Name: Rohan Dhoyda

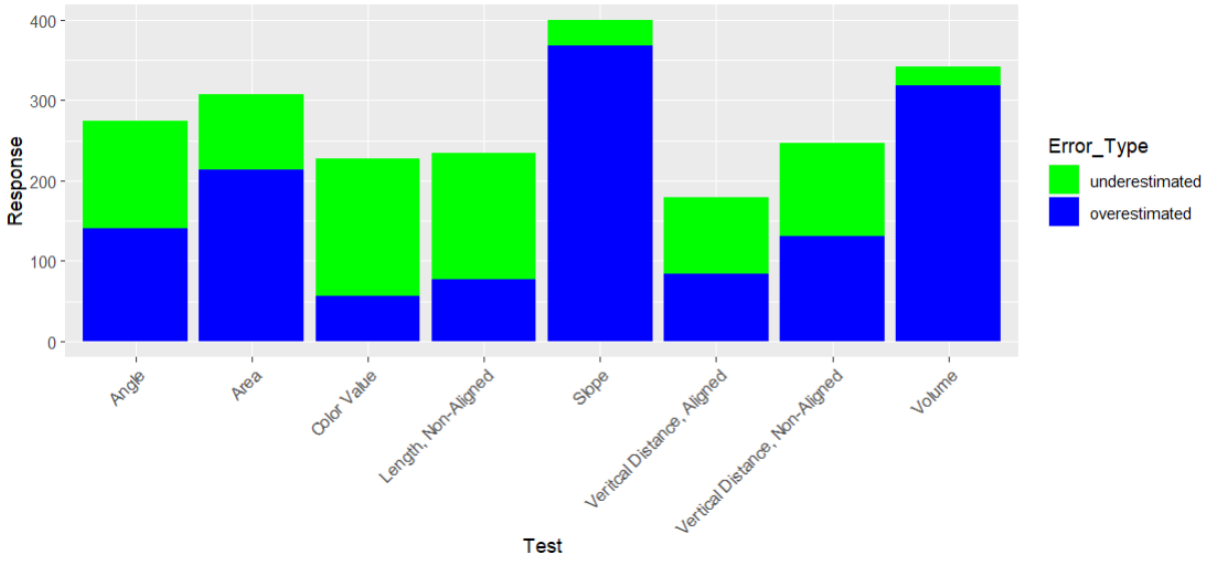
DATA VISUALISATION Homework 3

1) (20 pts) This problem will not only give you practice creating visualizations, but requires you to follow carefully a somewhat complicated specification of experimental data and use visualization for problem solving. Recall the perception experiment from our first week. You saw a sequence of slides each with four encoded values, marked A, B, C and D. You were supposed to write down the values for B, C and D as a proportion of A. On each slide the encodings (e.g. aligned bar, volume, etc.) changed, and each encoding was repeated. The data file for this problem, PerceptionExperiment.csv, contains the results from 92 previous students. (For those interested in experimental design, note that the order of the slides was changed fordifferent classes.)

1A)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.

Ans:

```
df_perception$Error = df_perception$Response - df_perception$TrueValue
> df_perception = df_perception %>% mutate(Error_Type = cut(Error, breaks=c(-Inf, -0.0001,0.00
001, Inf), labels=c("underestimated","equal","overestimated")))
> perception_plot_2 <- df_perception %>%
+   ggplot(aes(x = Test, y = abs(Error))) +
+   geom_boxplot(color = "blue",”Green”) +
+   geom_sina()
> print(plot_2)
```

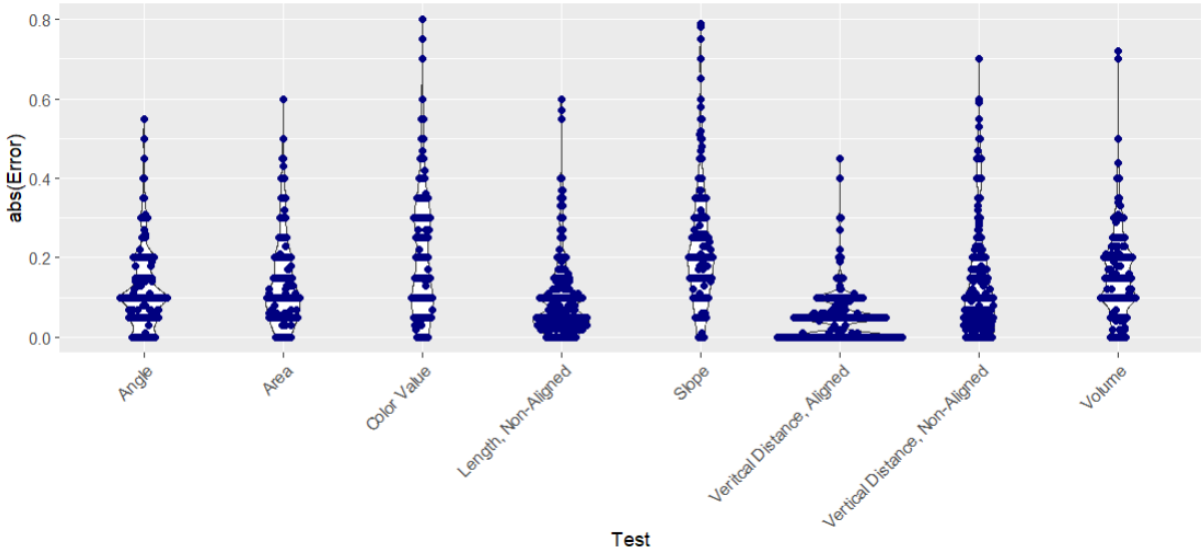


A column chart can be used to visualize the difference between how much people overestimate and underestimate their answers on each test. This chart shows that the "Slope" and "Volume" tests are the tests where people overestimate their answers the most.

1B) 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?

Ans:

```
 perception_plot_2 <- df_perception %>%
+   ggplot(aes(x = Test, y = abs(Error))) +
+   geom_violin() + theme(axis.text.x=element_text(angle=45,hjust = 1))+
+   geom_sina(color = "navy")
 perception_plot_2
```

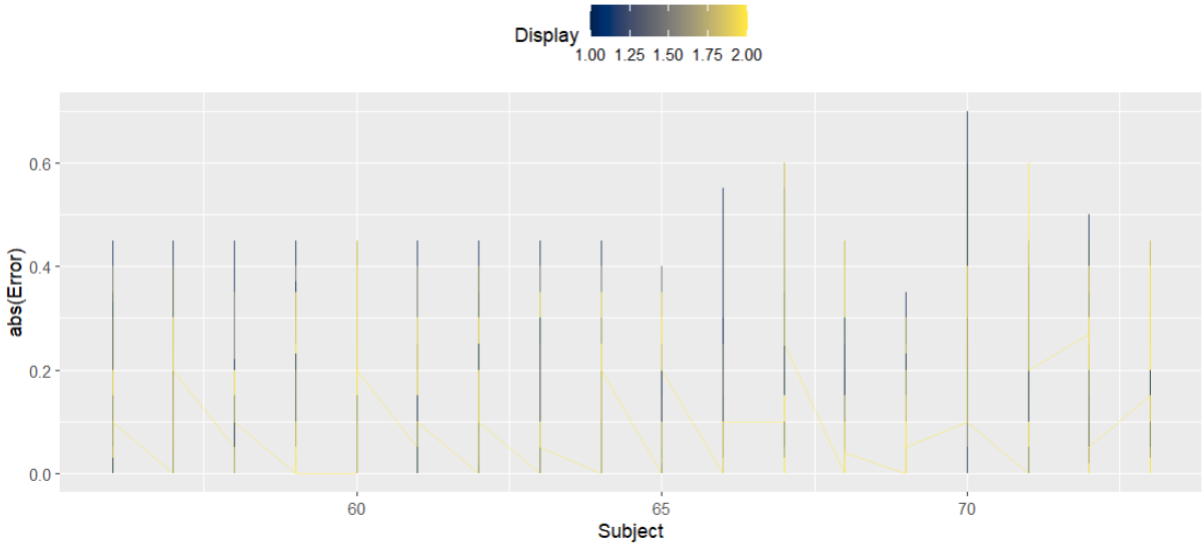


Every test has anomalous data points, and most tests have fairly similar faults with the exception of "Vertical Distance, Aligned," where errors are notably lower.

1C) 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?

Ans:

```
perception_plot_3 <- df_perception %>%
+   filter(Subject >= 56 & Subject <= 73) %>%
+   ggplot(aes(x = Subject, y = abs(Error), color = Display)) +
+   geom_line(alpha = 0.5) +
+   theme(legend.position = "top") +
+   scale_color_viridis(option = "E")
   perception_plot_3
```



The x-axis of the graph shows the subject number, from 56 to 73.

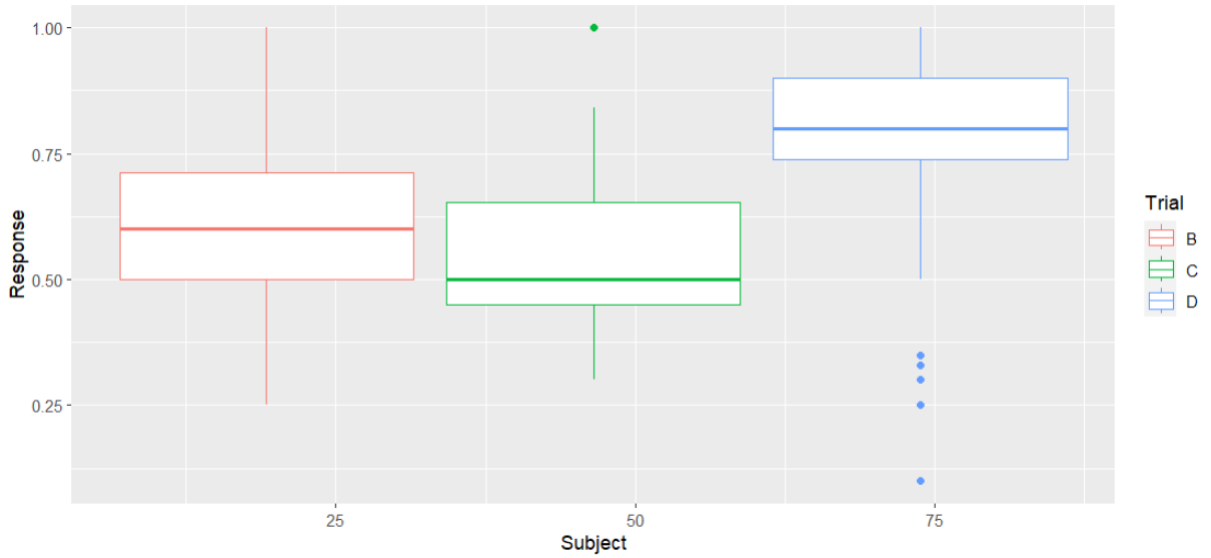
The y-axis of the graph shows the average absolute error for each subject.

The two lines on the graph represent the average absolute error for Display 1 (blue line) and Display 2 (golden line).

The blue line is generally higher than the golden line, which indicates that the participants had a slightly higher average absolute error for Display 1. This suggests that the participants may have found Display 1 to be more difficult to judge than Display 2.

1d. 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 is; they are only for a small subset of subjects.
Ans:

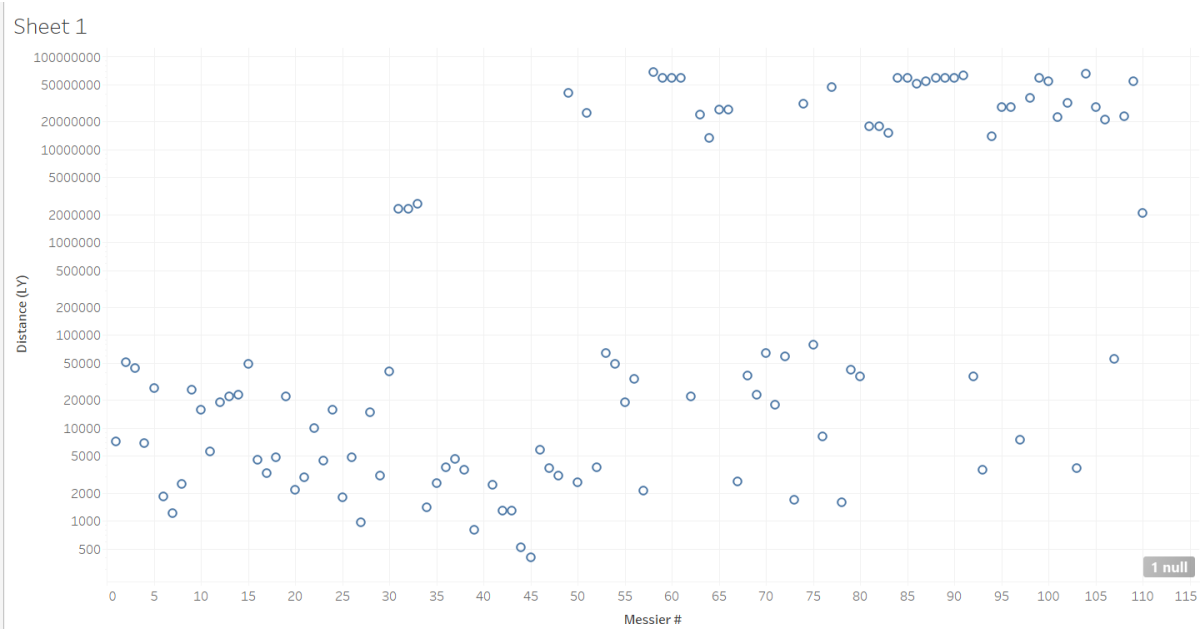
```
perception_plot_4 <- df_perception %>%
+   filter(Test =="Vertical Distance, Non-Aligned" & Display ==1) %>%
+   ggplot(aes(x = Subject, y = Response, color = Trial)) +
+   geom_boxplot()
> perception_plot_4
```



Box plot of response distribution shows that the anomalous sequence of "1" responses is more pronounced for subjects above 55.

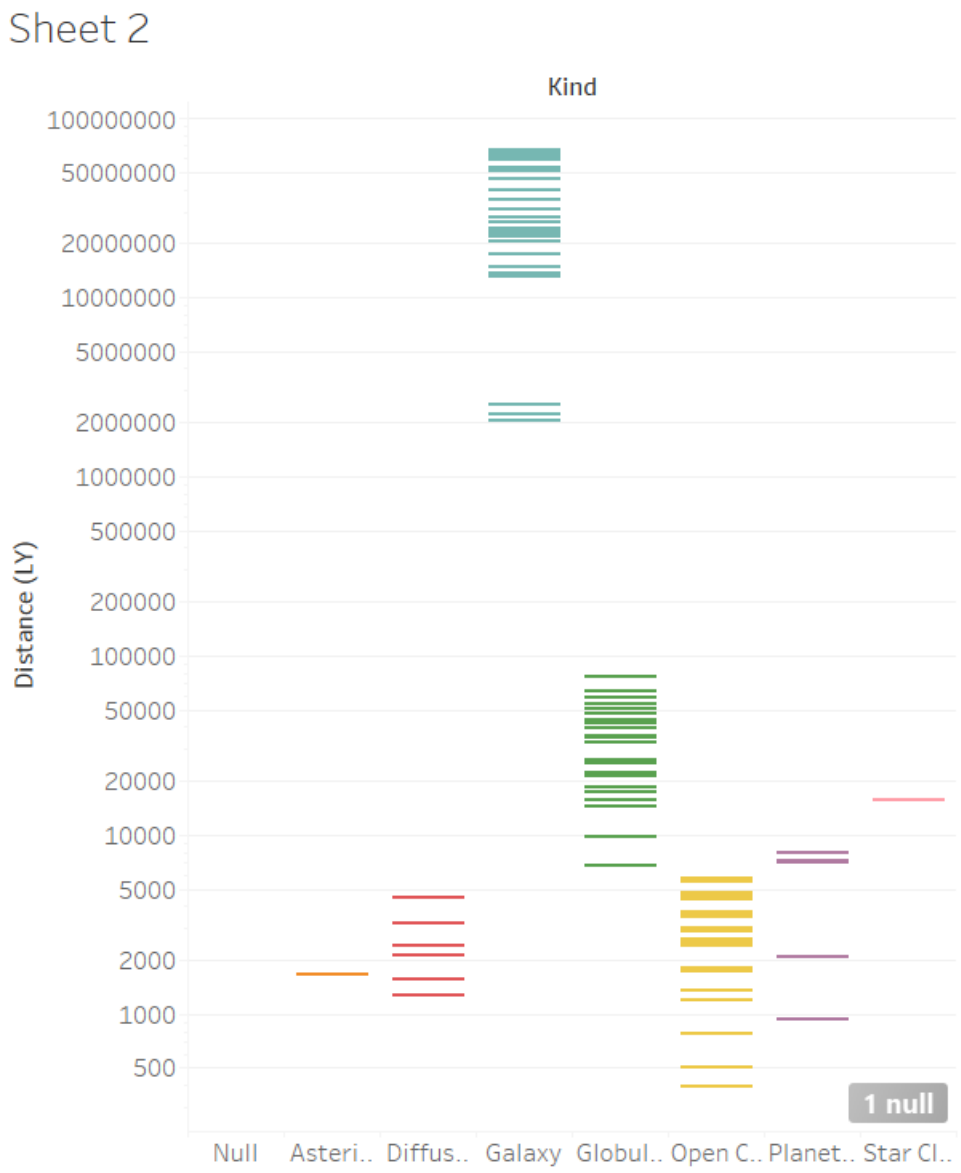
2)(20pts) Download the astronomical data for the Messier objects. These are objects that can be seen in a dark sky with binoculars or a telescope that Charles Messier cataloged in France in the 18th century so that they wouldn't be confused with comets. Some of these are clusters of stars or great clouds 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. 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. In particular, for distances, the scale should show the "order-of-magnitude" of the distance in light years (10, 100, 1000, etc.) clearly.

2a) 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?

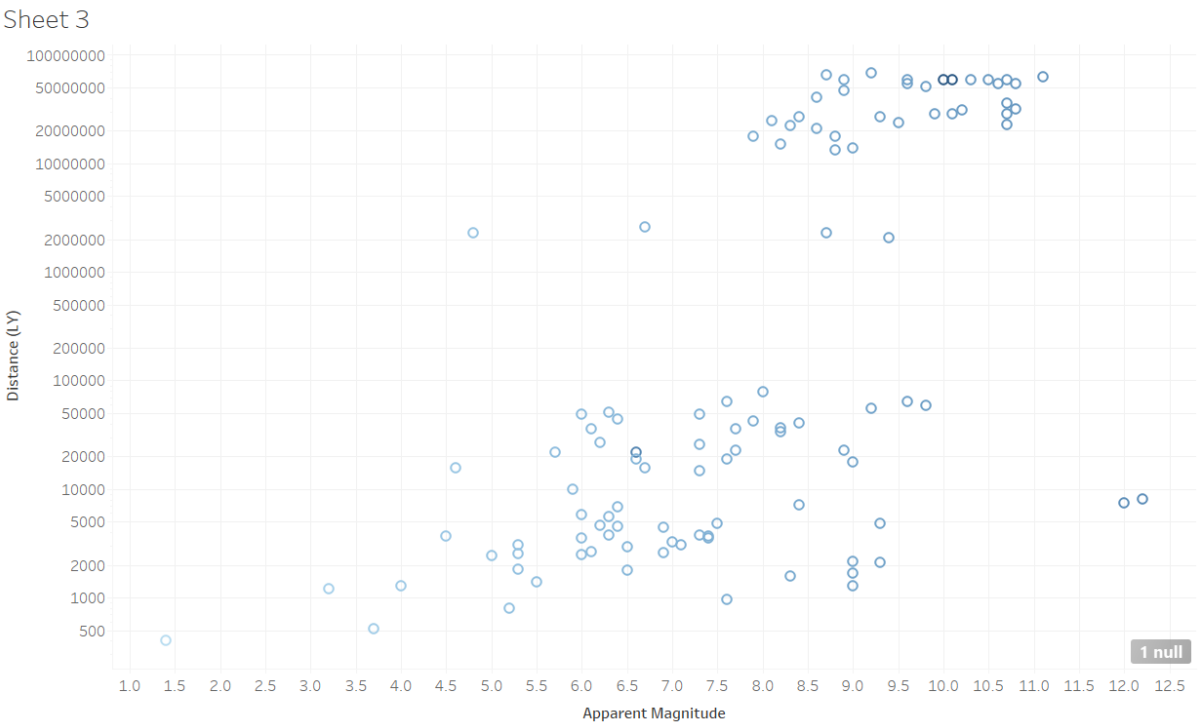


The data shows that there is a small but positive relationship between the Messier number and the Distance of the objects.

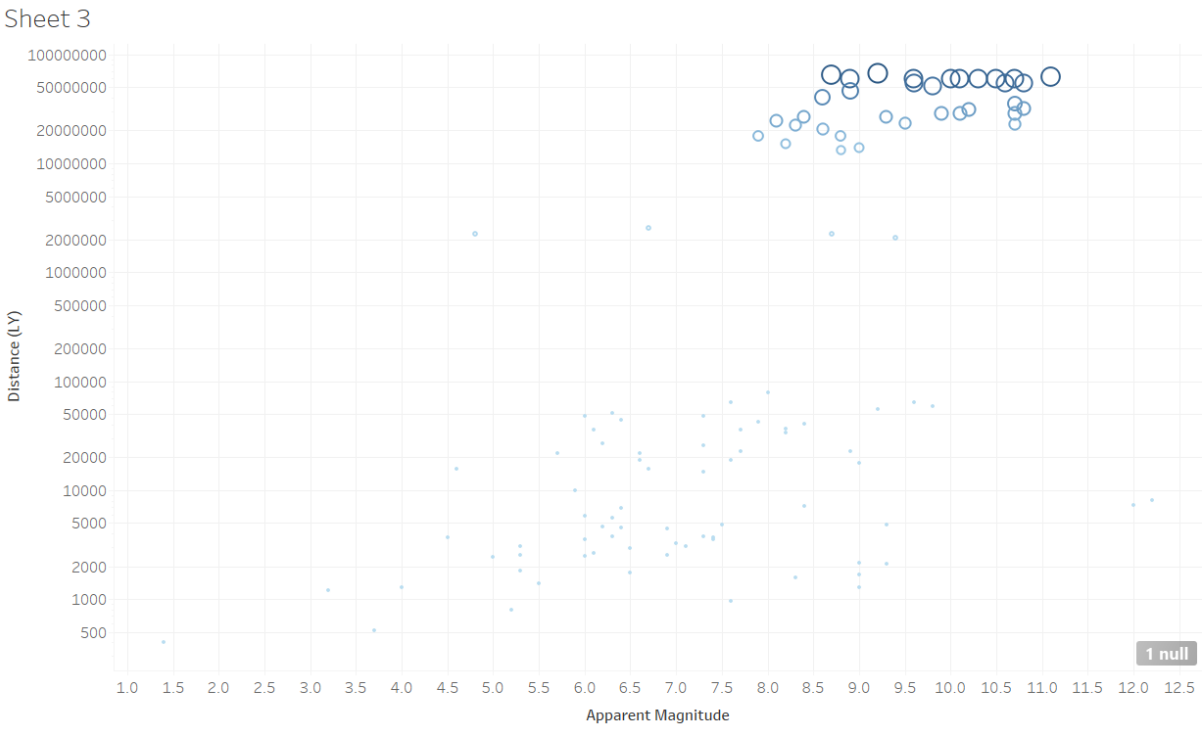
2b. Create a visualization that compares the distributions of the distances to the objects in each Kind. Note that the Type variable is a very different category and is really a subcategory of Kind. Do not use that here. Sort the distribution displays in a way that makes the relationship clear.



2c. 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… backward 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.



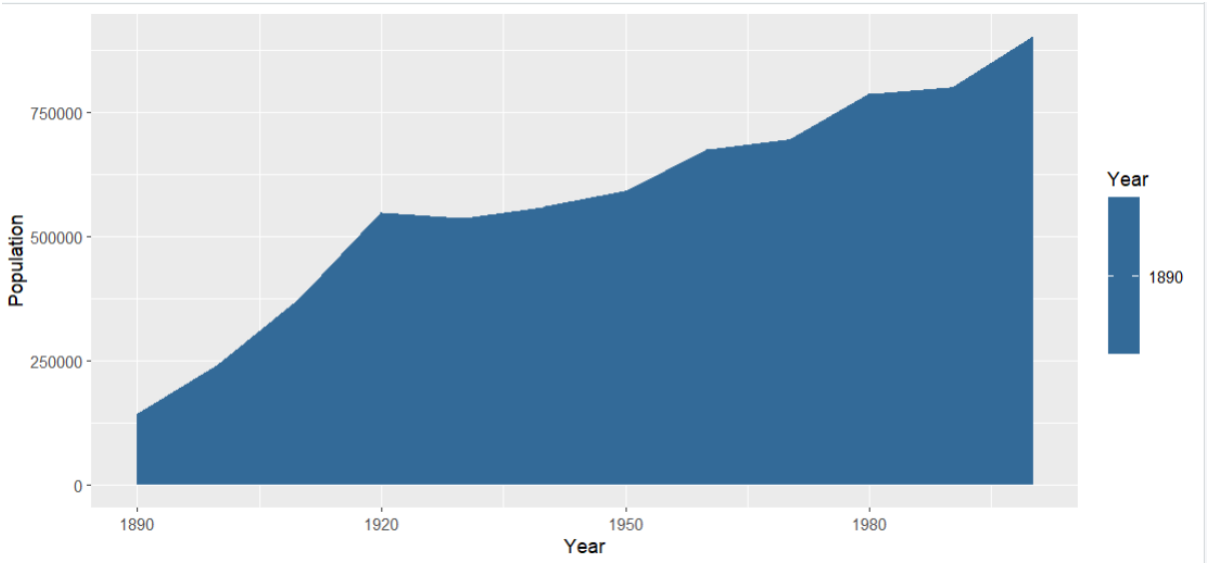
2d. Augment the visualization in (c) by adjusting the size of the points in the scatter-plot based on the angular Size of the objects in the sky. Evaluate how easy it is to analyze all encoded aspects of the data from this graph and give a suggestion on how you might modify the graph to display all this information more readably.



It was not a tough task to visualize the Apparent magnitude with distance. We can see that the size of the scatter plot is visible when the distance is near the apparent magnitude but increases as it goes higher.

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.

3a. How many times has the population doubled since 1890?
Ans:




```
montana_plot_1 <- ggplot(df_montana, aes(x = Year, y = Population, fill = Year)) + geom_area()

montana_plot_1
```

The graph shows that the population of the world has doubled four times, from 1890 to 1920. This means that the population has grown exponentially over the past century.

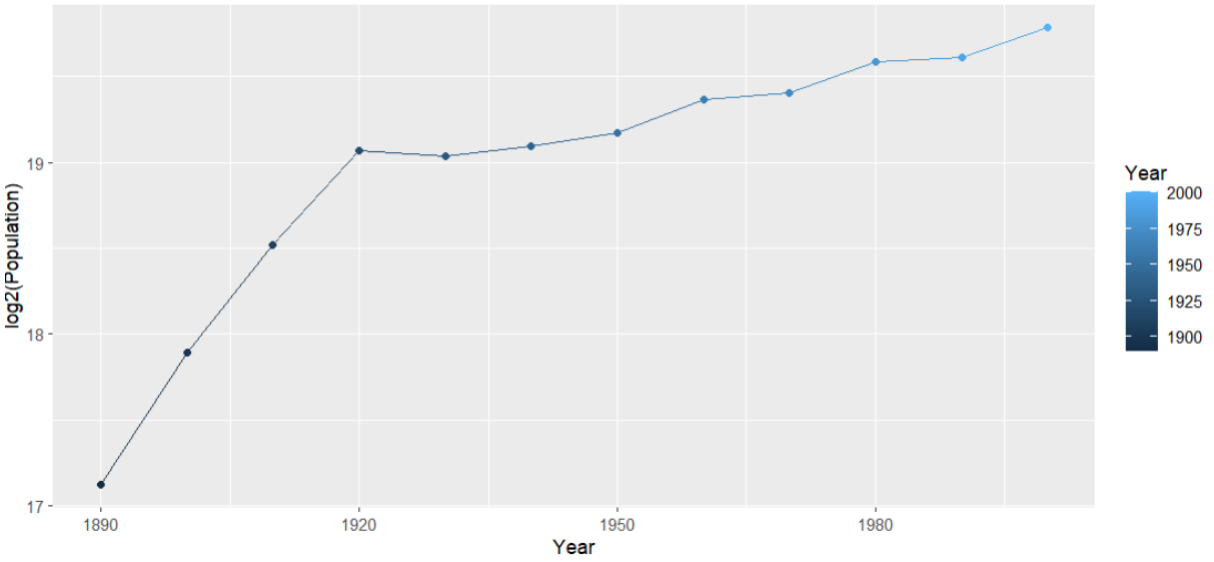
3b. Has the percentage rate of change in the population increased or decreased over the years? What years had the greatest increase in population %-wise?
Ans:

```
colors <- c("#E69F00", "#56B4E9", "#009E73", "#0072B2", "#D55E00")

montana_plot_1 <- ggplot(df_montana, aes(x = Year, y = log2(Population), color = Year)) + geom_line() + geom_point()

montana_plot_1 + scale_color_manual(values = colors)

print(montana_plot_1)
```



The percentage rate of change in population has declined since 1920 as seen from the graph

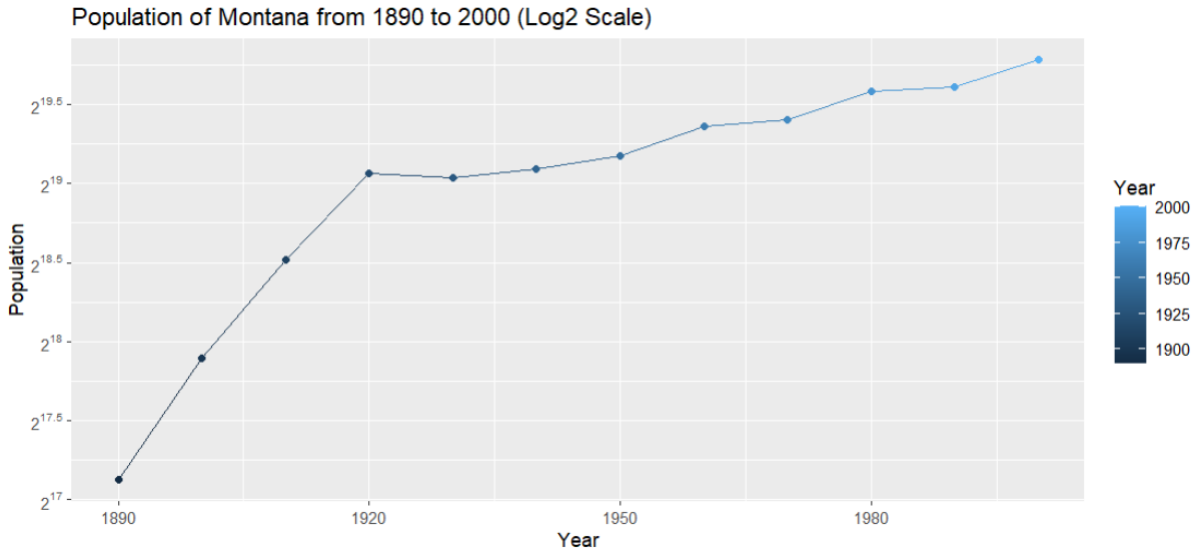
3c. What years was the population percentage increase greater than 15%?

```
colors <- c("#E69F00", "#56B4E9", "#009E73", "#0072B2", "#D55E00")

montana_plot_1_filtered <- ggplot(df_montana_filtered, aes(x = Year, y = Population, color = Year)) + geom_line() + geom_point() + scale_y_continuous(trans = "log2", breaks = trans_breaks("log2", function(x) 2^x), labels = trans_format("log2", math_format(2^.x))) + ggtitle("Population of Montana from 1890 to 2000 (Log2 Scale)")

montana_plot_1_filtered + scale_color_manual(values = colors)


print(montana_plot_1_filtered)
```



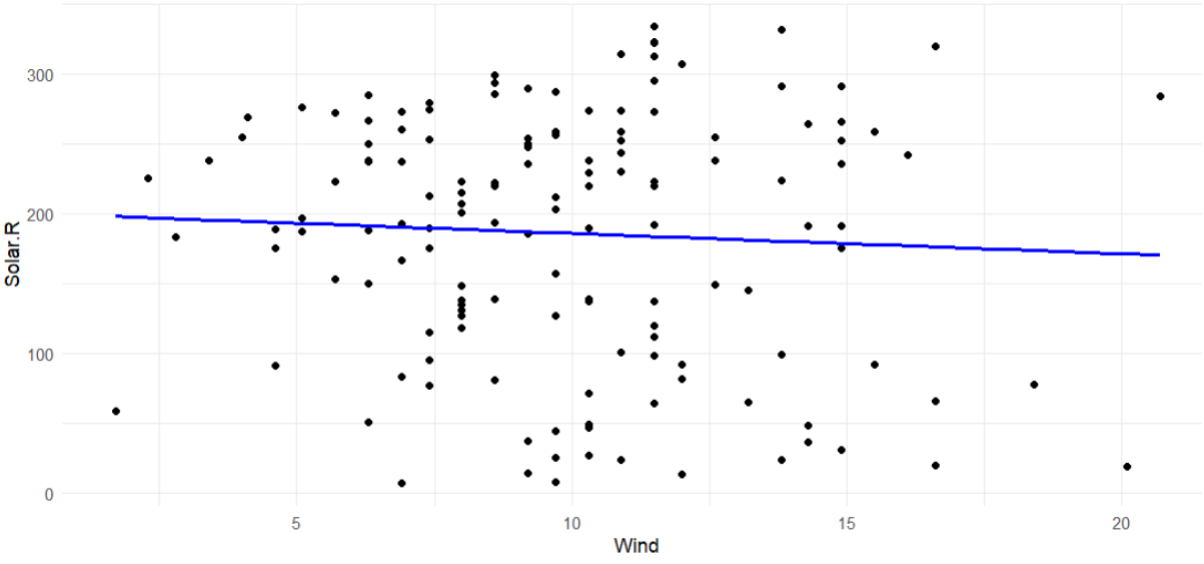
The population of Montana increased by more than 15% in the periods 1890-1920 and 1950-1980.

4. (20 pts) We will look at data on air quality, captured from May to September in New York. This is actually built into R, but not as a data frame. There is a copy on the D2L site

4a. Use a scatterplot 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.

Ans:

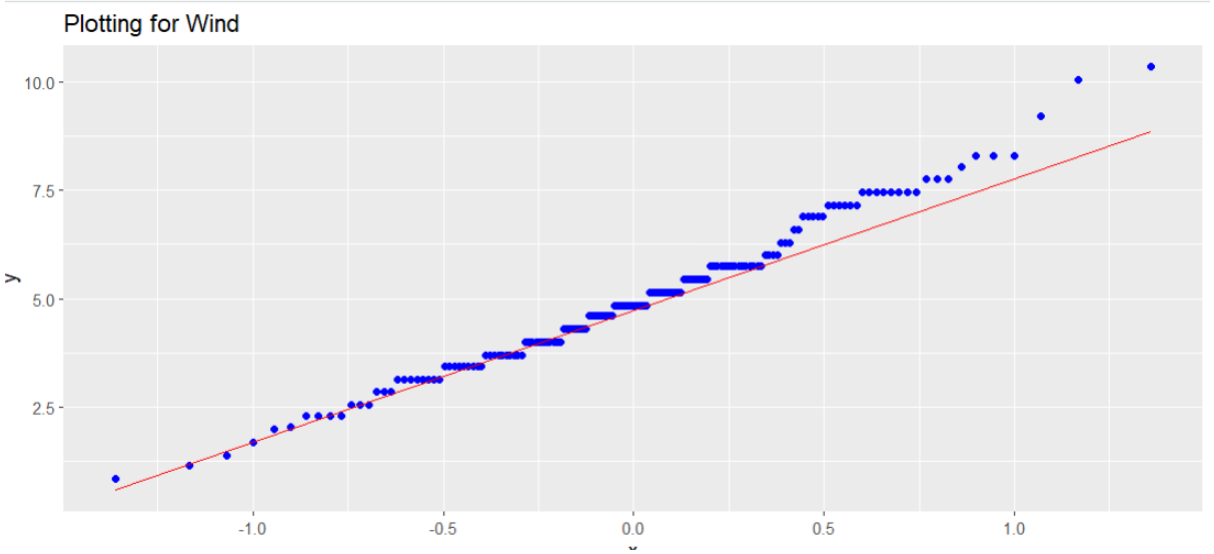
```
df_aq <- read.csv("AirQuality.csv")
 ggplot(df_aq, aes(x = Wind, y = Solar.R)) +
 geom_point() +  # Add scatterplot points
 geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add a linear regression trendline
  labs(x = "Wind", y = "Solar.R") +  # Label the axes
 theme_minimal()
```



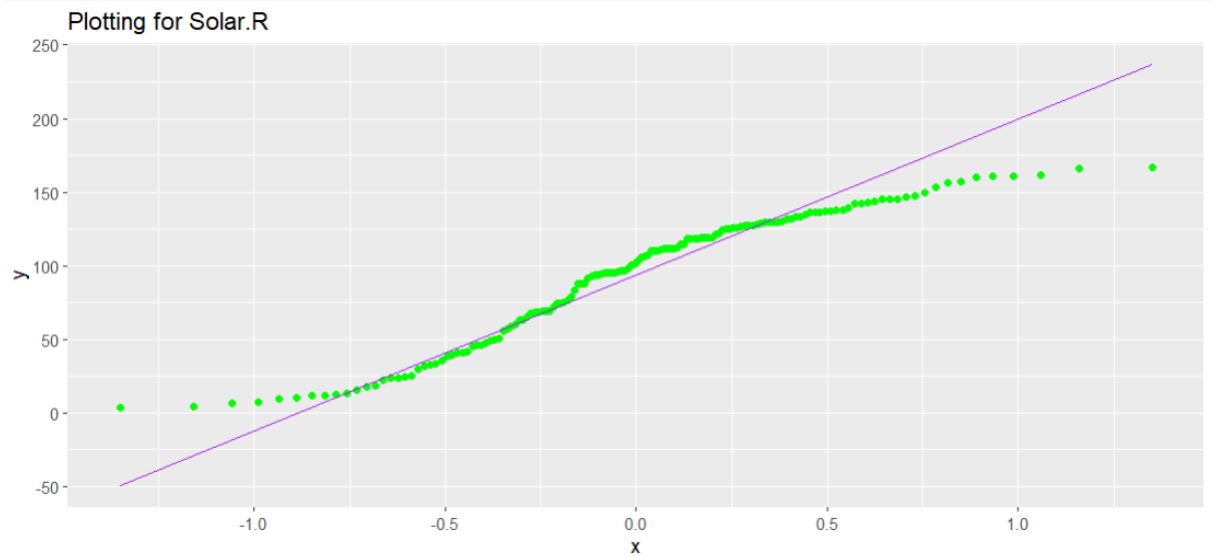
4b. Use a plot that will show the distributions of Wind and Solar.R and allow you to compare with fine detail.

Ans:

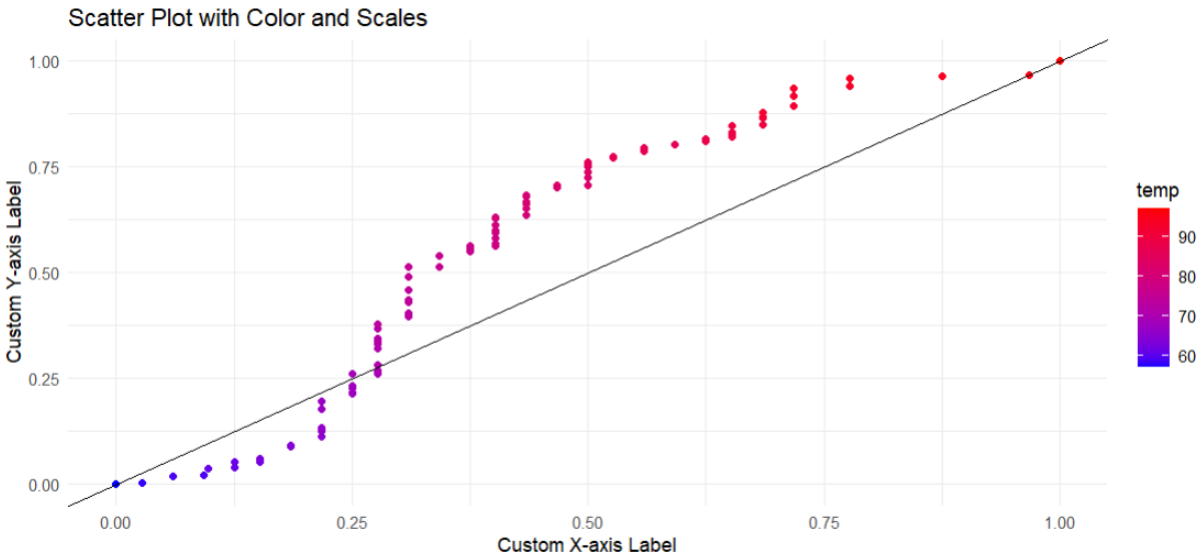
```
ggplot(df_aq, aes(sample = Wind)) +
geom_qq(color = "blue") +
geom_qq_line(color = "red") +
scale_x_continuous(labels = scales::number_format(scale = 0.5)) +
scale_y_continuous(labels = scales::number_format(scale = 0.5)) +
labs(title = "Plotting for Wind")
```




```
ggplot(df_aq, aes(sample = Solar.R)) +
geom_qq(color = "green") +
geom_qq_line(color = "purple") +
scale_x_continuous(labels = scales::number_format(scale = 0.5)) +
scale_y_continuous(labels = scales::number_format(scale = 0.5)) +
labs(title = "Plotting for Solar.R")
```



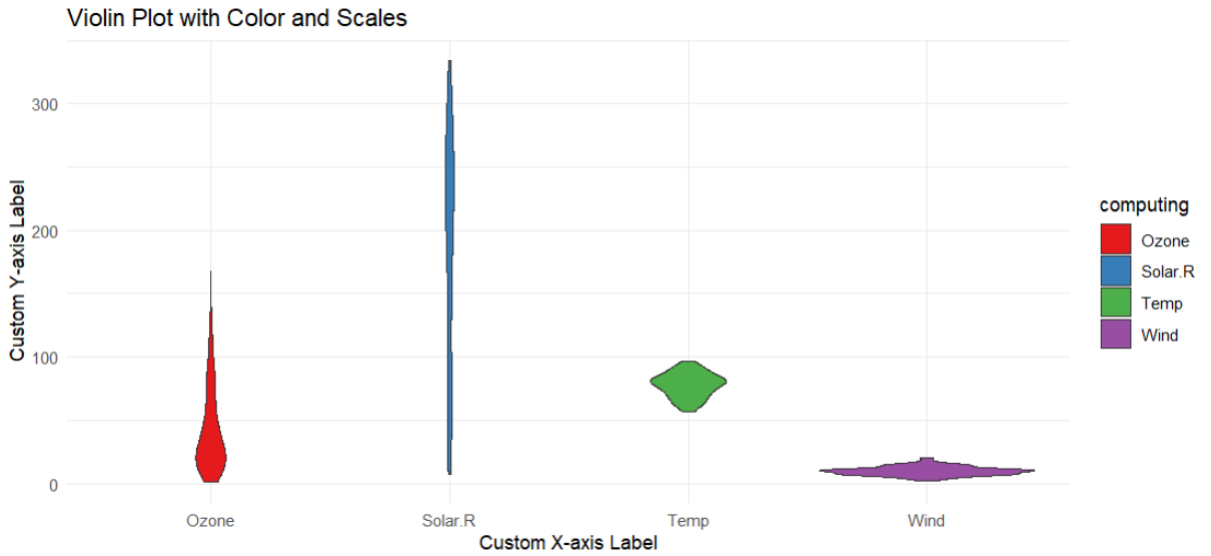
```
bq_plot_2 <- ggplot(bq.qq, aes(x = WIND, y = SOLAR, color = temp)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0) +
  scale_color_gradient(low = "blue", high = "red") +
  scale_x_continuous(name = "Custom X-axis Label") +
  scale_y_continuous(name = "Custom Y-axis Label") +
  labs(title = "Scatter Plot with Color and Scales") +
  theme_minimal()
print(bq_plot_2)
```



4c.Finally, show these distributions in context of the rest of the variables by using a technique for comparing multiple distributions.
Ans:

```
aqQuality=df_aq %>% pivot_longer(-c(Month, Day), names_to="computing", values_to="estimate")
head(aqQuality)
```

```
  aqQuality %>%
  ggplot(aes(x = computing, y = estimate, fill = computing)) +
  geom_violin() +
  scale_fill_brewer(palette = "Set1") +  # Using a color palette
  labs(x = "Custom X-axis Label", y = "Custom Y-axis Label", title = "Violin Plot with Color and Scales")
+  theme_minimal()
```



4d. For extra credit, compare Wind and Solar.R again with a QQ plot. What does this tell you?
Ans:

```
cq_plot_4 <- ggplot(cq.qq, aes(x = WIND, y = SOLAR, color = temp)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  scale_color_gradient(low = "green", high = "red") +
  scale_x_continuous(name = "WIND", limits = c(0, 1)) +
  scale_y_continuous(name = "SOLAR", limits = c(0, 1)) +     labs(title = "Scatter Plot with Color, Scale, and Regression Line") +    theme_minimal()
print(cq_plot_4)
```