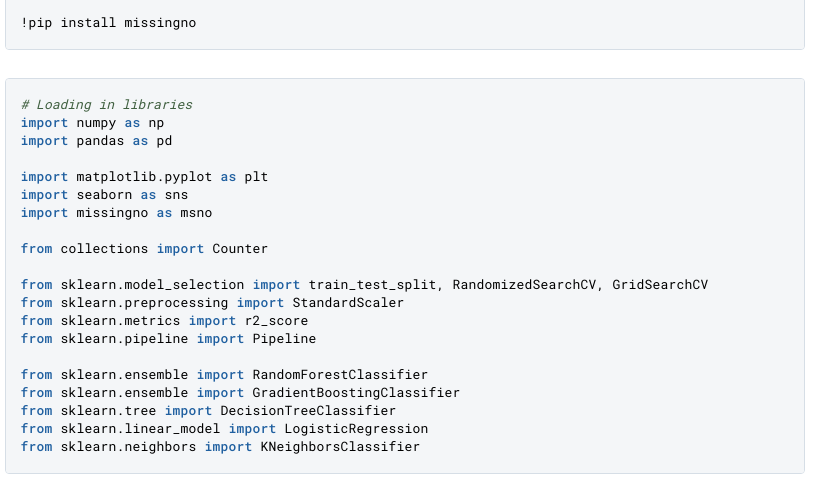
**WATER QUALITY ANALYSIS**

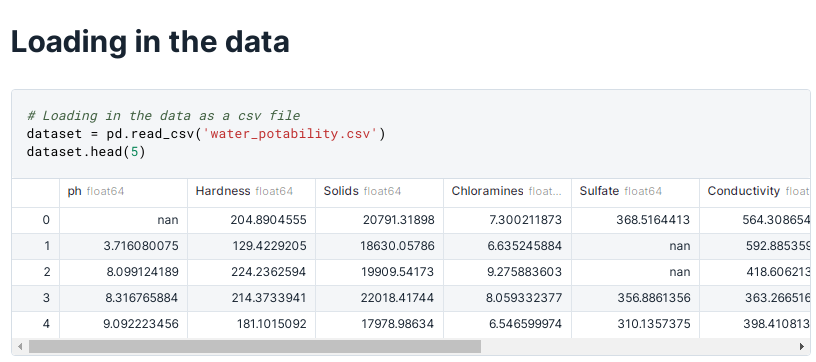
**Phase 3: Development Part 1**

**Name of the student: Hariharan**

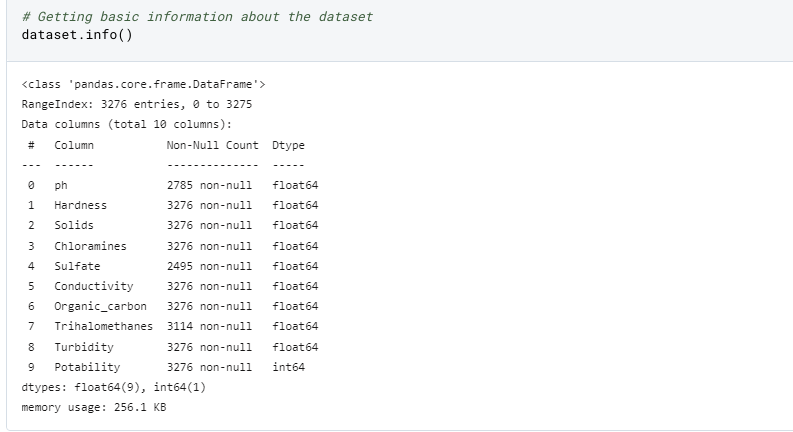
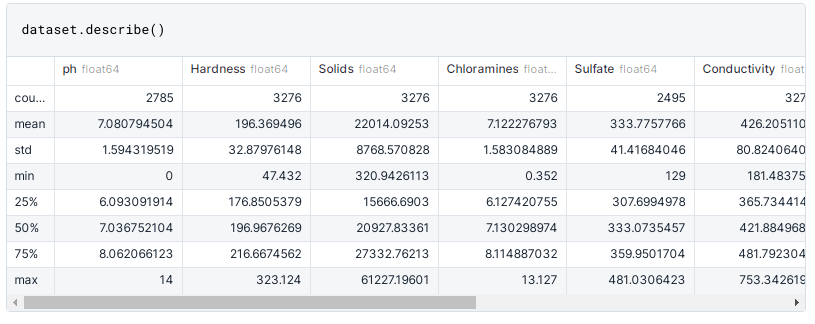
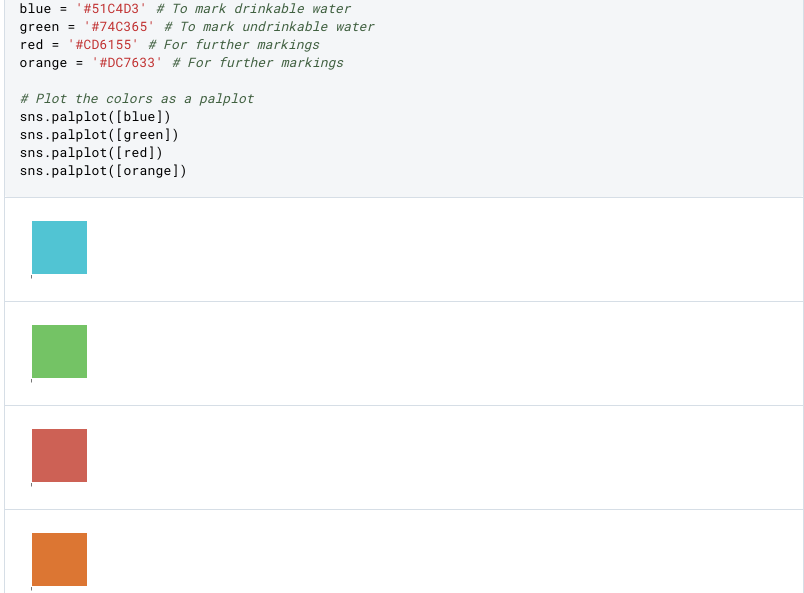
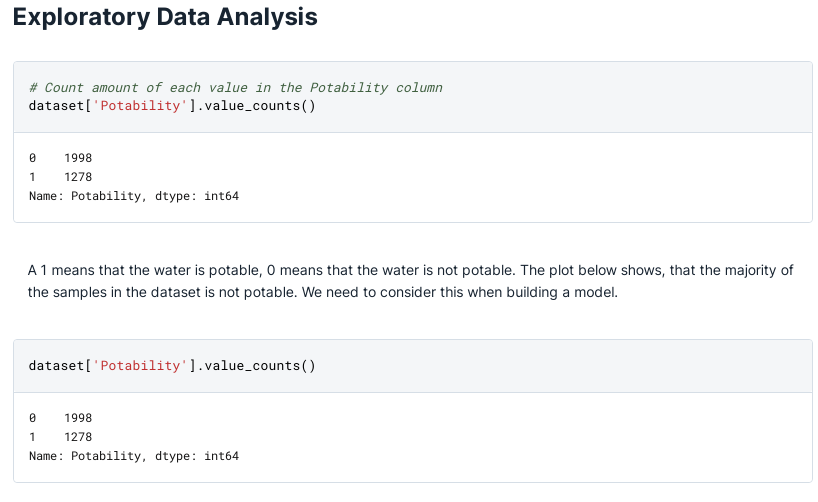
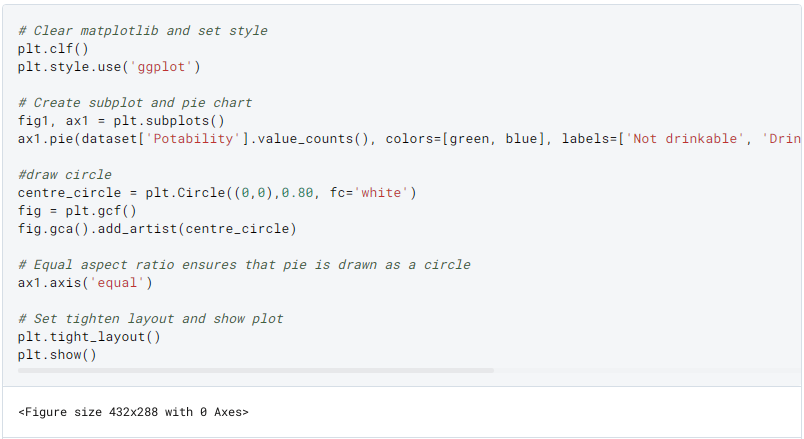
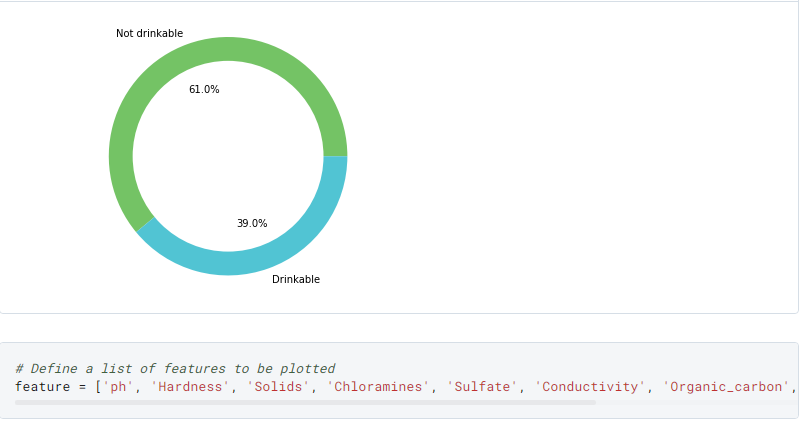
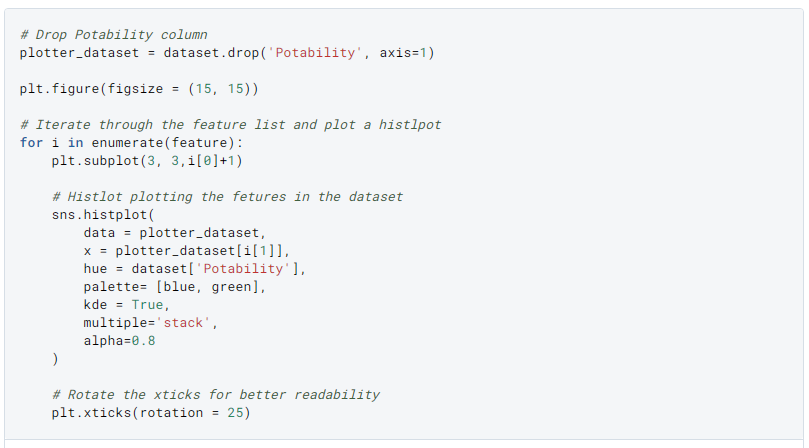
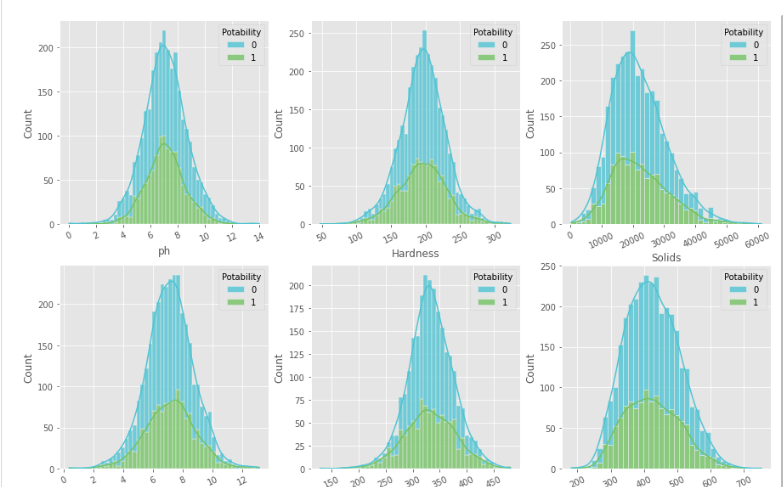
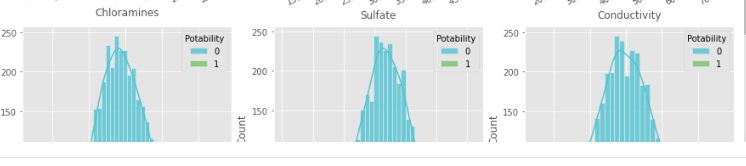
Water is the base for life as we know it. Every human needs water to survive. In many countries however, people do not have access to clean drinking water. The water quality dataset provides data on more than 3000 water samples. The data contains information about chemical components and wether or not the water ist acutally drinkable.

This notebook will determine what makes clean, drinkable water. To accomplish this, supervised machine learning models are used.

The dataset contains 9 columns in total.



* **ph**: The ph value of the water, which describes the acid-base balance of the water. A ph-value of 6.5 - 8.5 is recommended by the who.
* **Sardness**: The more calcium and magnesium the water contains, the harder the water is. Though these minerals are not harmful to consume, they could have an impact on the potability of the water.
* **Solids (Total dissolved solids - TDS)**: Measurement of how many organic and inorgnaic materials are contained in the water. Desirable limit for TDS is 500 mg/l, maximum limit is 1000 mg/l.
* **Chloramines**: Chlorine and chloramine are a common disinfectant. Chlorine of 4 mg/L or 4 parts per million (ppm) are considered save.
* **Sulfate**: Naturally occuring mineral, that is much more higher in seawater than in feshwater.
* **Conductivity**: Measurement of how conductivce the water is, meaning how well energy flows through it. According to WHO standards, the electric conductivity (EC) should not be higher than 400 μS/cm.
* **Organic carbon**: Total Organic Carbon (TOC) is the result of decaying organic matter in water. According to US EPA < 2 mg/L of TOC is considered drinkable water.
* **Trihalomethanes**: A chemical that occurs in water treated with chlorine. Levels up to 80 ppm are considered safe.
* **Turbidity**: Depends on the amount of solid matter in the water. The WHO recommends a value of 5.00 NTU.
* **Potability**: States wether water is safe. 0 = not safe to drink, 1 = safe to drink.

The colors below are going to be used throughout the entire notebook. Blue will show dinkable, or potable, water. The muddy green color will be used in water that is not save for consumption.        The WHO recommends ph-values of drinking water to be within the range of 6.5 and 9.5. My homecountry Germany also mandates that tap water should have a ph value between 6.5 and 9.5. With that in mind, let's see how many of our samples are within that range.

In a guideline for drinking water published by the WHO, the ph-value of water alone is not sufficient when it come to the potability of water. However, the ph-value is very important to ensure water disinfection and clarification. Therefore, pipes made out of copper or certain steels should not be used to transports water with very low ph-values, because the water is too acidic and would dissolve some of the metal in the pipe, making the water unsave due to the high amounts of metal in the water. The ph-value could therefore be correlated to other quality factors of the water.

## As the EU considers everything below 250 mg/L of Sulfate as save, the majority of the samples would not be considered Potable. The WHO however would see all of the samples as just fine, because they set the limit at 500 mg/L.

**DATA PREPAREATION: Dealing with missing data**

The ph, Sulfate and Trihalomethanes columnes contain a lot of missing values. Dropping all of these values would mean that we lose a considerable amount of data. Instead, we will look into replacing missing values with the mean of median.

### Median and mean are very similar. Missing values will be replaces by the median value. Checking for correlations

## Building a model

**Randomized, cross validated search**

Time to build a machine learning model. To determine what makes, potable water, five different machine learning algortihms are going to be used. To get (relatively) good results, a cross validated random search is used on every model.

The random search randomly seelcts parameters from a predefined grid and to find the best results within these parameters. Computing the results of all the possible paramters would be computationally expensive. Doing a random search saves both time and computational power. After all models have gone through a randomized search, the best model will be selected. To get the absolute best results from that model, a grid search is performed, where all possible cominations of parmateres are tested to find out the best parameters.

