Shawn Hwang

Project Proposal: Water Quality

Water is a vital resource necessary for all life. However, water quality is easily susceptible to a multitude of threats, such as acidity, pH, and pollution. This is why water quality prediction is so important, the sample data used in this data set can be used to predict water quality in the future. Being able to accurately predict water quality is important because it allows us time to identify threats and take precautionary measures to protect water quality in specific regions. The results of this project may be used to assist environmental agencies with their decision making about water quality. This may range anywhere from enacting new laws and regulations to simply identifying areas of focus.

Specific conductance: Specific conductance, also referred to as electrical conductivity, is the nature of water to conduct an electrical current based on salts and ions in the water. A higher specific conductance means that a pollutant may have entered the water. A higher temperature means that there is a higher specific conductance hence why we observe both.

pH: pH refers to power of hydrogen, and is a measure of how acidic or basic water is. It scores from 0-14 with 7 being neutral, 0 being basic, and 14 being acidic. The US Environmental Protection Agency (EPA) states that a pH score from 6.5-9 is good, and that any score outside of that range can lead to issues.

Dissolved Oxygen: Dissolved oxygen is a measure of how much oxygen is dissolved in water, which makes the water sustainable for life. A high dissolved oxygen level is better in water quality as it allows for life and fresh, drinkable water.

We observed a data set from the National Science Foundation that contained historical data of water quality using factors such as pH, dissolved oxygen, temperature, and specific conductance. This data set of collected samples from 36 sites, provides adequate information about pH in Georgia. We will perform EDA on this set of data in 5 different visuals while analyzing trends observed in the visuals as well as perform Ransac's linear regression, and K-NN regression.

What would we like to know: We would like to know the day to day water quality of the state of Georgia, and be able to use it to accurately build a model to predict the water quality for the coming days.

Data set: https://archive.ics.uci.edu/dataset/733/water+quality+prediction-1

Techniques:
Ransac linear regression

Explanation: Linear regression is a common tool used to create a fitted line to best predict the data set based on set variables. In a project where the goal is to predict the quality of water this is essential.

K-NN regression

Explanation: K-NN regression makes use of KNN algorithm in order to predict a line of best fit. In our case, we are using it on pH min, specific conductance, temperature mean, and dissolved oxygen mean, to predict pH max. This is because of the effects of a pH that is too high or too low.