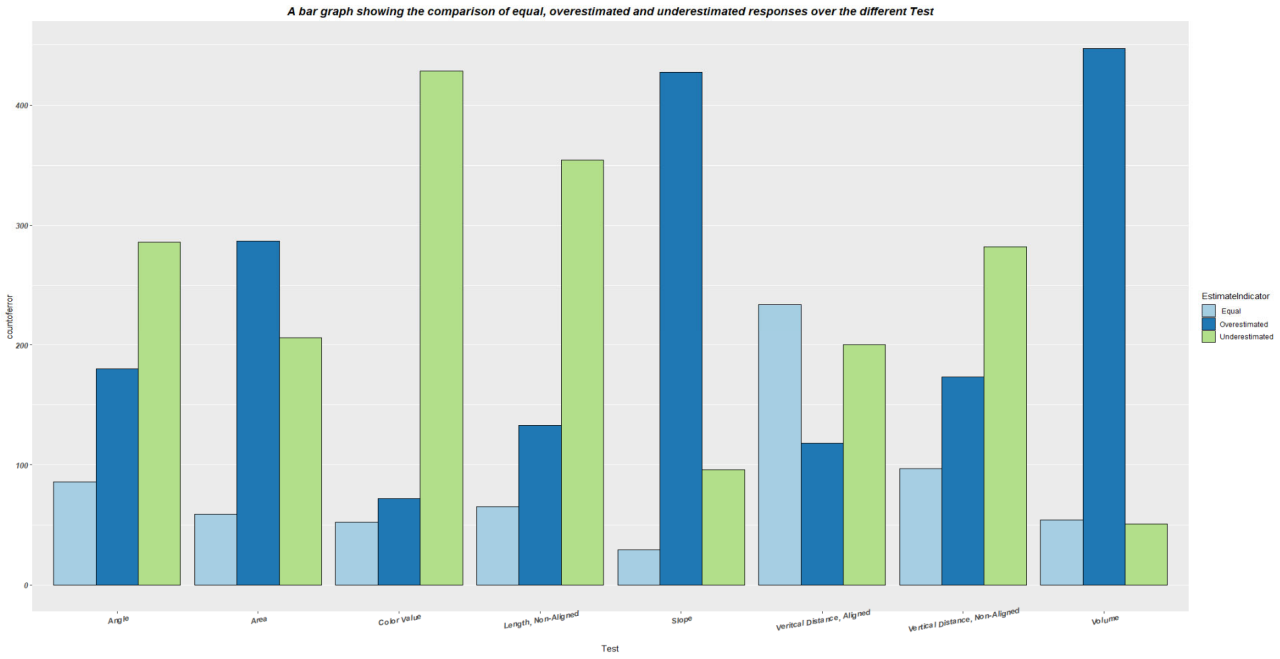
DSC 465 HomeWork3

Student ID: 2020637

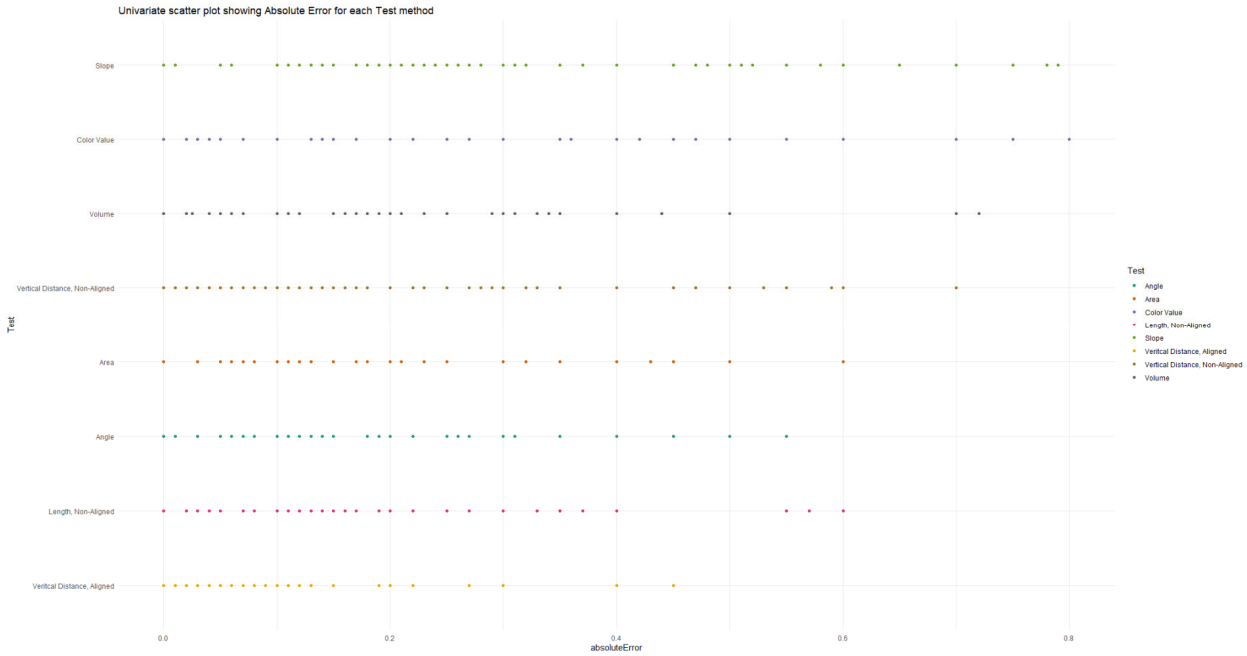
Name: Ronaldlee Ejalu

1) Perception Experiment data set
    a. A bar chart showing the comparison of equal, overestimated, and underestimated responses over the different Tests



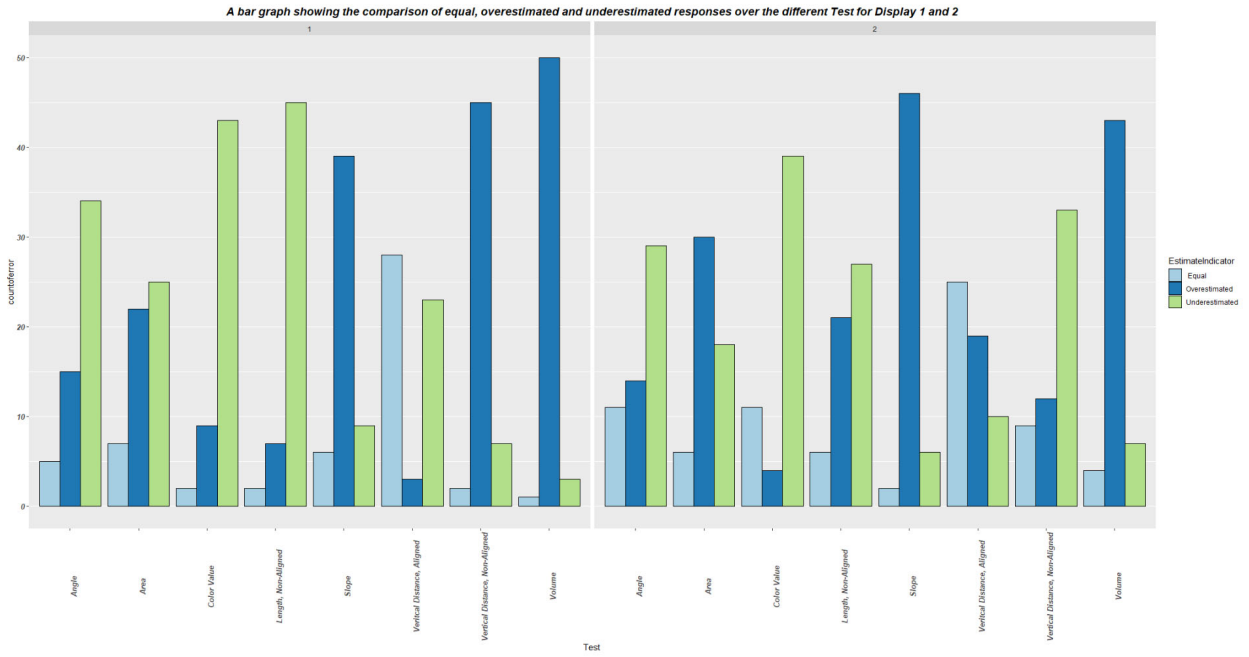

Yes, they were Tests where subjects either gave an overestimated or underestimated response. The above graph shows the proportion of equal, overestimated, and underestimated responses for the Tests in the data set. I derived a calculated field, EstimateIndicator using Error field where Error equal to 0.0 where defined as Equal responses, Error less than 0.0 were defined as underestimated responses and Error greater than 0.00 were defined as Overestimated responses so using the Test, count of error and EstimateIndicator. Using grouped bars, I can visualize the count of overestimated, underestimated and equal responses for each Test as this visualizes them as proportions of multiple categories of the EstimateIndicator amongst the values of the Test.

b. Univariate scatter plot showing the Absolute Error for each type of Test in R



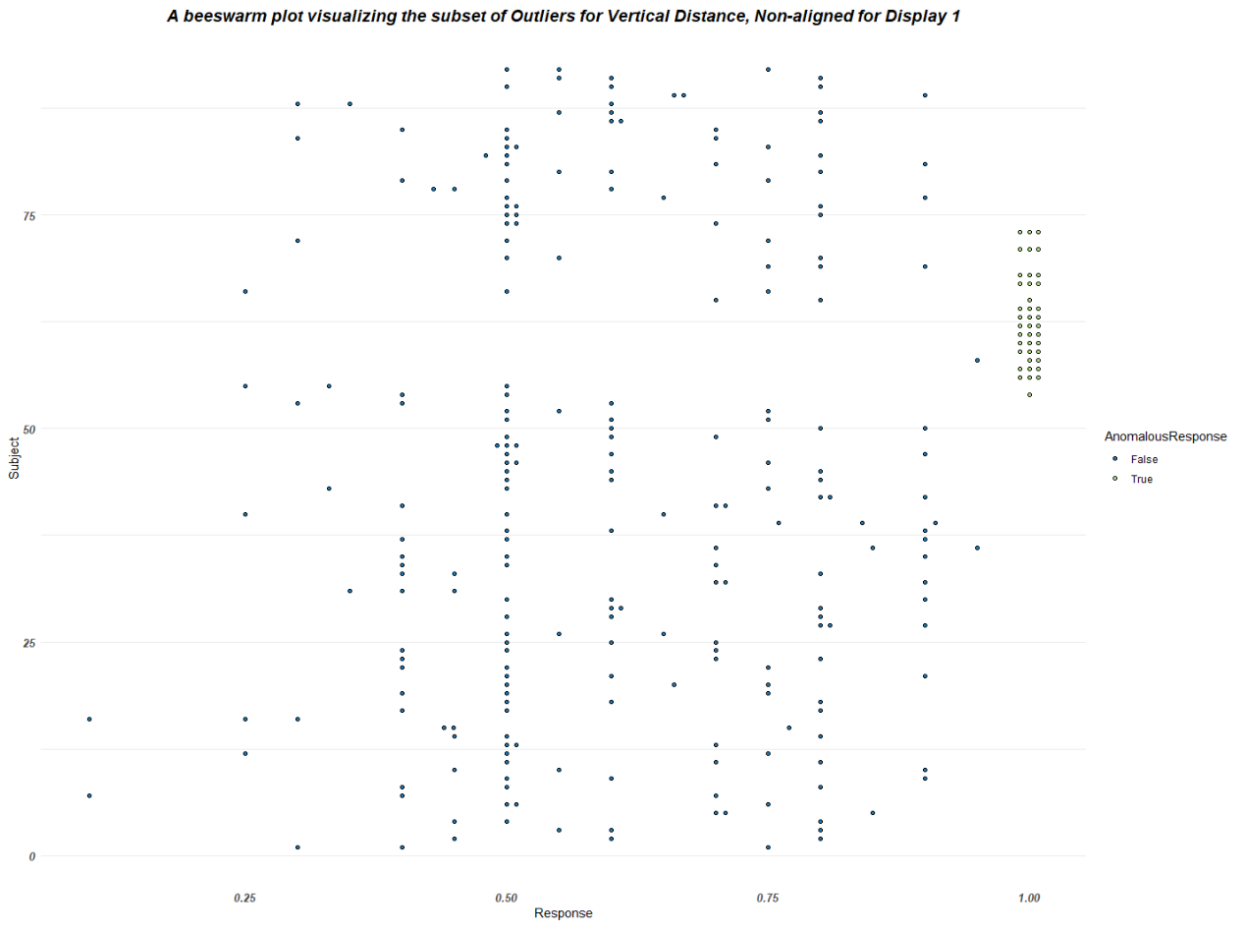

Yes, there is a noticeable clumping of the responses in the following Test Types: Slope; Vertical Distance, Non-Aligned; Length, Non-Aligned. Using a calculated variable in R, I derived the absolute error using the absolute error function with derived Error field. The distribution of the various Test methods can be seen from the above univariate chart: Color Value and volume have some evenly spread out responses. Vertical Distance Non-Aligned has more responses between 0.0 and 0.3. Vertical Distance Aligned, Length Non-Aligned, Angle, Area have more responses between 0.0 and 0.2. Slope has more responses between 0.1 and 0.3.

c. Comparing the data displays for 1 and 2 for subjects 56-73 in R




Yes, we can a noticeable difference between Display 1 and Display 2. The two different graphs side by side visualize the count of error for the different Test for Display 1 and Display 2 after filtering the subjects 56-73. It is shown that Display 2 has more equal responses for the Angle; Color Value; Length Non-Aligned; Vertical Distance, Non-Aligned and Volume whereas Display 1 has more equal responses for Area, Slope, and Vertical Distance Aligned. Yes, we can Display 2 was better when compared with Display 1 as students gave correct responses for Display 2.

d. Visualization of the raw scores in a way that highlights these values and makes their anomalous nature clear.

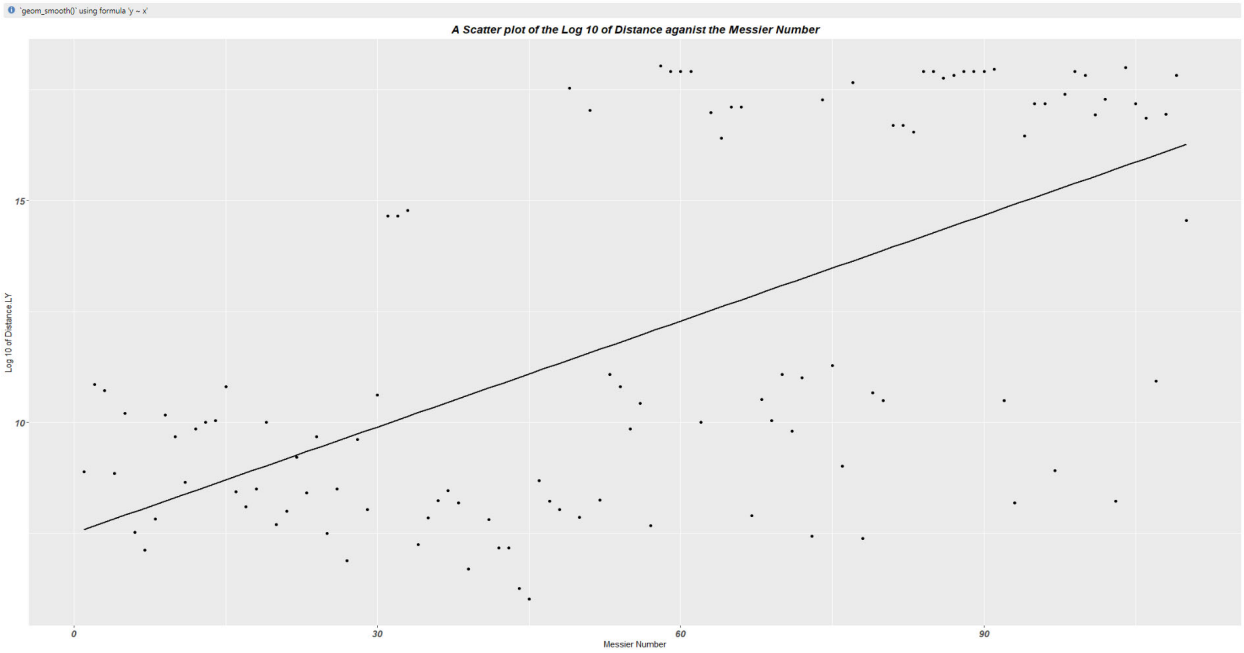



The above beeswarm plot visualizes the small subset of responses from the subjects as outliers in green color using a calculated field AnomalousResponse where the values of Responses marked as 1.0 are shown as outliers. This chart was put together using a filtered data set Test field value of Vertical Distance, Non-aligned for the Display 1. Also, I used the Subject and Response variables to come up with the beeswarm plot where the AnomalousResponse variable was used to differentiate the outliers from the regular responses and this was used for coloring as shown above.

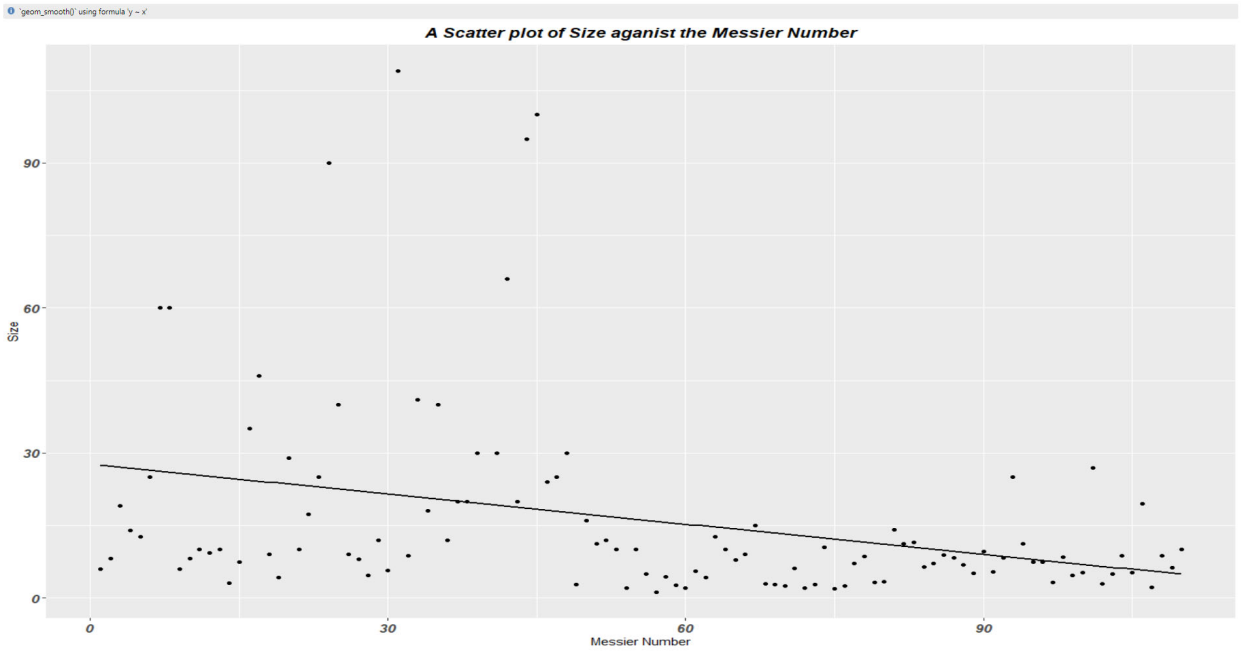
2) A astronomical data set

   a. Graphing one or more properties of the objects against the Messier Number

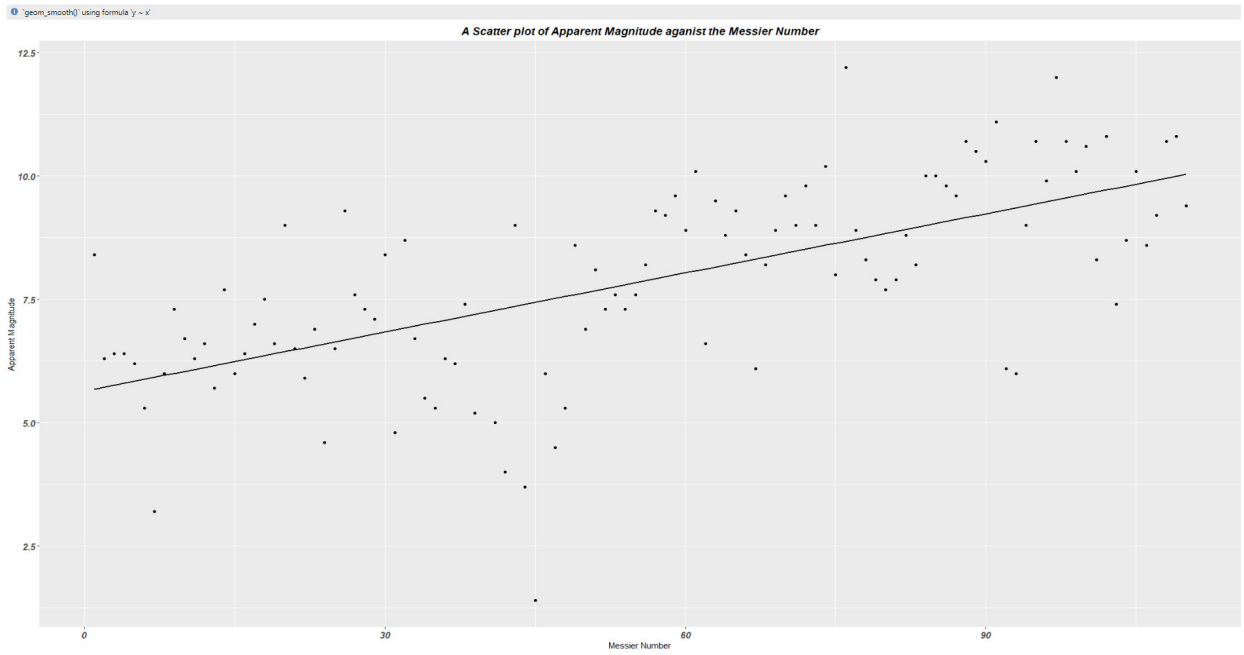
- A scatter plot of the log 10 of Distance against the Messier Number




- A scatter plot of Size against the Messier Number



- A scatter plot of Apparent Magnitude against the Messier Number




The above three scatter plots define a relationship:

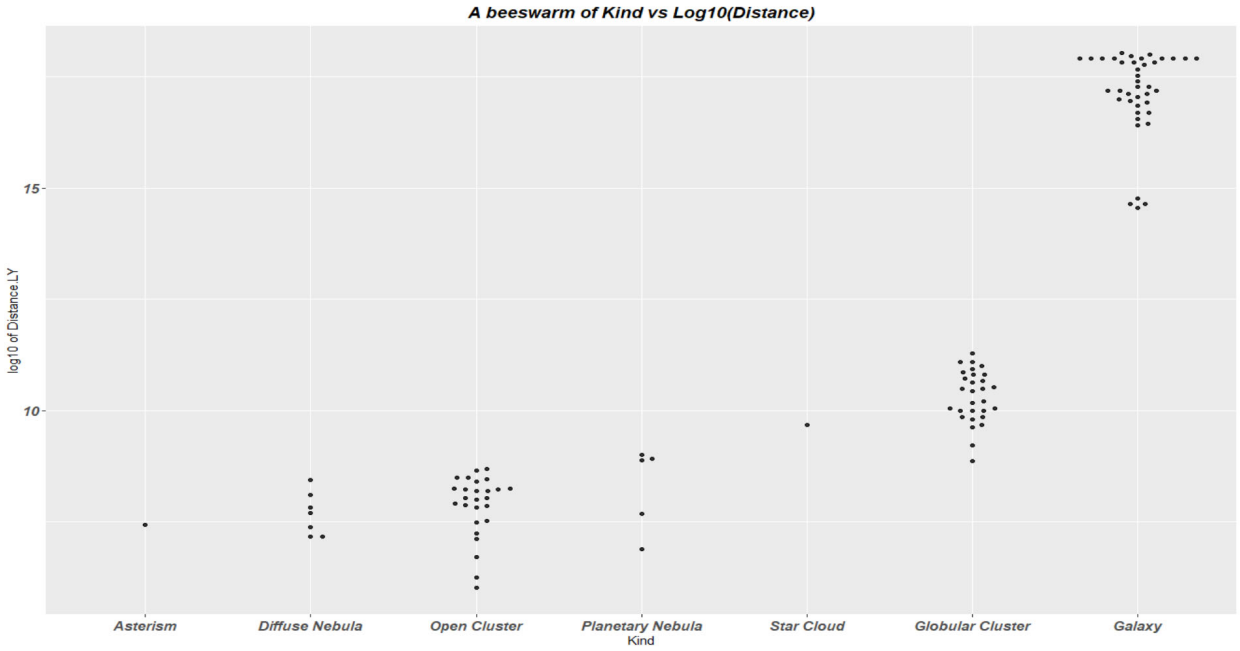
i. between Size and Messier Number

ii. between Apparent Magnitude and Messier Number

iii. between log 10 of Distance and Messier Number.

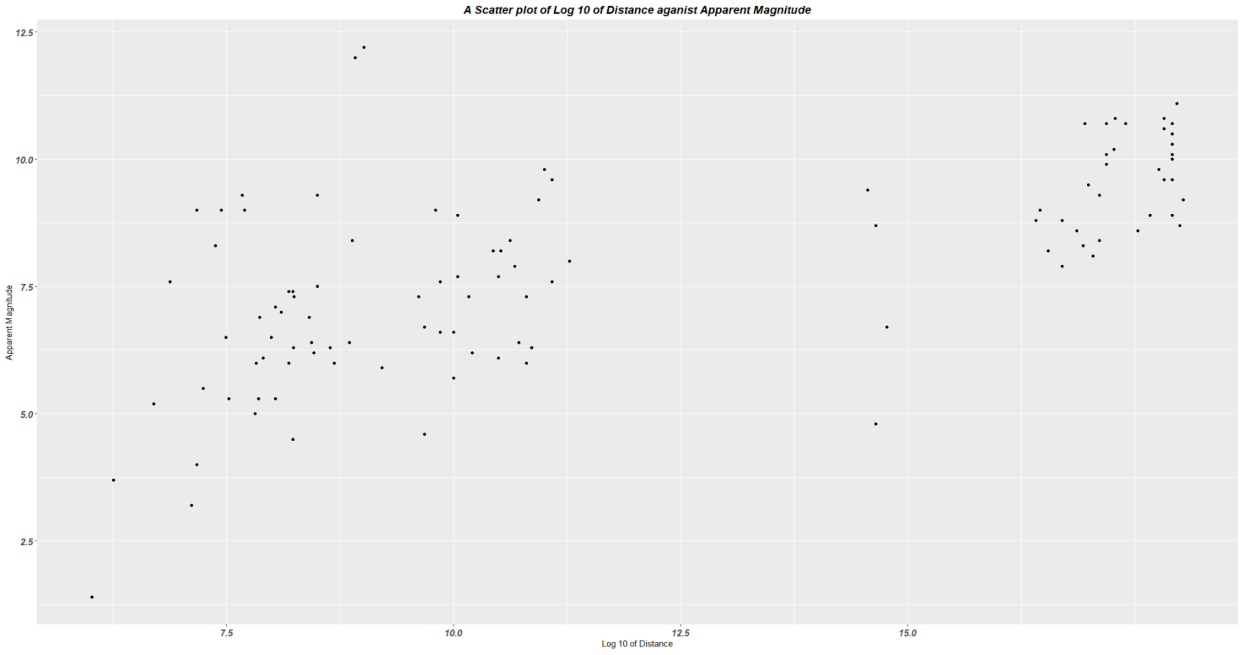
Amongst all the three properties, Apparent Magnitude exhibits a pattern with respect to the ordering in Messier's list. This is exhibited by the points which are around the regression line in the last scatter plot of Apparent Magnitude against the Messier Number.

b. A beeswarm of Kind Vs the log 10 of Distance.LY




The above chart is a Bee swarm graph derived out of R studio for Kind Vs. the Log 10 of Distance.LY. The graph visualizes the distribution of the distance in log 10 of the different Kinds of Messier objects.

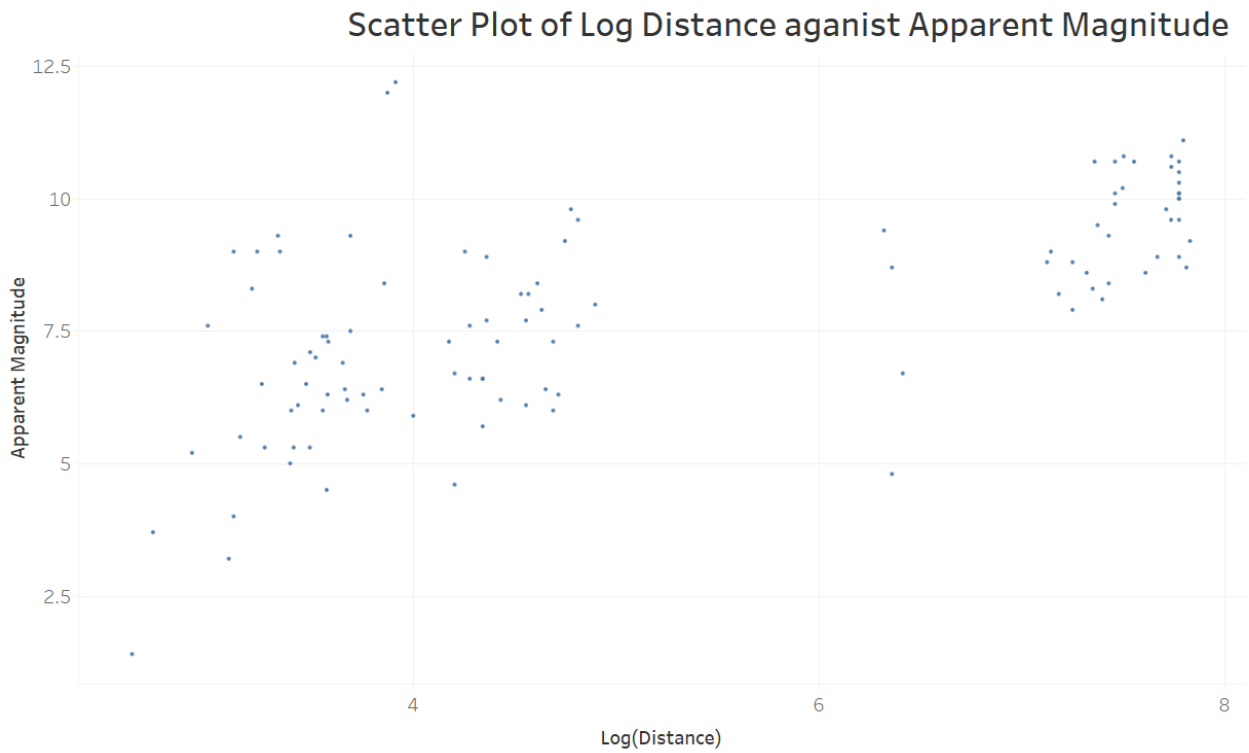
c. A scatter plot of log 10 of Distance against the Apparent Magnitude.




The chart above shows a scatterplot describing the relationship between the Log 10 of Distance Vs Apparent Magnitude of the Messier objects. As the distance

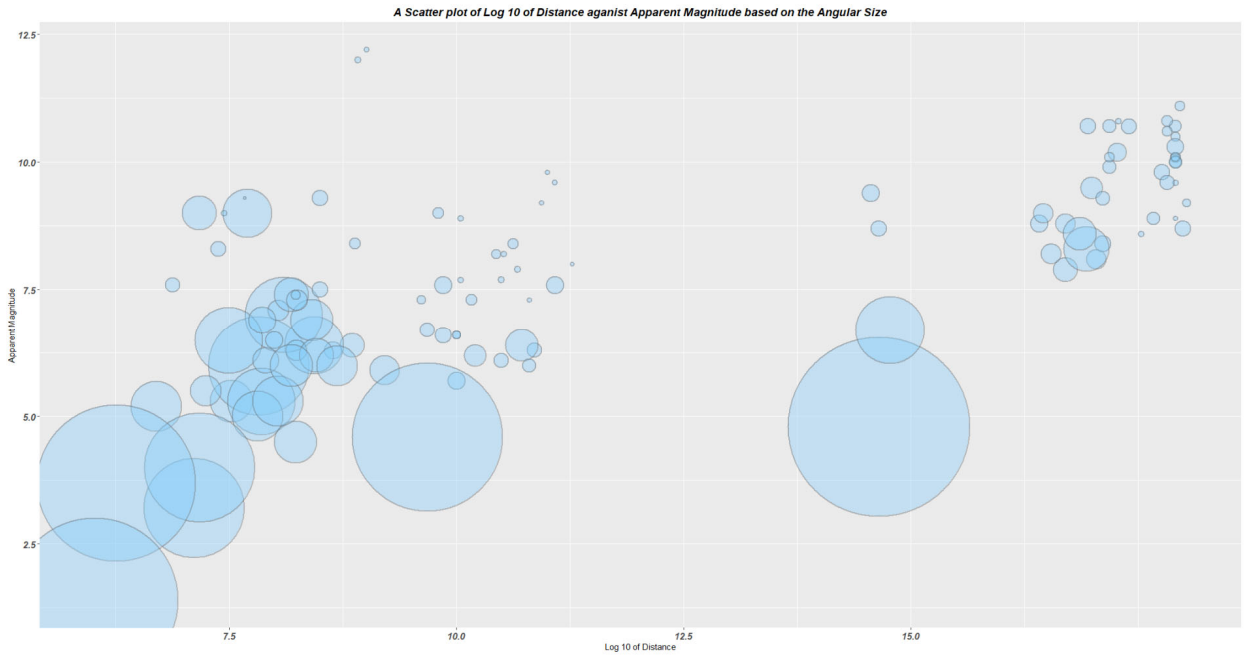
increases the objects become less clearly visible. As demonstrated above, the objects that are far may seem smaller in magnitude.

- **The same graph above in Tableau is shown below**:



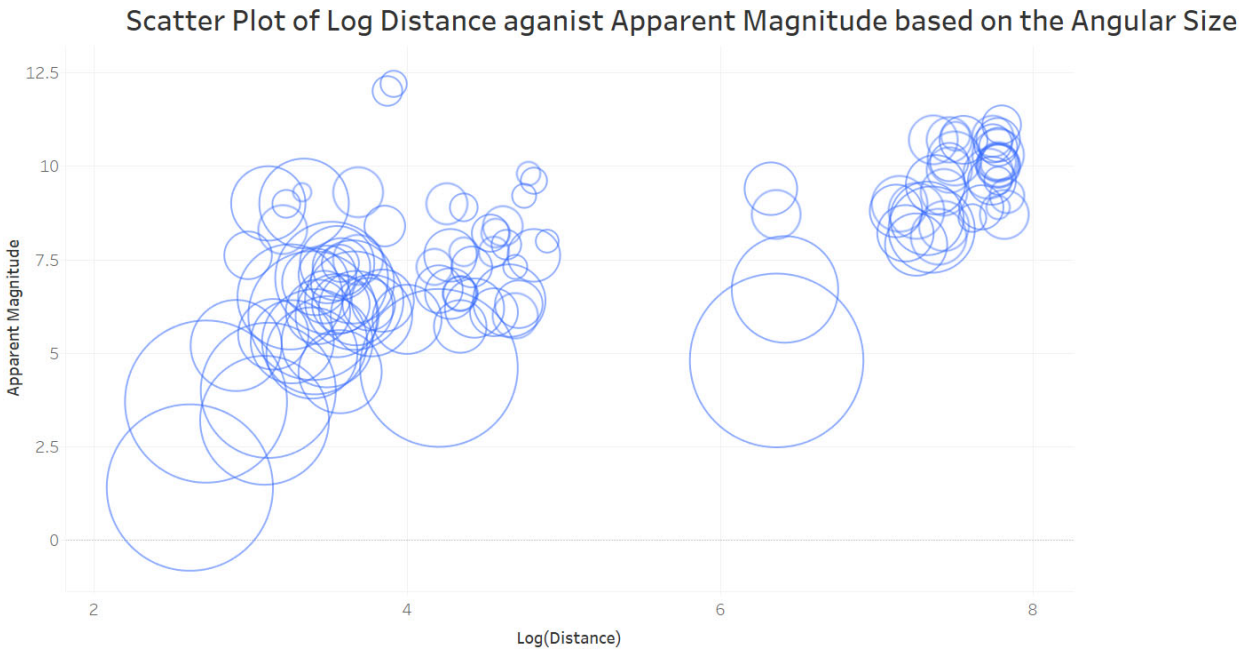

The above scatter plot was derived using a calculated field of log(Distance) and Apparent magnitude and this scatter plot describes the relationship between Log(Distance) and Apparent Magnitude; as the distance increases the objects become less clearly visible.

d. A scatter plot of log 10 of Distance against the Apparent Magnitude based on the Angular Size.



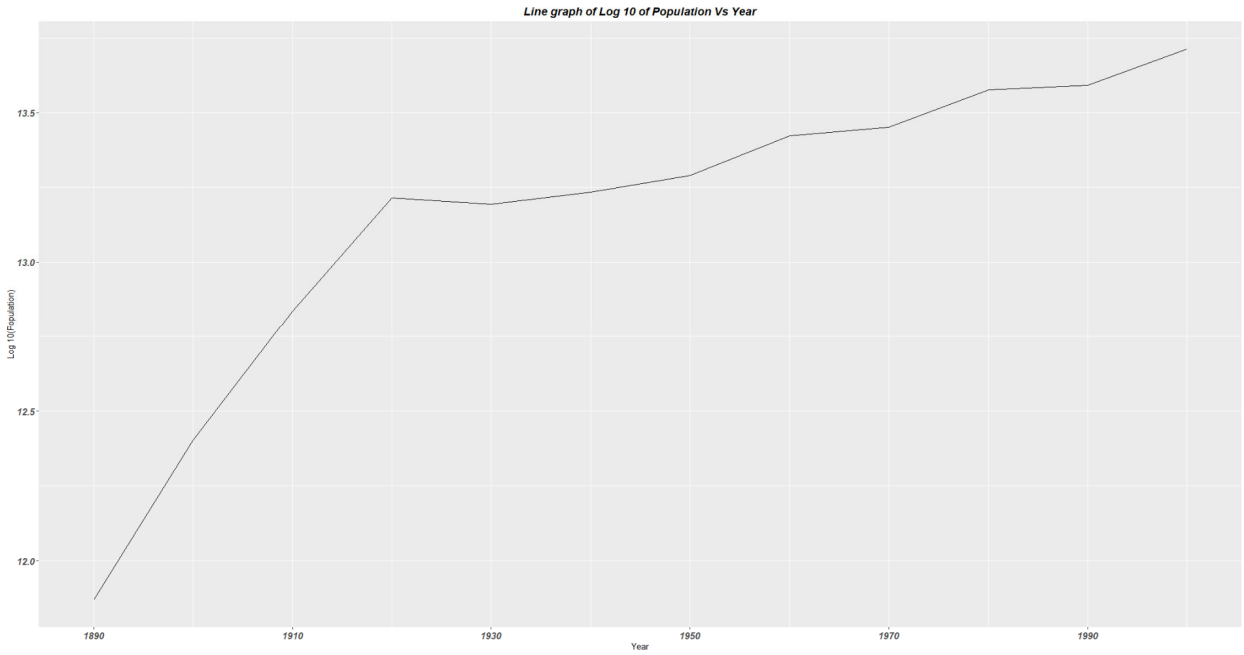

The above chart is a visualization of the log 10 of Distance against Apparent Magnitude with the representation of size using the "Size" attribute for representing the angular size of the Messier objects. Deriving the scatter plot, I used light blue color for filling the circles and also light black for the borders of the circles. A majority of the objects seem to be cluttered on the left, which make it difficult to understand the size of the circles and it looks like as the log of the distance increases, that is around 17.5, the size of circles are shrinking. Maybe changing the scale on the x-axis might be beneficial in visualizing the size of the differences of the objects.

- **The same graph above in Tableau is shown below**:



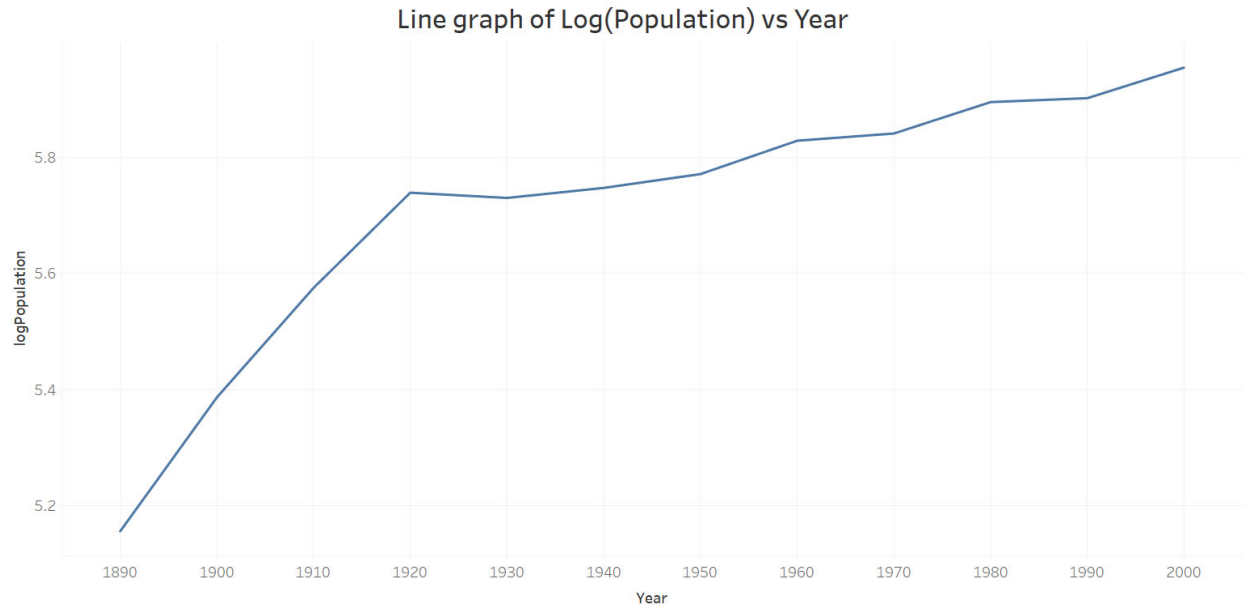

The above chart is a visualization of the log 10 of Distance against Apparent Magnitude with the representation of size using the "Size" attribute for representing the angular size of the Messier objects.

3) Montana Population data set
    a. Line graph of Log 10 of Population Vs Year



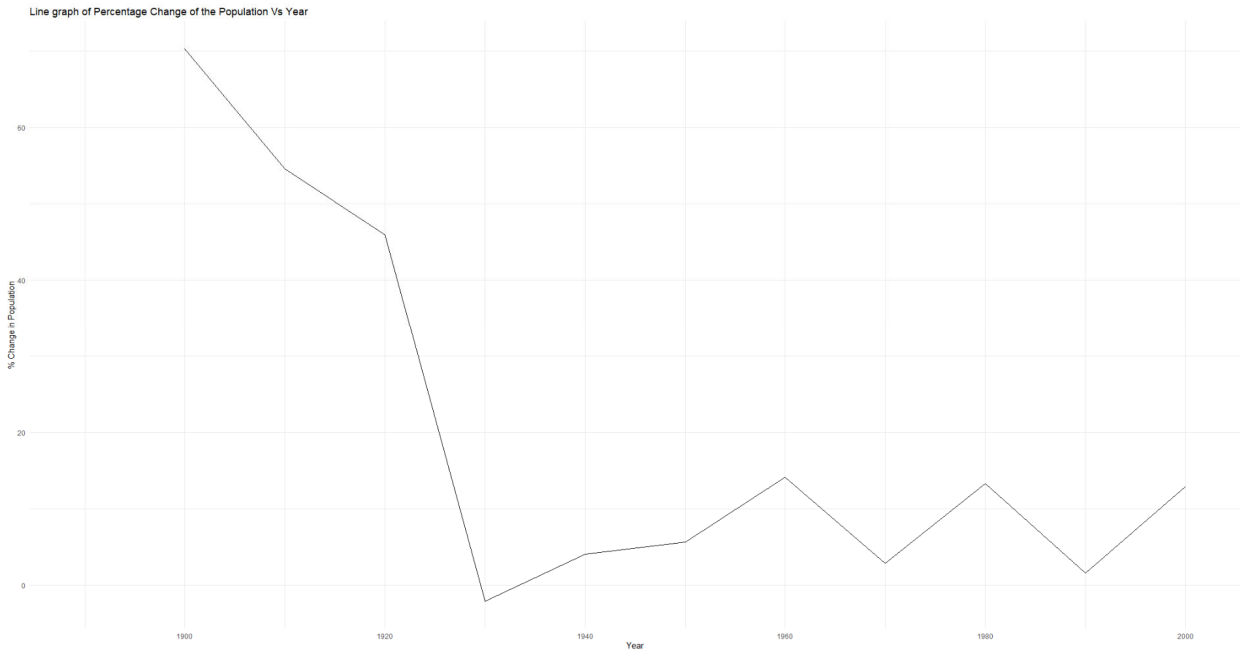

The above line graph was derived using the log 10 of Population on the Y-axis and Year on the X-axis. The graph exhibits an exponential growth in the population of Montana. Looking at the line chart, Montana population doubled 2 times.

- **The same graph in Tableau is shown below:**



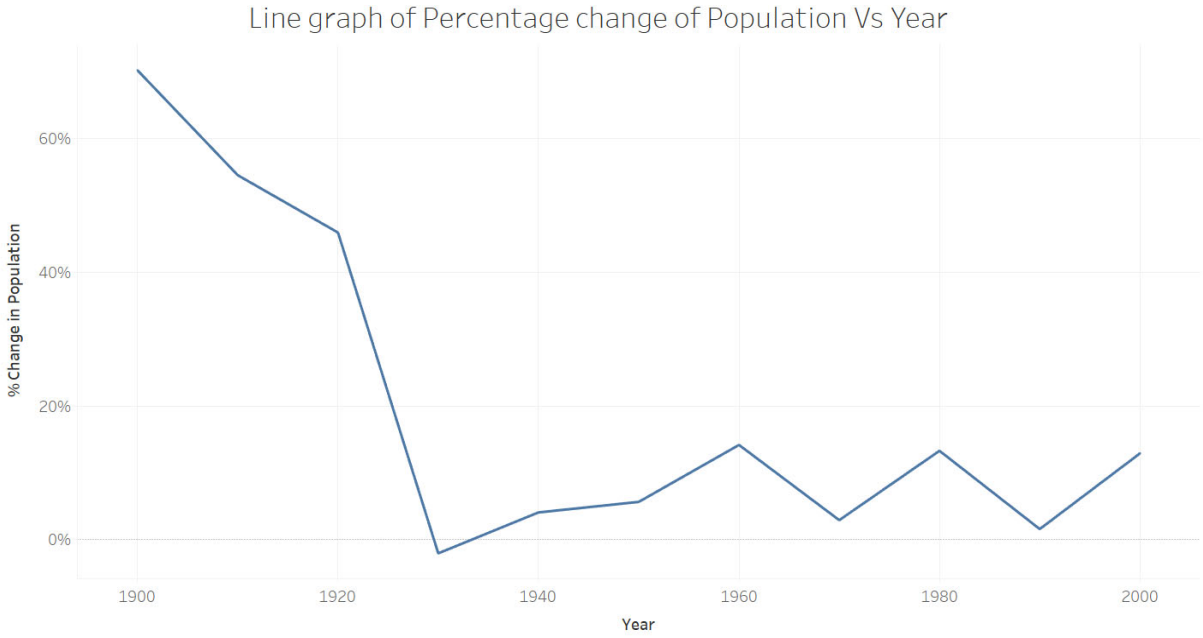

I derived the above graph using a calculated field on the y-axis which was created using the log function with Population field and Year on the X- axis.

b. Line graph of Percentage Change of the Population Vs Year in R




The above line chart shows that the greatest increase in the population was between the 1900 to 1910 and from 1920 to 1930

- **The same graph above in Tableau is shown below**:

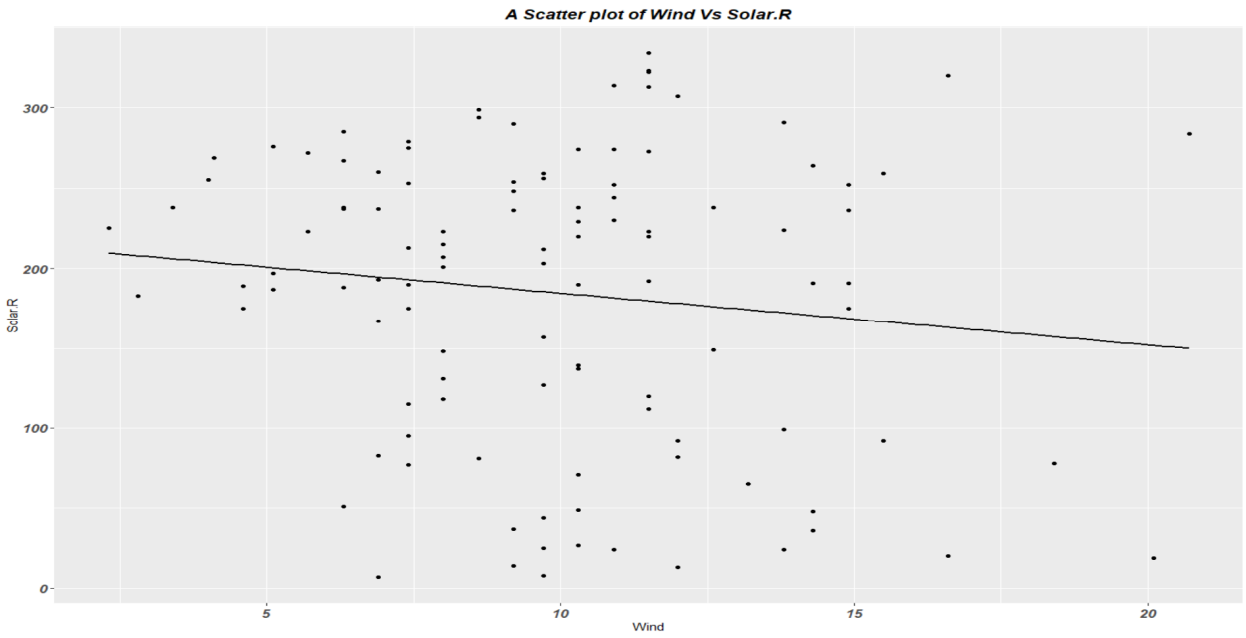


To come up with this graph, I had to right click on the SUM(population) pill in the masks card and select quick Table calculation then Percent Difference which gives us a percent difference for each year.

c. The year between 1900 and 1910 and 1920 to 1930 had the population increase greater than 15%.

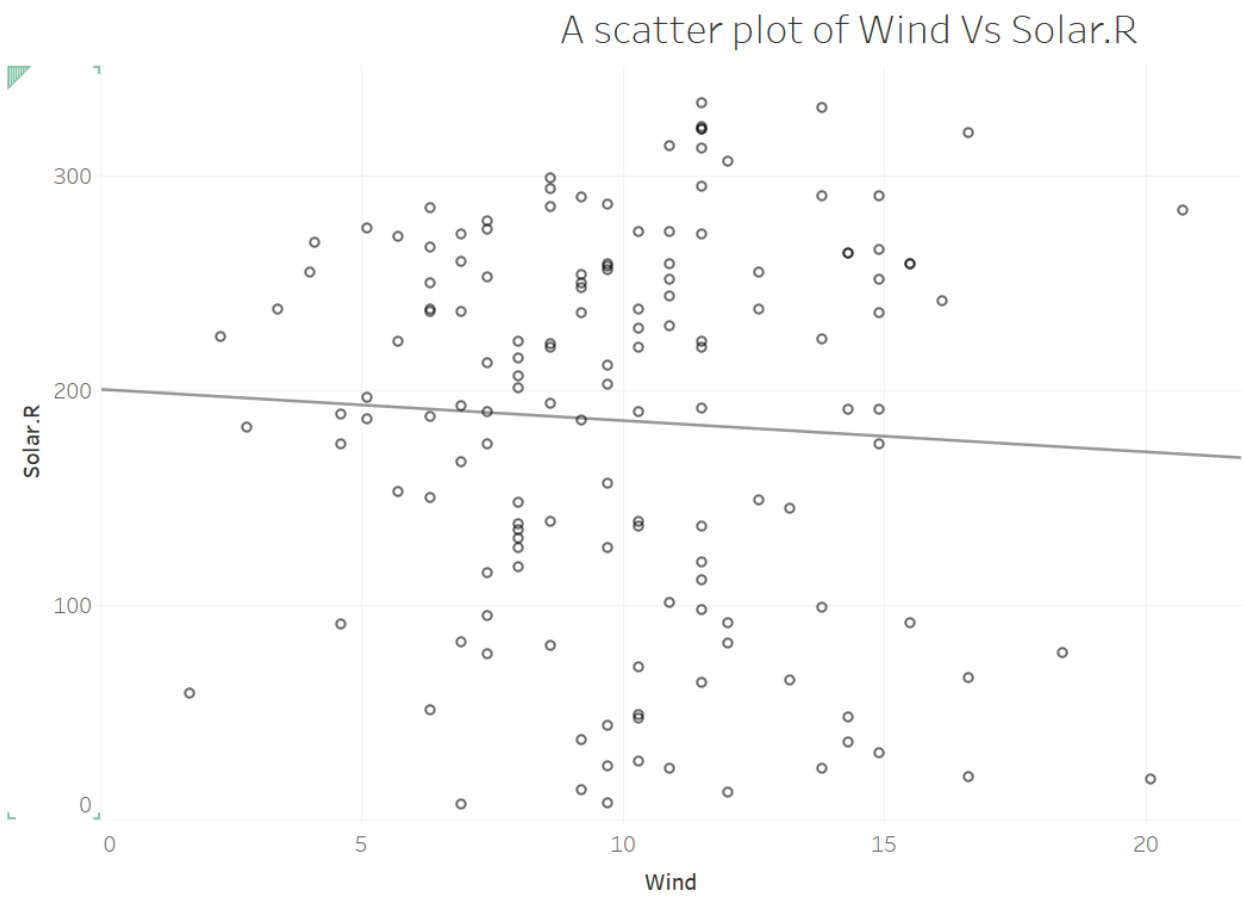
4) Air Quality data set

a. A scatter plot of Wind Vs Solar.R



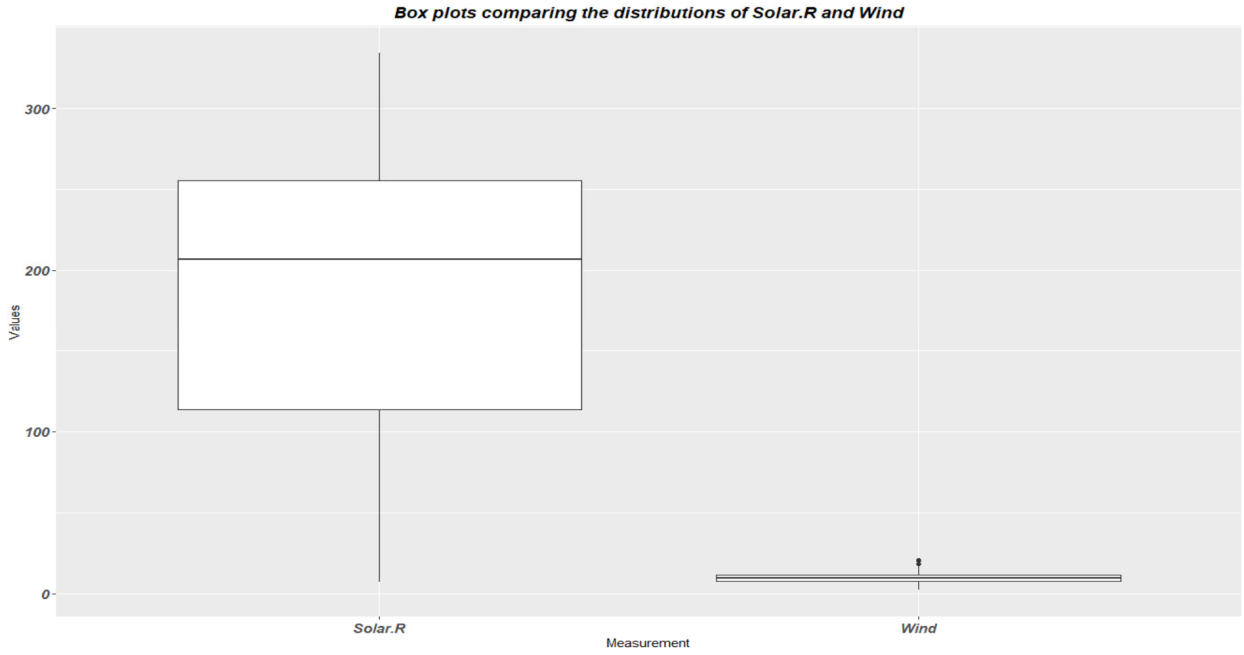

The scatter plot chart above with a trend line describes the relationship between Wind and Solar.R variables, they are both negatively correlated meaning that as Wind increases, Solar.R increases in the opposite direction.

- **The same graph in Tableau is shown below**:




The above scatter plot shows the relationship between Solar.R and Wind. They are both negatively correlated.

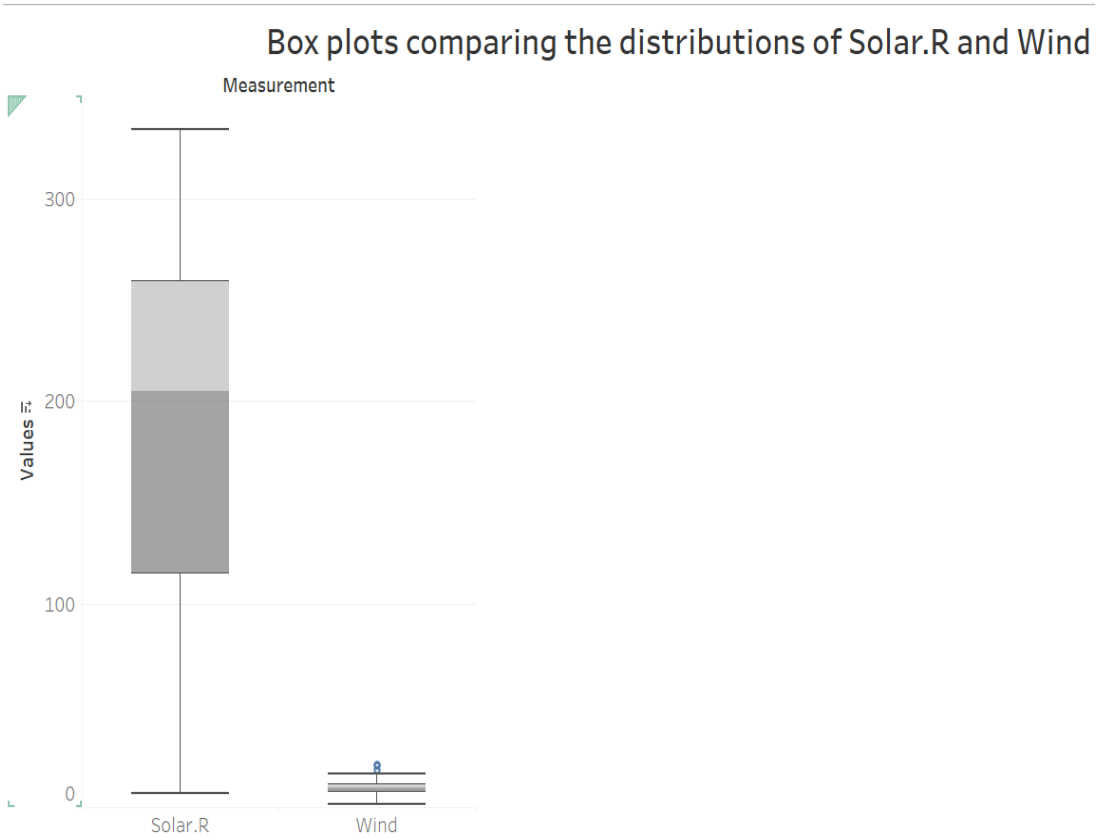
b. Box plots comparing the distributions of Solar.R and Wind.




Before deriving the box plot, I had to transform the data in R by using the Pivot_longer() function then filtered the data set using the Measurement levels of Solar.R and Wind. I used the Measurement values of Solar.R and Wind and their values on the y axis to graph the box plots which compares the distribution of these variables.

Both Box plots for Solar.R and Wind show that their median are on the above the mean meaning that the data range is negatively skewed. Also, it is demonstrated that the Wind measurement has some potential outliers, which need to be investigated and removed.

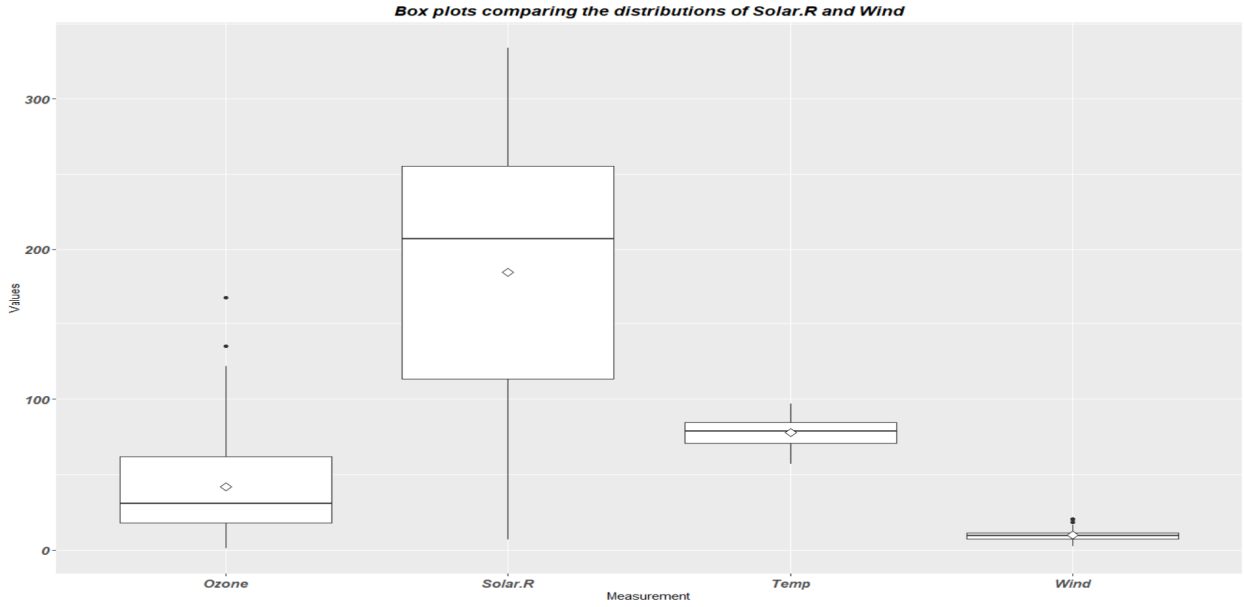
- **The same graph in Tableau is shown below:**




Before graphing these box plots in Tableau, I had to Pivot the following variables: Solar.R, Wind, Temp and Ozone, which ended up creating both the Measurement dimension and Values measure variables.

I used the Measurement on the columns axis and the Values on the rows axis to graph these box plots. I also filtered the data set by the Measurement Dimension to focus on the values of Solar.R and Wind. I removed the null values by filtering on Values measure by focusing on non null values. As you can see, the whiskers extend at most 1.5 interquartile range beyond the box. The boxplot for Wind has individual points, which are regarded as outliers. Furthermore, both box plots for Solar.R and Wind indicate that the median is skewed to the top meaning that their data ranges are negatively skewed.

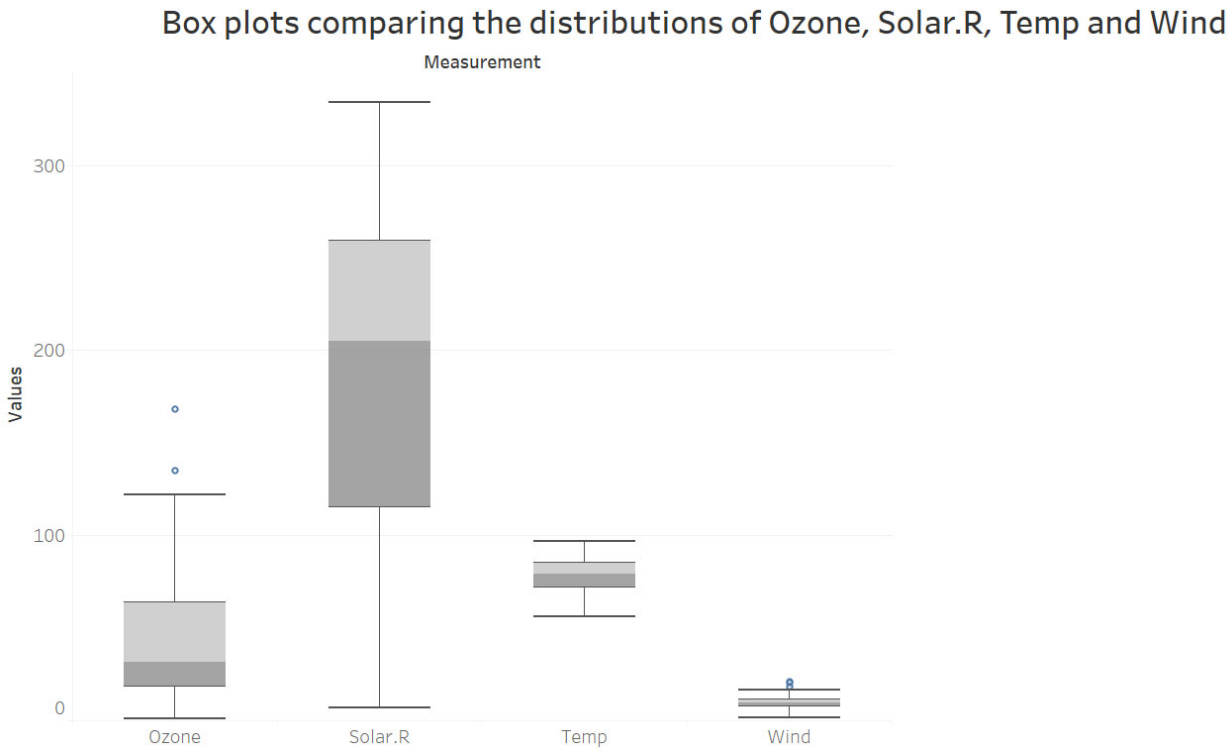
c. Box plots comparing the distributions of Ozone, Solar.R, Temp and Wind




The above box plot shows the comparison of the distribution of Ozone, Solar.R, Temp and Wind. The white diamond symbol within the box plots is the mean.

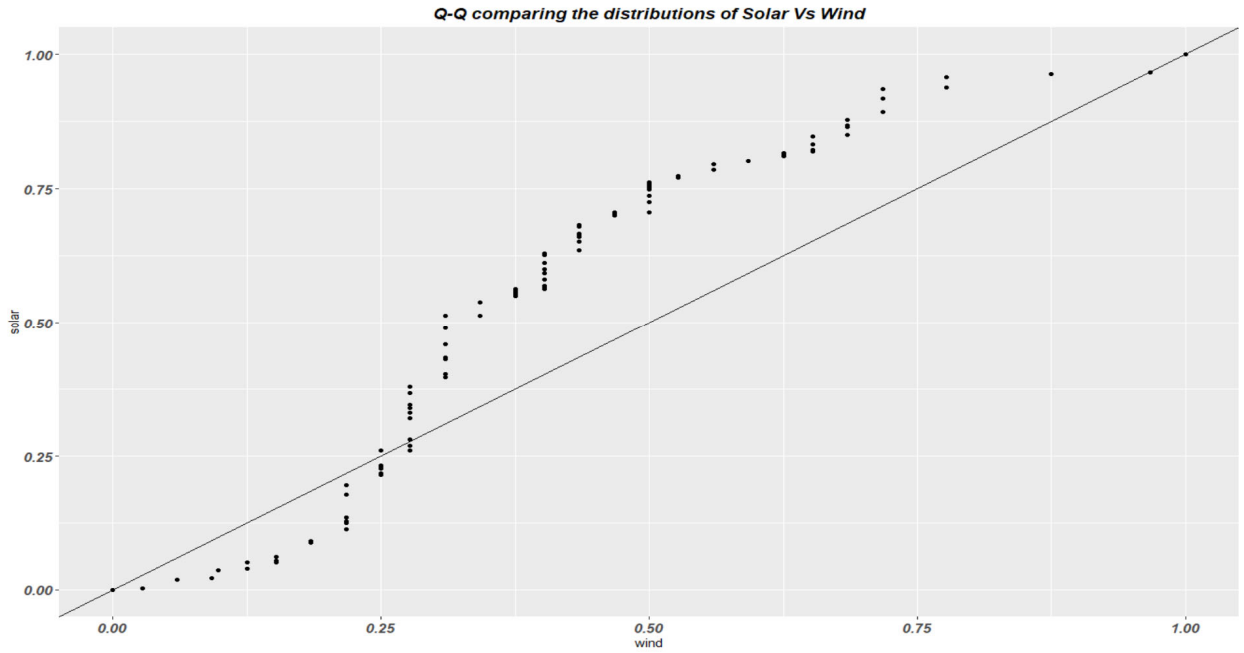
The median for the Ozone box plot is on lower end meaning that the data range is positively skewed, and it also has potential outliers, which need to be investigated and removed. The Box plot for Solar.R shows that the median is on the above end meaning that the data range is negatively skewed. The same that applies to the box plot of Temp and Wind, their median is above the mean meaning that both their data range is negatively skewed though it shown that the Wind Measurement has some outliers which need to be investigated and removed.

- **The same graph in Tableau is shown below**:




We learn that both Ozone and Wind measurements have Individual points which are considered as outliers and these points need to be investigated and removed.

d. Comparison of Wind and Solar using the QQ plot

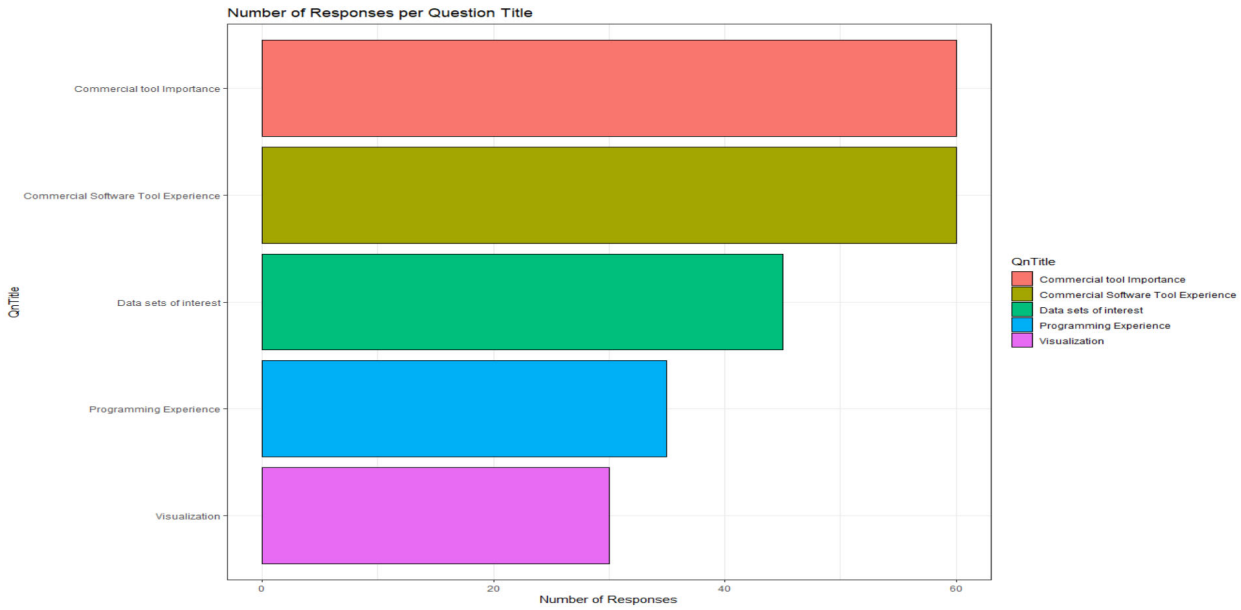



The Q-Q plot does not support normality because it seems to me that the distribution of samples are more skewed to the left.

5) Student Skills Survey data set

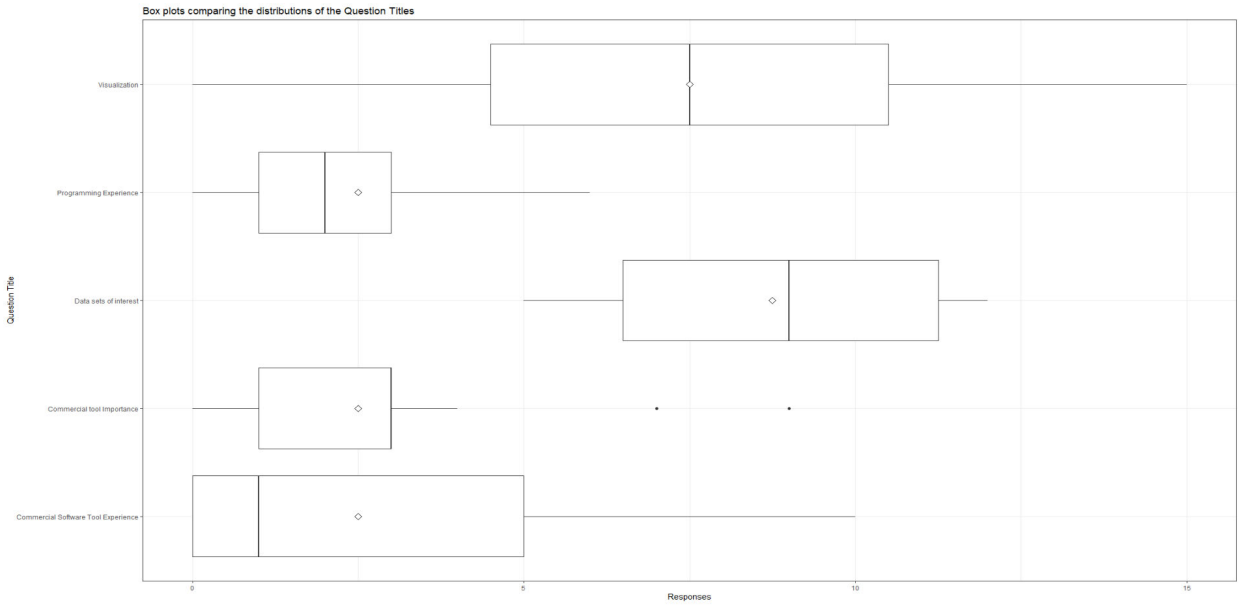
First, I cleaned up the data set by transforming the column names into meaningful names then I went ahead to drop Section, Bonus, Difficulty variables from the data set since their records contained missing values. Also, the first four records corresponding to the Question Text of Visualization had null values in the QnTitle variable so I had to re
After cleaning up my dataset, I came up with the following visualizations:

- Bar chart showing the number of responses per Question Title




The above bar chart shows that amongst all question titles, both Commercial tool importance and Commercial Software Tool Experience had the highest number of responses compared to Data sets of Interest , Programming Experience and Visualization.

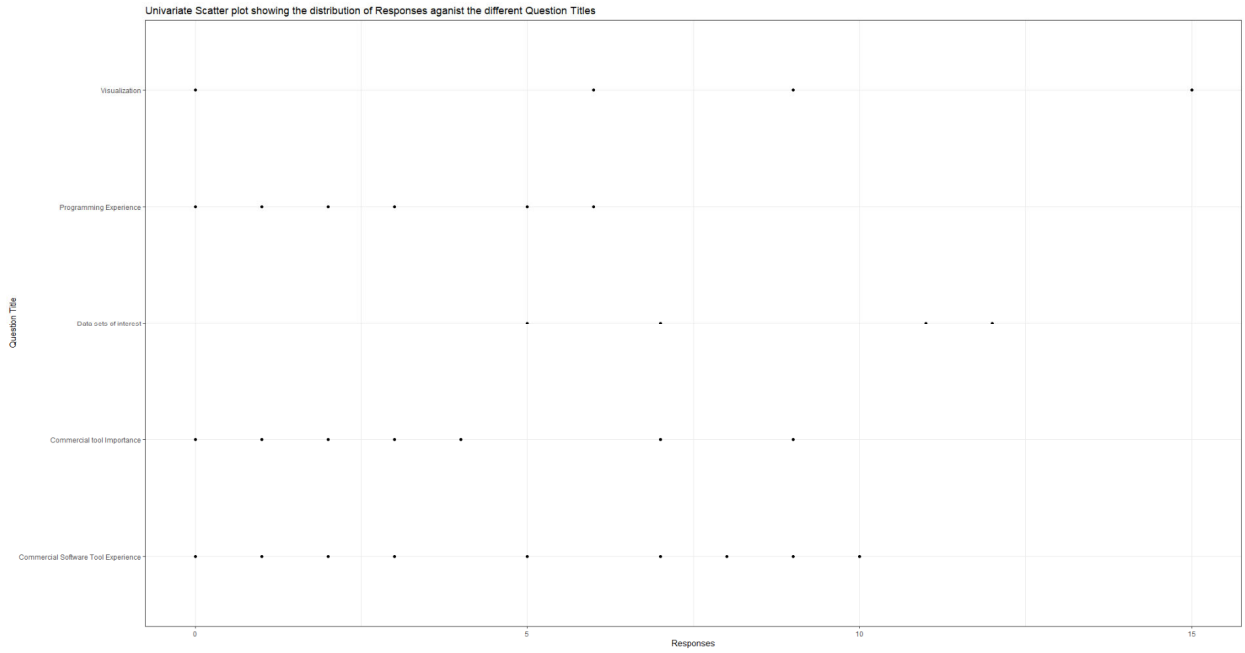
- Box plots comparing the distributions of Question Titles




The diamond symbol in the box plot is the mean. These box plots provide a visual summary of the data enabling us to quickly identify the mean values, the dispersion of the data set and any signs of skewness. Looking at the Visualization boxplot, both its mean and the median are in the center showing a perfect normal distribution. Both programming experience and Commercial Software Tool Experience, the box plot's median is skewed to the left (the bottom), meaning that their distribution is positively skewed.

Both Data sets of interest and Commercial tool importance, their box plot's median is closer to the top meaning that their distribution is negatively skewed. We also notice that Commercial tool importance has some outliers, which need to be investigated and removed.

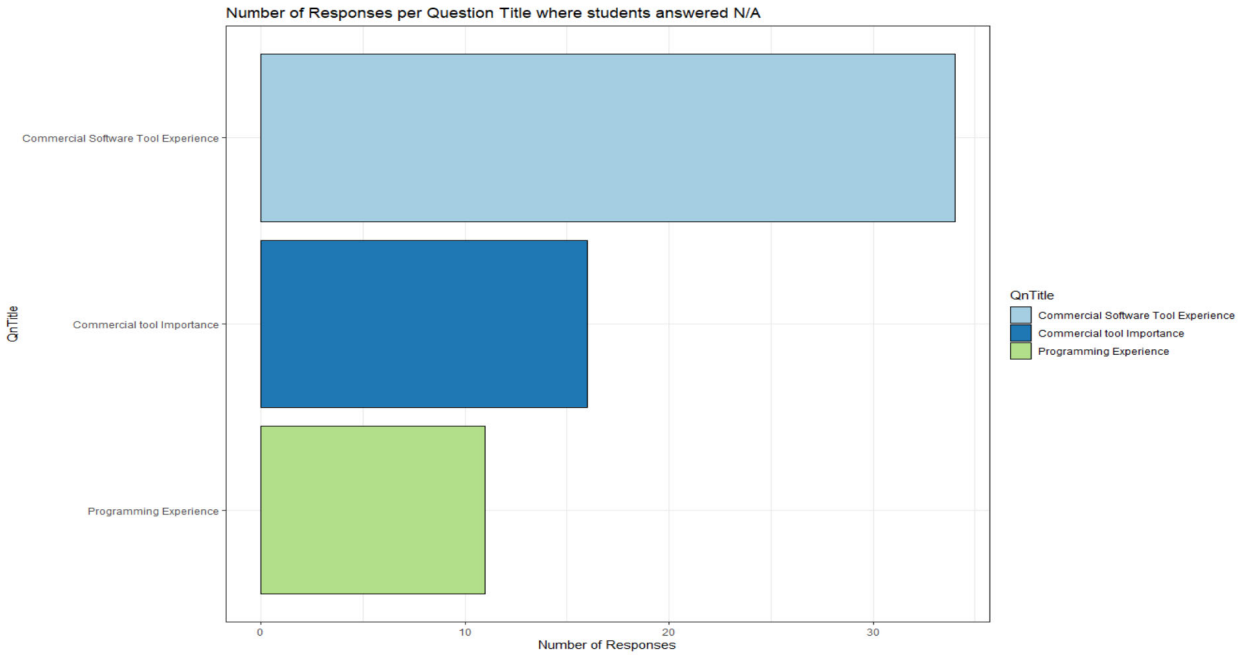
- Univariate Scatter plot showing the distribution of Responses against the different Question Title




Here we see that the Responses for Question Titles of Visualization were evenly distributed.

The responses for both Commercial tool importance and Commercial software Tool Experience range from 0 and 10. The responses for Programming Experience range from 0 to approximately 6 and that of data set of interest range 5 to approximately 11.

- Bar chart showing the number of responses per Question Title where students responded with N/A meaning that they didn't have any have experience.




From the above chart, we learn that a majority of students didn't have any experience with Commercial Software Tools since they were 34 responses with N/A and this is seconded by Commercial tool importance and Programming Experience comes in the last position.