1. **Changes Made to Initial Design:**

   According to the feedback we got from project 2B, we made some changes in some of the plots. Specifically, we changed the pH scale's color in the scatter plot of the Hardness and Solids with pH. The color scale was changed to make each point clearly distinguished at each pH level. In the box plot, we added the statistical values to represent the median, Q1, Q3, minimum, and maximum by using the Potability and pH column. Finally, we changed the line graph of pH and Turbidity compared to Potability to the Scatter plot of pH and Turbidity compared to Potability, since the line makes the graph more confusing for the audience.

2. **Reason for Changing Initial Dataset:**

   We are using the same dataset that was used in Project 2A, "Water Potability."

3. **Most and Least Useful Aspects of Course:**

   Most Useful:

   - One of the most beneficial things we learned from this class is working with a new dataset and learning how to represent the variables and observances in a graph visualization based on the data type. The hands-on experience using Tableau to represent our visualizations on a public platform was also helpful.
   - The opportunities we were given to get experience with different visualization tools such as RAW, Tableau, and ggplot2 in R.
   - The ontologies of visualization and some analytical frameworks for data visualization such as user tasks, data value types, visual complexity (1-d, 2-d, 3-d, temporal, multi-dimension, network, tree or structured.

   Least Useful:

   - All the course material concepts are linked with each other, so we can not pick anything that is least useful.

Dataset Description:

We are using a dataset about drinking water's potability in different areas from Kaggle. There are 10 variables and 3,276 observations. The dataset provides data about the potability of water in different areas over years. Variables included in the dataset to decide if the water is potable include pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability. The content was taken from various websites like data.world, kaggle and some features are updated from those sites. The variables we are going to use with their statistical descriptions are: pH is numerical and mean is 7.081, hardness is numerical and mean is 196.37, solids is numerical and mean is 22,014.1, chloramines is numerical and mean is 7.122, sulfate is numerical and mean is 333.8, conductivity is numerical and mean is 426.2, organic carbon is numerical and mean is 14.28, trihalomethanes is numerical and mean is 66.396, turbidity is numerical and mean is 3.967, and lastly potability is binary.

We believe visualizing and analyzing the potability of the drinking water may interest water quality testers, community water suppliers, and the consumers themselves. We think the water quality testers are the ones who are most familiar with this topic because their daily job is to assess the quality of water at its source. They will be familiar with all the attributes used to test water quality such as pH, hardness, solid particles, sulfate particles, and more. They may know the standards to judge whether the water is good to drink, potable. However, they may not know the relationships between those water testing attributes. They may be interested in the relationship between those attributes, such as higher pH, how low the hardness is in the water, or they may also be interested in the comparisons of water potability in different pH scales, in different hardness elements, etc.

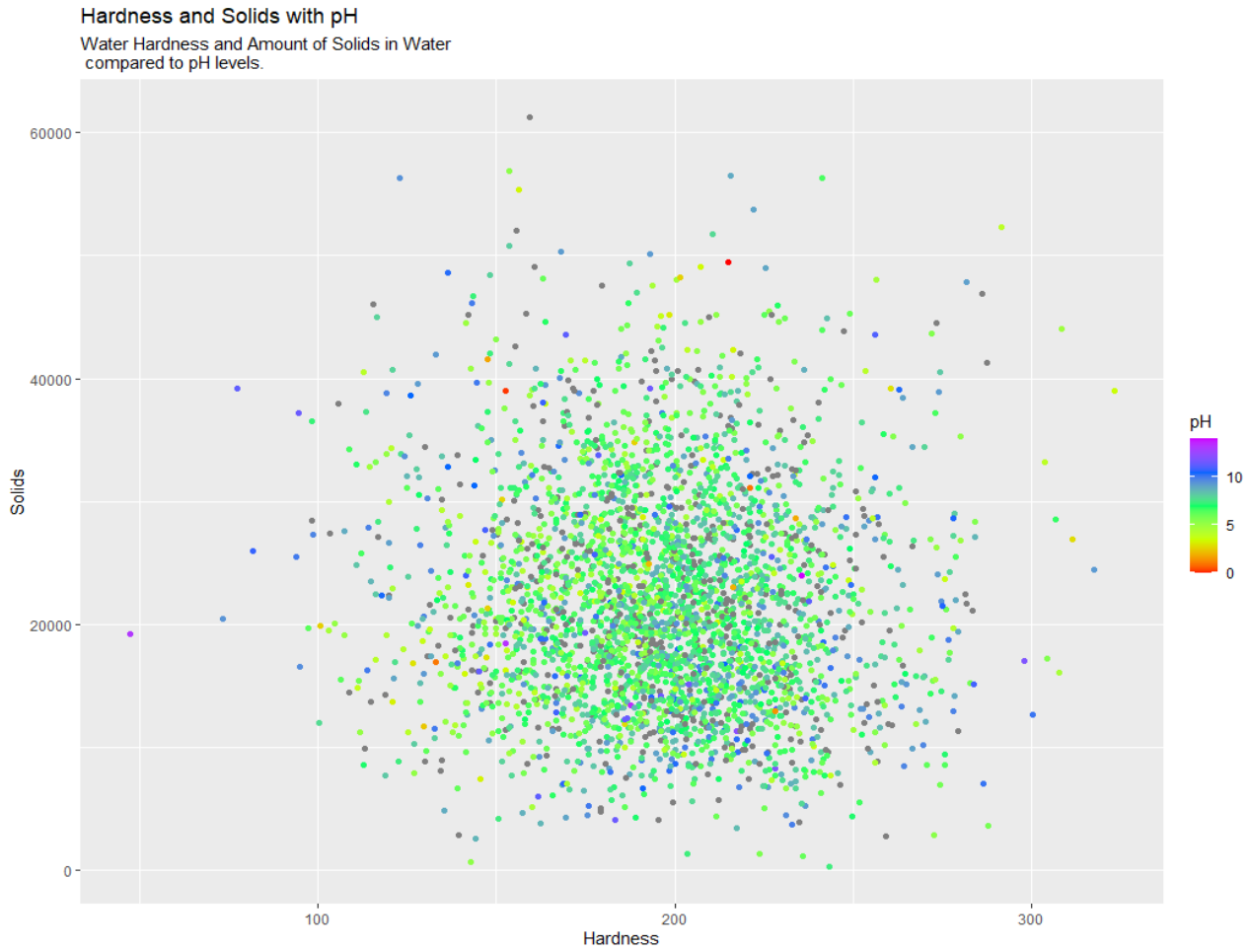
Community water suppliers may acknowledge different attributes used to qualify the drinking water. They may know the required standards for making water safe to drink. However, like the water quality testers, the community water suppliers may not acknowledge the relationships between the water testing attributes or the potability of the drinking water in different pH scales, in different hardness elements, and more.

Consumers usually do not know or care much about what makes the water safe to drink. As long as the community water suppliers and experts say it is safe to drink, the consumers will assume it is safe. However, consumers are directly affected if the water is not safe to drink. Thus, we think the comparisons of potability of the drinking water in different pH scales or different chemical elements amounts will interest them. Although consumers can receive an annual report from their community water supplier on their local drinking water quality, we believe that the story can be conveyed to them in a much easier and more interesting way with visualization.

Visualizations:

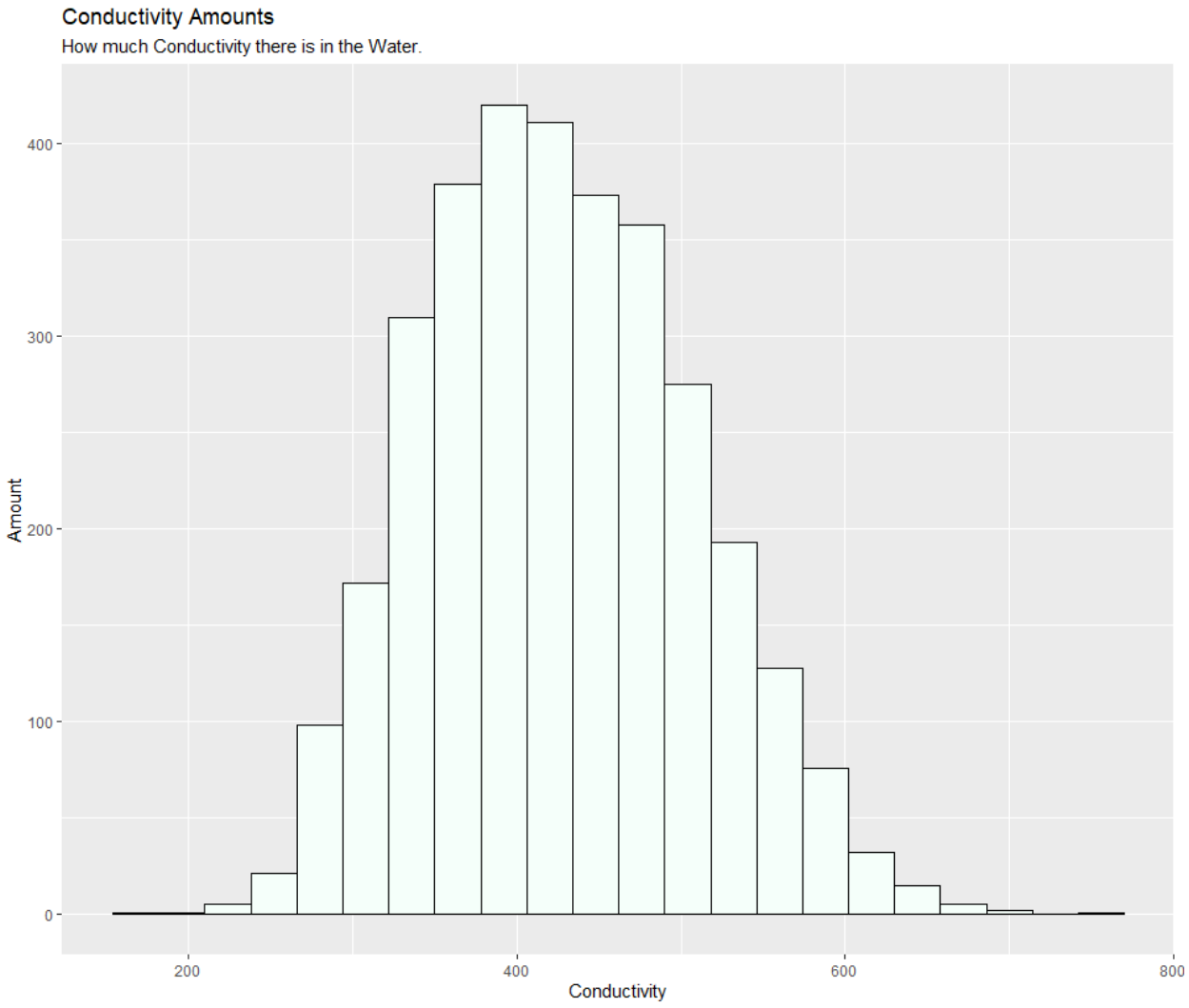
All visualizations were made in Rstudio. The visualizations are two scatterplots, a histogram, a pie chart, and a boxplot.

1. Scatter plot of the Hardness and Solids with pH



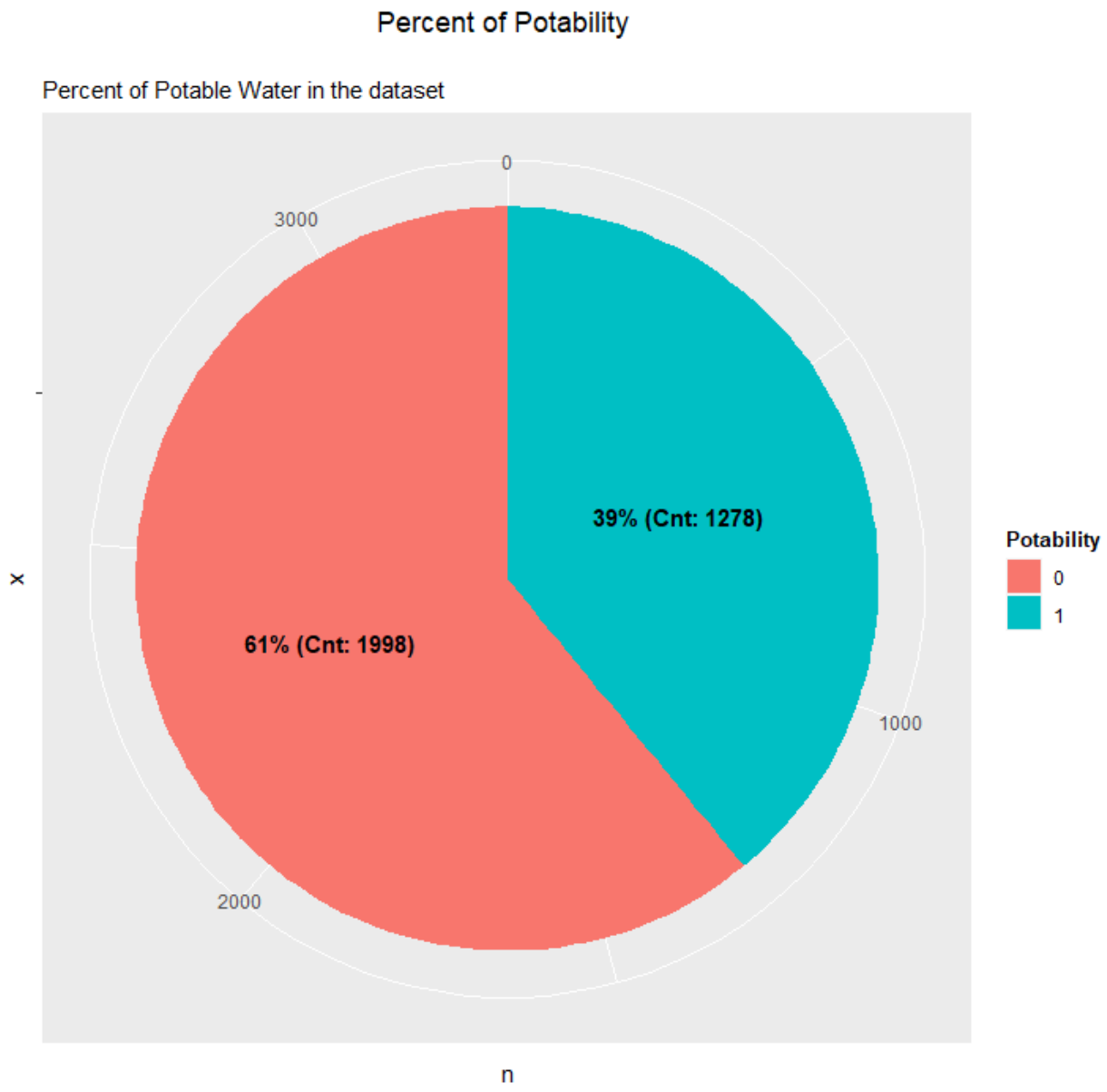
This scatter plot represents the water hardness against the number of solids in different pH levels. Looking at the plot, we can see that there are many points located at the center where the value of the Hardness is in the range 150 to 250 and the value for Solids in the range 1000 to 3000. The color of each point represents its pH level. The hotter the color (red, orange) the lower the pH, the more acidic the water is, and vice versa, the cooler the color (blue, purple), the higher the pH level, and the more alkaline the water is. The green range represents pH in range 6-7. Most of our points are green meaning most of our data points have pH in the range 6-7.

2. Histogram for the Conductivity Amounts



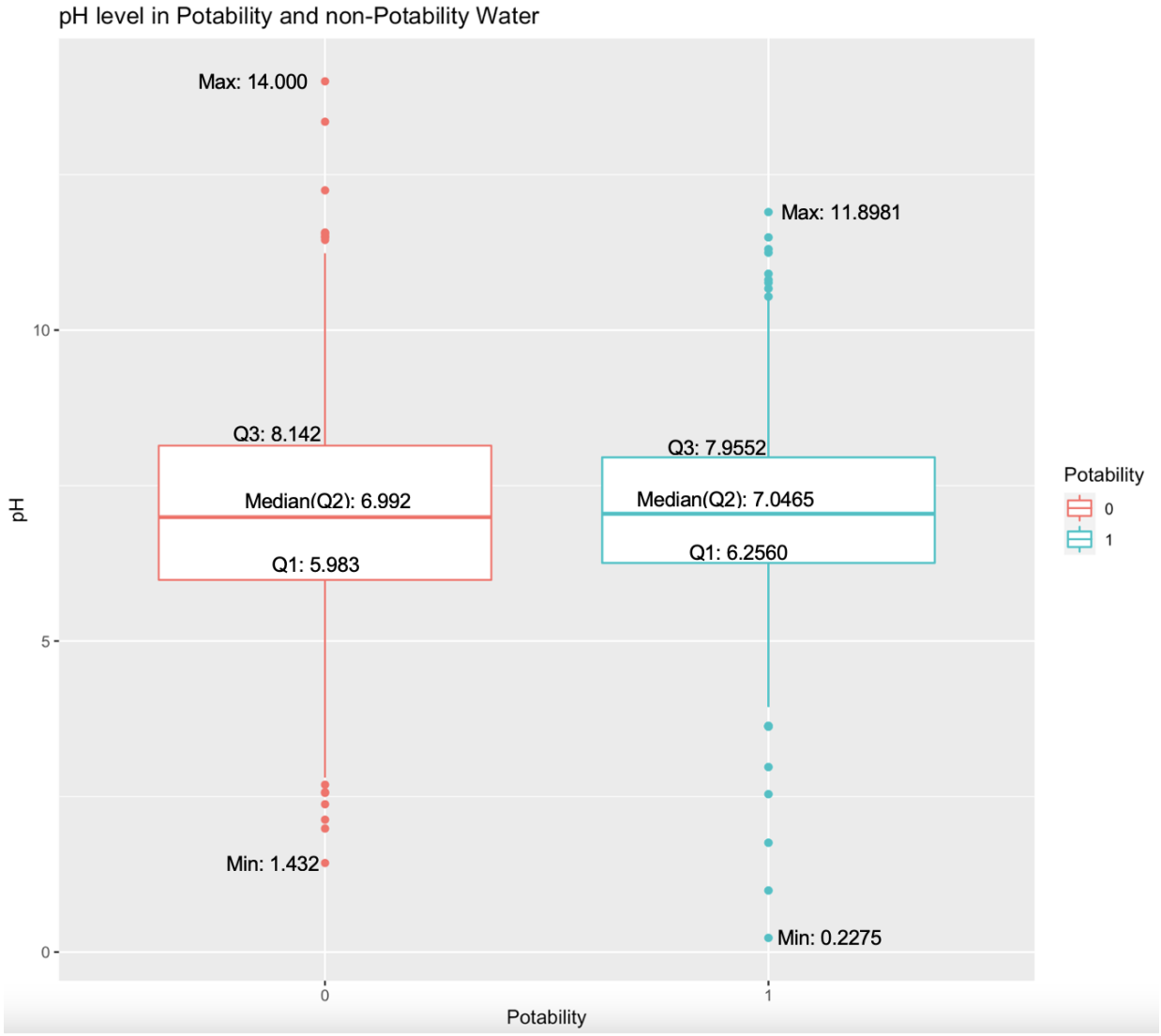
This is a histogram graph representing the conductivity in water that is affected by the presence of inorganic dissolved solids such as chloride, nitrate, sulfated, and phosphate anions. It is a measurement of the water. To pass the electrical current and a higher conductivity value means more chemicals dissolved in the water that's the reason it is important in the water dataset. Regular drinking water contains 200 to 800 µS/cm which tells us the dataset we have is safe/good for drinking water.

3. Pie chart of Potability Percentages



This pie chart graph represents the percentage of potability in the dataset divided into two categories: 0, which means the water is not potable, and 1, which is potable. From the chart, we can see that 61% of the dataset showing that the water is not potable to drink, only 39% of the dataset showing that the water is potable for drinking. The 'Cnt' represents the number of data in each category. There are 1998 data points belonging to the not potable group (group 0) corresponding to 61%, and there are 1278 data points belonging to the potable group (group 1) corresponding to 39%.

4. Boxplot of pH levels and Potability



This graph represents the distribution of pH for different potability of water. Potability 0 means the water is not potable, and 1, means the water is potable. According to the graph, there is not too much difference in the pH distribution for the 'potable' and the 'non-potable' groups. The median of the pH is the same between the two groups. The box of the 'potable' group (potability = 1) is narrower, meaning the pH distribution of this group is more condensed than non-potable water. Other than that, the pH scale for the 'potable' and 'non-potable' groups is quite similar in this dataset. It is understandable since water potability depends not only on the pH level but also on other factors such as hardness, solids, turbidity, etc.

5. Scatterplot of pH Levels and Turbidity compared to Potability

This scatter plot shows the pH levels compared to turbidity for each observation. Along with that, the color of the point shows whether or not it is potable. The blue point represents that the water is potable. The peach color point represents that the water is not potable. This graph has more peach color points than blue which means that most of the water tested for potability is not safe to drink. The higher the turbidity is the more cloudy the water is, and based on this graph most of the data points average to around 4, which would mean that the water needed to be filtered before it is safe to drink. For water to be safe to drink based on pH levels, it would be around 7 and most of the data observances are around 7 which means the pH levels are mostly neutralized.

Works Cited

Kadiwal, Aditya. “Water Quality.” *Kaggle*, 25 Apr. 2021, www.kaggle.com/adityakadiwal/water-potability.