ASSIGNMENT 3

DSC 465 DATA VISUALIZATION

Name: Vismay Bhavinkumar Patel

Date: 05/13/20

Question 1

Part a:

Visualization for overestimation and underestimation of the data is given as below:

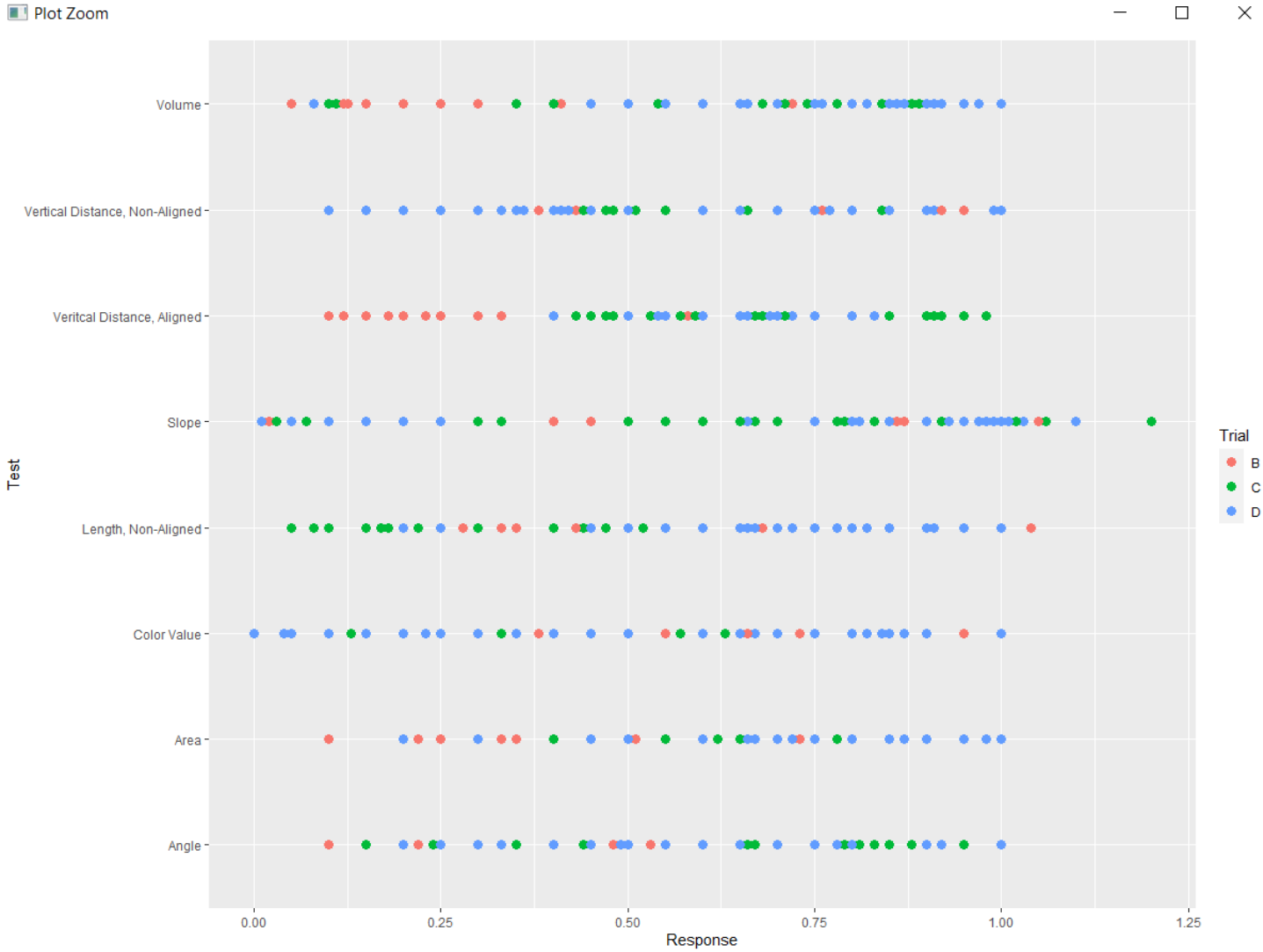

Figure 1.0

The above graph is generated by R code, plotting Test and Response data on x-axis and y-axis respectively. The reason to plot test and response is that, number of tests carried out and the number of responses based on the tests, can help us determine whether a specific test was overestimated or underestimated. This can be distinguished if the values are very close to the lowest value or highest value on the Response axis. Observing the current visualization, we can say that **slope** and **color value**, are amongst the variables which

are underestimated, since some of the values here are very close to '0' and are zero. Also, it can be noted that **slope** is overestimated since its value is close to 1.25 on the x-axis (response variable).

However, we can also determine which category of Trial variable was overestimated or underestimated, this is because the values are categorized based on colour. Trial D is the underestimated for **color value,** and Trial D, B are underestimated and Trial C is overestimated for **slope.**

Part b:

Graph for univariate scatterplot for Absolute error vs Test data:

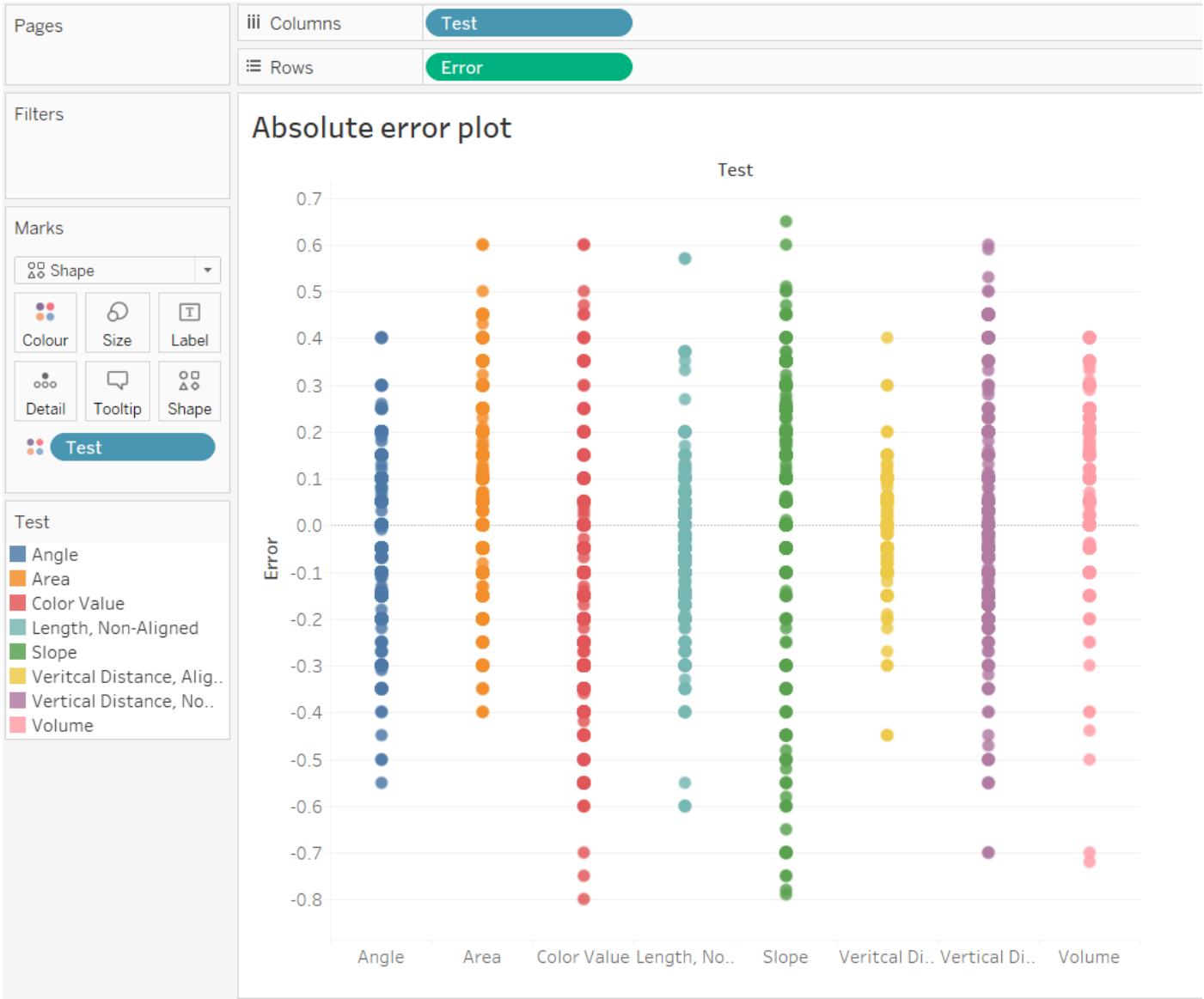

Figure 1.1

The above visualization is created in TABLEAU, Error value is plotted on y-axis and Test categories are plotted on x-axis. The errors here are highlighted by color of different test values, so that the higher absolute error in any of the test values can be interpreted and distinguished. Taking a close look at the plot, we can say the least negative Absolute errors have occurred in **Area**, while the least positive errors have occurred in **Angle** and **Vertical Distance Aligned** data. The filter box at the left shows the distinct colours of all the categories with labels, so that the values on the axis can be read clearly for interpretation. Also, we can say that most of the other Test attributes consist of moderate absolute error.

Part c:

Graph for the comparison of Display value 1 with the subjects between 56-73 and the response variable:

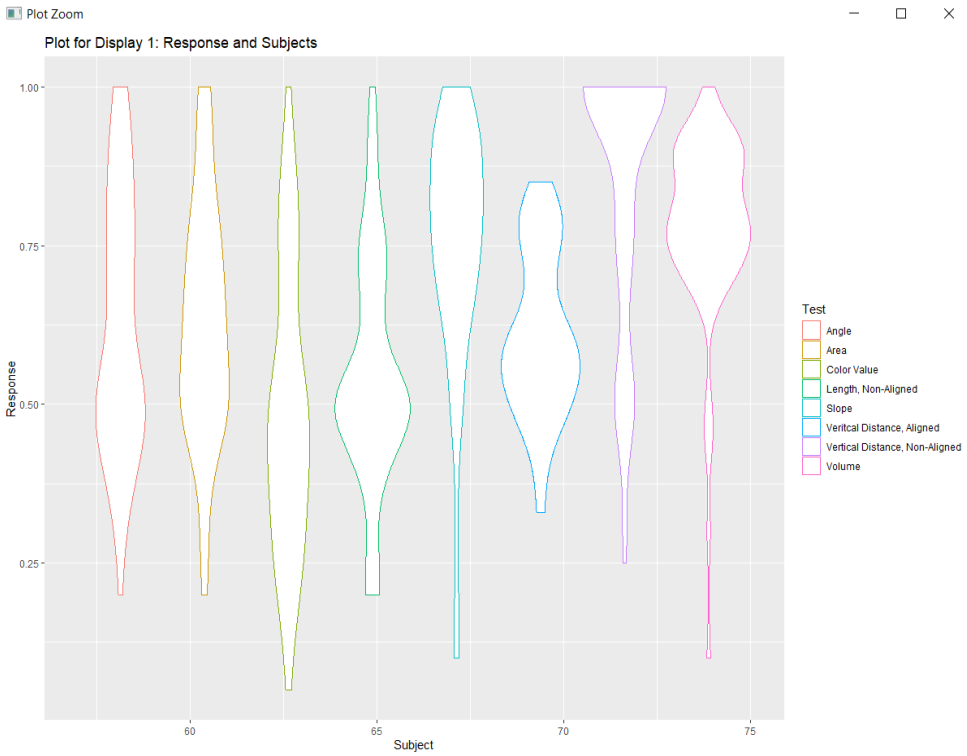

Figure 1.2

Graph for display value 2 with subjects and responses:

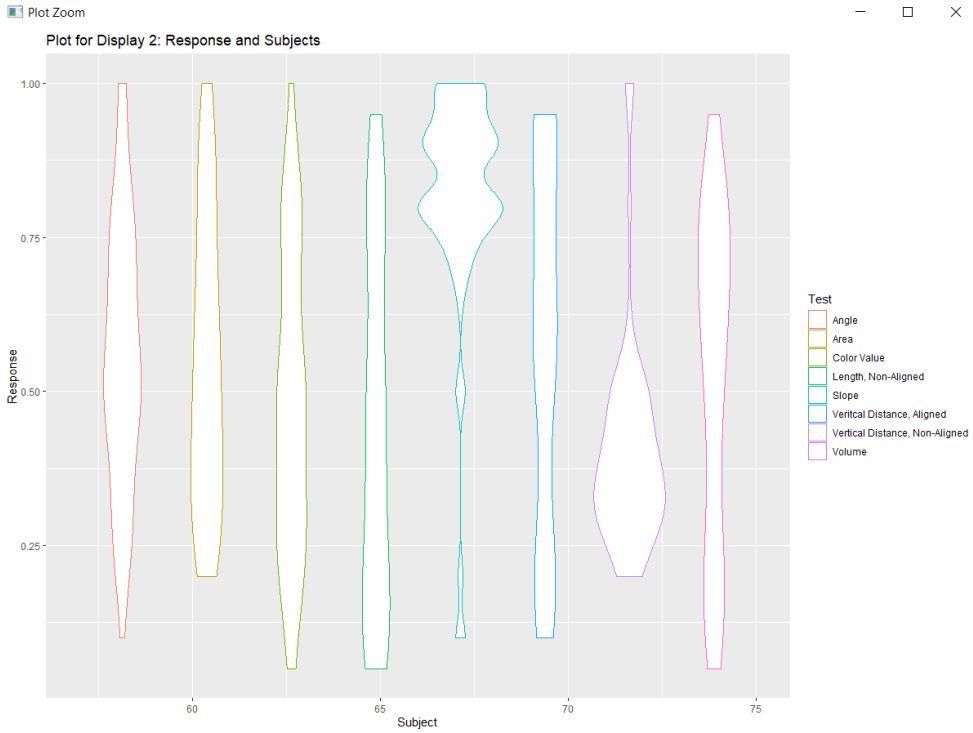

Figure 1.3

By visualizing and analysing the two plots, we can clearly see that the results obtained by the first set of Displays is better than that of second Displays. In the first display consists of lowest response value of around 0.25, while second display has lowest values below 0.25. It can also be noted that most of the values of first response have been between or around 0.35 to 1, which is a good response value. We can say that they did not get a good response in the second attempt as they got in the first attempt.

Part d:

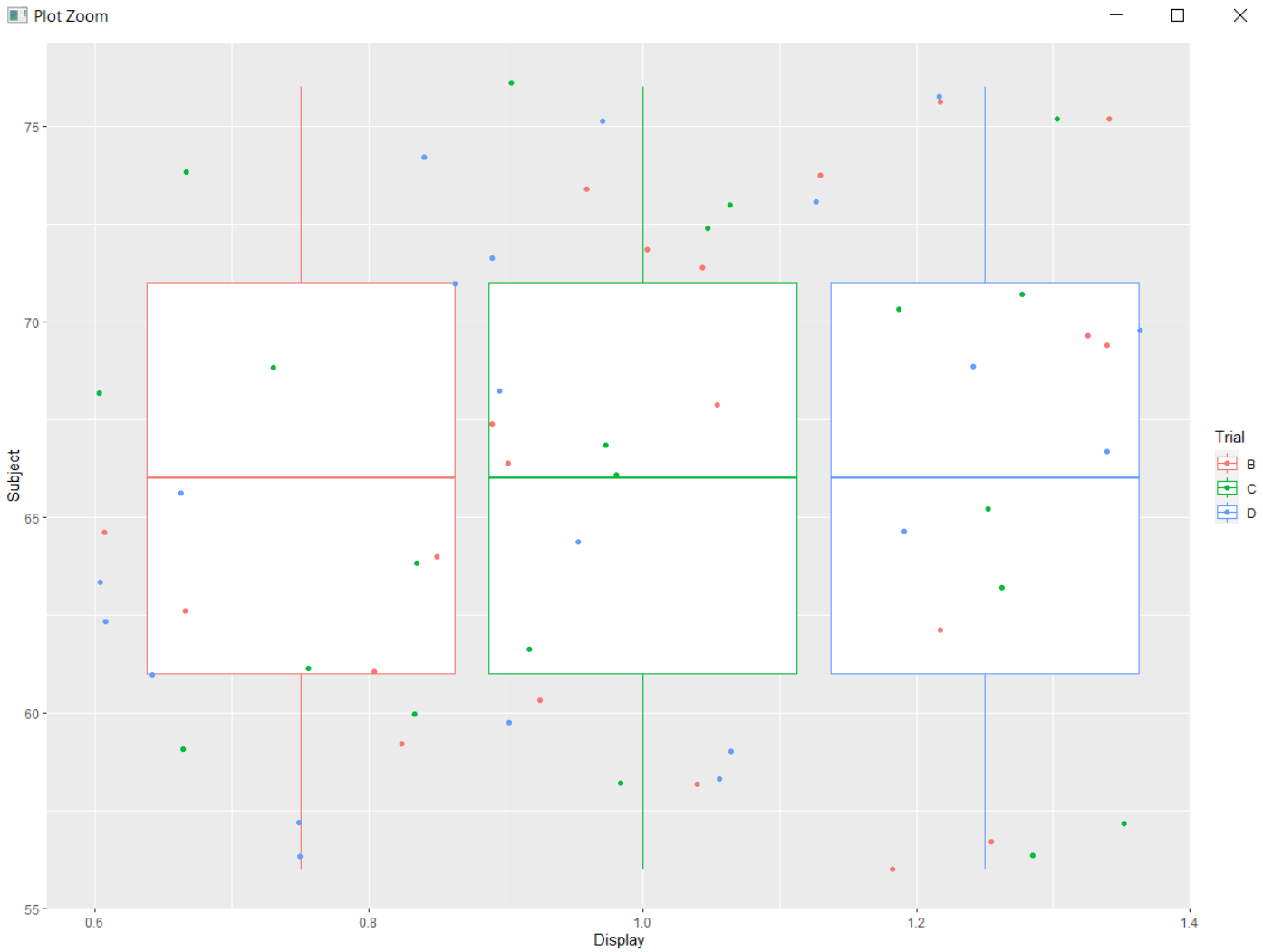

Figure 1.4

With the help of boxplot and jitter, we can distinguish the overlapping or erroneous data of Vertical distance, non-aligned data. The reason why boxplot and jitter can help is because, box plot shows the region covered by the values, and leaves the outliers at the end of the plot. The use of jitter plot is useful here because Jitter plot groups the points together in columns, which allows us to find and point out uneven pattern. From the pattern above we can say that most of the true values lie between 59-73.

Question 2

Part a:

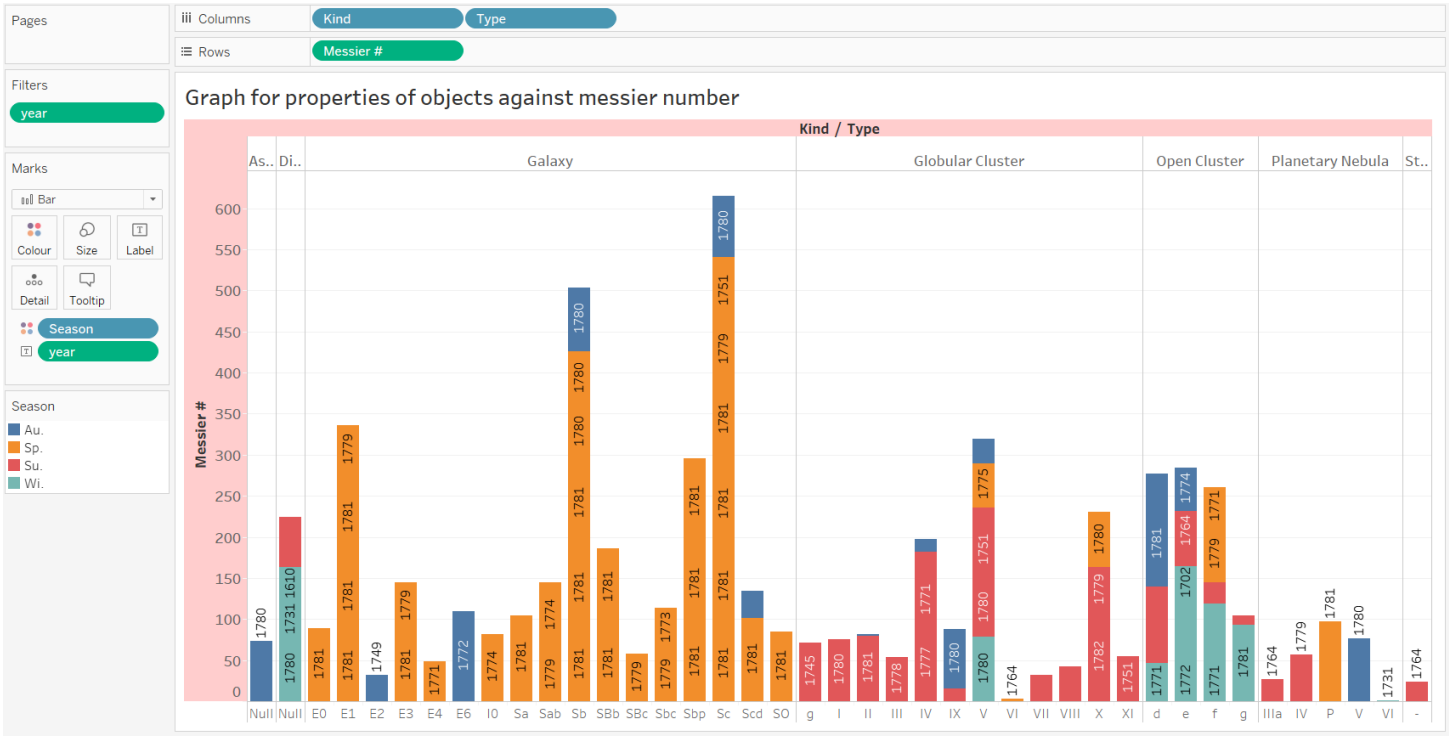

Figure 2.0

From the above representation which is a histogram, representing messier number and a few objects, namely kind and type. These are further specified or categorized by the season in which the type / kind of galaxies or clusters or stars were found out. From the above visualization we can interpret that, most of the galaxies were found in spring season, highlighted by orange colour and listed in the filter at the left corner. While, Globular clusters were mostly found out in summer season, which are highlighted by red colour. However, we do not see more discoveries in winter season, possible reason could be due to the rain or cloudy weather. However, in the Autumn season, we can identify that there are few discoveries made in mostly all the categories. One thing to note here is that, since most of the discoveries are made in the 18th century or between the 1740's to 1785/1790, we cannot find a pattern of discoveries year wise for different Types or Kinds of space discoveries.

Part b:

Graph for visualization that compares distances with each kind.

Figure 2.1

From the above relationship we can determine that, farthest kind is Galaxy (light blue-green colour), followed by Globular clusters (green colour), star clouds (pink) and Planetary nebula (purple). Moreover, Open Clusters, Diffusion Nebula and Asterism mostly come in similar distance range compared to that of other planetary or galactical objects.

Part c:

Scatter plot for distance to messier object and their Apparent Magnitude.

Figure 2.2

As we can see in the above visualization, the points are distinguished by colour scale that fades from dark to light. The light colour points determine higher magnitude values than that of the dark colour pattern. The fainter object in the graph that we can see is having Magnitude of 12, and lies at the distance of 1,000 to 10,000 light years.

Part d:

Augmented graph for Magnitude and distance with changes in size parameter:

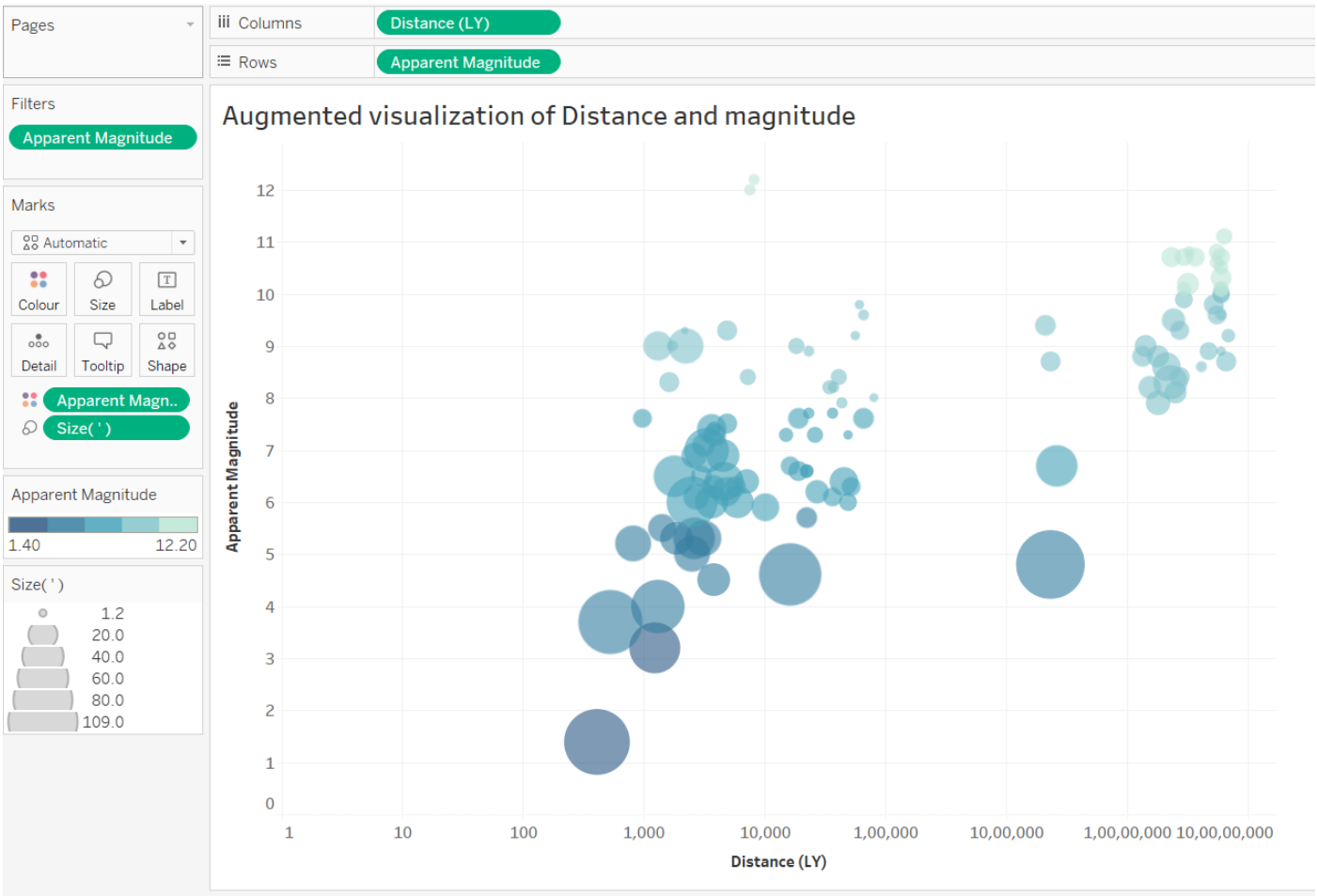

Figure 2.3

From the above visualization, we can say that changing the size can make it easier for the viewer to interpret the distance and message that the graph tries to convey. The lighter the objects, the higher the magnitude and farther (less size) they are from other objects. To improve the visualization however, colour contrast can be changed to a more appropriate scale, including the opacity of the colour which plays an essential aspect to determine the overlapping objects and makes the interpretation clearer.

Question 3

Part a:

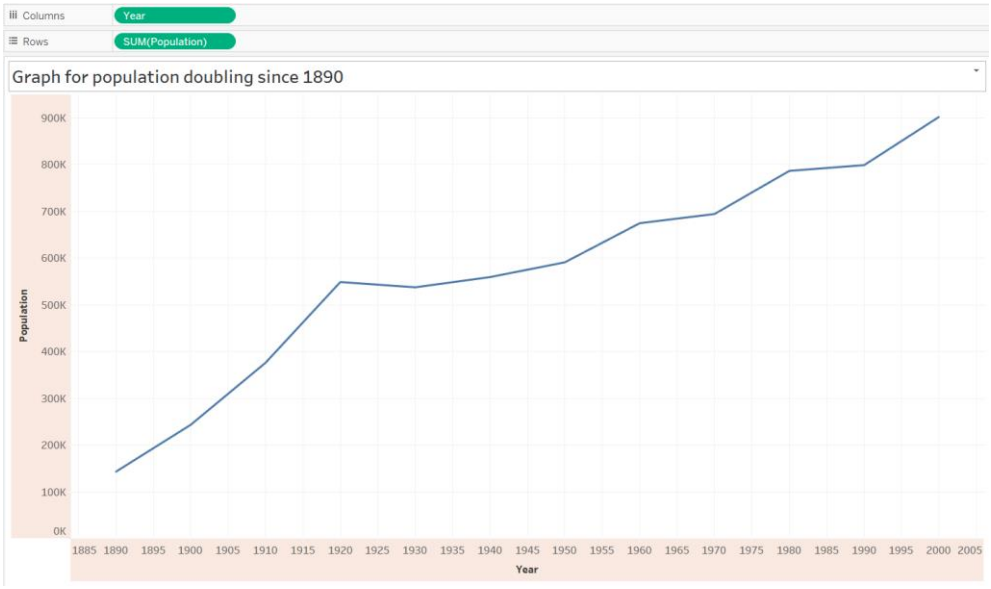

Figure 3.0

Form the above representation which displays the population growth of a Montana, we can say that the city/county has seen growth in population between 1960 to 1980 and from 1980 to 2000. Before that, there was a huge spike in population growth from 1890 to 1920, followed by a slight decrease and slight increase till 1960. After that the population has increased in a moderate pattern.

Part b:

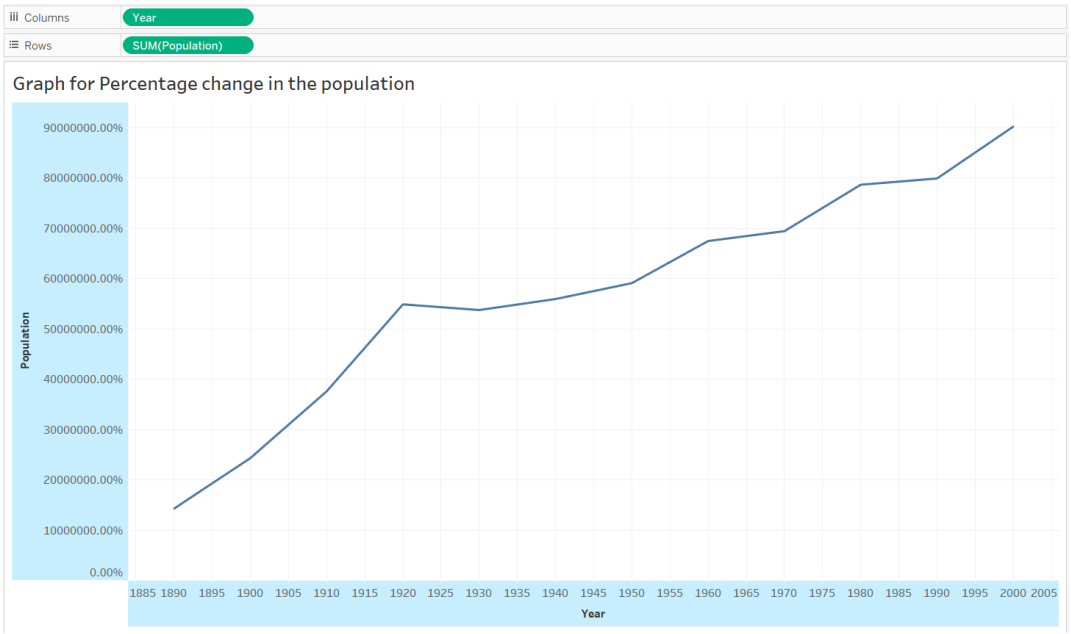

Figure 3.1

The highest spike in the percentage increase, about 35% approx. is seen in population between years 1890 to 1920. After that, there was a slight decrease of about 15% with between years 1920 to 1960. Since 1960 the percentage of population has been moderately increasing in a uniform rate of about 10-12% every 5 years till 2000.

Part c:

Years 1890 to 1920, have experienced the percentage increase of more than 15 percent.

Question 4

Part a:

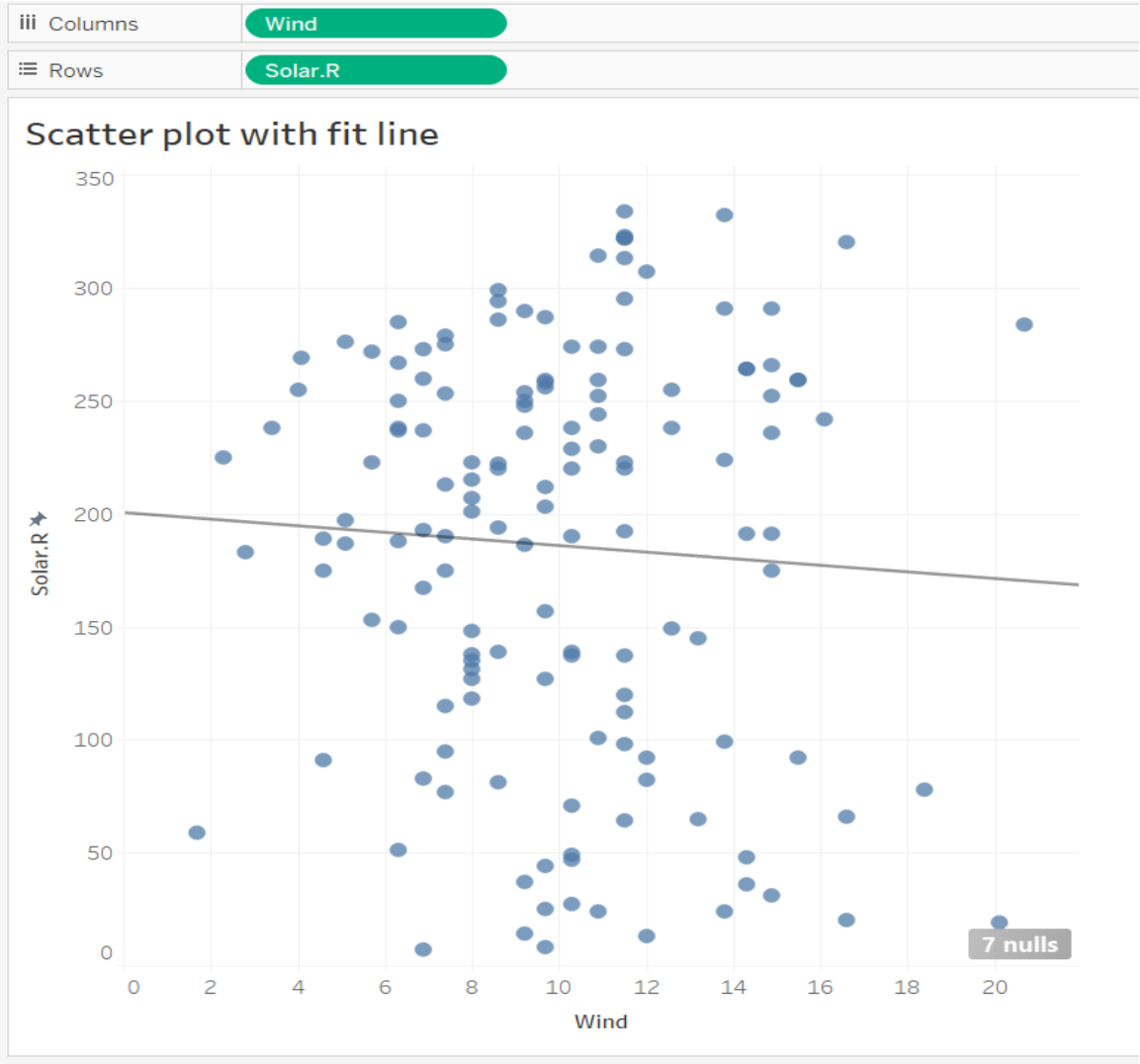

Figure 4.0

The above figure shows the scatter plot between wind and Solar.R variables. The points are scattered over the plot not displaying a significant pattern, also the fit line passes through the points at 200. This does not allow us to find a specific pattern.

Part b:

Bar plot between Wind and Solar.R

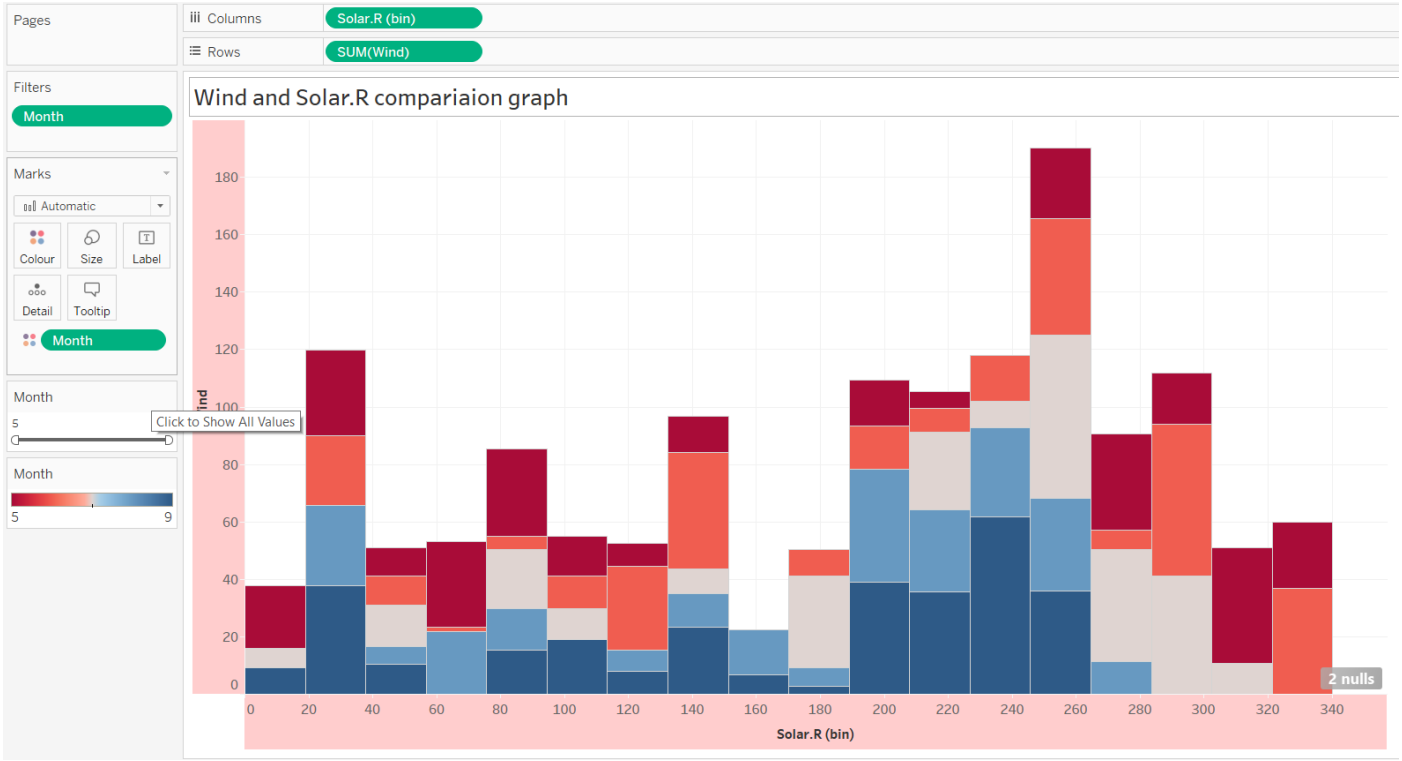

Figure 4.1

The above figure shows the relation between Wind and Solar.R variables, the graph is plotted with categorization of month variable from 5th month to 9th month. From the visualization we can draw the insight, that during the 5th month the wind speed is mostly higher, when the Solar radiation levels are between 20-40, 80-100, 137-150, and 190-300. Similar pattern can be observed during the 6th month, represented by orange colour. On the contrary in the 9th month the wind speed only increases between the radiation levels of 20 to 40 and 190 to 262, rest of the time the speed in less.

Part c:

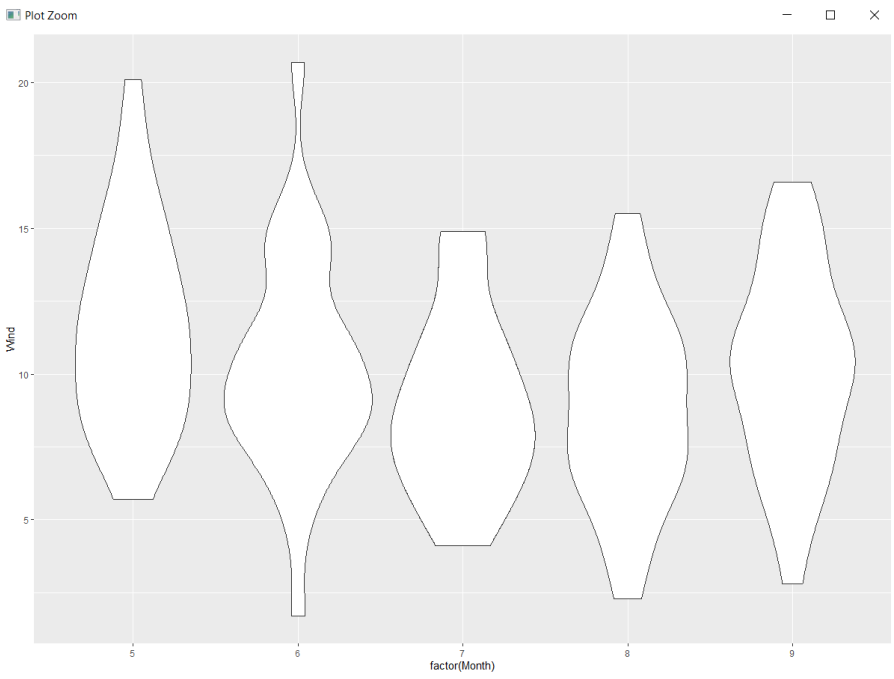

Figure 4.2

Part d: EXTRA CREDIT

QQ plot for Wind:

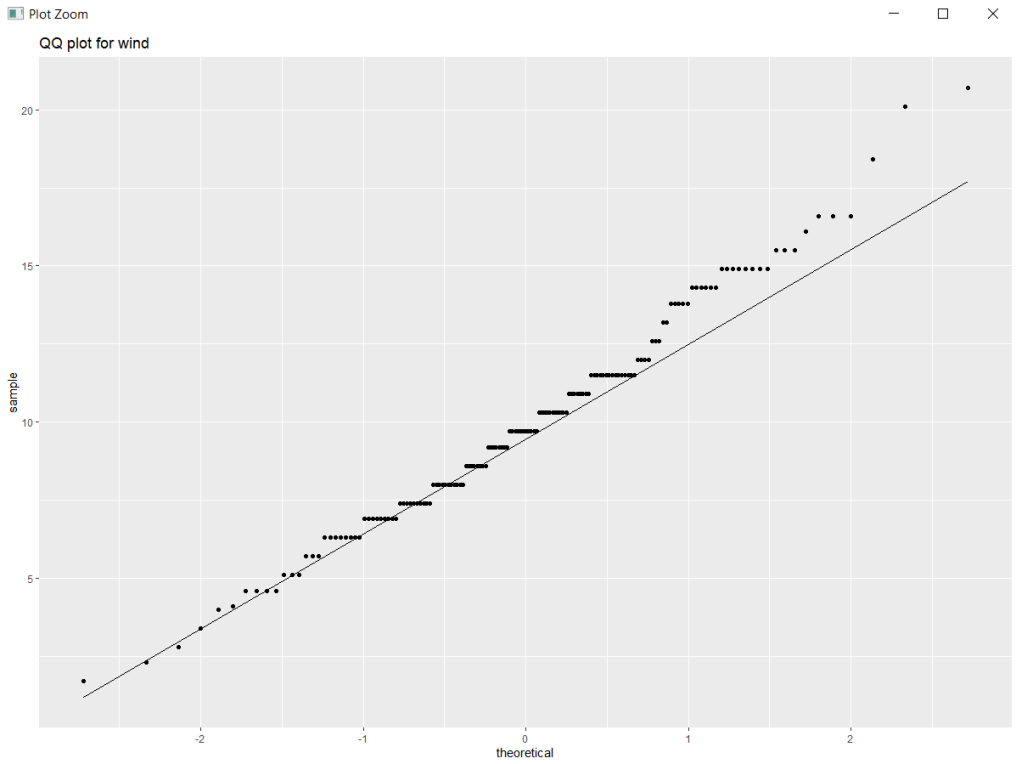

Figure 4.3

QQ plot for Solar.R (Solar radiation):

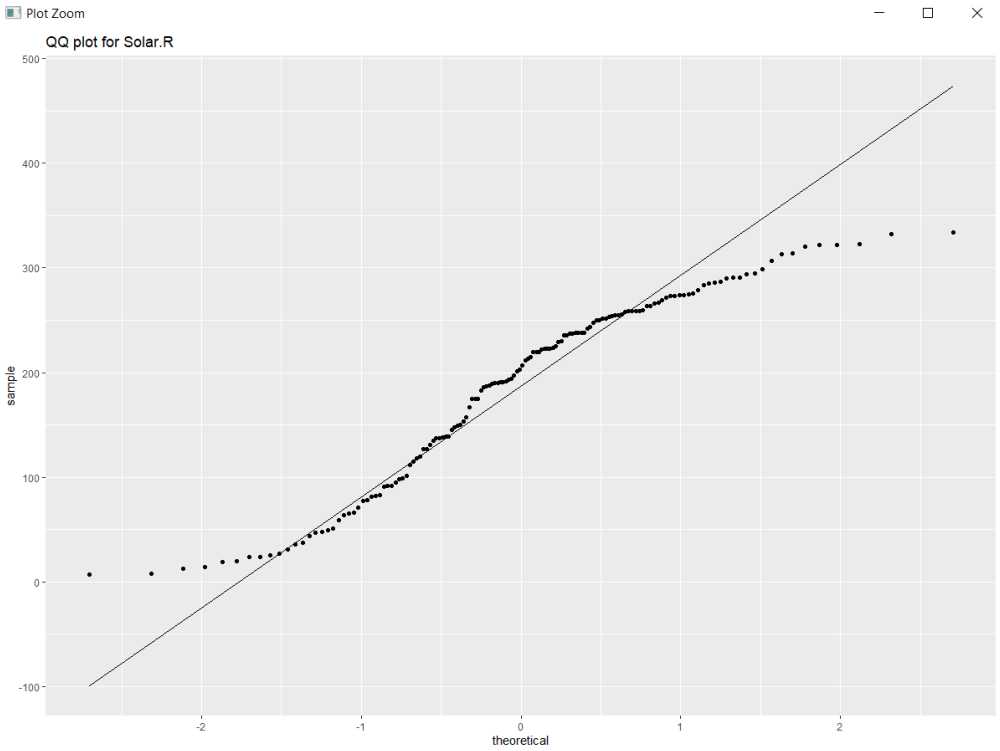

Figure 4.4

From both the plots we can draw an insight that, wind is more aligned towards the fit line, in comparison to Solar Radiation. Solar Radiation displays a pattern where, after crossing 250 the points start moving away from the line of fit. Similar pattern can be found in wind but when the wind speed crosses 17 kmph or mph.