Water Quality Analysis-An ML approach

One of the main areas of research in machine learning is the analysis of water quality. It is also known as water potability analysis because our task here is to understand all the factors that affect water potability and train a machine learning model that can classify whether a specific water sample is safe or unfit for consumption.

1.Importing Dataset

```
# This is formatted as code
```

+ Code + Text

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np

data = pd.read_csv("water_probability.csv")
data.head()
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890456	20791.31898	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.05786	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.54173	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.41744	8.059332	356.886136	363.266516	18.436525	100.341674	4.628771	0
4	9.092223	181.101509	17978.98634	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

2.Preprocessing Dataset

Preliminary view of Dataset shows presence of "nan" or missing values which must be cleaned

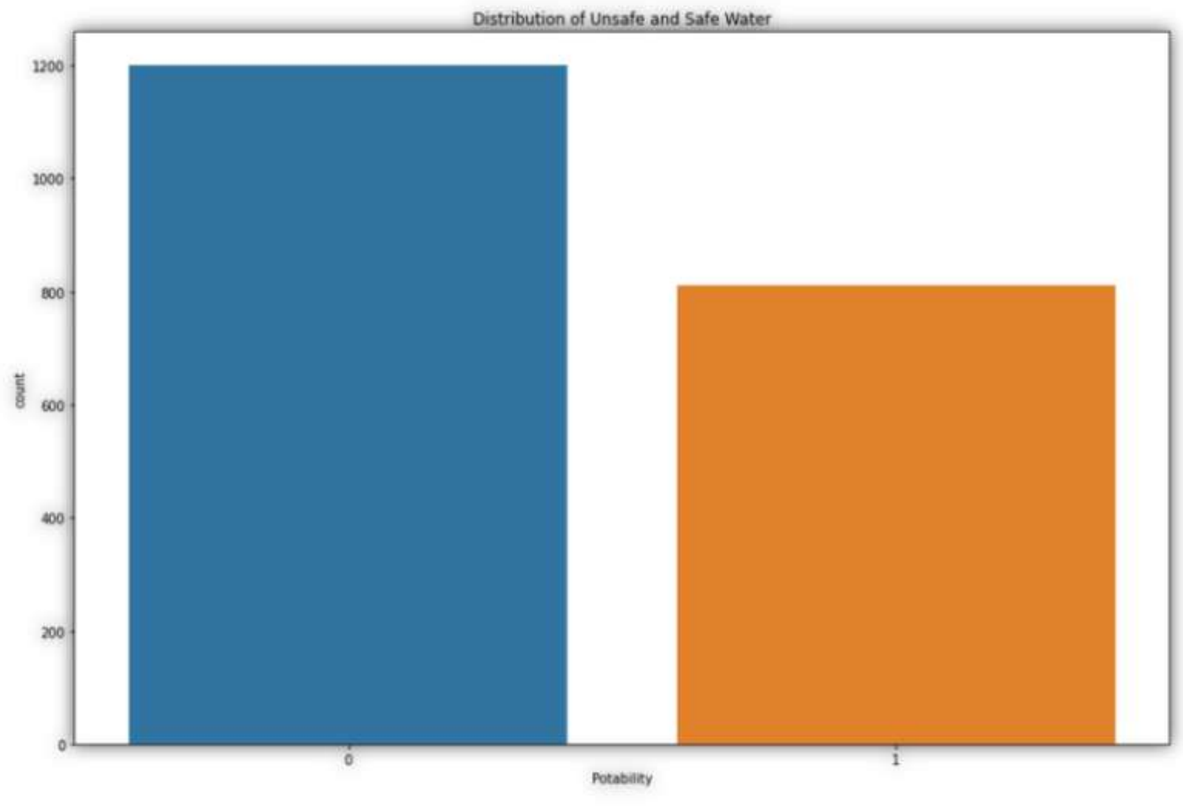
```
data = data.dropna()
data.isnull().sum()
```

```
ph                0
Hardness          0
Solids            0
Chloramines       0
Sulfate           0
Conductivity      0
Organic_carbon    0
Trihalomethanes   0
Turbidity         0
Potability        0
dtype: int64
```

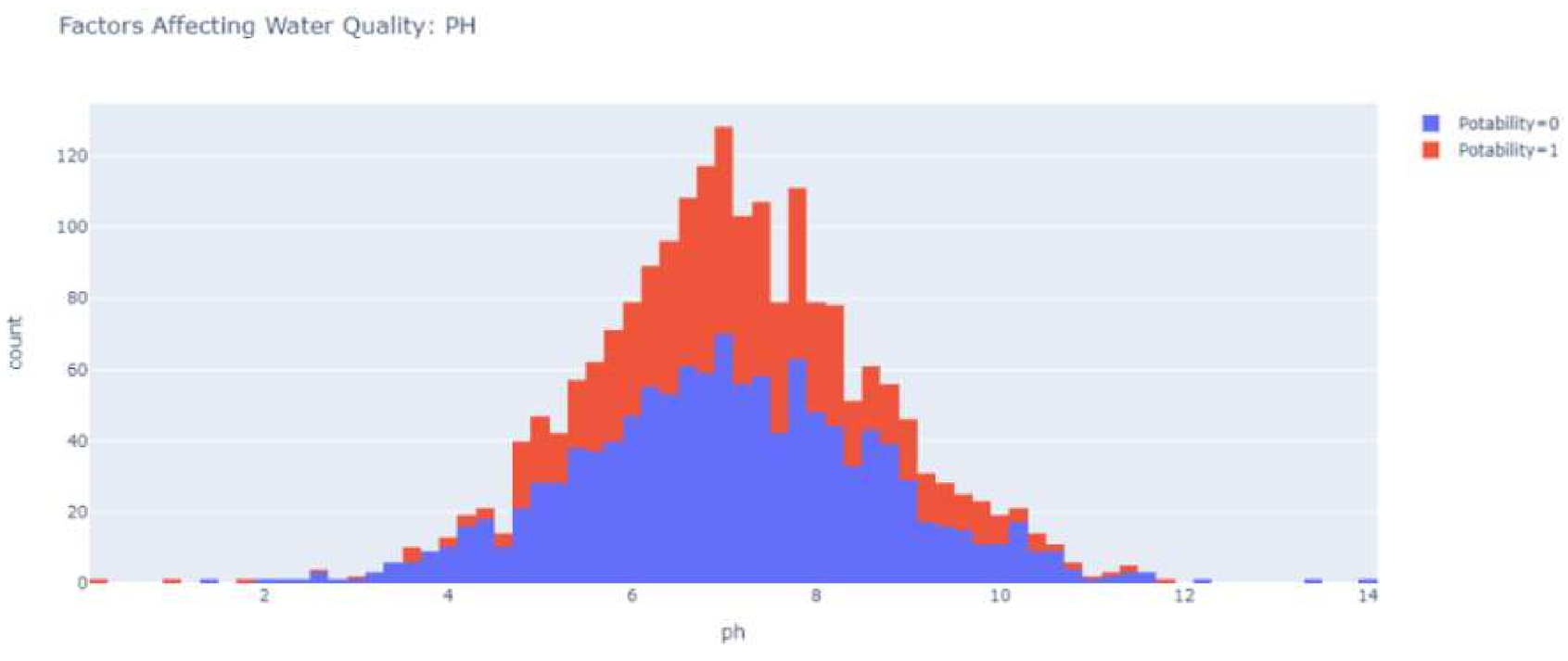
Exploratory Data Analysis

The Potability column of this dataset is the column we need to predict because it contains values 0 and 1 that indicate whether the water is potable (1) or unfit (0) for consumption. So let's see the distribution of 0 and 1 in the Potability column:

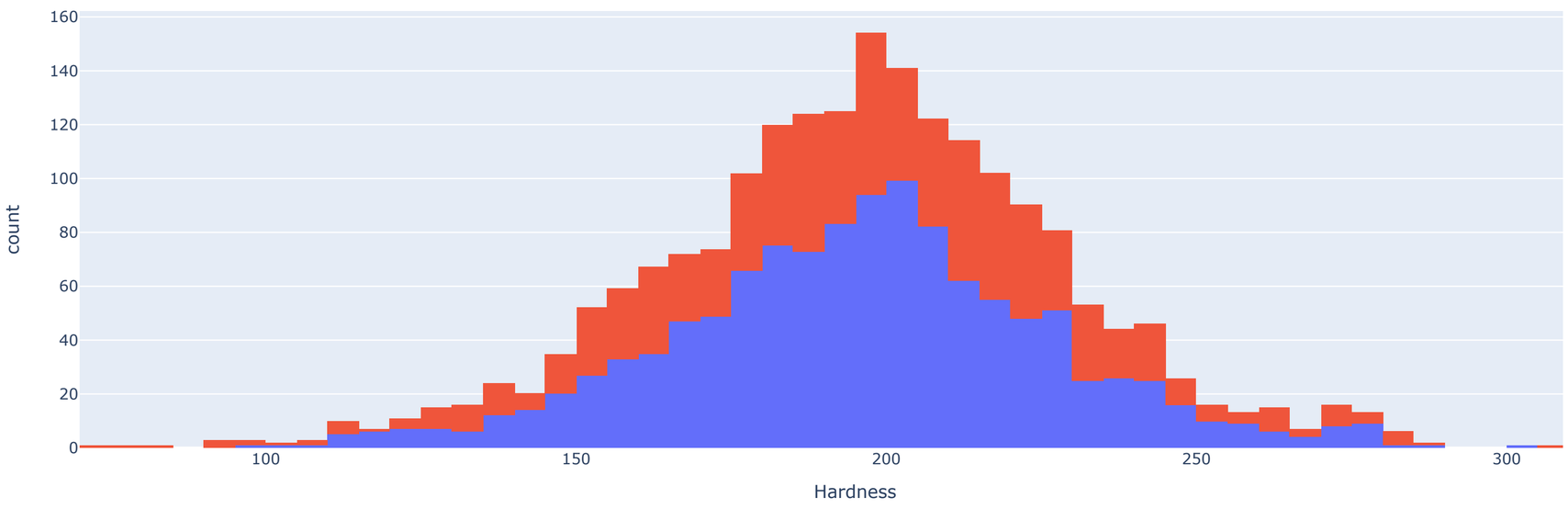
```
plt.figure(figsize=(15, 10))
sns.countplot(a)
plt.title("Distribution of Unsafe and Safe Water")
plt.show()
```



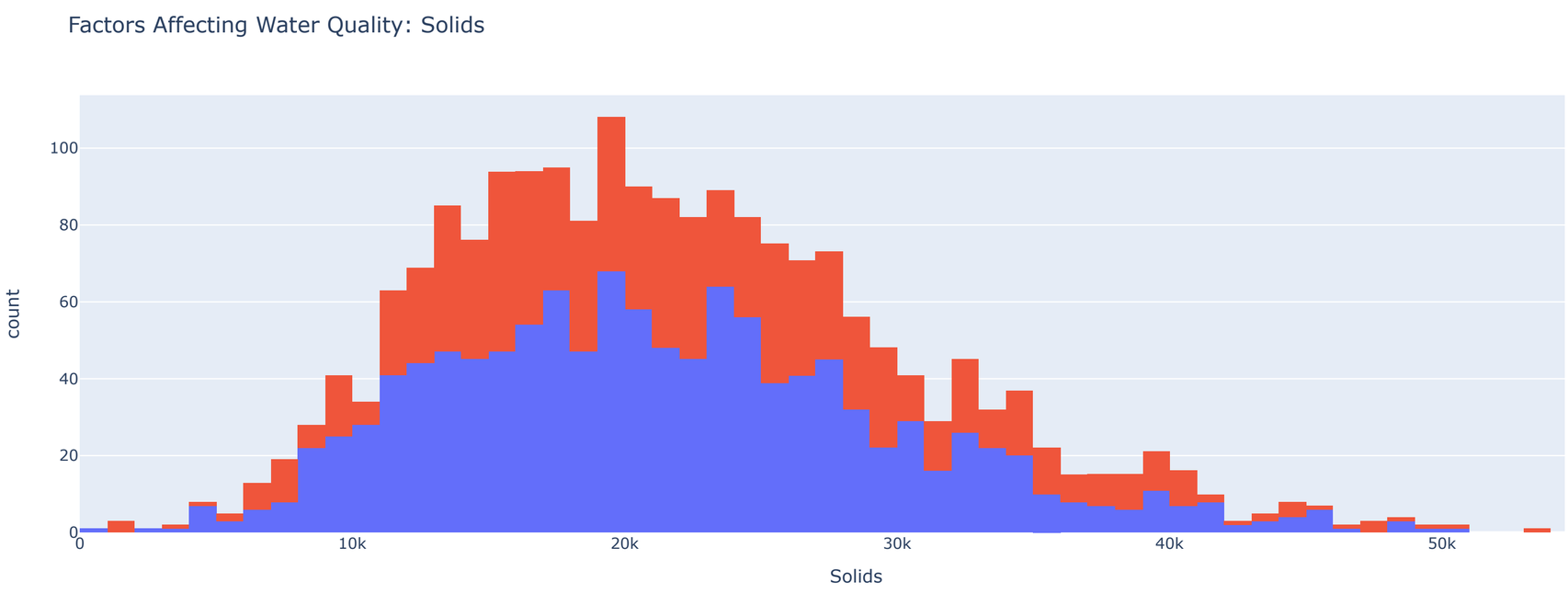

```
import plotly.express as px
data = data
figure = px.histogram(data, x = "ph",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: PH")
figure.show()
```



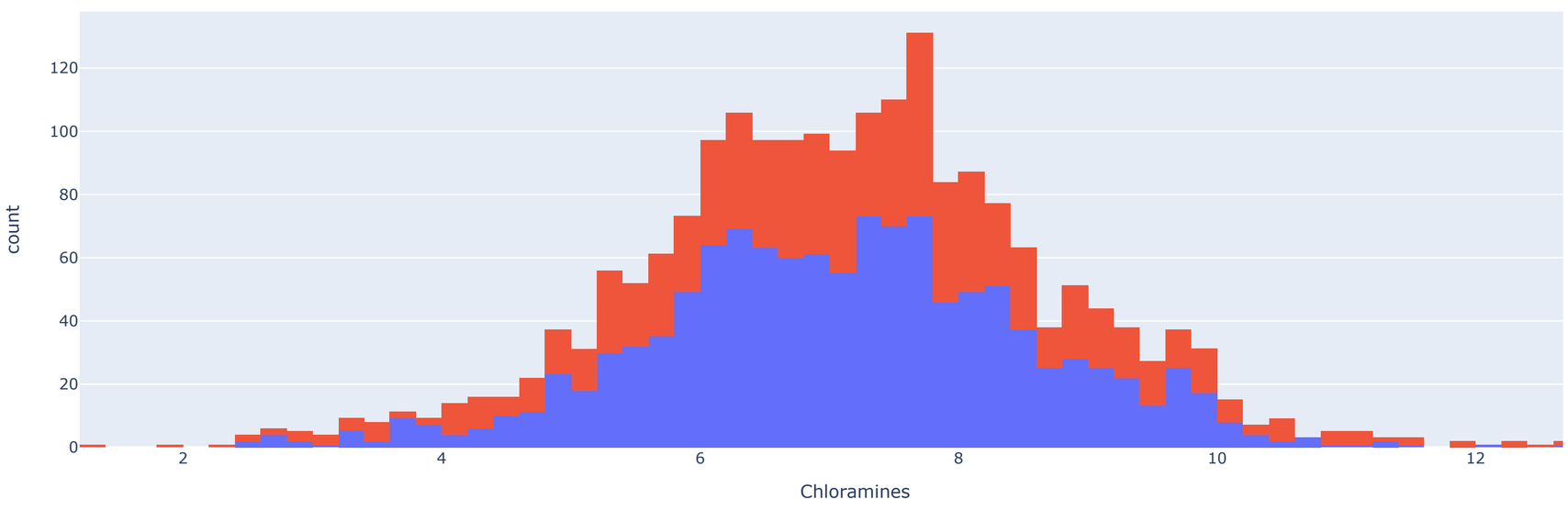
```
figure = px.histogram(data, x = "Hardness",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Hardness")
figure.show()
```



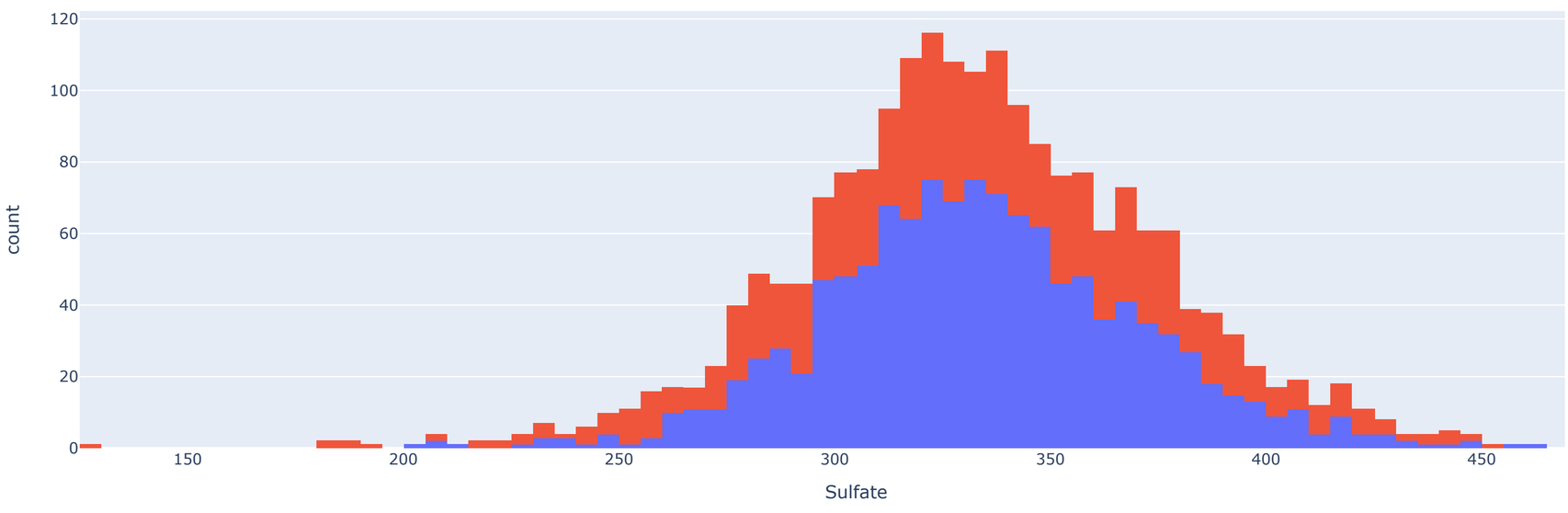

```
figure = px.histogram(data, x = "Solids",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Solids")
figure.show()
```



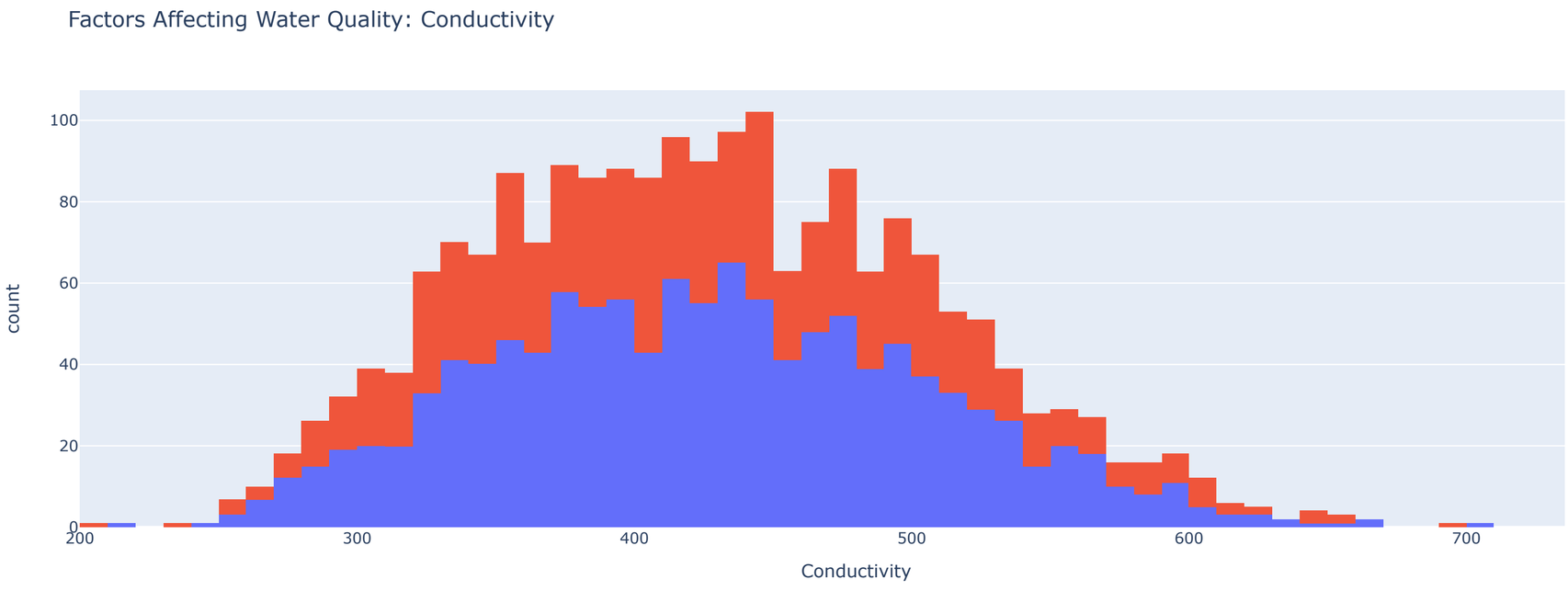

```
figure = px.histogram(data, x = "Chloramines",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Chloramines")
figure.show()
```



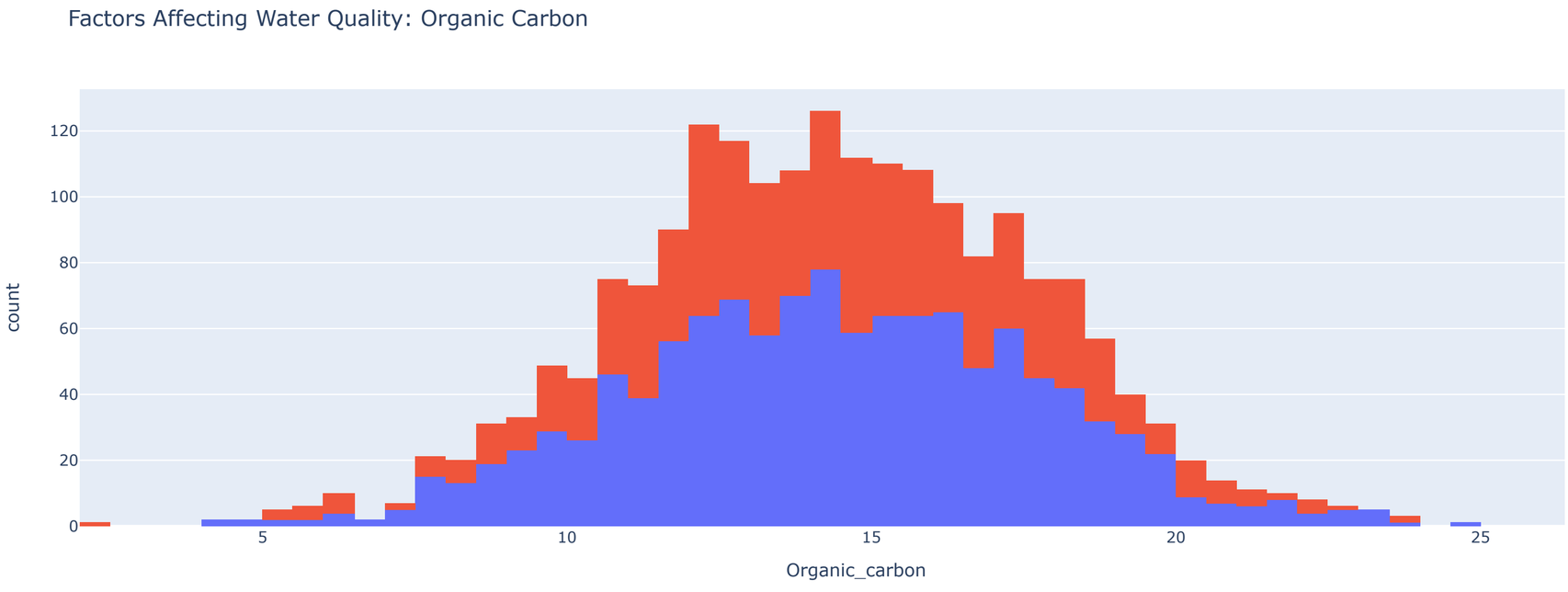

```
figure = px.histogram(data, x = "Sulfate",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Sulfate")
figure.show()
```



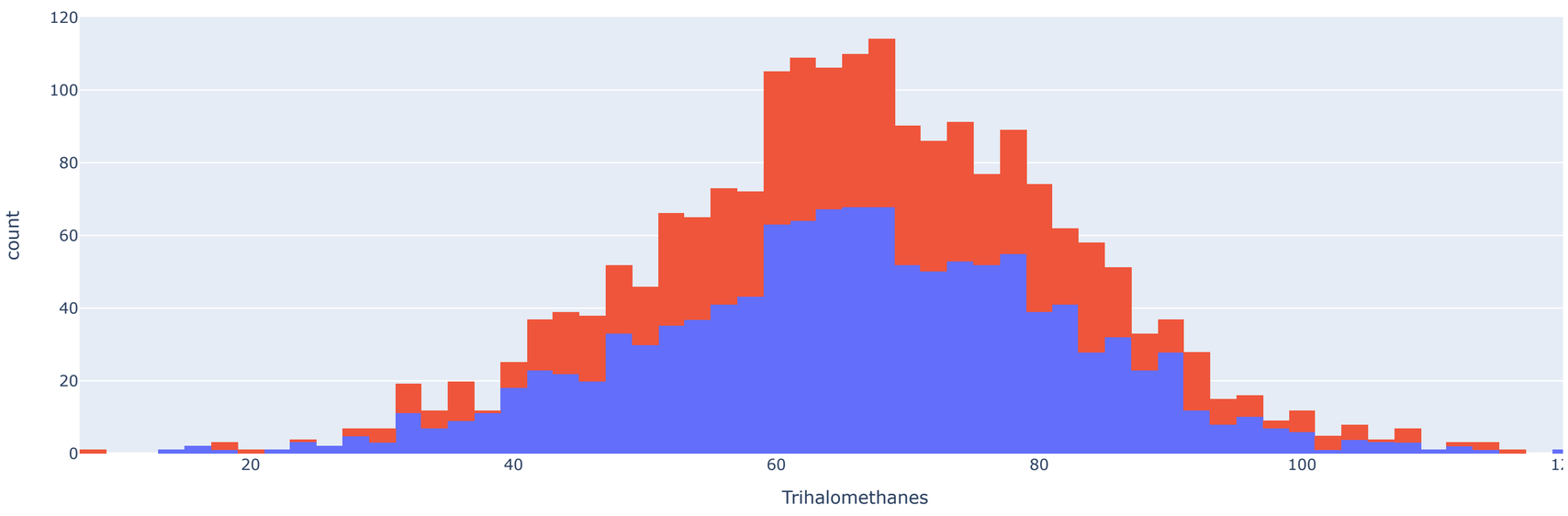
```
figure = px.histogram(data, x = "Conductivity",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Conductivity")
figure.show()
```



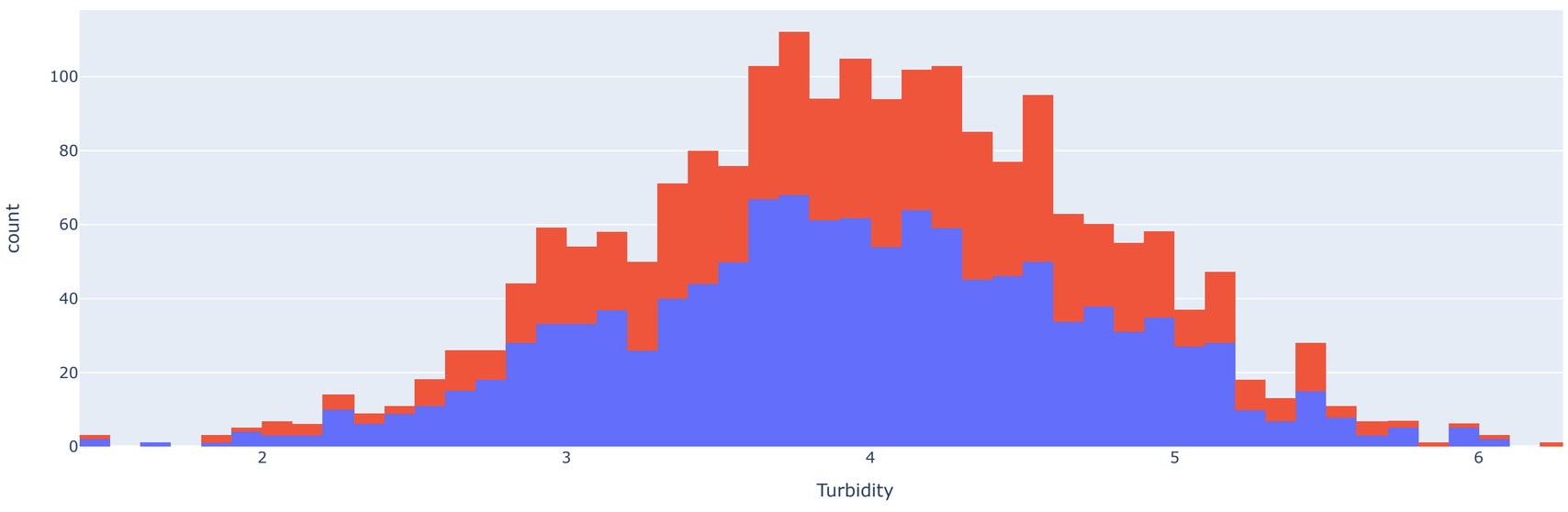

```
figure = px.histogram(data, x = "Organic_carbon",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Organic Carbon")
figure.show()
```




```
figure = px.histogram(data, x = "Trihalomethanes",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Trihalomethanes")
figure.show()
```




```
figure = px.histogram(data, x = "Turbidity",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Turbidity")
figure.show()
```



Building Machine Learning Model for predicting Potability

Inference from EDA

Above histograms indicate strict difference in distribution of potable and non potable quality based on each feature which indicate relevance of each feature and statistical ability to form machine learning models predicting potability from above features

Analyzing Correlation

```
correlation = data.corr()
correlation["ph"].sort_values(ascending=False)
```

```
ph                 1.000000
Hardness           0.108948
Organic_carbon     0.028375
Trihalomethanes    0.018278
Potability         0.014530
Conductivity       0.014128
Sulfate            0.010524
Chloramines       -0.024768
Turbidity         -0.035849
Solids            -0.087615
Name: ph, dtype: float64
```

Analyzing best ML model for the dataset

```
from pycaret.classification import *
clf = setup(data, target = "Potability",  session_id = 786)
compare_models()
```

	Description	Value
0	Session id	786
1	Target	Potability
2	Target type	Binary
3	Original data shape	(2011, 10)
4	Transformed data shape	(2011, 10)
5	Transformed train set shape	(1407, 10)
6	Transformed test set shape	(604, 10)
7	Numeric features	9
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10
14	CPU Jobs	-1
15	Use GPU	False
16	Log Experiment	False
17	Experiment Name	clf-default-name
18	USI	4272

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.6802	0.6956	0.3952	0.6778	0.4977	0.2870	0.3100	0.5270
rf	Random Forest Classifier	0.6780	0.6844	0.4040	0.6696	0.5024	0.2854	0.3063	1.4460
qda	Quadratic Discriminant Analysis	0.6745	0.7091	0.3866	0.6795	0.4879	0.2746	0.3013	0.0590
gbc	Gradient Boosting Classifier	0.6489	0.6554	0.3581	0.6232	0.4505	0.2186	0.2397	0.9290
lightgbm	Light Gradient Boosting Machine	0.6432	0.6658	0.4869	0.5719	0.5232	0.2416	0.2453	0.3400
xgboost	Extreme Gradient Boosting	0.6389	0.6581	0.4744	0.5629	0.5129	0.2301	0.2331	0.5720
nb	Naive Bayes	0.6212	0.6280	0.2506	0.5728	0.3474	0.1344	0.1581	0.0960
ridge	Ridge Classifier	0.5984	0.0000	0.0282	0.6267	0.0534	0.0137	0.0499	0.0900
lda	Linear Discriminant Analysis	0.5970	0.5189	0.0299	0.5867	0.0564	0.0115	0.0421	0.0630
dummy	Dummy Classifier	0.5970	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0600
dt	Decision Tree Classifier	0.5956	0.5784	0.4902	0.4981	0.4927	0.1570	0.1576	0.1030
knn	K Neighbors Classifier	0.5423	0.5226	0.3262	0.4122	0.3625	0.0145	0.0145	0.2420

According to the above result, the random forecast classification algorithm is best for training a machine learning model for the task of water quality analysis. So let's train the model and examine its predictions:

```
model = create_model("et")
predict = predict_model(model, data=data)
predict.head()
```

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6525	0.6544	0.3509	0.6250	0.4494	0.2238	0.2437
1	0.6879	0.7289	0.3860	0.7097	0.5000	0.3009	0.3304
2	0.6596	0.6682	0.3158	0.6667	0.4286	0.2279	0.2602
3	0.6950	0.7311	0.4035	0.7188	0.5169	0.3188	0.3472
4	0.6454	0.6211	0.3333	0.6129	0.4318	0.2055	0.2257
5	0.6525	0.7076	0.4211	0.6000	0.4948	0.2422	0.2510
6	0.7163	0.7422	0.4737	0.7297	0.5745	0.3758	0.3956
7	0.7143	0.7067	0.4107	0.7667	0.5349	0.3548	0.3909
8	0.7071	0.7133	0.4821	0.6923	0.5684	0.3574	0.3708
9	0.6714	0.6825	0.3750	0.6562	0.4773	0.2628	0.2847
Mean	0.6802	0.6956	0.3952	0.6778	0.4977	0.2870	0.3100
Std	0.0259	0.0364	0.0522	0.0521	0.0495	0.0595	0.0612

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extra Trees Classifier	0.9040	0.9738	0.8126	0.9414	0.8723	0.7961	0.8016

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability	prediction_label	prediction_sc
3	8.316766	214.373398	22018.417969	8.059333	356.886139	363.266510	18.436525	100.341675	4.628770	0	0	
4	9.092223	181.101517	17978.986328	6.546600	310.135742	398.410828	11.558279	31.997993	4.075076	0	0	
5	5.584086	188.313324	28748.687500	7.544869	326.678375	280.467926	8.399734	54.917862	2.559708	0	0	
6	10.223862	248.071732	28749.716797	7.513409	393.663391	283.651642	13.789696	84.603554	2.672989	0	0	
7	8.635849	203.361526	13672.091797	4.563009	303.309784	474.607635	12.363816	62.798309	4.401425	0	0	

Summary

Access to safe drinking water is one of the essential needs of all human beings. From a legal point of view, access to drinking water is one of the fundamental human rights. Many factors affect water quality, it is also one of the major research areas in machine learning.