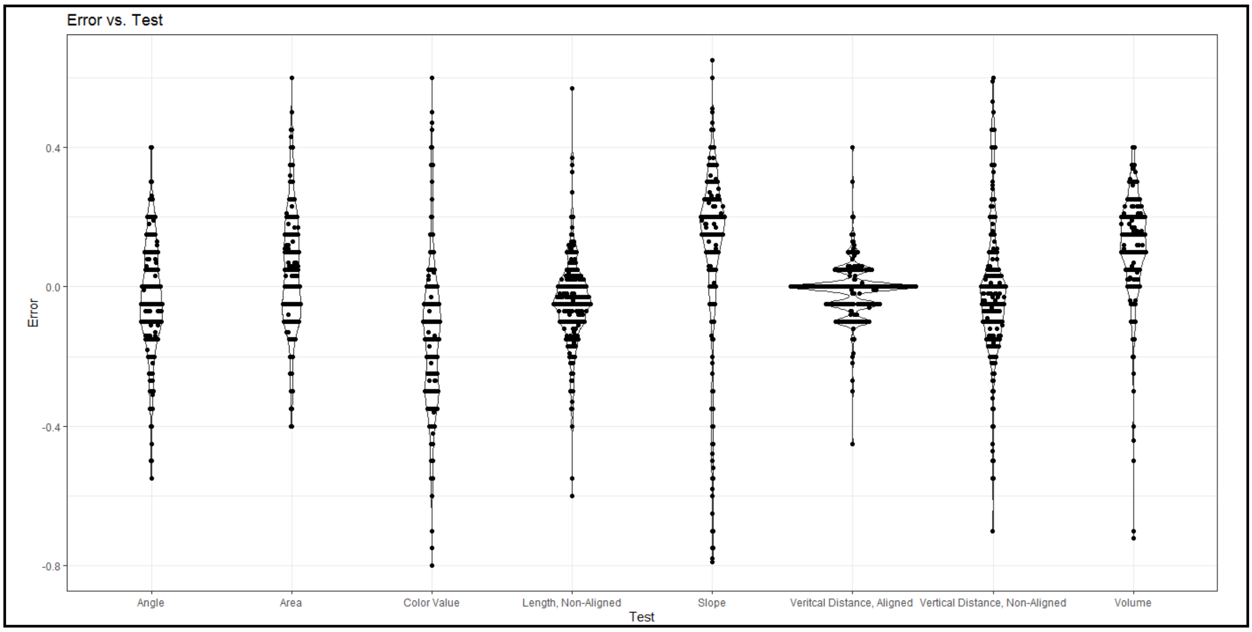
**Error vs. Test**

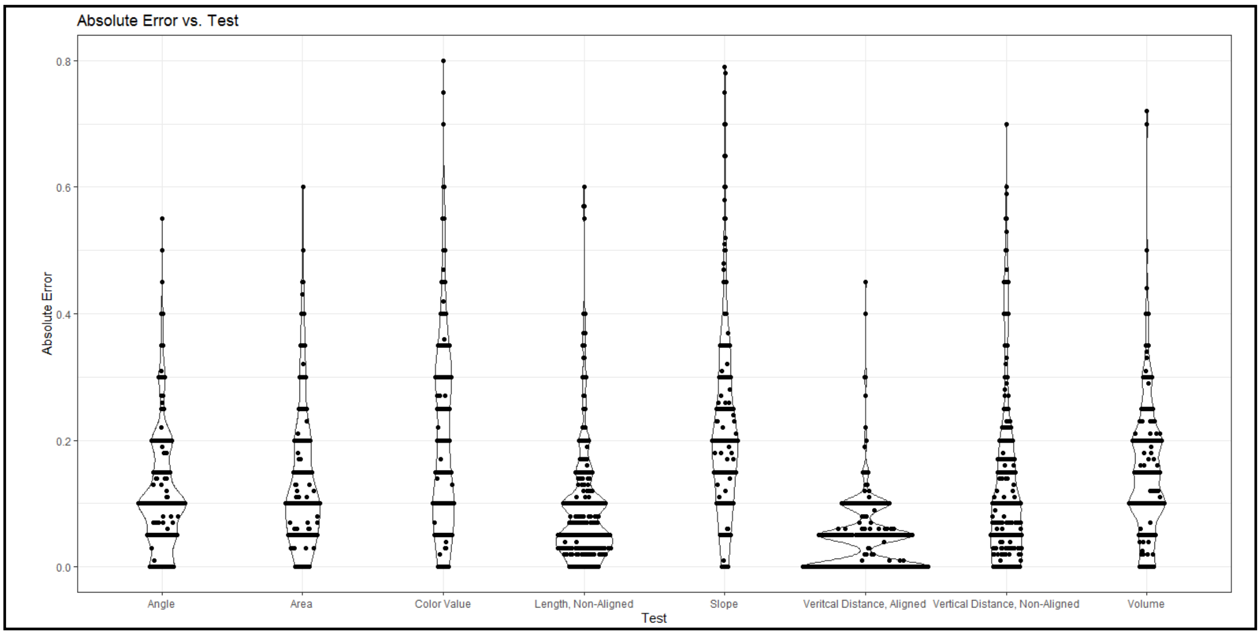
```
> perplot <- PerceptionExperiment %>%
+   ggplot(aes(x=factor(Test), y=Error))

> perplot  + geom_violin() + geom_sina() + theme_bw() +
+ labs(x = "Test",
+      y = "Error",
+      title = "Error vs. Test")
```




**Conclusion:** The above graph shows the error distribution for various tests.
As can be seen in the violin plot, the values are plotted from negative to positive. The distribution curve can help us determine if the users overestimated or underestimated the data.
For Angle, Color Value, Length Non Aligned and Vertical distance non aligned tests the graphs are skewed heavily on the negative side. Therefore we can say the users have underestimated the response.
For Area, Slope and Volume tests the graphs are inclined towards the positive side. Hence we can say users have overestimated the values.

**Absolute error vs. Test**

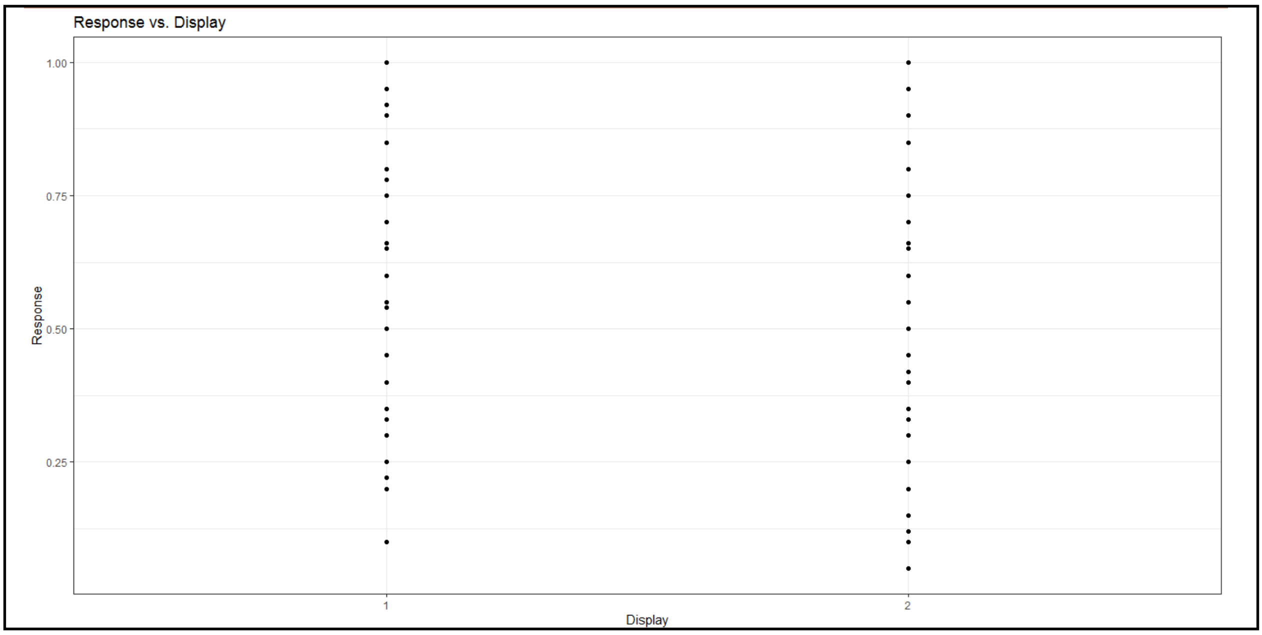
```
> perplot2 <- PerceptionExperiment %>%
+   ggplot(aes(x=factor(Test), y=AbsError))

> perplot2  +   geom_violin() + geom_sina() + theme_bw() +
+   labs(x = "Test",
+        y = "Absolute Error",
+        title = "Absolute Error vs. Test")
```




**Conclusion:** The above graph shows the absolute error distribution for various tests.
As can be seen in the violin plot, the data distribution helps us determine the error rate for each tests.
For Angle, Area, Length Non Aligned, Vertical Distance Aligned and Vertical distance Non Aligned tests, most of the errors are between 0 to 0.1.
As for the Color Value, Slope and Volume tests, the errors are clamped between 0 and 0.2.
The Color Value and Slope tests have maximum error rate while Vertical Distance Aligned the error rate is minimum.

**Response vs. Display**

```
> FilteredPerceptionExperiment <- PerceptionExperiment %>% filter(between(
Subject, 56, 73))

> perplot5673 <- FilteredPerceptionExperiment %>%
+   ggplot(aes(x=factor(Display), y=Response))

> perplot5673 + geom_point()+ theme_bw() +
+   labs(x = "Display",
+        y = "Response",
+        title = "Response vs. Display")
```

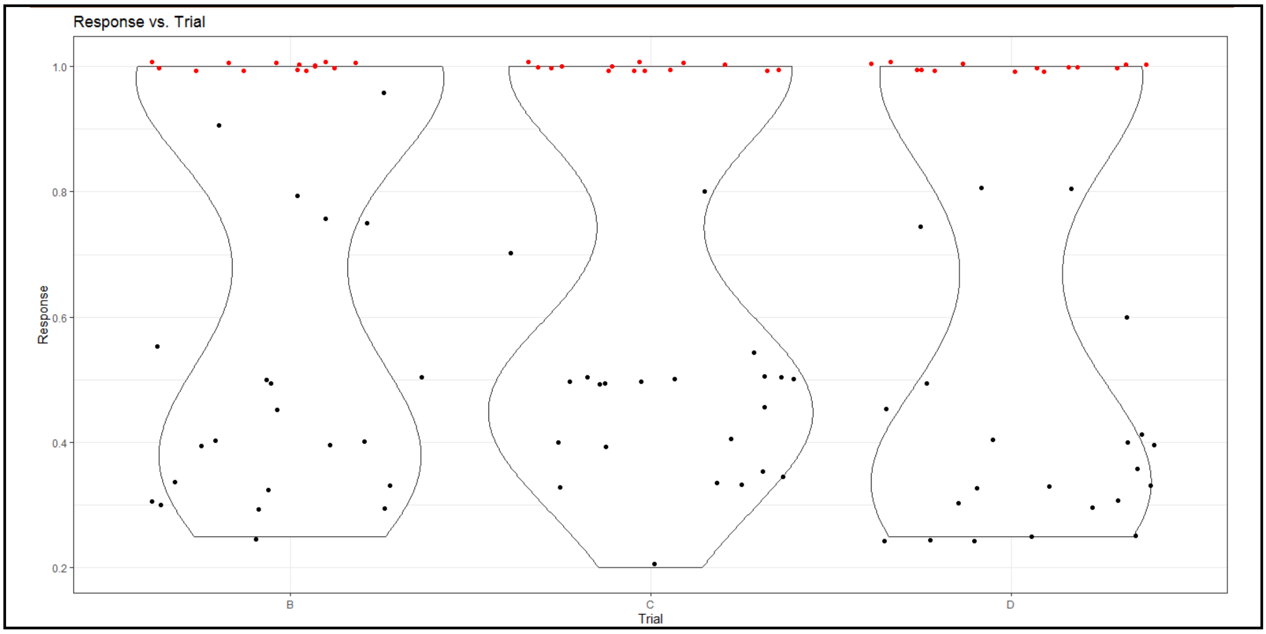



**Conclusion:** The above dot graph shows the response value against display.
As can be seen in the graph, the response data is clamped together for display 1 while the data is spread out for display 2.
From the overall standpoint, the data shows users have a better prediction response after taking the test the first time. The graph shows users have a broader response during the second test.

**Trial vs. Response**

```
> FilteredPerceptionExp <-
+   PerceptionExperiment %>% filter(Test == "Vertical Distance, Non-Aligne
d") %>% filter(between(Subject, 56, 73))

> FilteredPerceptionExp %>%
+   ggplot(aes(x = factor(Trial), y = Response)) + geom_violin() + geom_ji
tter(col = ifelse(between(FilteredPerceptionExp$Subject, 56, 73) & Filtere
dPerceptionExp$Response == 1, "red", "black"),show.legend=TRUE) + theme_bw
() +
+   labs(x = "Trial",
+        y = "Response",
+        title = "Response vs. Trial")
```

As analyzed in the dataset, most of the responses with value 1 are for subjects 56 to 73. Hence the above code filters all the subjects from 56 to 73 and plots the violin and jitter plot for all trials.



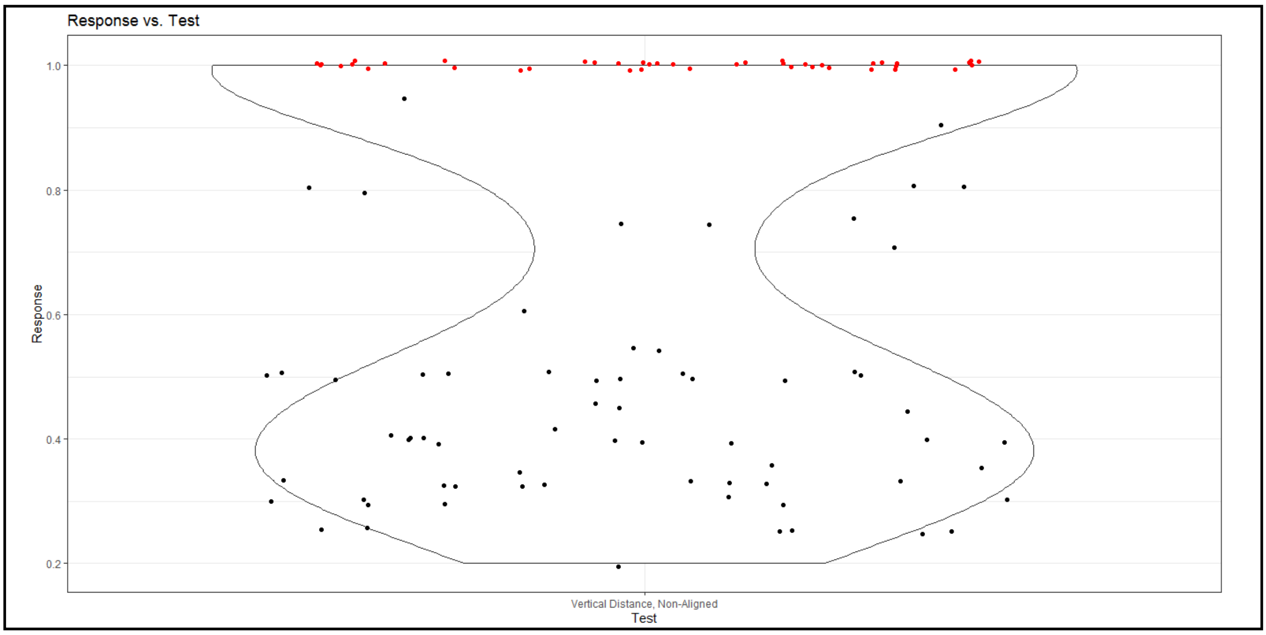

**Conclusion:** The above graph shows the data distribution for various trials and their responses. The violin plot shows a bimodal distribution because of these outliers. All the erroneous responses are for subjects between 56 and 73.

The above dataset has been filtered for Test = 'Vertical Distance, Non-Aligned' and subjects between 56 and 73.

The outliers with response value as 1 are highlighted in 'red'. As can be seen in the above graph, none of the other responses are near 1.

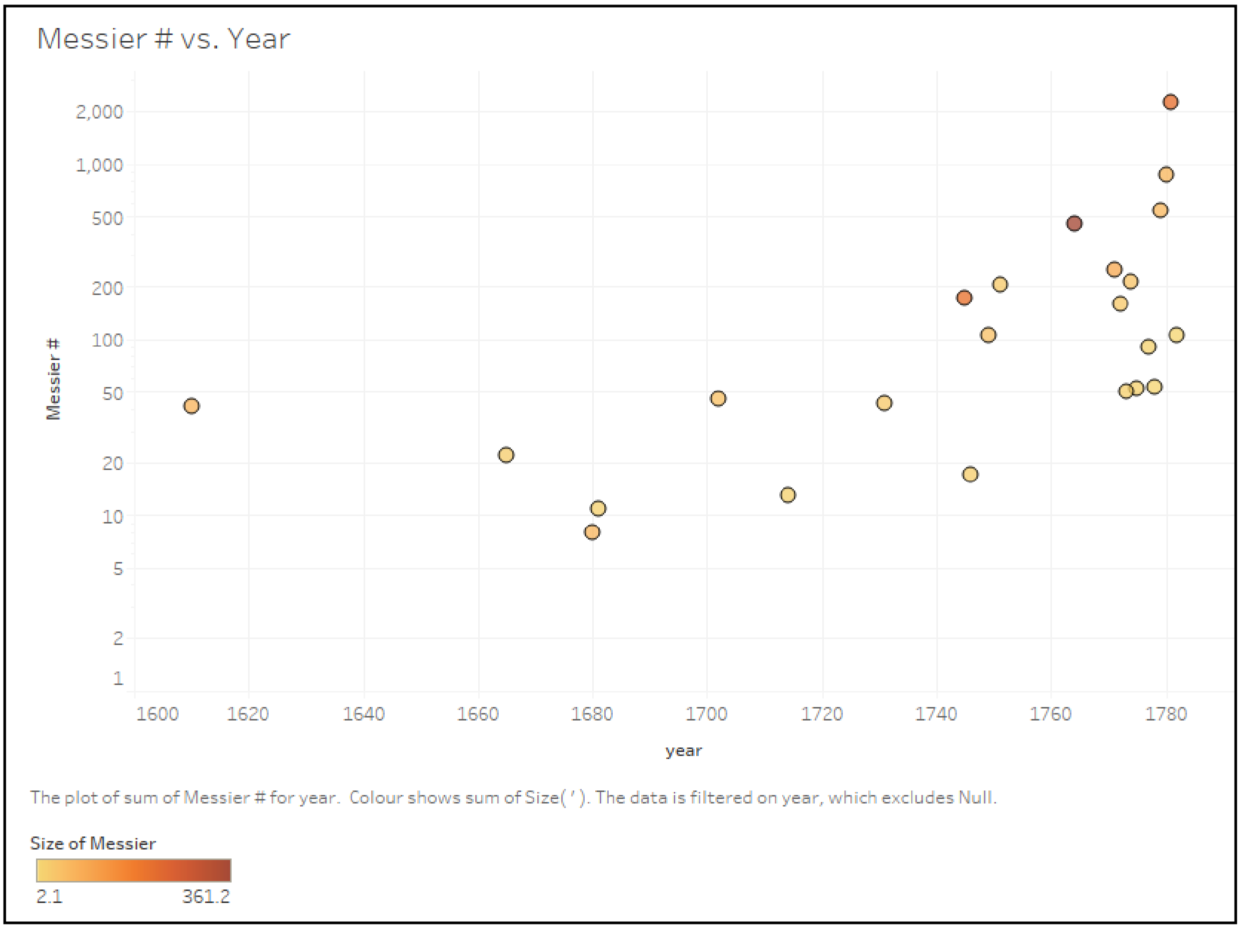
Merging all the trials together

```
> FilteredPerceptionExp %>%
+   ggplot(aes(x = factor(Test), y = Response)) + geom_violin() + geom_jit
ter(col = ifelse(between(FilteredPerceptionExp$Subject, 56, 73) & Filtered
PerceptionExp$Response == 1, "red", "black"),show.legend=TRUE) + theme_bw(
) +
+   labs(x = "Test",
+        y = "Response",
+        title = "Response vs. Test")
```



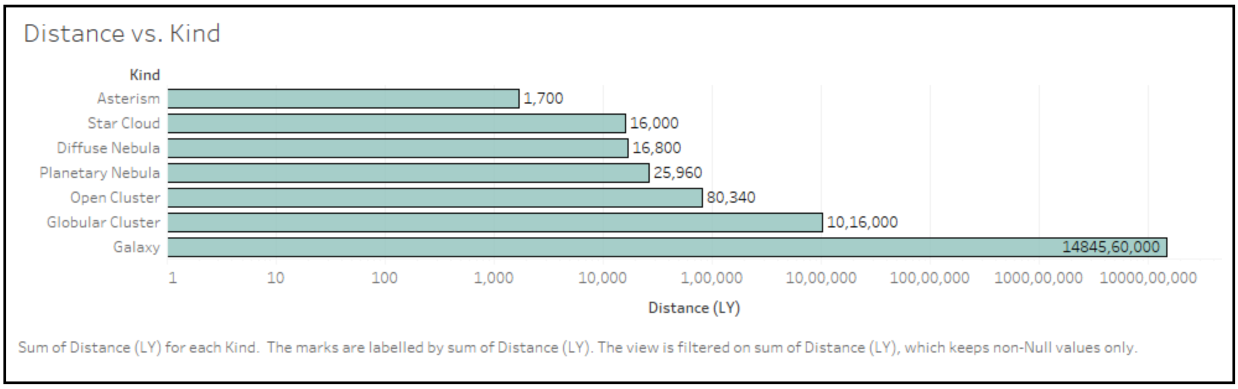

**Conclusion:** The above graph shows the data distribution for Vertical Distance, Non-Aligned test and their responses.
The violin plot shows a bimodal distribution because of these outliers. All the erroneous responses are for subjects between 56 and 73.
The above dataset has been filtered for Test = ‘Vertical Distance, Non-Aligned’ and subjects between 56 and 73.
The outliers with response value as 1 are highlighted in ‘red’. As can be seen in the above graph, none of the other responses are near 1. Most of the data is below 0.8.

**Messier vs. Year**



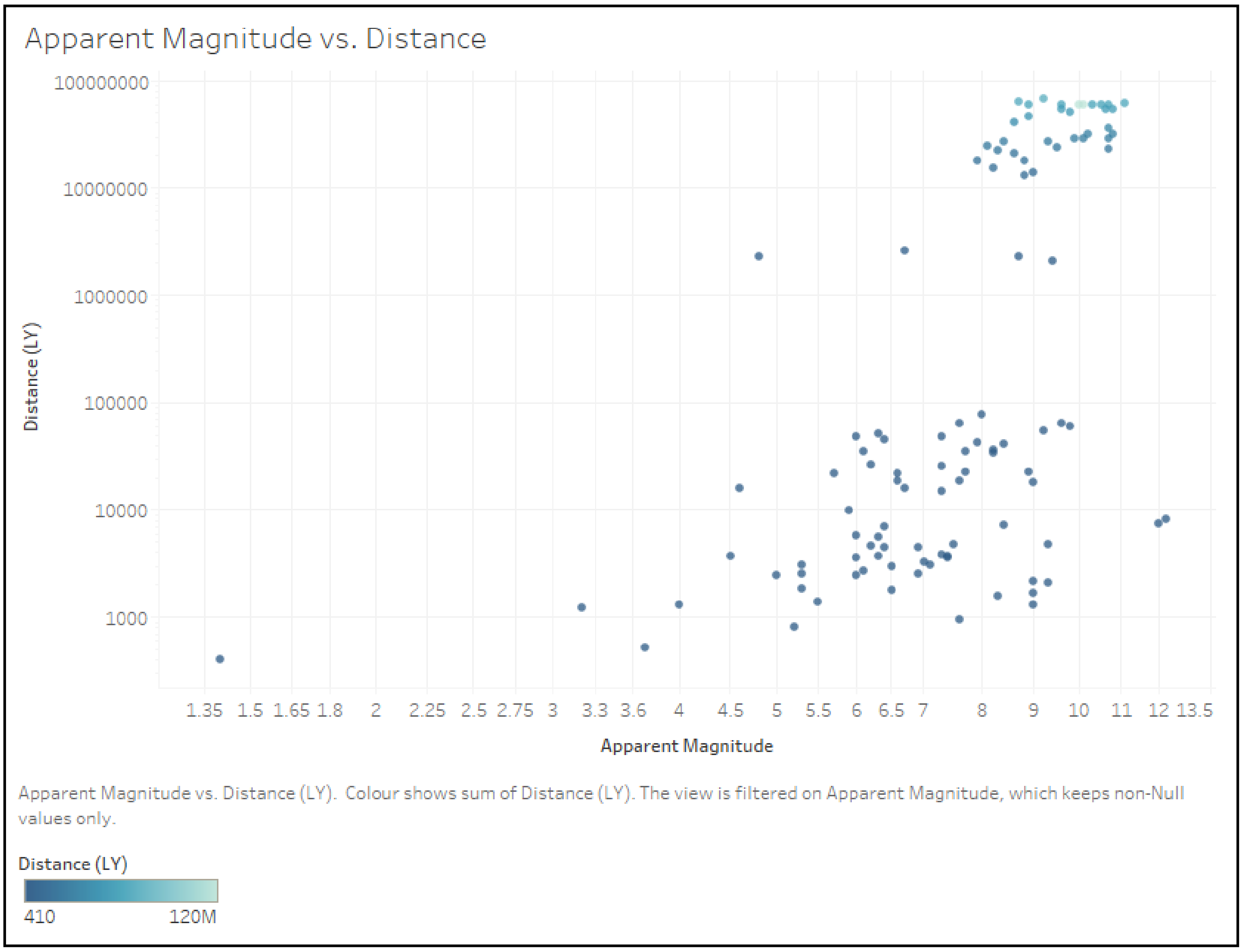

**Conclusion:** The above scatter plot shows the data between Number of messier against the Year. As can be seen in the graph, the number of Messiers appeared have increased over the years. The graph shows in the later years we were able to detect huge Messier and at a distance.

**Distance vs. Kind**




**Conclusion:** The above bar graph shows the kind of Messier against the distance of them appearing. As can be seen in the graph, Galaxies appears the farthest while Asterism appears the nearest.

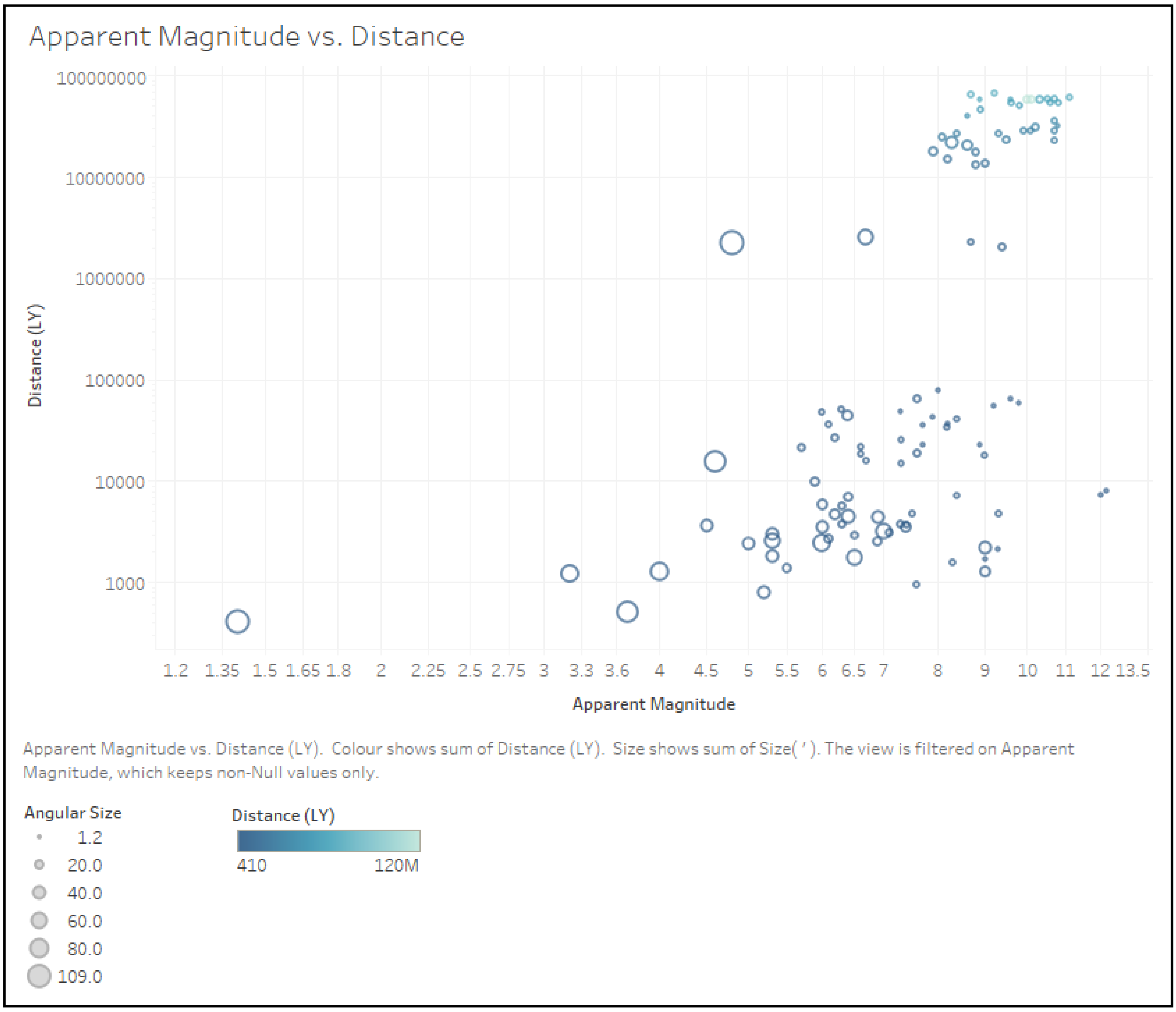
**Apparent Magnitude vs. Distance**




**Conclusion:** The above scatter plot shows the values between Apparent Magnitude vs. Distance. As can be seen in the graph, the apparent magnitude values between 3 and 10 have a distance between 1000 and 100000. The records having the apparent magnitude above 8 have a distance greater than 10000000.
As the apparent magnitude increases, the distance decreases.

The objects are fainter as the distance increases. The color of the dots lightens up as the distance increases.

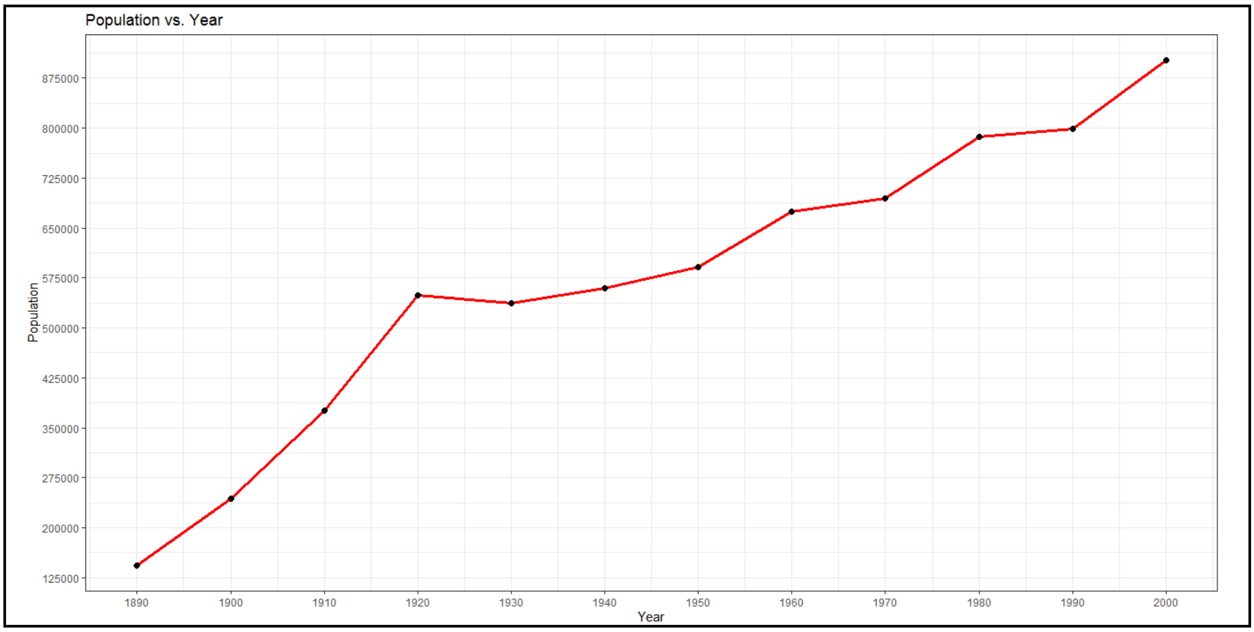
**Apparent Magnitude vs. Distance with Angular Size**




**Conclusion:** The above scatter plot shows the values between Apparent Magnitude vs. Distance.
As can be seen in the graph, the apparent magnitude values between 3 and 10 have a distance between 1000 and 100000. The records having the apparent magnitude above 8 have a distance greater than 10000000.
As the apparent magnitude increases, the distance decreases.
The objects are fainter as the distance increases. The color of the dots lightens up and the size of the object decreases as the distance increases.
Most of the huge messier appears at a lower distance and apparent magnitude.

**Population vs. Year**

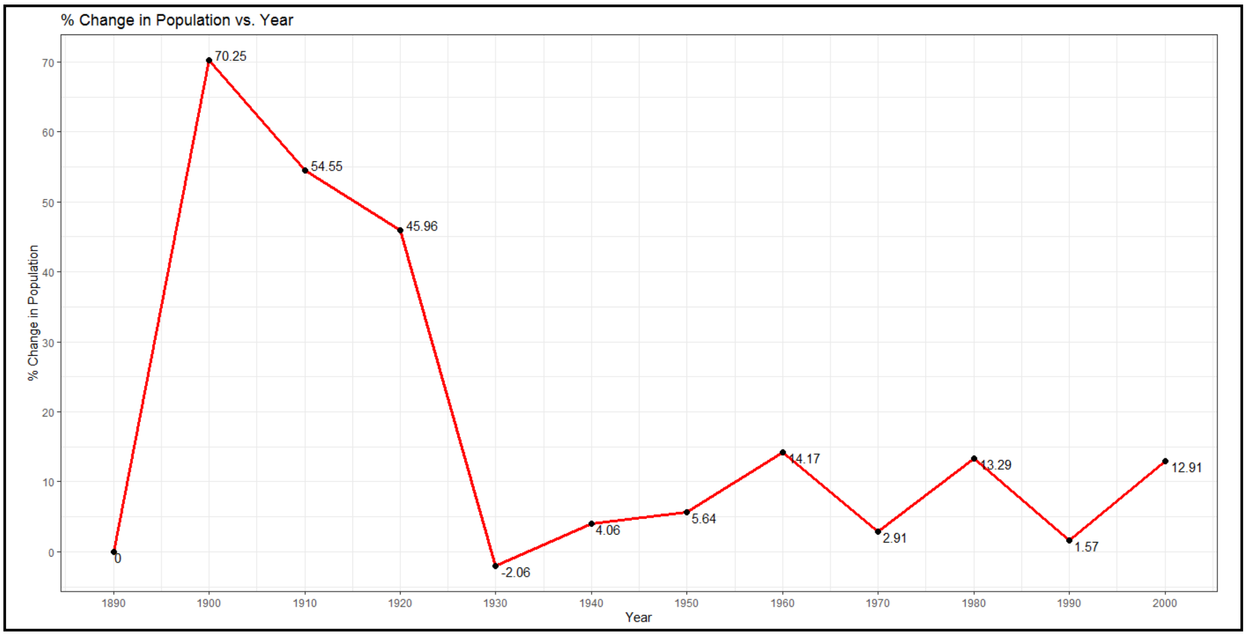
```
> popplot <- MontanaPopulationData %>%
+   ggplot(aes(x = Year, y = Population))

> popplot  + geom_line(size = 1.3,col='red') +
+   geom_point(size = 2.2) +
+   theme_bw() +
+   scale_y_continuous(breaks = seq(125000 , max(MontanaPopulationData$Pop
ulation) , by = 75000)) +
+   scale_x_continuous(breaks = seq(1880 , max(MontanaPopulationData$Year)
, by = 10)) +
+   labs(x = "Year",
+        y = "Population",
+        title = "Population vs. Year")
```



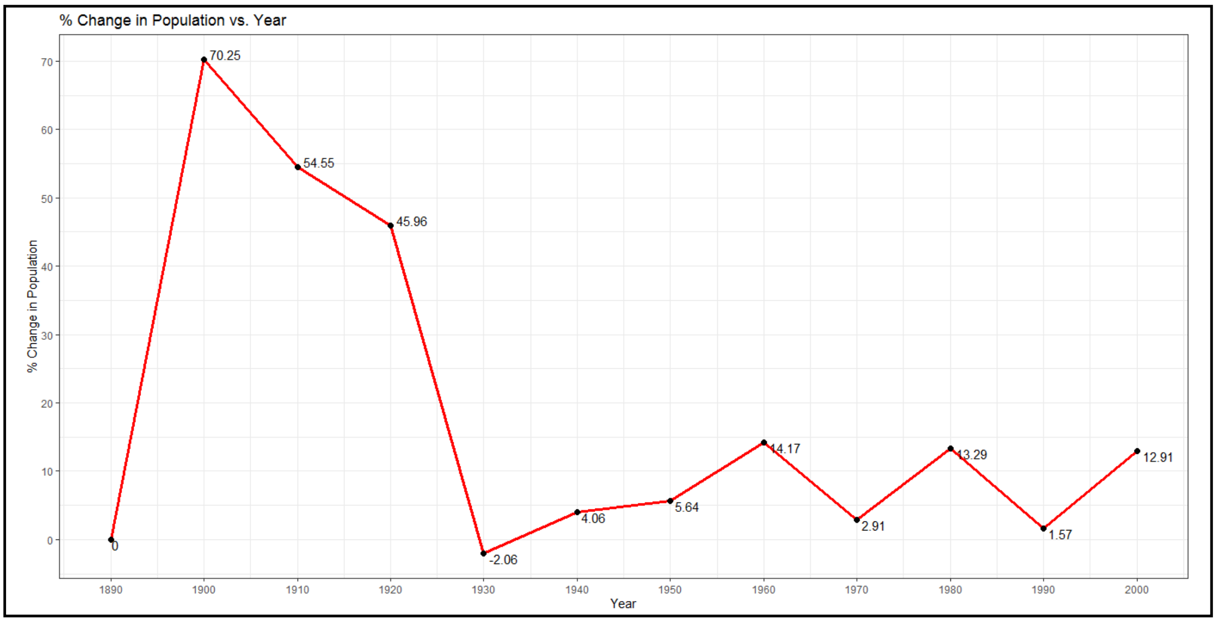

**Conclusion:** The line graph plots the population for Montana across the years.
The line graph shows the value of population around 140000 for 1890 and the value increases to 900000 in 2000.
Population has doubled from 1890 – 1910 having a value of 140000 to 375000.
There is another spike from 1910 – 1980 with an increase from 375000 to 780000.

**% Change in Population vs. Year**

```
> growth_rate = MontanaPopulationData %>%
+   # first sort by year
+   arrange(Year) %>%
+   mutate(
+     Diff_year = Year - lag(Year),# Difference in time (just in case ther
e are gaps)
+     Diff_growth = Population - lag(Population),# Difference in route bet
ween years
+     percent = (Diff_growth) / lag(Population) * 100 # growth rate in per
cent
+   )

> growth_rate$percent[is.na(growth_rate$percent)] = 0 # Assigning value 0
to NA

> popplot2 <- growth_rate %>%
+   ggplot(aes(x = Year, y = percent))

> popplot2  + geom_line(size = 1.3,col='red') +
+   geom_point(size = 2.2)  +
+   theme_bw()  +
+   scale_y_continuous(breaks = seq(0 , 100 , by = 10)) +
+   scale_x_continuous(breaks = seq(1880 , max(MontanaPopulationData$Year)
, by = 10))+
+   labs(x = "Year",
+        y = "% Change in Population",
+        title = "% Change in Population vs. Year") +
+   geom_text(aes(label = round(percent, 2)),
+             vjust = 'outward', hjust = -0.2,
+             show.legend = FALSE)
```

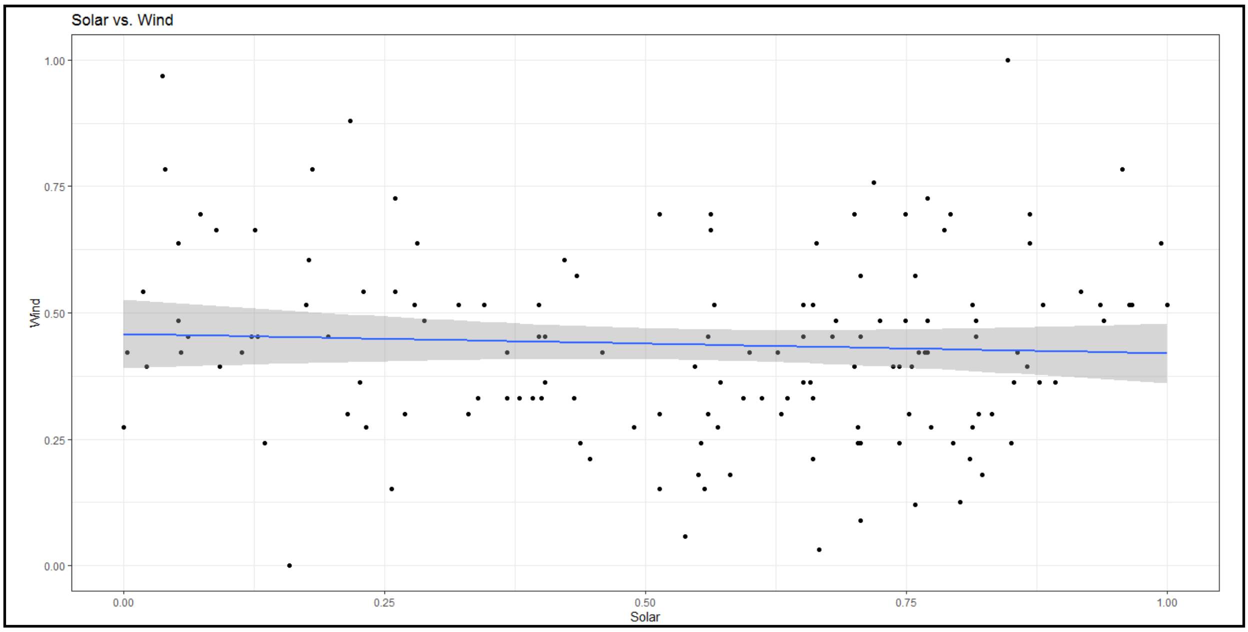



**Conclusion:** The line graph plots the % change in population for Montana across the years. As can be seen in the line graph, the rate of change in population flattens out in the later years. The line graph shows an increase in rate of population change of 70% in the year 1890 to 1900. The rate of population change plummets from 46% to -2% in the year 1920 to 1930.



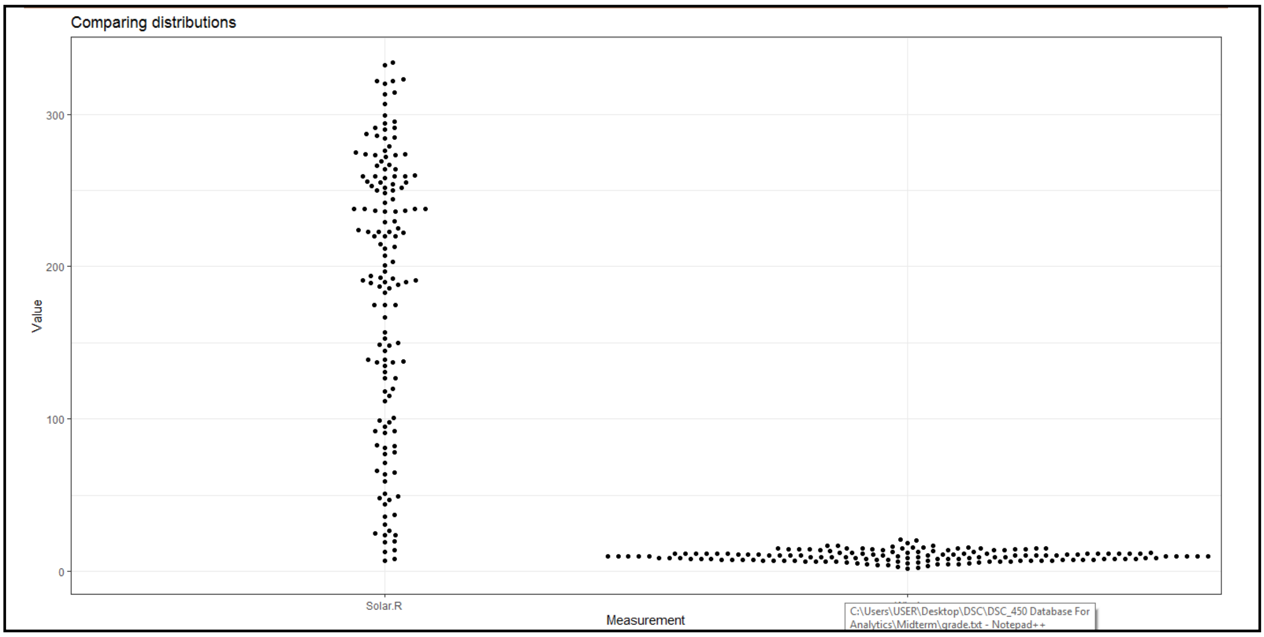

**Conclusion:** The line graph plots the % change in population for Montana across the years.
As can be seen in the line graph, the rate of change in population flattens out in the later years.
The line graph shows an increase in rate of population change of 70% in the year 1890 to 1900.
The rate of population change plummets from 46% to -2% in the year 1920 to 1930.
The years 1890 – 1900, 1900 – 1910 and 1920 – 1930 show a difference of more than 15% on the negative or positive side.
The percentage increase of more than 15% was on the positive side in the year 1890 – 1900.

```
> AirQuality %>%
+   mutate(Wind=rescale(Wind, to=c(0,1))) %>%
+   mutate(Solar=rescale(Solar.R, to=c(0,1))) %>%
+   ggplot(aes(Solar,Wind)) +
+   geom_point() +  geom_smooth(method=lm) + theme_bw()+
+   labs(x = "Solar",
+        y = "Wind",
+        title = "Solar vs. Wind")
```



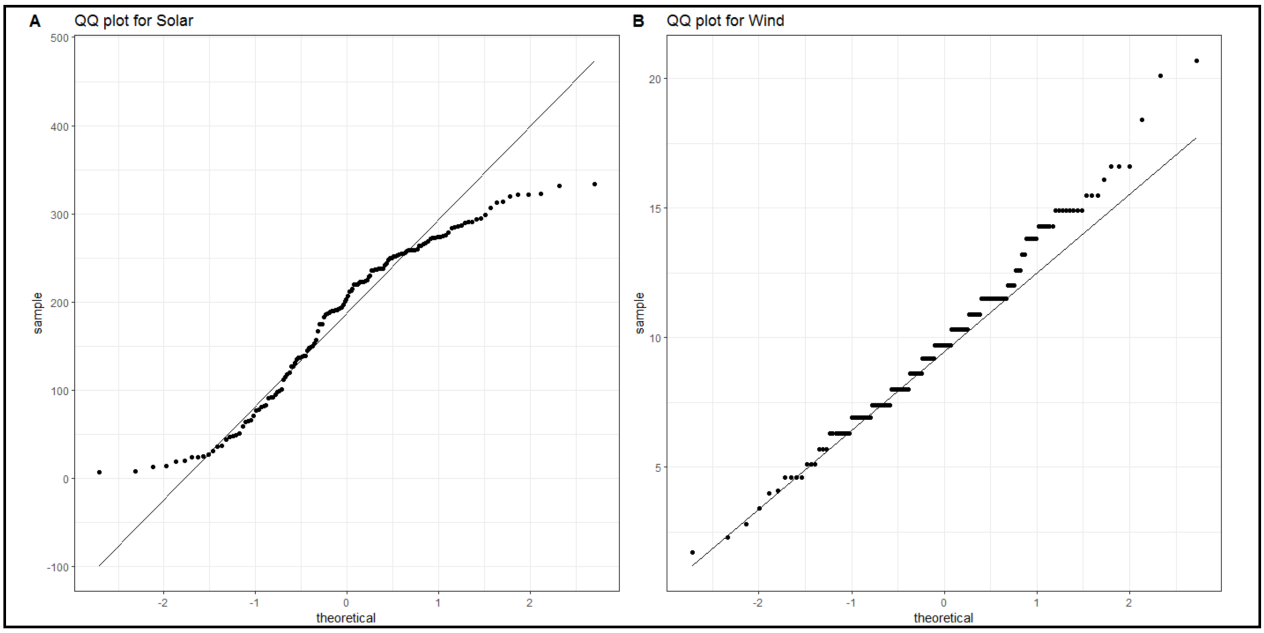
**Conclusion:** The scatter plot shows the data points for Solar against Wind.
The fit line for the data points is either flat or slightly on the negative side.
As can be seen in the graph, we do not see a particular pattern with the points plotted.

```
> AirQuality_pivot <- AirQuality %>% select(Ozone, Solar.R, Wind, Temp, Da
y, Month) %>% gather(key, value, -Day, -Month)

> filter_key <- c("Solar.R", "Wind")

> AirQuality_pivot %>% filter(key %in% filter_key) %>%
+   ggplot(aes(key, value)) +
+   geom_beeswarm() +
+   theme_bw() +
+   labs(x = "Measurement",
+        y = "Value",
+        title = "Comparing distributions")
```

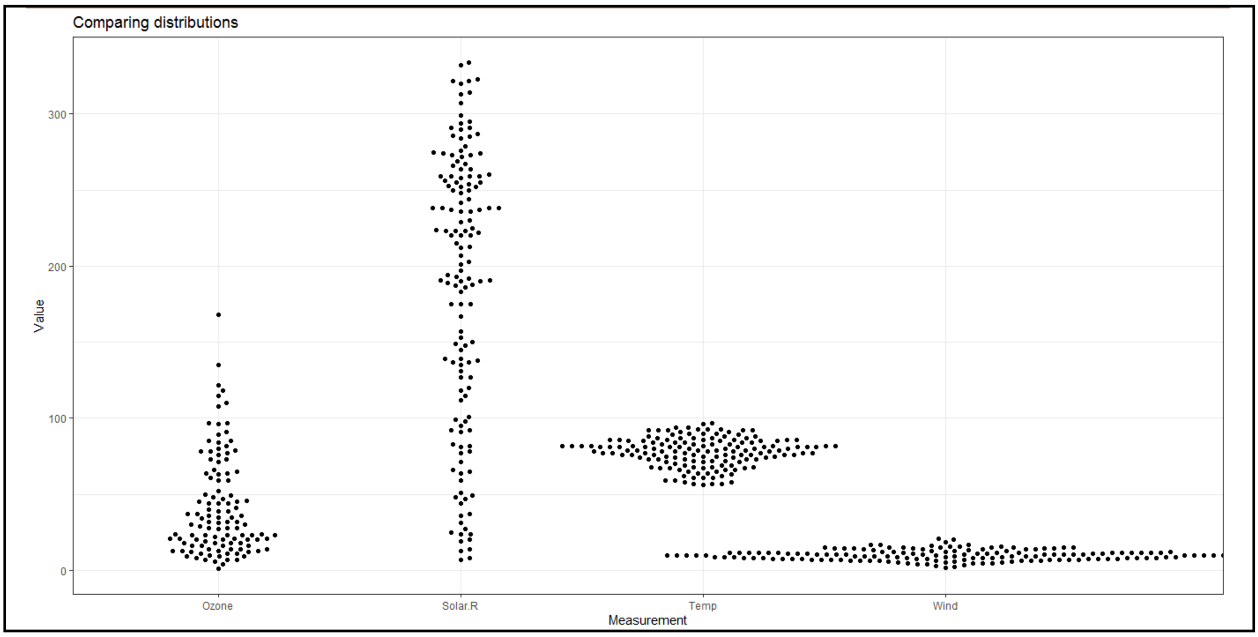



**Conclusion:** The beeswarm plot shows the distribution of the data points for Solar and Wind.
The data distribution for Solar is vertical ranging from 0 to 350. The data is all confined without any spread.
The data distribution for Wind is flat with most of the data ranging from 0 to 20.

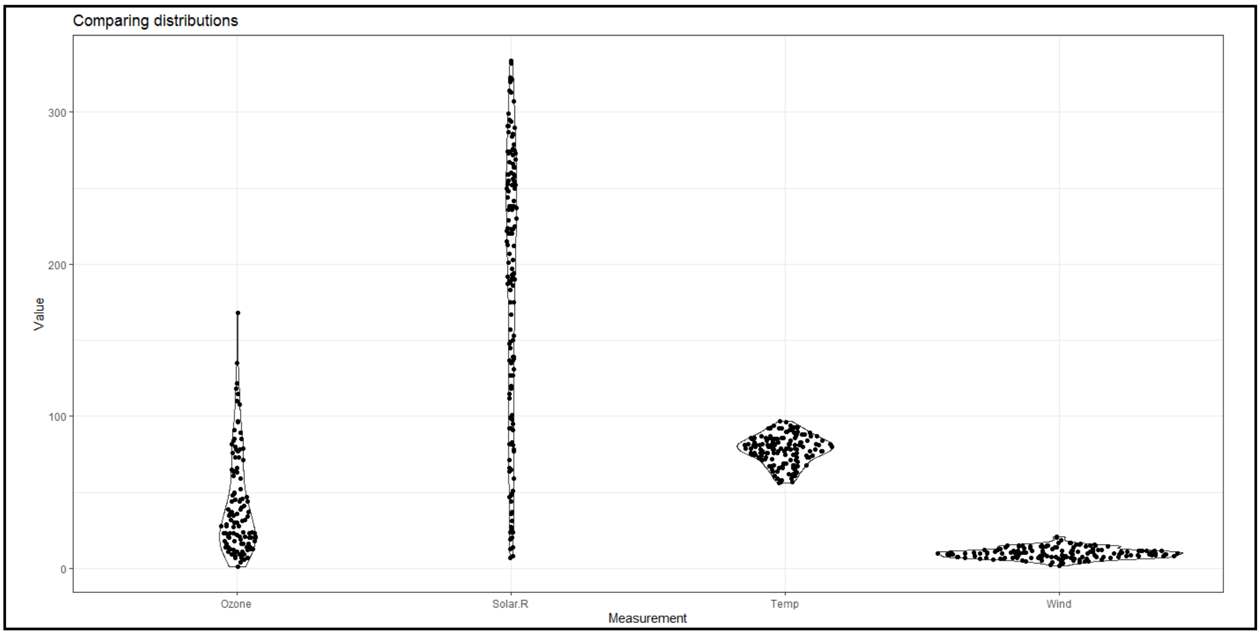
```
> qqSolar <- ggplot(AirQuality, aes(sample=Solar.R)) +
+   geom_qq() +
+   geom_qq_line() +
+   theme_bw() +
+   labs(title = "QQ plot for Solar")

> qqWind <- ggplot(AirQuality, aes(sample=Wind)) +
+   geom_qq() +
+   geom_qq_line() +
+   theme_bw() +
+   labs(title = "QQ plot for Wind")

> plot_grid(qqSolar, qqWind, labels = "AUTO")
```




```
> AirQuality_pivot %>%
+   ggplot(aes(key, value)) +
+   geom_beeswarm() +
+   theme_bw() +
+   labs(x = "Measurement",
+        y = "Value",
+        title = "Comparing distributions")
```

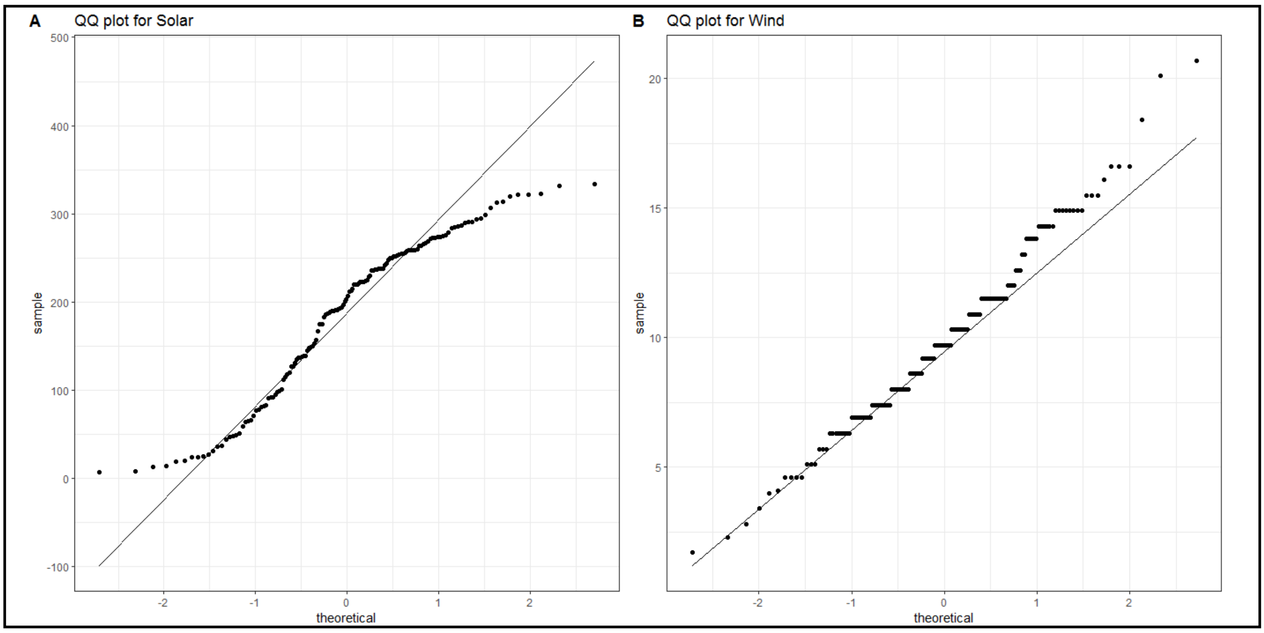



```
> AirQuality_pivot %>%
+   ggplot(aes(key, value)) +
+   geom_violin() +
+   geom_sina() +
+   theme_bw() +
+   labs(x = "Measurement",
+        y = "Value",
+        title = "Comparing distributions")
```




**Conclusion:** The beeswarm plot shows the distribution of the data points for various categories.
The data distribution for Solar is vertical ranging from 0 to 350. The data is all confined without any spread.
The data distribution for Wind is flat with most of the data ranging from 0 to 20.
The data distribution for Temp is confined to a small space ranging from 55 to 100.
The value for Ozone is heavily skewed near 0 with values ranging until 170.

```
> qqSolar <- ggplot(AirQuality, aes(sample=Solar.R)) +
+   geom_qq() +
+   geom_qq_line() +
+   theme_bw() +
+   labs(title = "QQ plot for Solar")

> qqWind <- ggplot(AirQuality, aes(sample=Wind)) +
+   geom_qq() +
+   geom_qq_line() +
+   theme_bw() +
+   labs(title = "QQ plot for Wind")

> plot_grid(qqSolar, qqWind, labels = "AUTO")
```




**Conclusion:** The qqplot plot shows the distribution of the data points for Solar and Wind.
The plot shows the data distribution for Solar and Wind mostly falls on the line from – 1 to 1.
The values for Solar start deviating on the negative end from 1 onwards while the values for Wind start deviating on the positive side.