Water Quality

EDA and Prediction

By: Omnia Elmenshawy
AI Engineering undergraduate student
Supervisor: Dr. Doaa Mohammed

Data science challenges
Reveal data secrets

Agenda:

- Definition and Task Statement
- Data Report
- Objective Data problems and Cleaning
- Exploratory Data Analysis
- Modelling and Accuracy
- Improvements

Data science challenges
Reveal data secrets

What is Water Potability?

Clean water is an essential element for quality life, no wonder why “Clean water and Sanitation” is the 6th Goal of the UN 17 Sustainable development goals

Data science challenges
Reveal data secrets

Why Potability?

It is our target, and it Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

Task Statement:

EDA for Water potability
Predict if water is safe for human consumption

Data science challenges
Reveal data secrets

Data Statistics

	count	mean	std	min	25%	50%	75%	max
ph	2785.000000	7.080795	1.594320	0.000000	6.093092	7.036752	8.062066	14.000000
Hardness	3276.000000	196.369496	32.879761	47.432000	176.850538	196.967627	216.667456	323.124000
Solids	3276.000000	22014.092526	8768.570828	320.942611	15666.690297	20927.833607	27332.762127	61227.196008
Chloramines	3276.000000	7.122277	1.583085	0.352000	6.127421	7.130299	8.114887	13.127000
Sulfate	2495.000000	333.775777	41.416840	129.000000	307.699498	333.073546	359.950170	481.030642
Conductivity	3276.000000	426.205111	80.824064	181.483754	365.734414	421.884968	481.792304	753.342620
Organic_carbon	3276.000000	14.284970	3.308162	2.200000	12.065801	14.218338	16.557652	28.300000
Trihalomethanes	3114.000000	66.396293	16.175008	0.738000	55.844536	66.622485	77.337473	124.000000
Turbidity	3276.000000	3.966786	0.780382	1.450000	3.439711	3.955028	4.500320	6.739000
Potability	3276.000000	0.390110	0.487849	0.000000	0.000000	0.000000	1.000000	1.000000

Potable water data Statistics:

	count	mean	std	min	25%	50%	75%	max
ph	1101.000000	7.073783	1.448048	0.227499	6.179312	7.036752	7.933068	13.175402
Hardness	1278.000000	195.800744	35.547041	47.432000	174.330531	196.632907	218.003420	323.124000
Solids	1278.000000	22383.991018	9101.010208	728.750830	15668.985035	21199.386614	27973.236446	56488.672413
Chloramines	1278.000000	7.169338	1.702988	0.352000	6.094134	7.215163	8.199261	13.127000
Sulfate	985.000000	332.566990	47.692818	129.000000	300.763772	331.838167	365.941346	481.030642
Conductivity	1278.000000	425.383800	82.048446	201.619737	360.939023	420.712729	484.155911	695.369528
Organic_carbon	1278.000000	14.160893	3.263907	2.200000	12.033897	14.162809	16.356245	23.604298
Trihalomethanes	1223.000000	66.539684	16.327419	8.175876	56.014249	66.678214	77.380975	124.000000
Turbidity	1278.000000	3.968328	0.780842	1.492207	3.430909	3.958576	4.509569	6.494249
Potability	1278.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Report

Overview

Dataset Statistics

Number of Variables	10
Number of Rows	3276
Missing Cells	1434
Missing Cells (%)	4.4%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	256.1 KB
Average Row Size in Memory	80.0 B
Variable Types	Numerical: 9 Categorical: 1

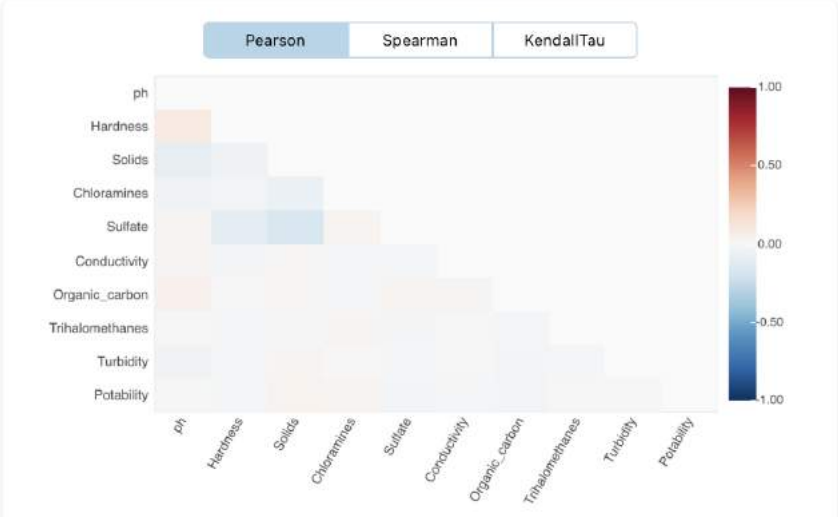
Dataset Insights

`ph` has 491 (14.99%) missing values	Missing
`Sulf…` has 781 (23.84%) missing values	Missing
`Trih…` has 162 (4.95%) missing values	Missing
`Pota…` has constant length 1	Constant Length

Dataset Insights

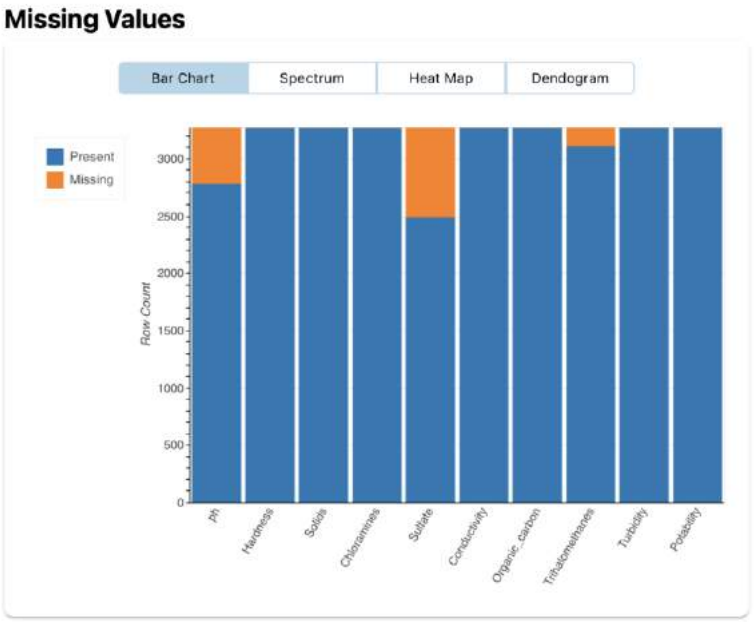
`ph` is skewed	Skewed
`Sulfate` is skewed	Skewed
`Potability` has constant length 1	Constant Length

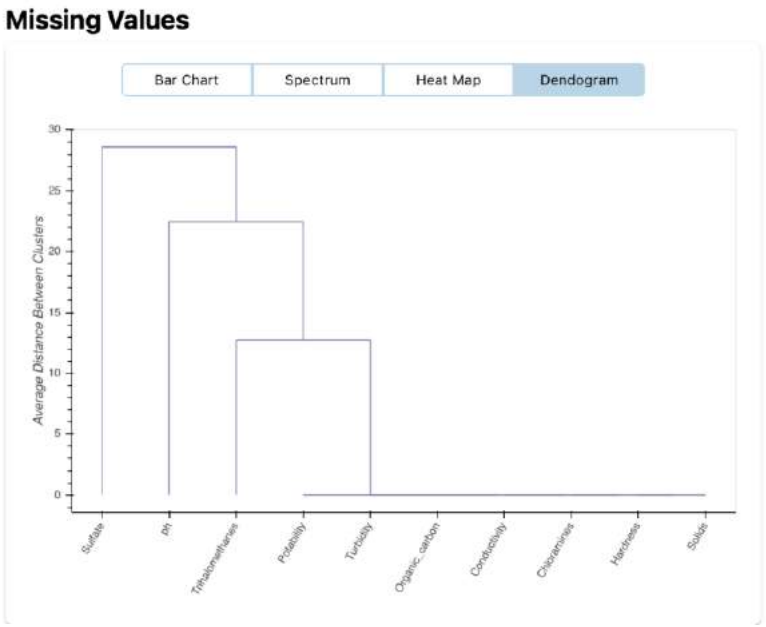
Correlations

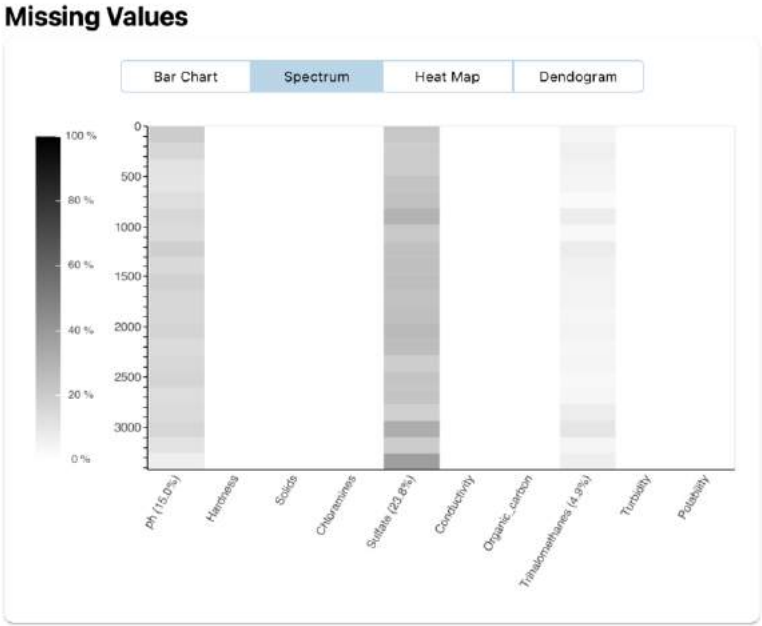


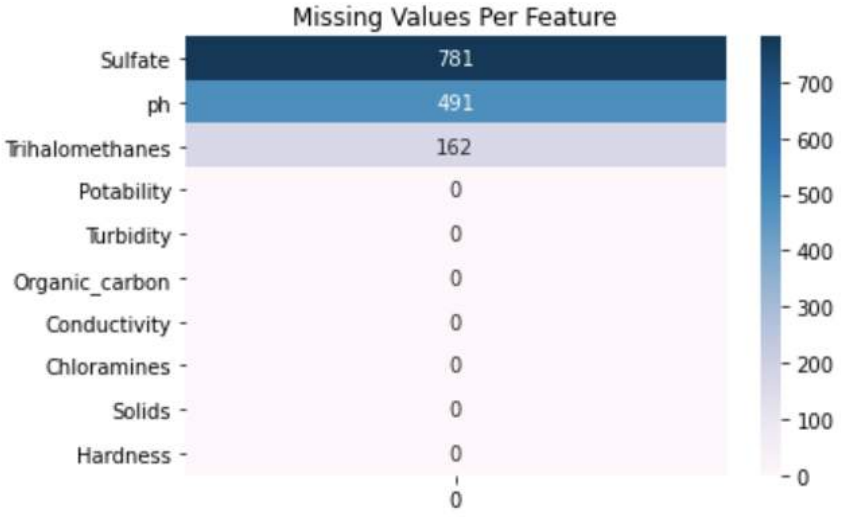
Data science challenges
Reveal data secrets

Missing Values:





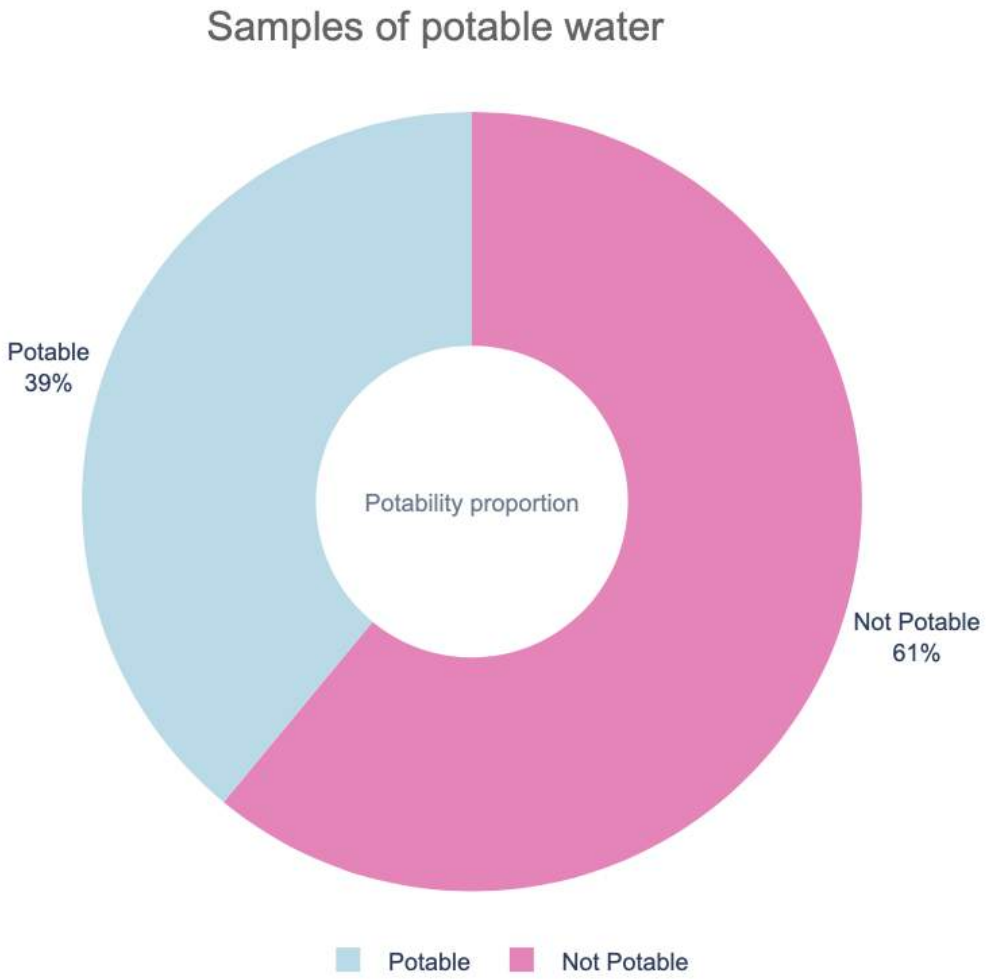




The dendrogram view shows how missing values are related across columns by using hierarchical clustering.

Missing Imputation Using KNN Imputer vs. Filling by the Meadian

Data science challenges
Reveal data secrets



Need to resample the data
to get a balanced dataset

39% of the data is potable and 61% is not potable

The Statistics shows that the data is imbalanced
So, we will fix it with SMOTE while Modelling.

Problem Statement:
More than 80% of the data which classified as Potable doesn't match the global standards for Potability.

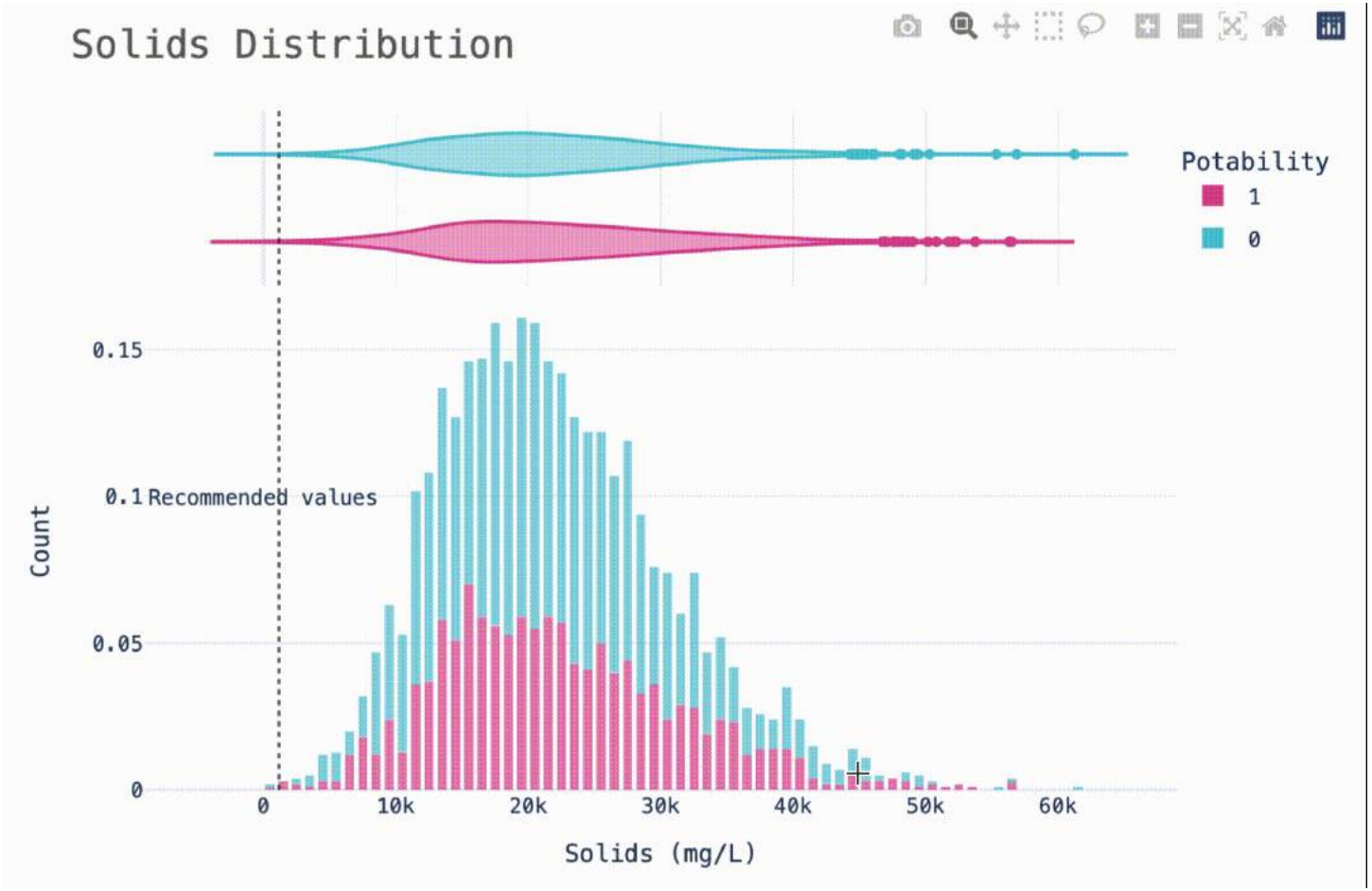
Data science challenges
Reveal data secrets

Feature Exploration and Feature Problems

1- Solids (Total dissolved solids - TDS):

Guideline: Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

Observation: The water samples which are above the acceptable limit are more than 98% of the data, which leaves us considering the majority of the potable data already unpotable, and this is one of our business problems which we are discovering.

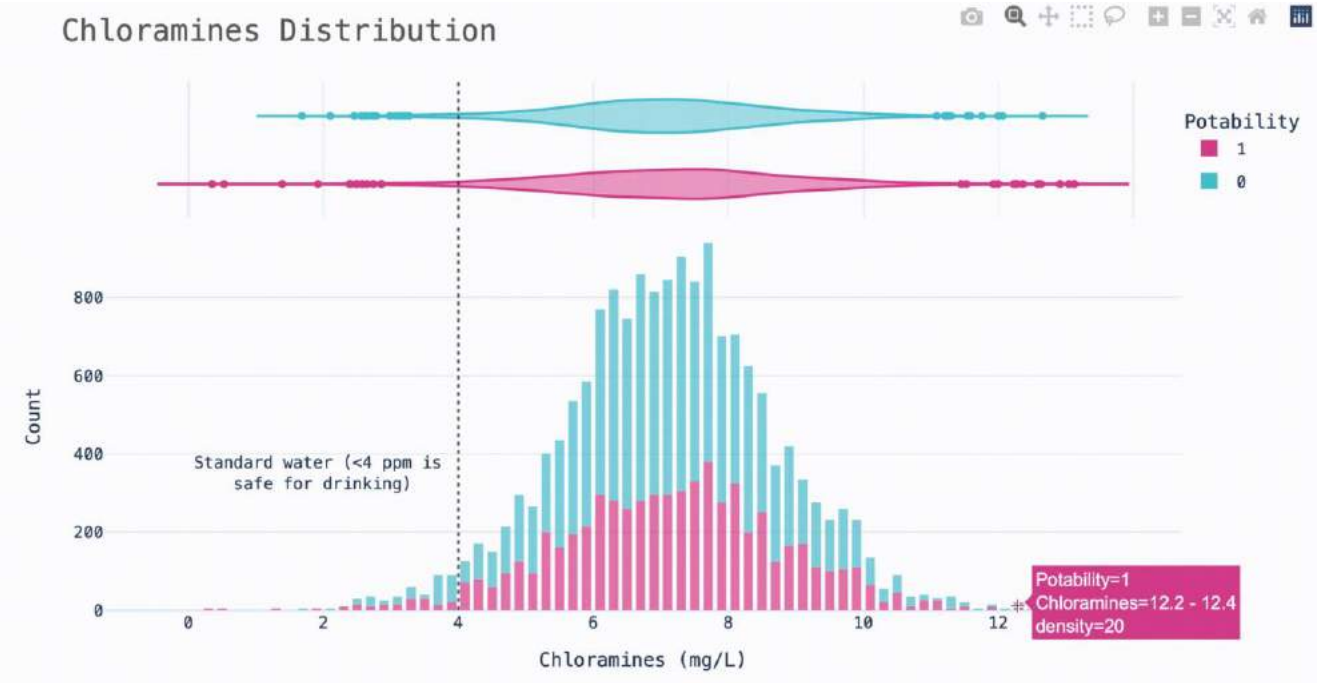


Data science challenges
Reveal data secrets

2- Chloramines:

Guideline: Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

Observation: The potable water samples with high level of Chloramines are more than 75% of the data, which again leaves us considering the majority of the potable data already unpotable, and this is one of our business problems which we are discovering.

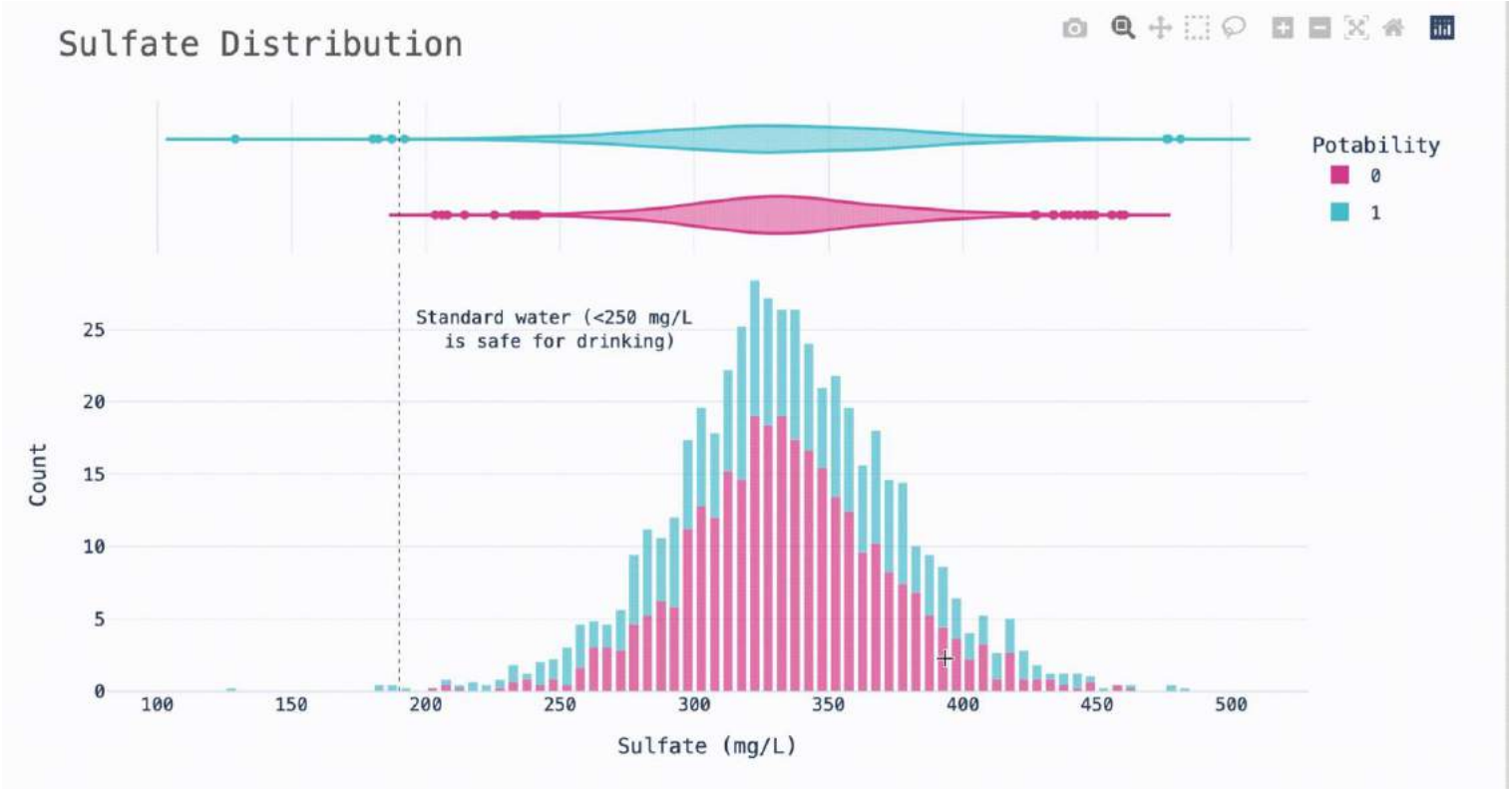


Data science challenges
Reveal data secrets

3- Sulfates:

Guideline: WHO: It is generally considered that taste impairment is minimal at levels below 250 mg/l. No health-based guideline value has been derived for sulfate.

Observation: Only about 1.8% of the water samples were safe in terms of Sulfate levels. Which again leaves us considering the majority of the potable data already unpotable, and this is one of our business problems which we are discovering.

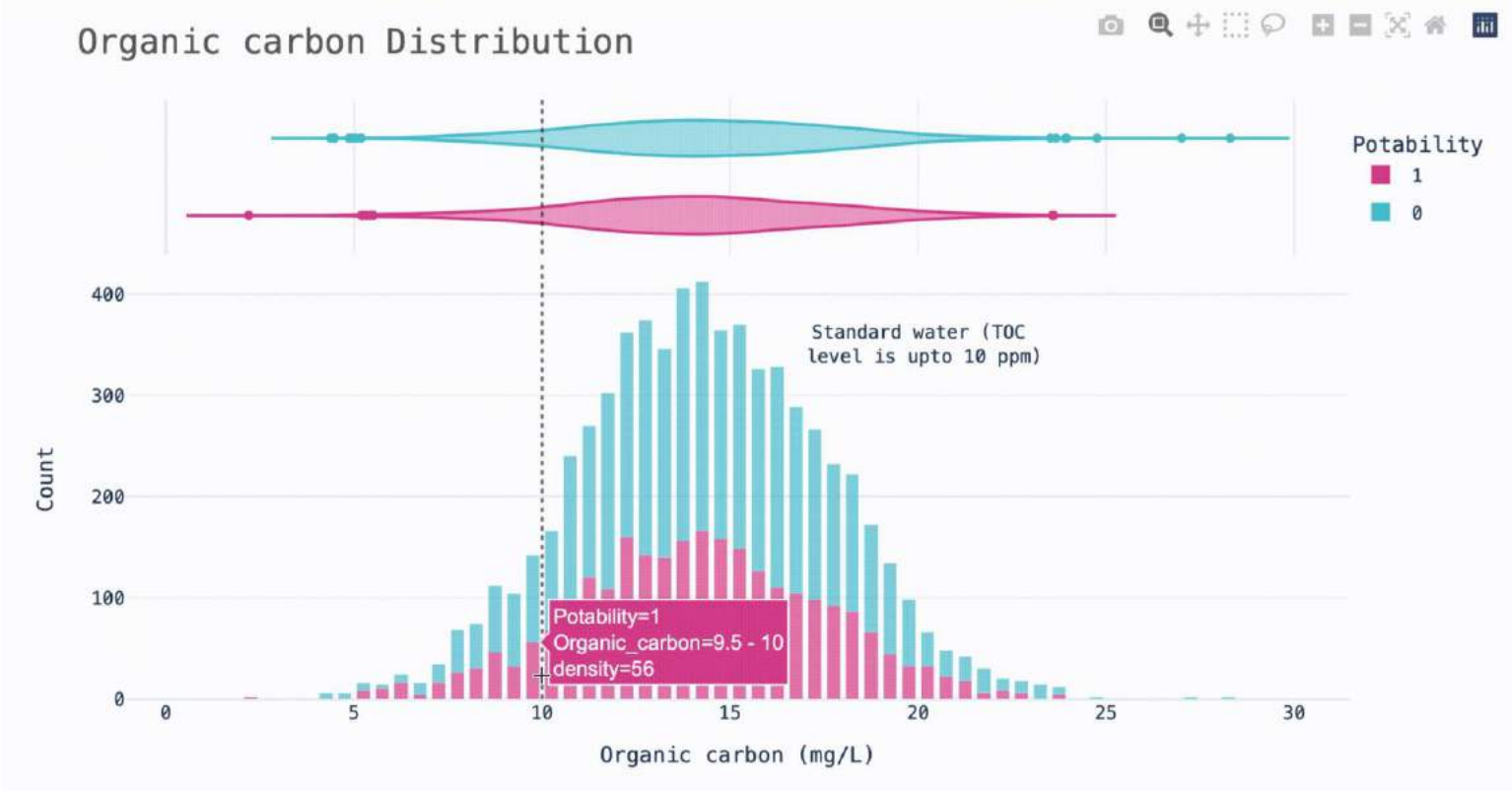


Data science challenges
Reveal data secrets

4-Total Organic Carbon (TOC):

Guideline: 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

Observation: Some Samples (above 65%) of the data classified as potable while they are under the US EPA standards for potable water.

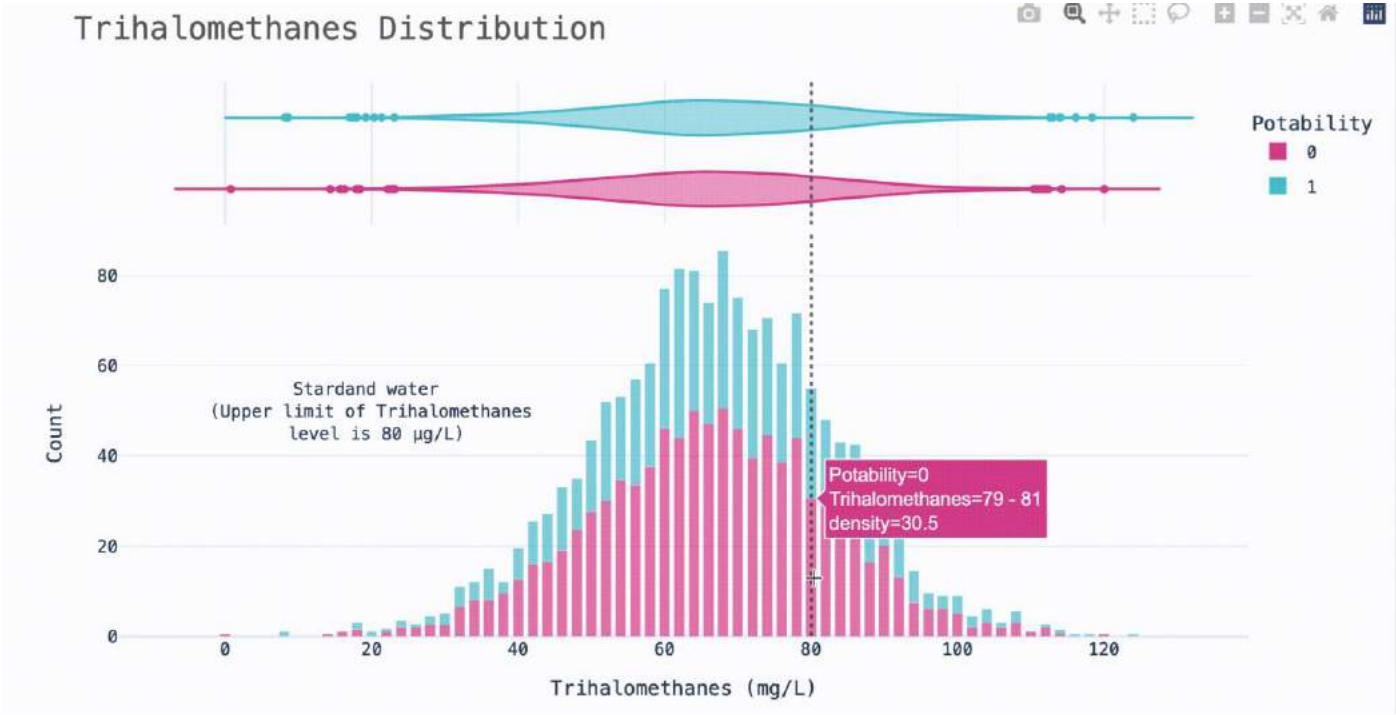


Data science challenges
Reveal data secrets

5- Trihalomethanes:

Guideline: - THM levels up to 80 ppm is considered safe in drinking water.

Observation: Some Samples (about 20% of the data) classified as potable while they exceeded the 80 ppm levels of potable water.

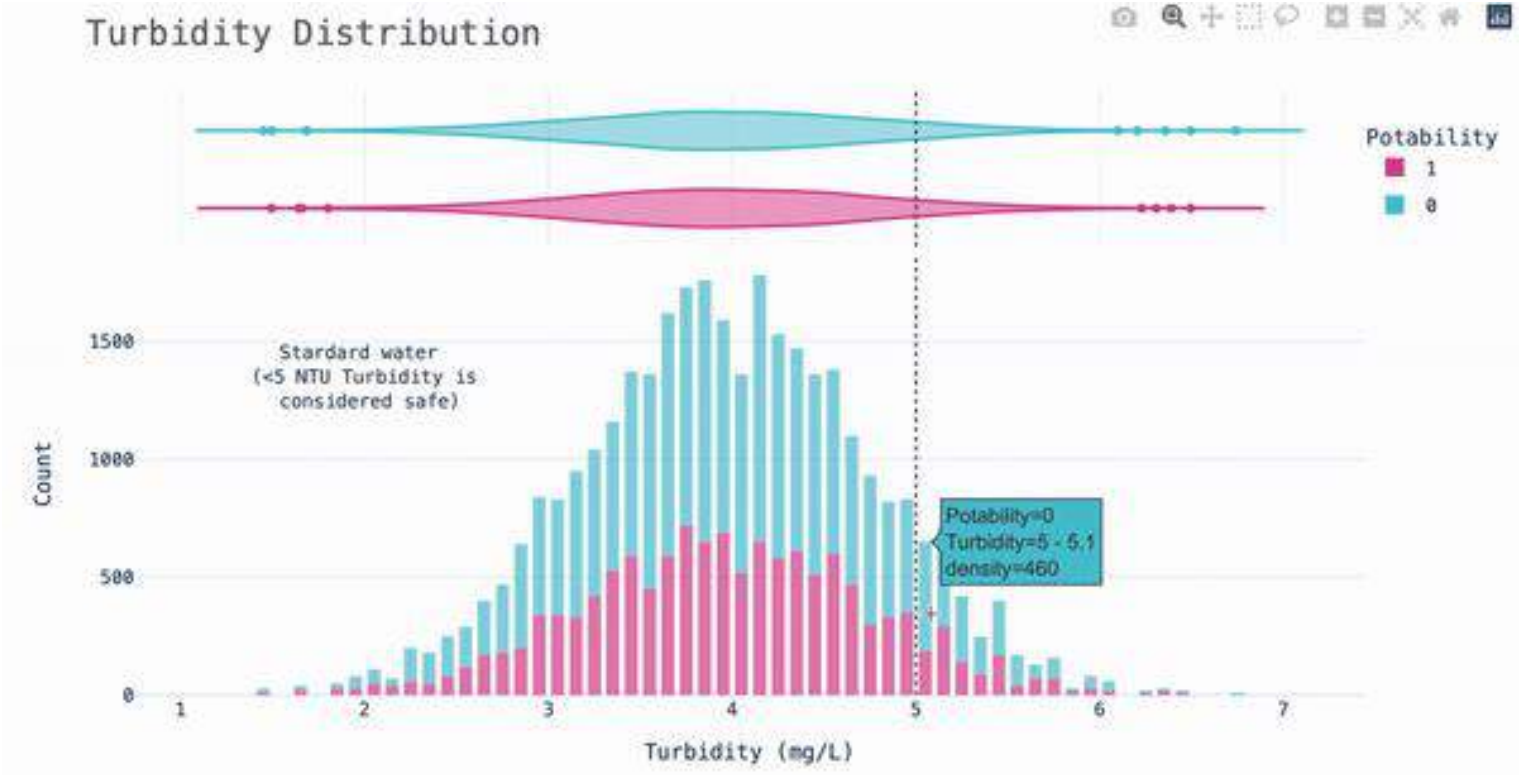


Data science challenges
Reveal data secrets

6- Turbidity:

Guidelines: WHO: recommended value below 5.00 NTU, ideally below 1 NTU

Observation: Some Samples (about 10% of the data) classified as potable while they exceeded the 5.00 NTU levels of potable water.

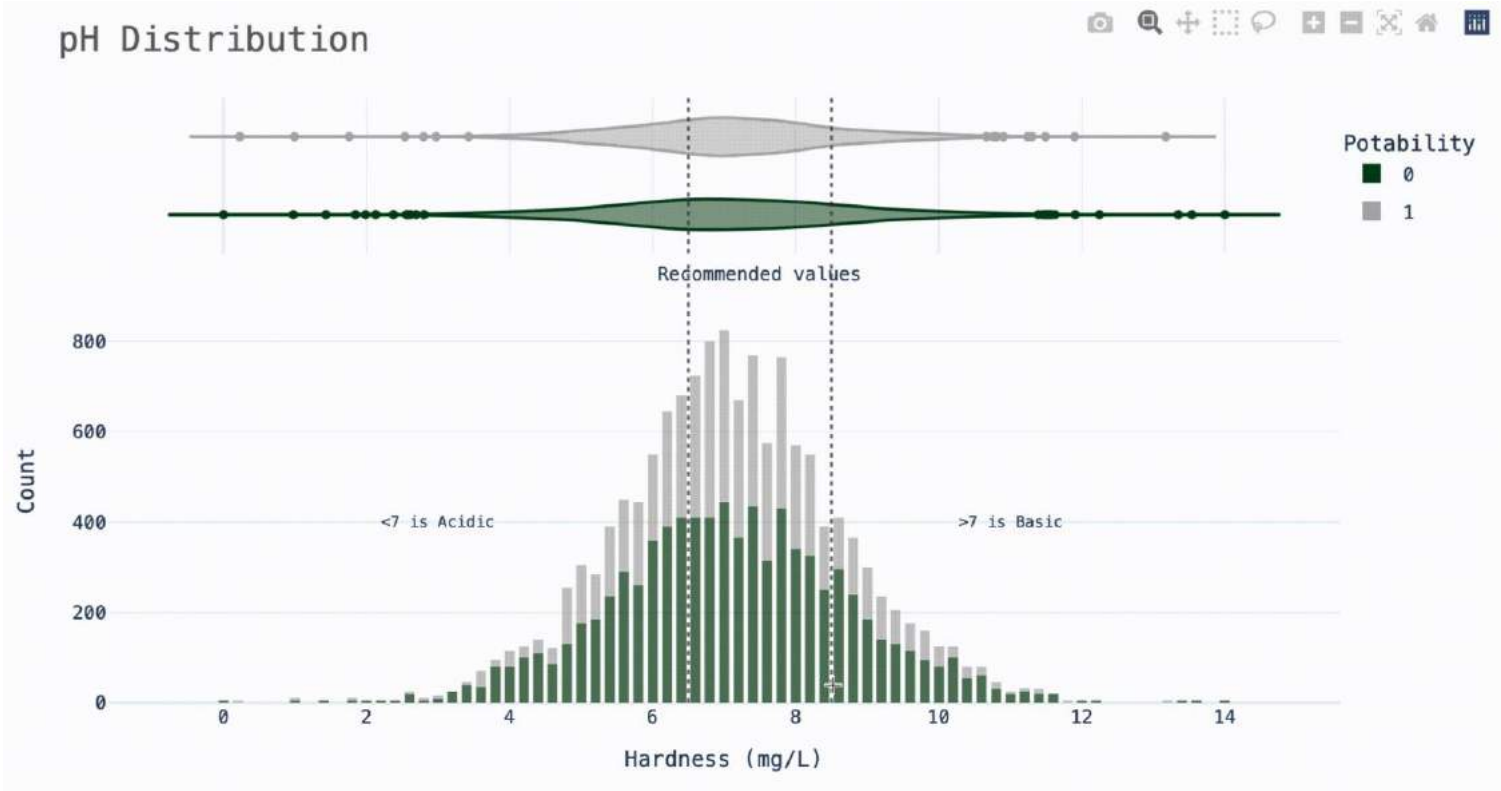


Data science challenges
Reveal data secrets

7- pH Value:

WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

Observation: Some Samples are outside the recommended limits, which may contain many harmful metals which is in the end against the potability of the water.

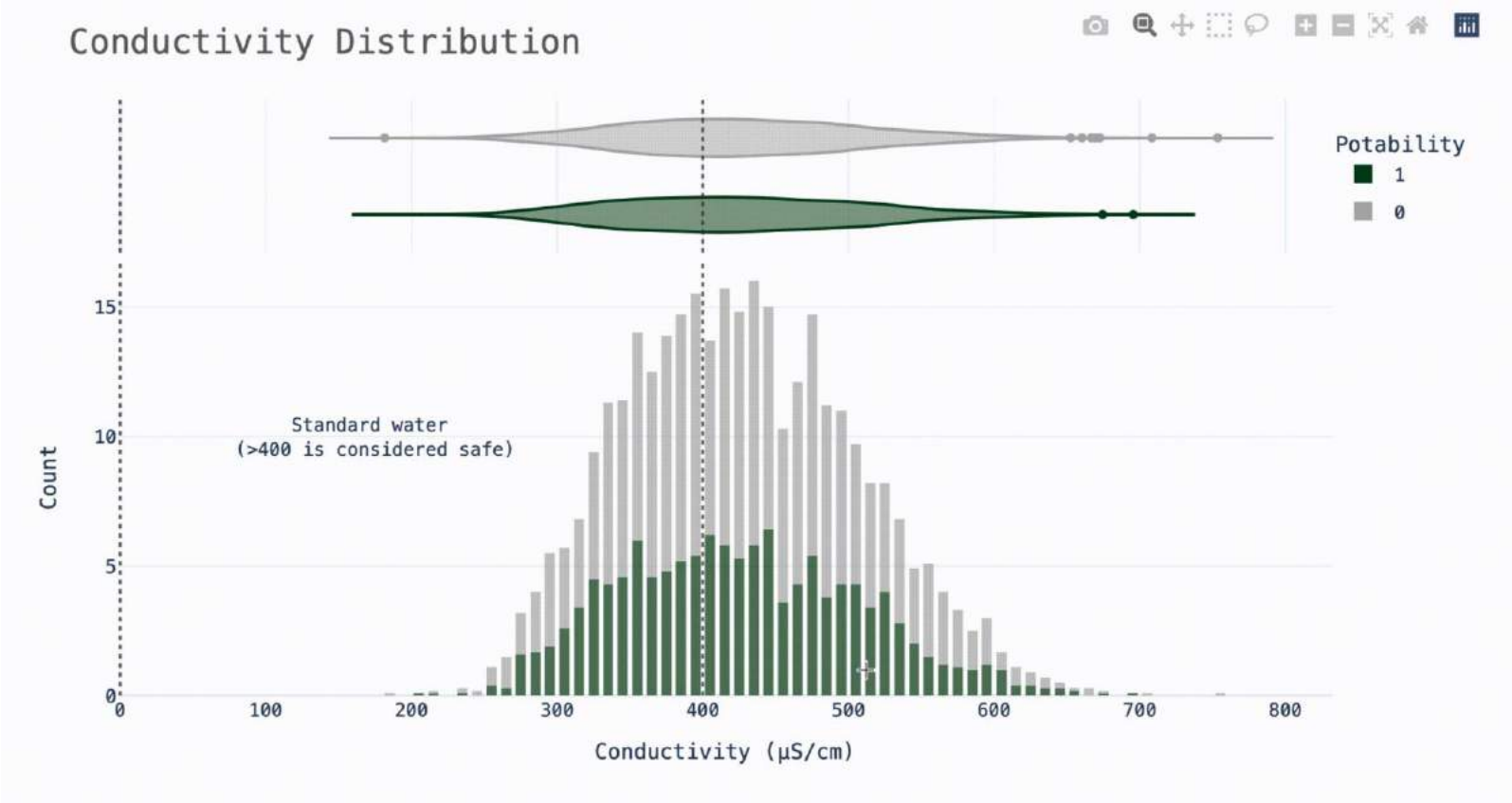


Data science challenges
Reveal data secrets

8- Conductivity:

According to WHO standards, EC value should not exceeded 400 µS/cm.

Observation: Almost half of the Samples exceeded 400 µS/cm. and still classifies as potable.

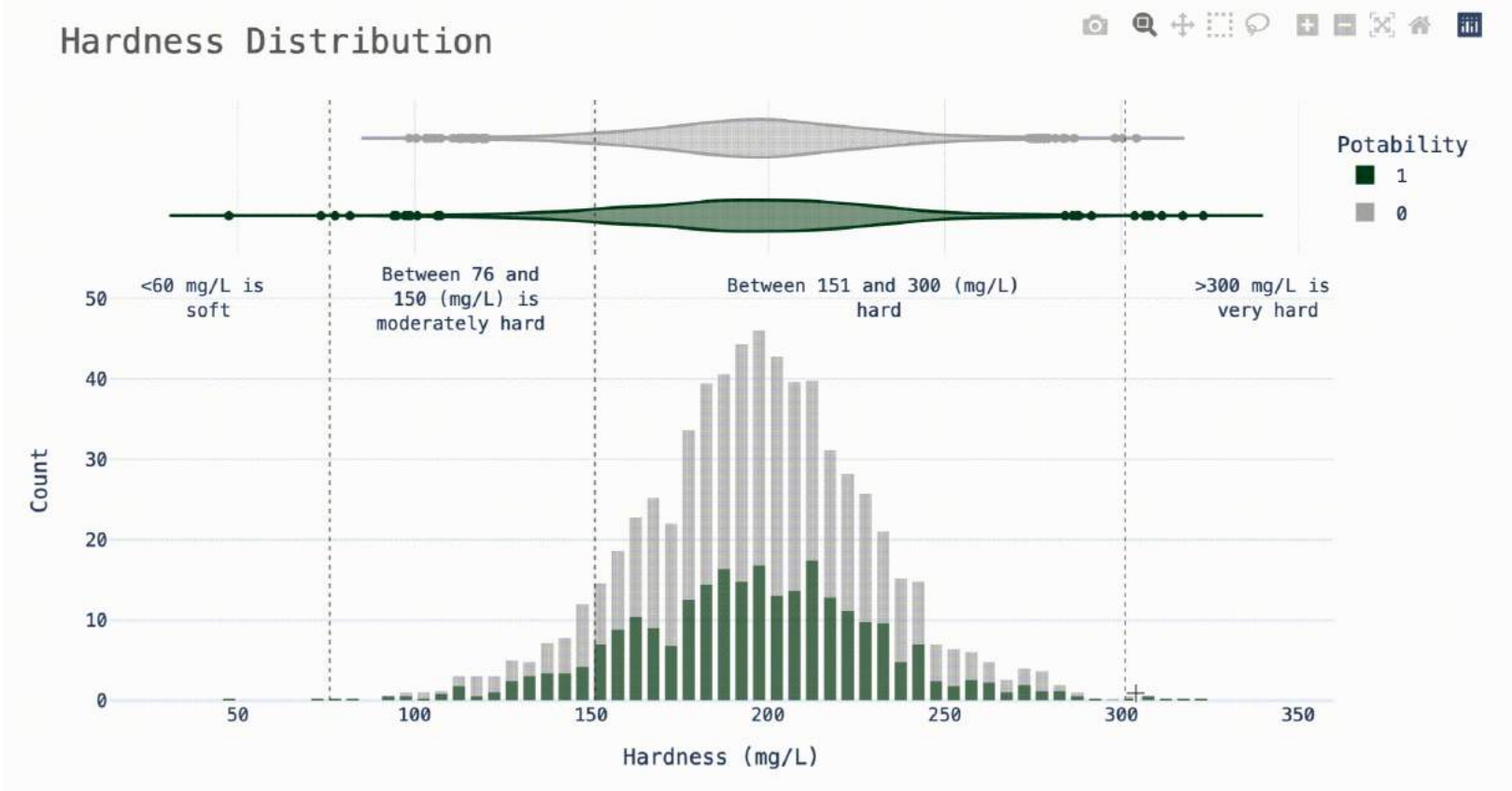


Data science challenges
Reveal data secrets

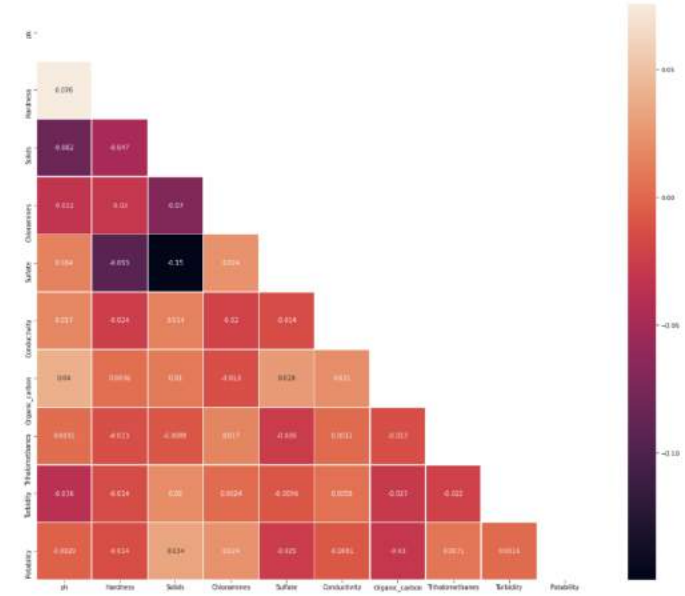
9- Hardness:

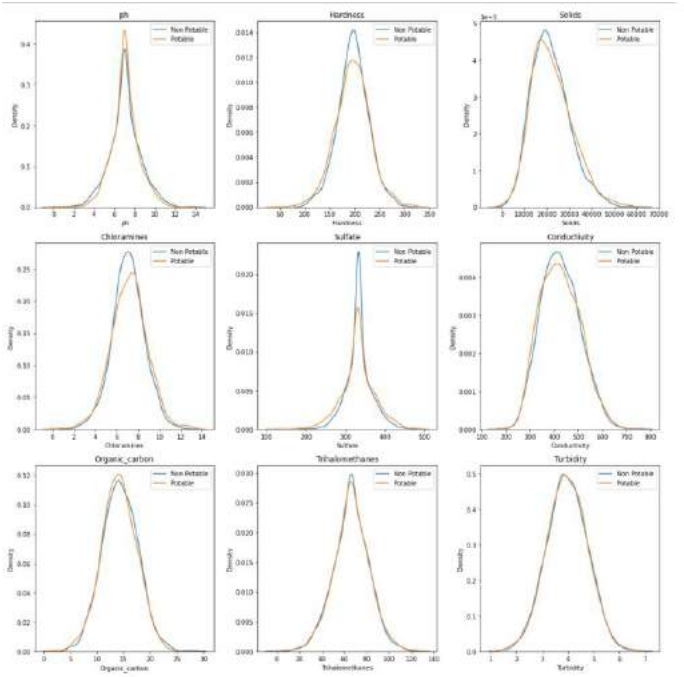
In some instances, consumers tolerate water hardness in excess of 500 mg/l.

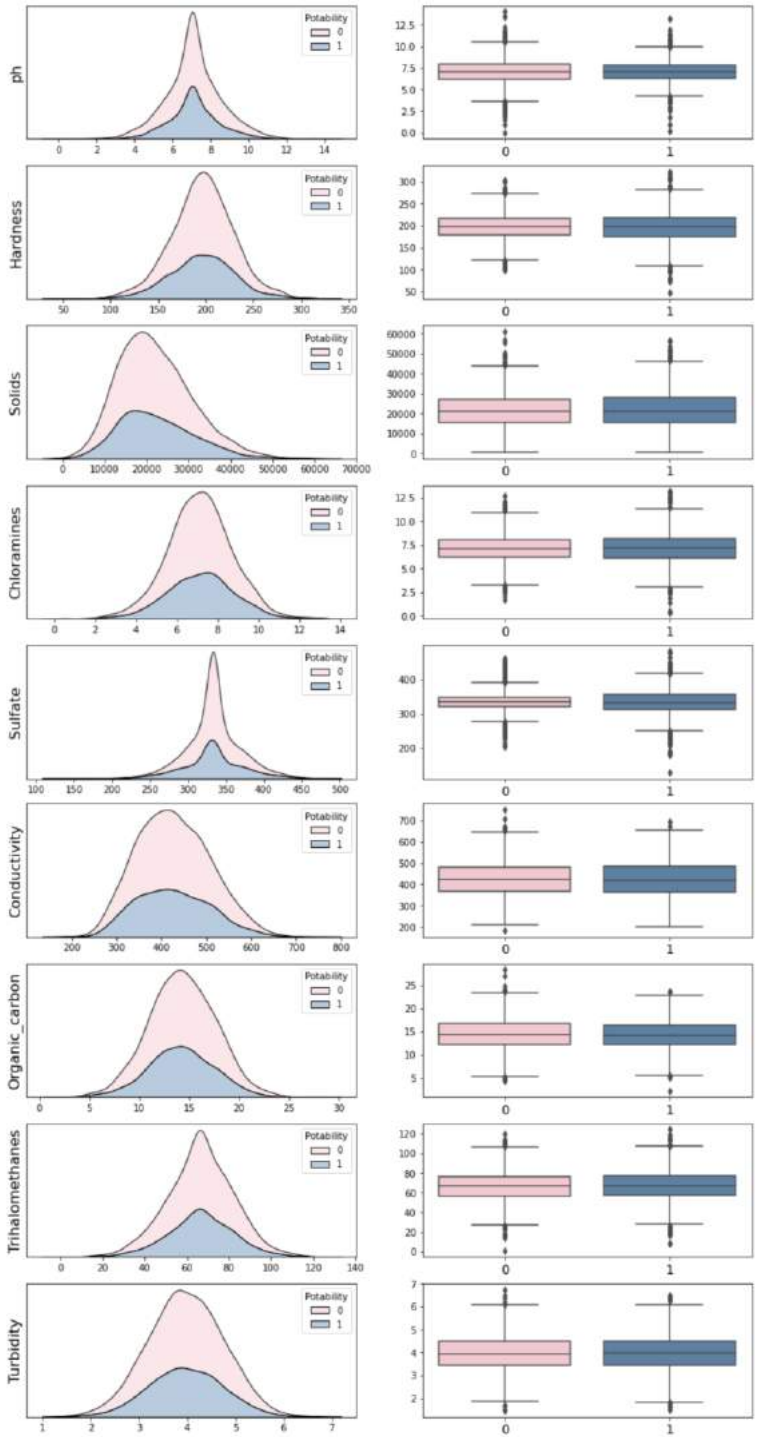
Observation: All of the data samples are in the acceptable range.



Data science challenges
Reveal data secrets







Correlation and Outliers:

Observation:

No linear relationship between the features because we have binary label and continuous features. So, Linear model may not work on this case.

The Boxplot and density distribution of different features by Potability show that the difference in mean values aren't big.

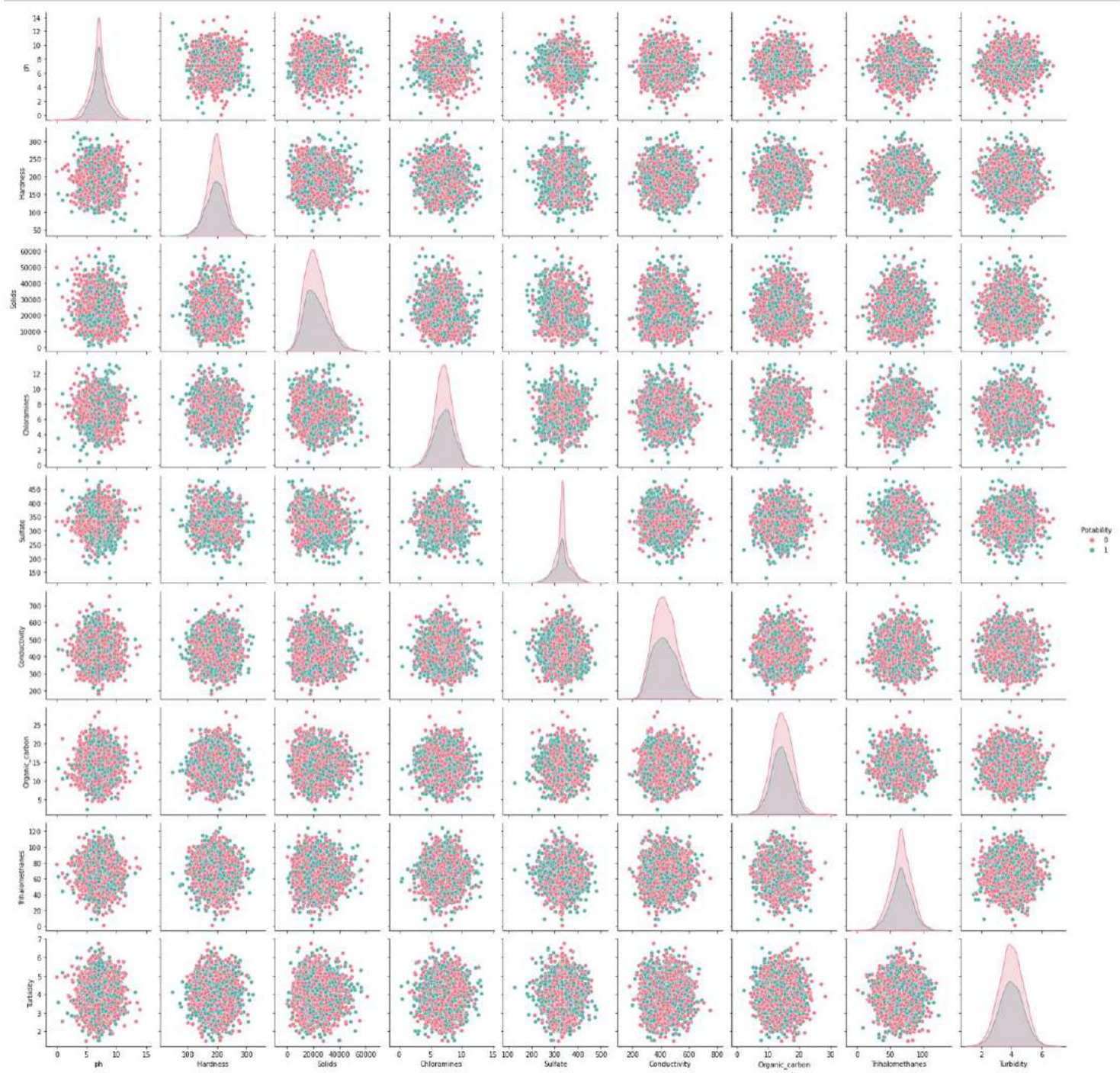
LOCAL FACTOR OUTLIER METHOD

There are some differences in the feature distribution among the potability, so we could get use of these differences while modelling. (ex: Sulfate, hardness, and chloramines have big impact on the water health).

Stats	Pearson	Spearman
	Pearson	**Spearman**
Highest Positive Correlation	0.076	0.105
Highest Negative Correlation	-0.15	-0.129
Lowest Correlation	0.001	0.001
Mean Correlation	-0.009	-0.008

Data science challenges

Reveal data secrets



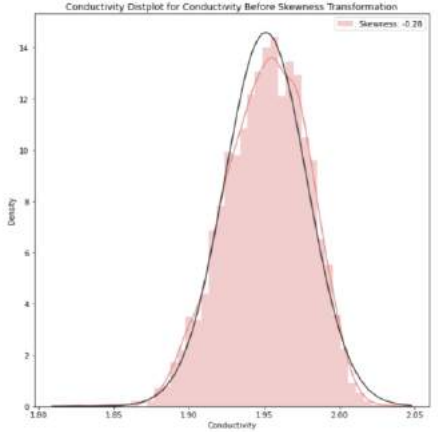
Skewness

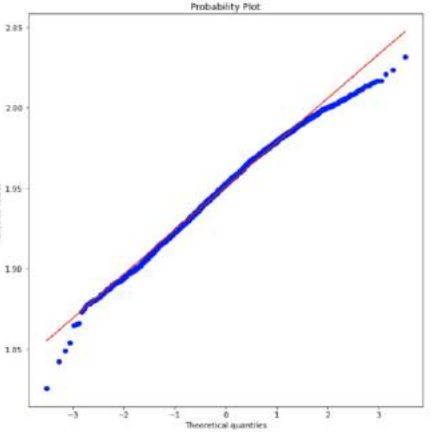
Skewness in numerical features:

	Skewness
Conductivity	-0.277978
Turbidity	-0.730482
Sulfate	-0.894808
Hardness	-1.024578
Organic_carbon	-1.108195
Solids	-1.593006
Chloramines	-1.908682
ph	-3.220965
Trihalomethanes	-4.283803

Skewness Before Transformation for Conductivity: -0.27810496

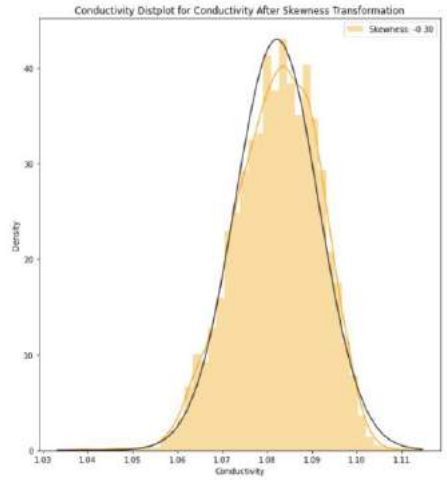
Mean before Transformation for Conductivity : 1.9511098861694336, Standard Deviation before Transformation for Conductivity : 0.02735399827361107

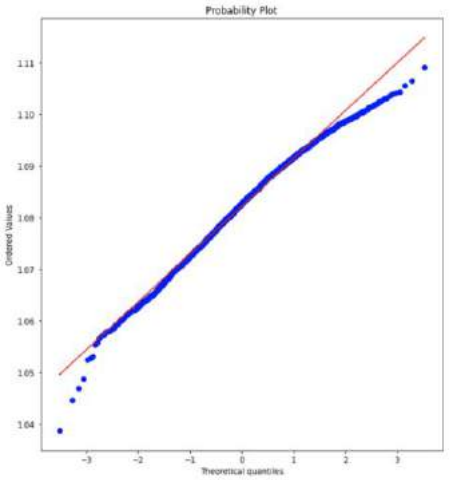




Skewness After Transformation for Conductivity: -0.30267483

Mean before Transformation for Conductivity : 1.0821382999420166, Standard Deviation before Transformation for Conductivity : 0.009281928651034832

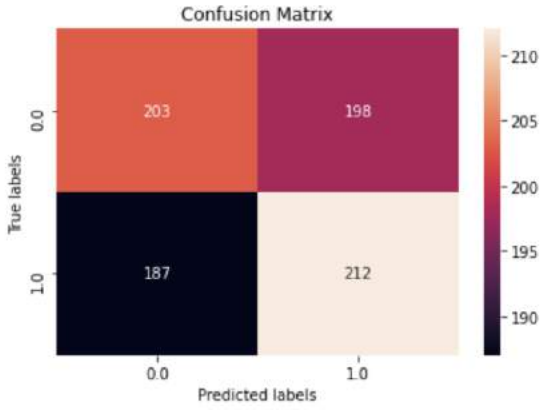




Choosing the best Model:

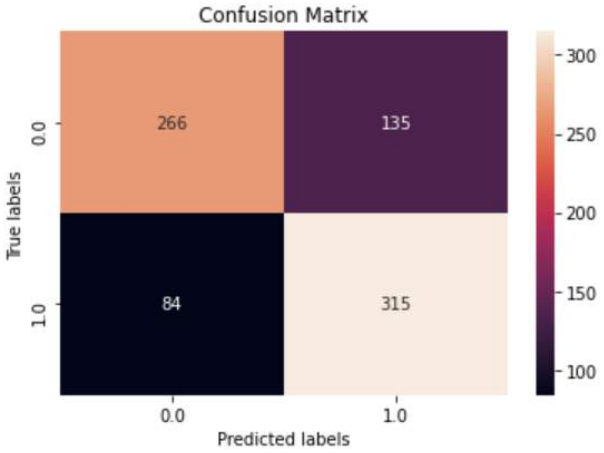
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.7677	0.8715	0.7622	0.6732	0.7140	0.5199	0.5238	0.5760
lightgbm	Light Gradient Boosting Machine	0.7508	0.8564	0.7213	0.6564	0.6870	0.4807	0.4825	0.7140
gbc	Gradient Boosting Classifier	0.7339	0.8550	0.8032	0.6154	0.6964	0.4670	0.4807	0.0880
xgboost	Extreme Gradient Boosting	0.7557	0.8526	0.7201	0.6651	0.6912	0.4896	0.4910	0.3630
rf	Random Forest Classifier	0.7579	0.8524	0.6931	0.6787	0.6851	0.4886	0.4894	0.0940
ada	Ada Boost Classifier	0.7001	0.8067	0.8045	0.5768	0.6712	0.4095	0.4294	0.0350
et	Extra Trees Classifier	0.6854	0.7212	0.4298	0.6263	0.5092	0.2894	0.3008	0.0900
dt	Decision Tree Classifier	0.7085	0.6994	0.6616	0.6074	0.6330	0.3921	0.3933	0.0100
qda	Quadratic Discriminant Analysis	0.6164	0.6726	0.6147	0.4976	0.5492	0.2222	0.2265	0.0070
knn	K Neighbors Classifier	0.5746	0.6021	0.5632	0.4522	0.5012	0.1381	0.1409	0.0150
nb	Naive Bayes	0.5261	0.5610	0.5644	0.4101	0.4743	0.0624	0.0652	0.0060
lr	Logistic Regression	0.5087	0.5069	0.4956	0.3857	0.4322	0.0115	0.0121	0.0080
lda	Linear Discriminant Analysis	0.5087	0.5069	0.4956	0.3857	0.4322	0.0115	0.0121	0.0090
svm	SVM - Linear Kernel	0.4735	0.0000	0.5879	0.3782	0.4515	-0.0071	-0.0094	0.0080
ridge	Ridge Classifier	0.5087	0.0000	0.4956	0.3857	0.4322	0.0115	0.0121	0.0070

1- Logistic regression



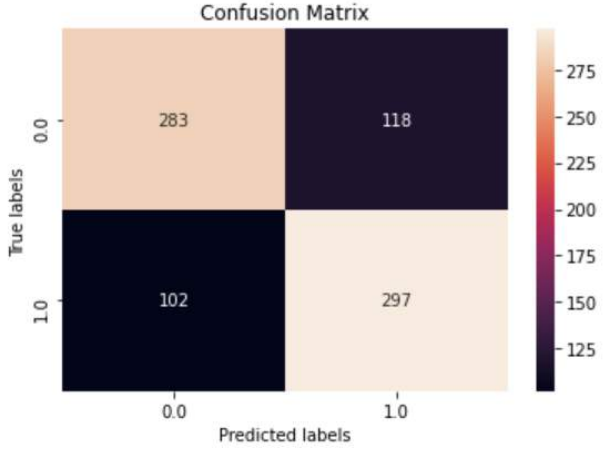

```
Confusion Matrix Done [ ✓ ]

Evaluating Model Performance [*]
Validation Accuracy is : 0.51875
Evaluating Model Performance [ ✓ ]

Applying K-Fold Cross Validation [*]
Accuracy: 48.97 %
Standard Deviation: 2.39 %
K-Fold Cross Validation [ ✓ ]

Complete [ ✓ ]

Time Elapsed :  0.3795661926269531 seconds
```

2- KNN




```
Confusion Matrix Done [ ✓ ]

Evaluating Model Performance [*]
Validation Accuracy is : 0.72625
Evaluating Model Performance [ ✓ ]

Applying K-Fold Cross Validation [*]
Accuracy: 71.56 %
Standard Deviation: 2.31 %
K-Fold Cross Validation [ ✓ ]

Complete [ ✓ ]

Time Elapsed :  0.3889470100402832 se
```

3- Decision Trees

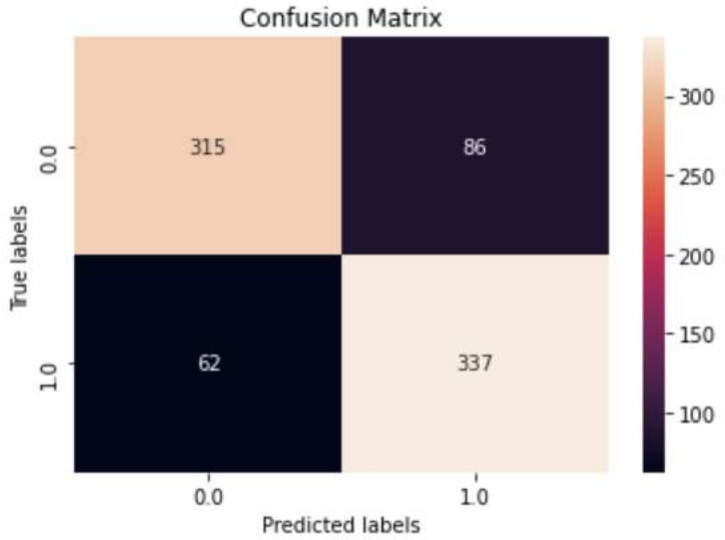



```
Confusion Matrix Done [ ✓ ]

Evaluating Model Performance [*]
Validation Accuracy is : 0.725
Evaluating Model Performance [ ✓ ]

Applying K-Fold Cross Validation [*]
Accuracy: 72.84 %
Standard Deviation: 2.45 %
K-Fold Cross Validation [ ✓ ]

Complete [ ✓ ]

Time Elapsed :  0.6468570232391357 second
```

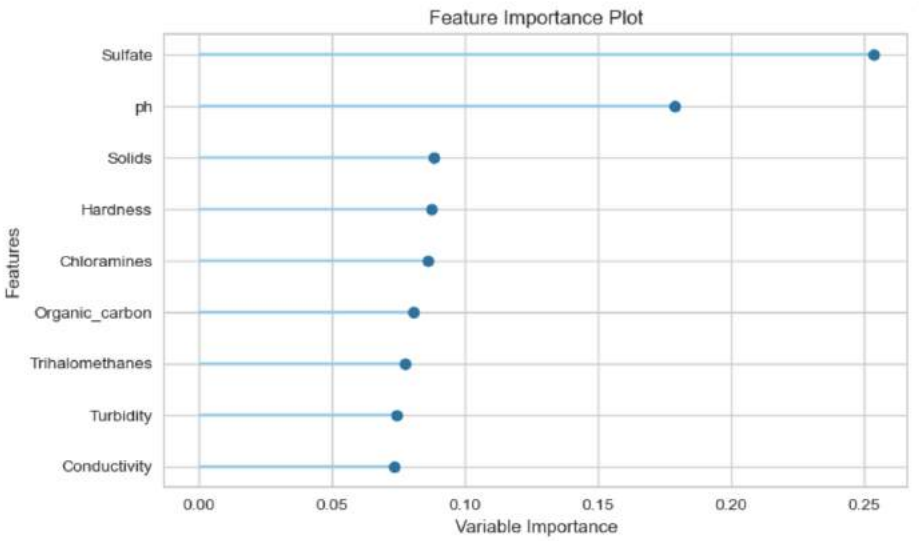
Data science challenges

Reveal data secrets

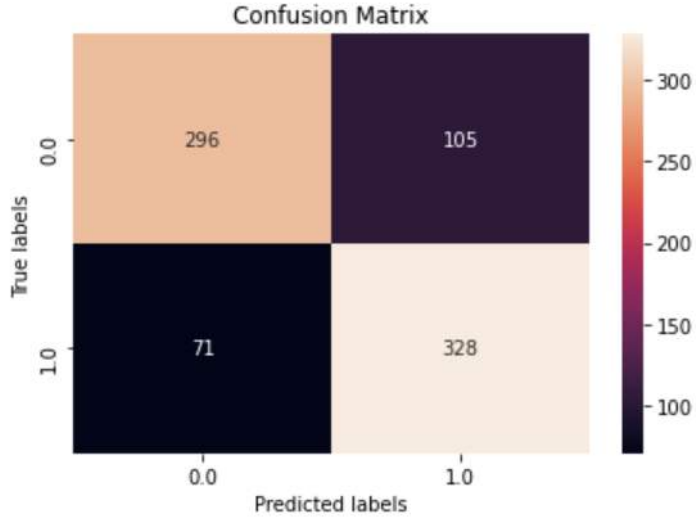
4- Random Forest Classifier:




```
Confusion Matrix Done [ ✓ ]

Evaluating Model Performance [*]
Validation Accuracy is : 0.815
Evaluating Model Performance [ ✓ ]

Applying K-Fold Cross Validation [*]
Accuracy: 81.04 %
Standard Deviation: 1.29 %
K-Fold Cross Validation [ ✓ ]

Complete [ ✓ ]

Time Elapsed :  9.177282094955444 seconds
```



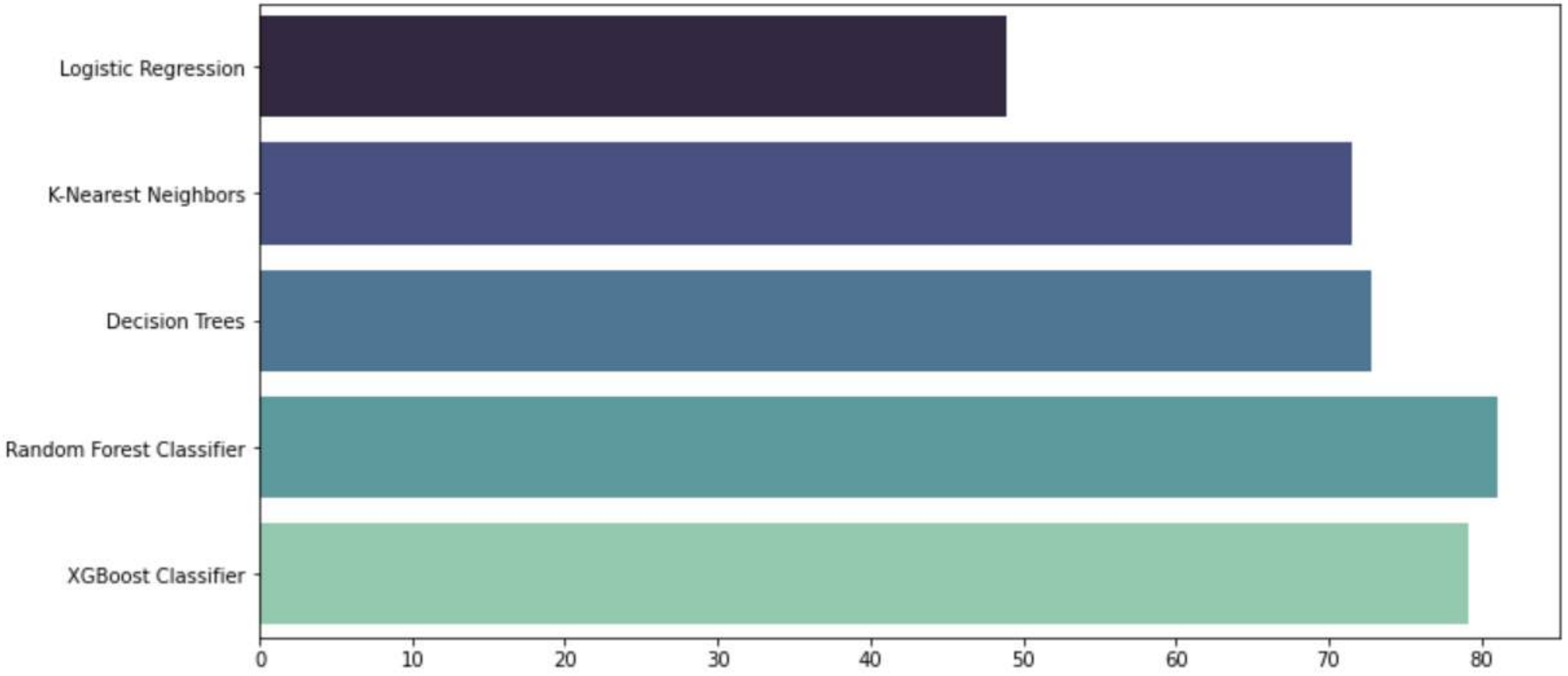
5- XGBOOST:




```
Confusion Matrix Done [ ✓ ]

Evaluating Model Performance [*]
Validation Accuracy is : 0.78
Evaluating Model Performance [ ✓ ]

Applying K-Fold Cross Validation [*]
Accuracy: 79.19 %
Standard Deviation: 1.43 %
K-Fold Cross Validation [ ✓ ]

Complete [ ✓ ]

Time Elapsed :  12.814713954925537 seconds
```

Data science challenges
Reveal data secrets

The model with highest K-Fold Validation Accuracy score is Random Forest Classifier with an accuracy of 81.04



Data science challenges
Reveal data secrets

Conclusion:

After Handeling the Missing Data, Fixing the Imbalanced data, Fixing the Skewness, and fixing the wrong classified potable water, we conclude that:

- Potability is super sensitive to some features like its pH level, and the Carboon levels.
- water could be classified as potable without being really drinkable, which is really dangerous for human use as we visuallized each feature with its reccomended limits.
- Some countries and places my recive drinkable water without filtering it while they should filter it and take a closer look to each feature which may effect the water health so that no harm could be done to the water users.
- Considering this data as a global water sample lets us understand that unpotable water exists more than the potable one, and that we should consider more sustainble ways to filter out our water instead of using fussil fuels.
- also, people should be more aware of consuming plastics, as plastic endes in water and decays to microplastics which effects the water carbon levels and lasts **between 450 and 1,000 years**.

Special Thanks to Dr. Doaa Mahmoud for her support, amazing help, and precious time.

Data science challenges

Reveal data secrets