Water Quality Analysis

One of the main areas of research in machine learning revolves around the analysis of water quality, also referred to as water potability analysis. The objective is to comprehend all the variables influencing water potability and develop a machine learning model capable of classifying whether a specific water sample is safe for consumption.

To undertake the water quality analysis task, I will utilize a Kaggle dataset encompassing data on the major factors impacting water potability. Given the significance of all these factors in determining water quality, a comprehensive exploration of each feature within this dataset is imperative before proceeding to train a machine learning model for predicting the safety or unsuitability of a water sample.

Let's initiate the water quality analysis by importing essential Python libraries and loading the dataset:

In [1]:
```
import warnings
warnings.filterwarnings('ignore')

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

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

Out[1]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993

In [2]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64  
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

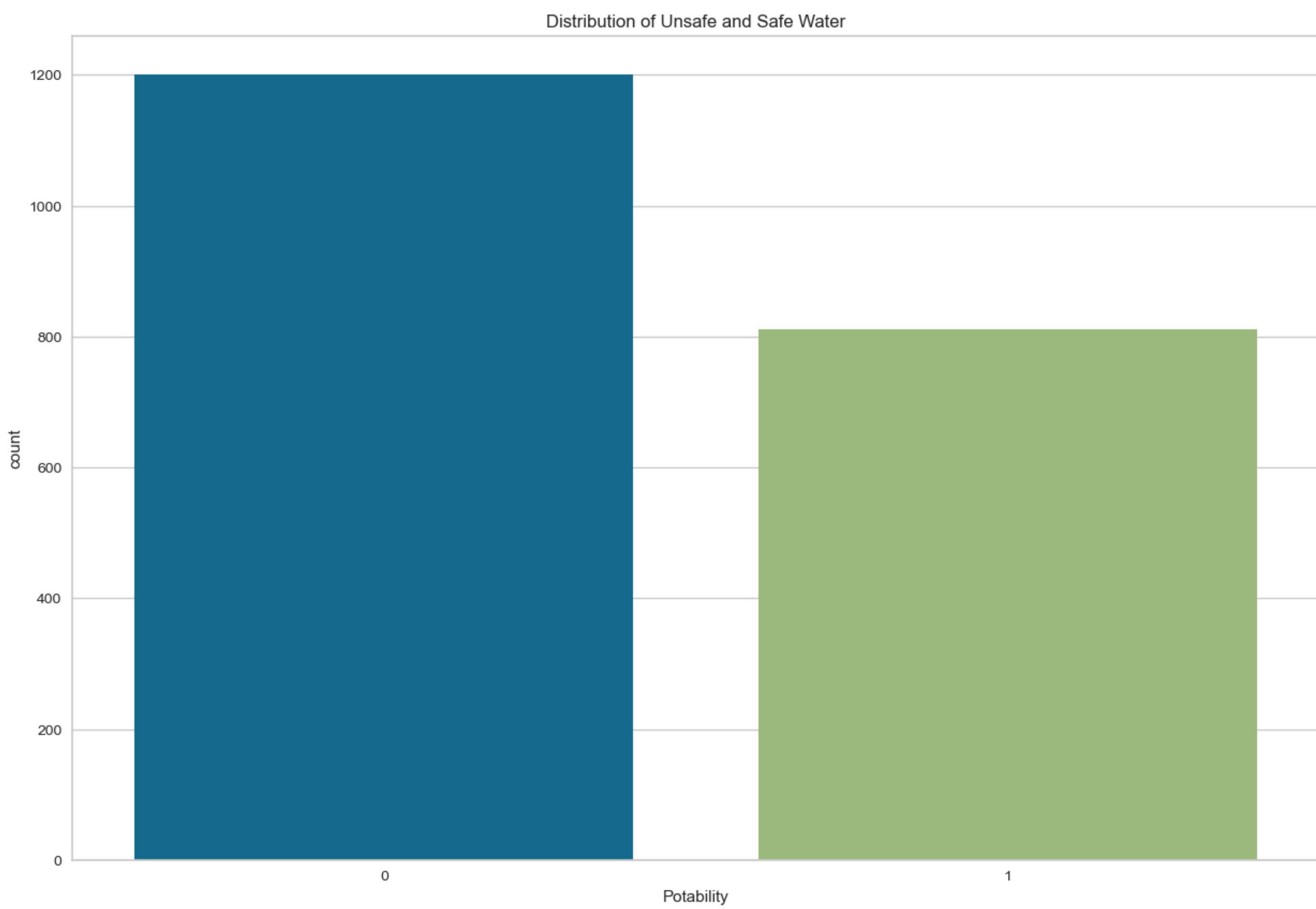
Null values are apparent in the initial glimpse of this dataset. Therefore, as a preliminary step, let's eliminate all rows containing null values before proceeding further:

In [27]:
```
data = data.dropna()
data.isnull().sum()
```

Out[27]:
```
ph                 0
Hardness           0
Solids             0
Chloramines        0
Sulfate            0
Conductivity       0
Organic_carbon     0
Trihalomethanes    0
Turbidity          0
Potability         0
dtype: int64
```

Now, let's examine the distribution of the Potability column in this dataset, as it contains values of 0 and 1 indicating whether the water is safe (1) or unsafe (0) for consumption.

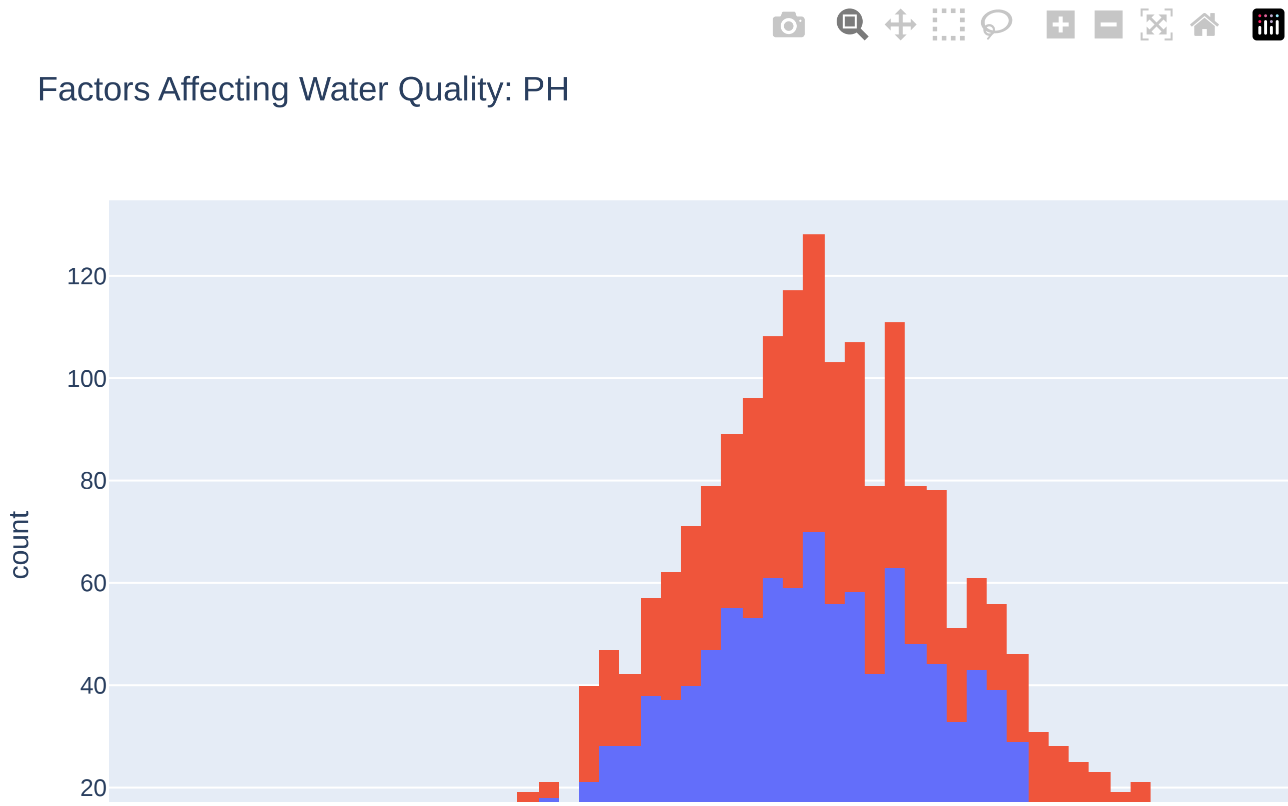
In [26]:
```
plt.figure(figsize=(15, 10))
sns.countplot(x=data.Potability,data = data)
plt.title("Distribution of Unsafe and Safe Water")
plt.show()
```



An important observation about this dataset is its imbalance, where the number of samples labeled as 0s exceeds those labeled as 1s.

Given there are no unignorable factors affecting water quality, we'll systematically explore all columns. Beginning with the 'ph' column:

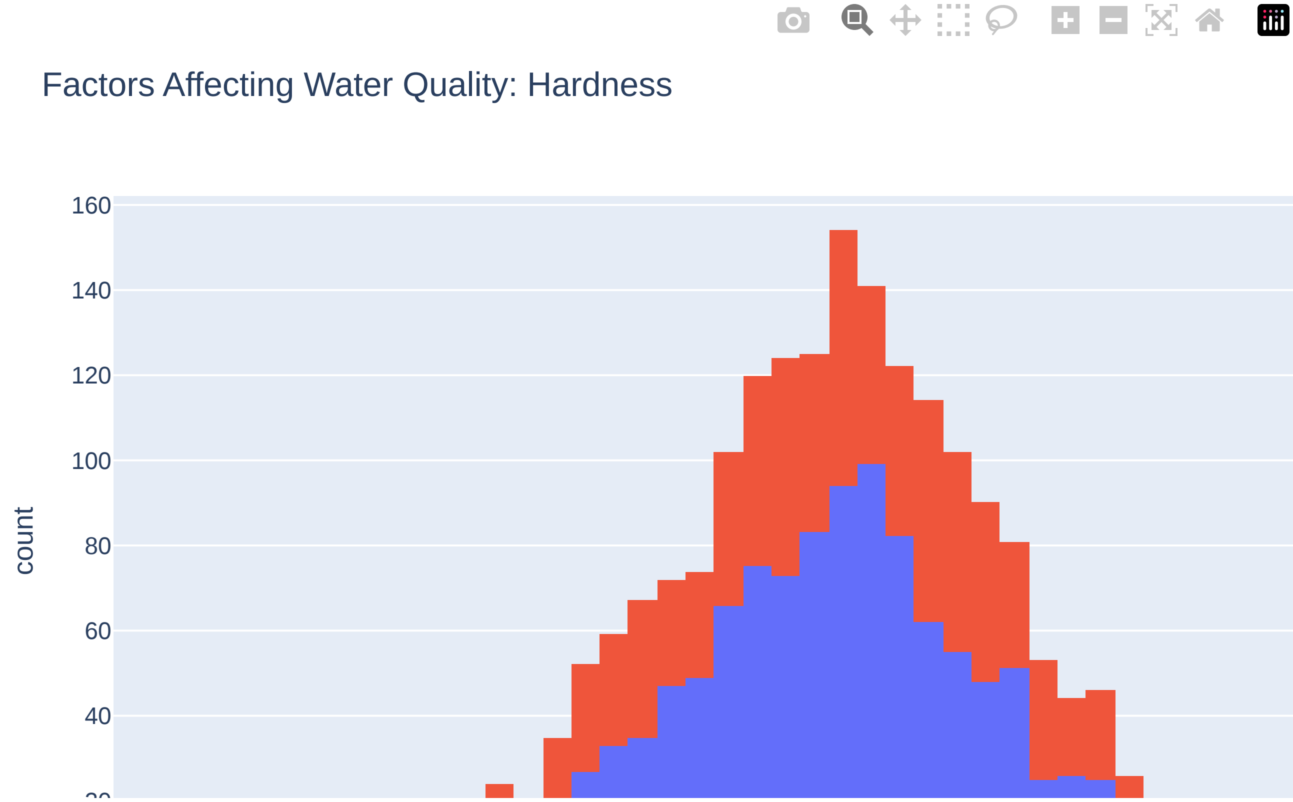
In [5]:
```
import plotly.express as px
data = data
figure = px.histogram(data, x = "ph",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: PH")
figure.show()
```



"The 'ph' column denotes the pH value, a crucial factor in assessing the acid-base equilibrium of water. The recommended pH range for drinking water is between 6.5 and 8.5. Now, let's proceed to examine the second determinant of water quality in this dataset:"

In [25]:
```
figure = px.histogram(data, x = "Hardness",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Hardness")
figure.show()
```

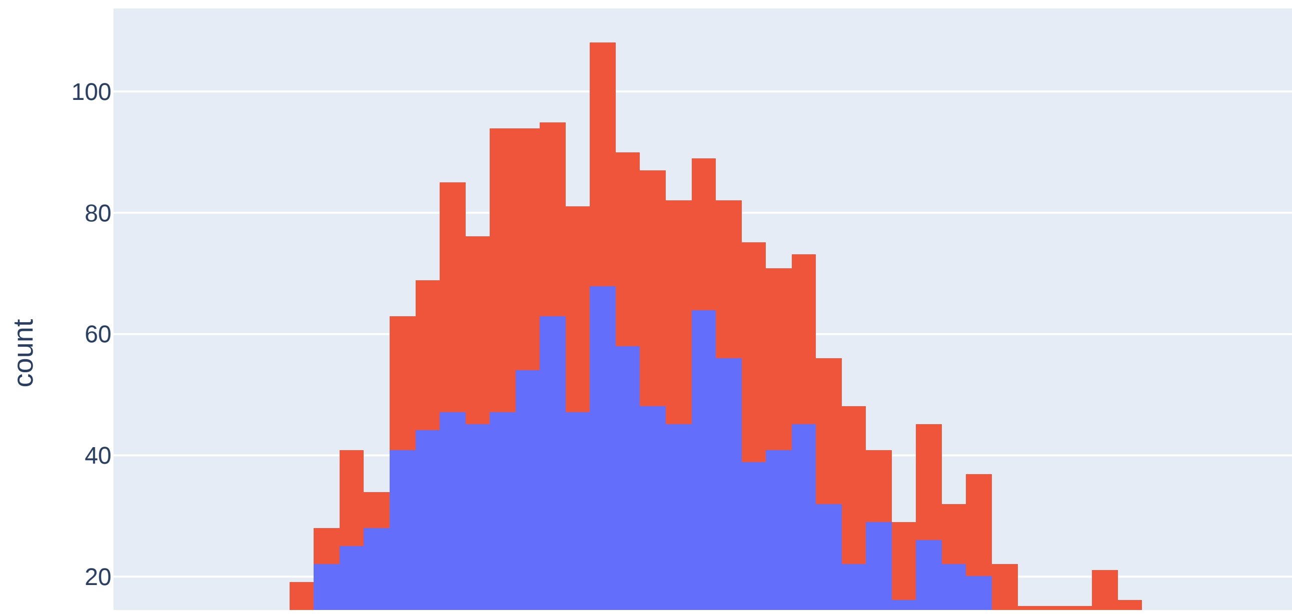
Loading [MathJax]/extensions/Safe.js



The depicted graph illustrates the distribution of water hardness within the dataset. Typically, water hardness varies based on its source, yet water with a hardness ranging between 120-200 milligrams is considered potable. Let's now delve into the subsequent factor influencing water quality.

In [24]:
```
figure = px.histogram(data, x = "Solids",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Solids")
figure.show()
```

Factors Affecting Water Quality: Solids



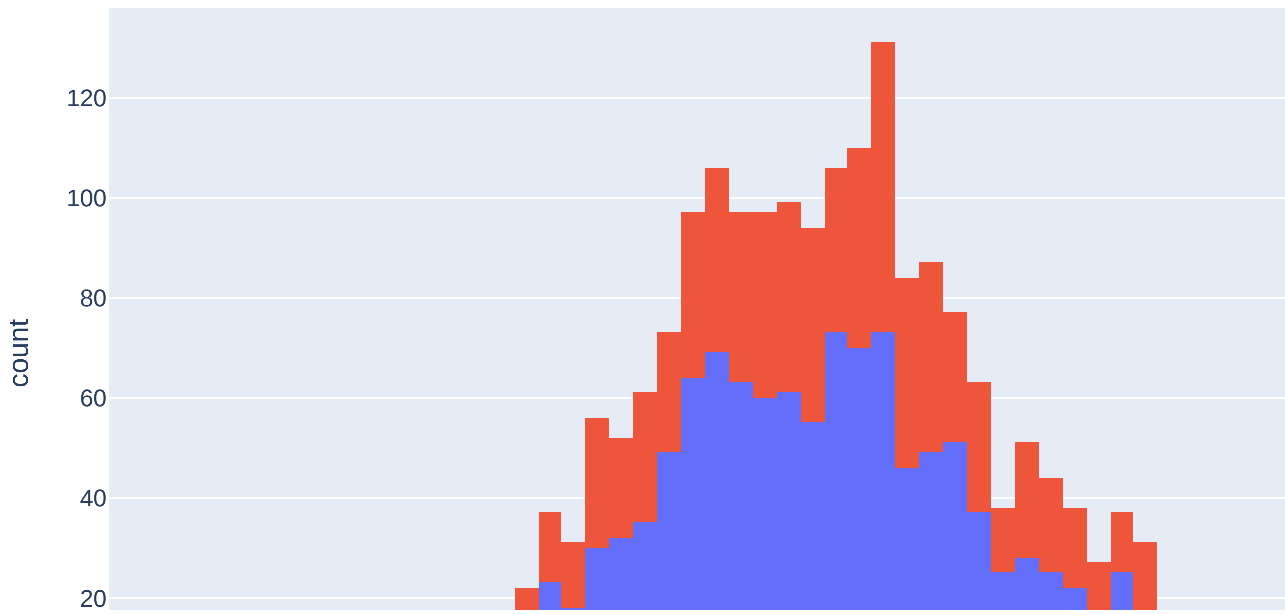
The diagram provided illustrates the dispersion of total dissolved solids within the water dataset. Total dissolved solids encompass both organic and inorganic minerals present within the water. Water with elevated levels of dissolved solids is often described as highly mineralized.

Now, let's delve into the subsequent factor influencing water quality:

In [23]:
```
figure = px.histogram(data, x = "Chloramines",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Chloramines")
figure.show()
```

Loading [MathJax]/extensions/Safe.js

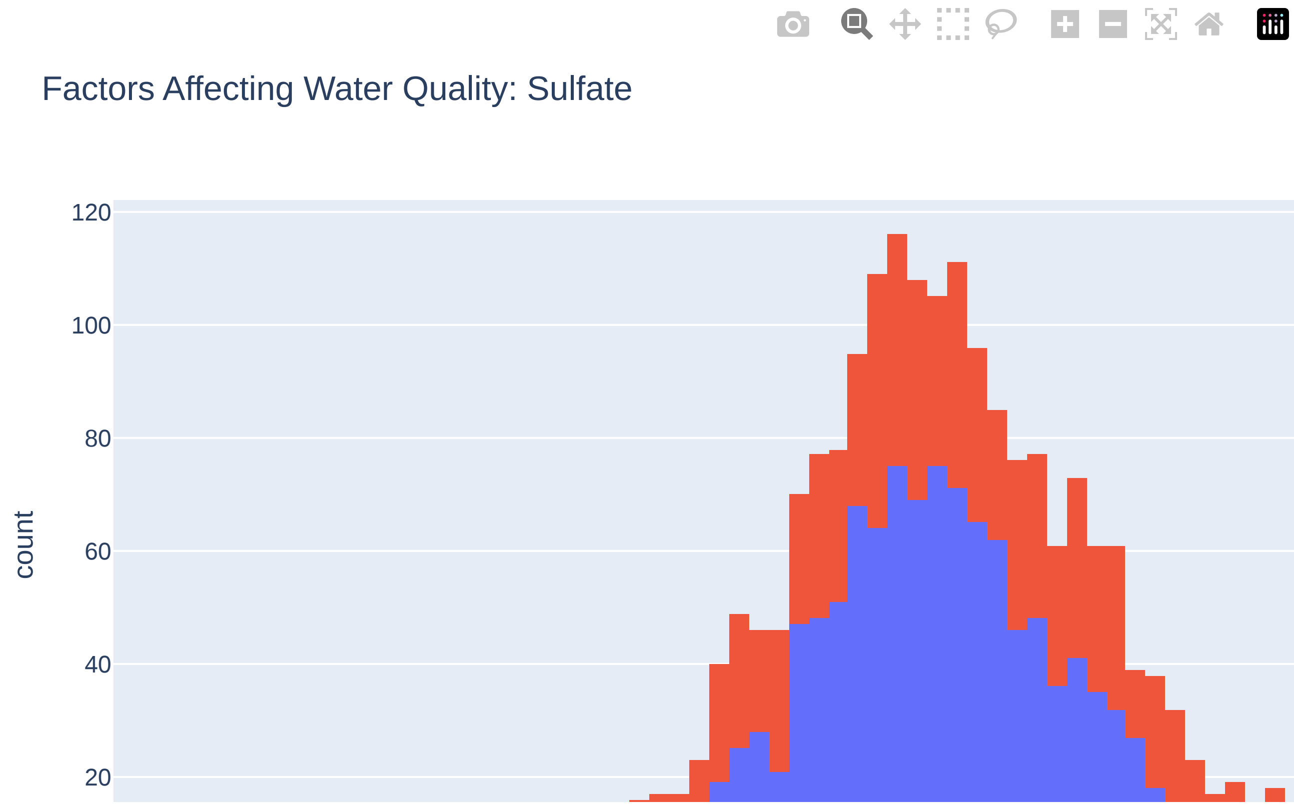
Factors Affecting Water Quality: Chloramines



The depicted graph showcases the distribution of chloramine within the water dataset. Chloramine, alongside chlorine, serves as a disinfectant employed in public water systems.

Now, let's shift our focus to the subsequent factor influencing water quality:

In [22]:
```
figure = px.histogram(data, x = "Sulfate",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Sulfate")
figure.show()
```

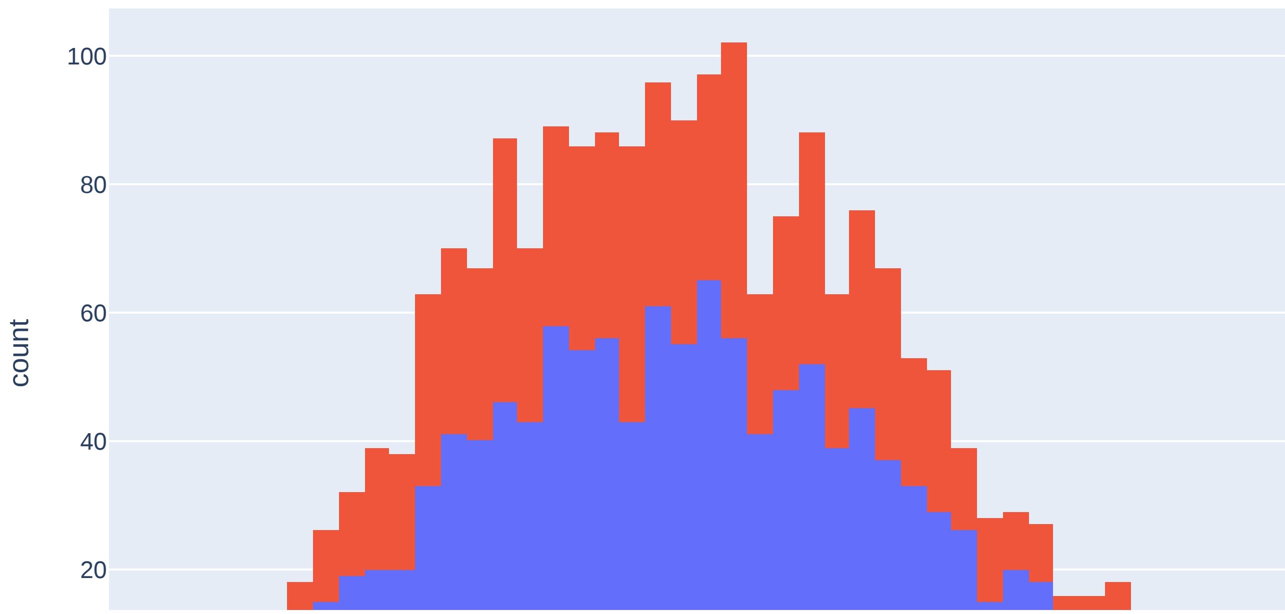


The presented diagram illustrates the dispersion of sulfate within the water dataset. Sulfate is naturally occurring in minerals, soil, and rocks. Water with sulfate levels below 500 milligrams is deemed safe for consumption.

Let's now proceed to examine the following factor:

In [21]:
```
figure = px.histogram(data, x = "Conductivity",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Conductivity")
figure.show()
```

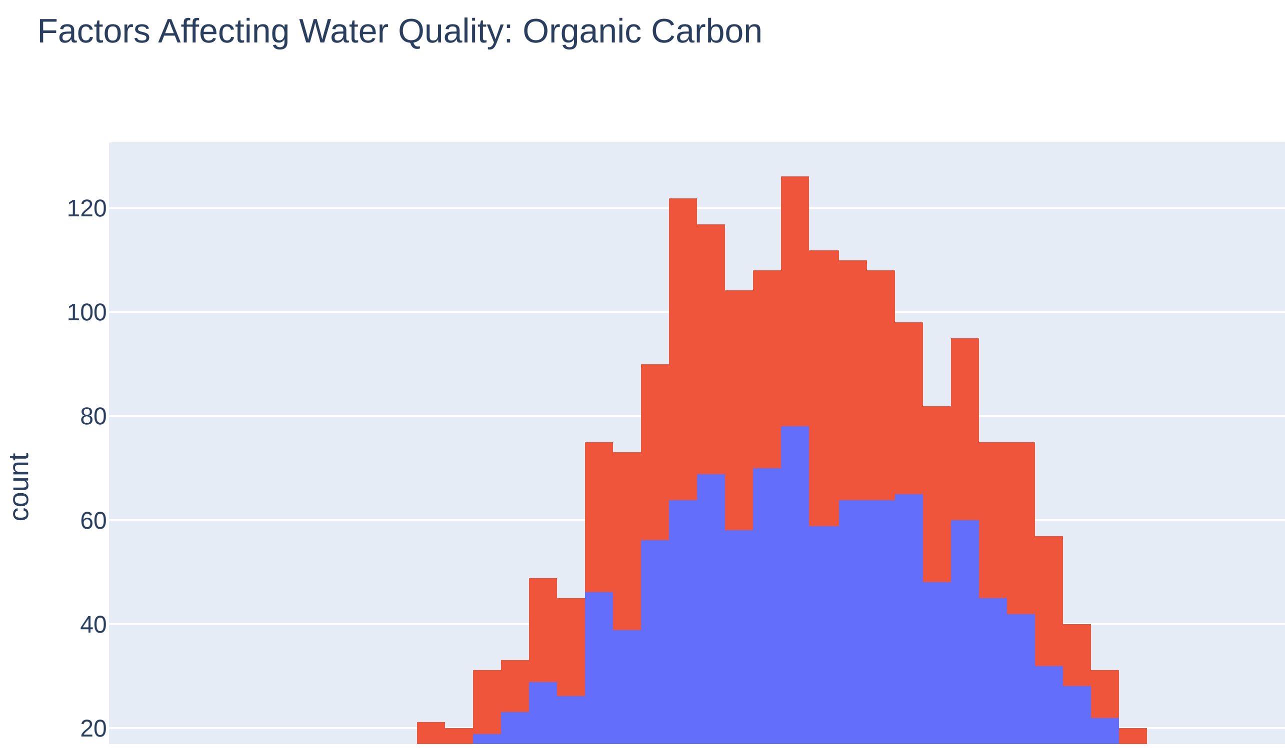
Factors Affecting Water Quality: Conductivity



The graph above displays the distribution of water conductivity within the dataset. While water is generally a conductor of electricity, pure water exhibits poor conductivity. Water with an electrical conductivity below 500 is considered potable.

Let's now move on to explore the next factor:

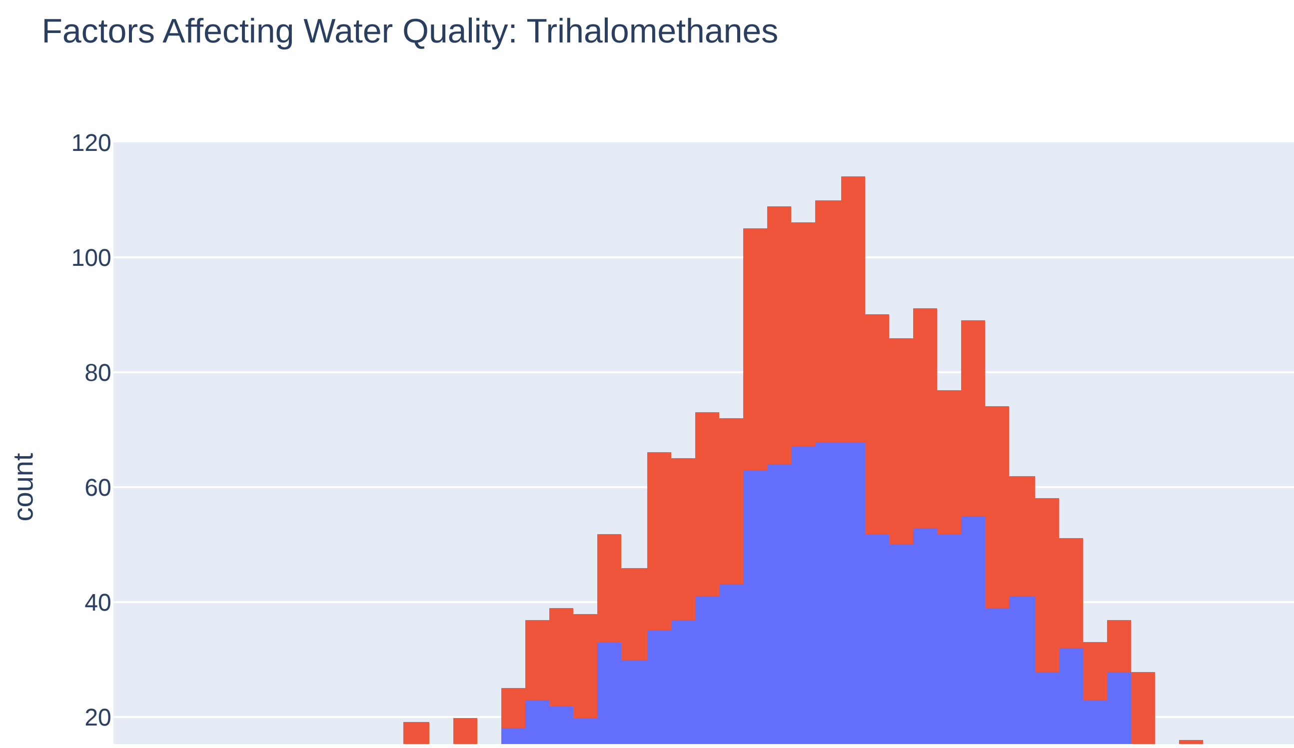
In [20]:
```
figure = px.histogram(data, x = "Organic_carbon",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Organic Carbon")
figure.show()
```



The depicted graph illustrates the distribution of organic carbon within the water dataset. Organic carbon originates from the decomposition of natural organic matter and synthetic sources. Water with organic carbon levels below 25 milligrams is deemed safe for consumption.

Let's now examine the subsequent factor influencing drinking water quality:

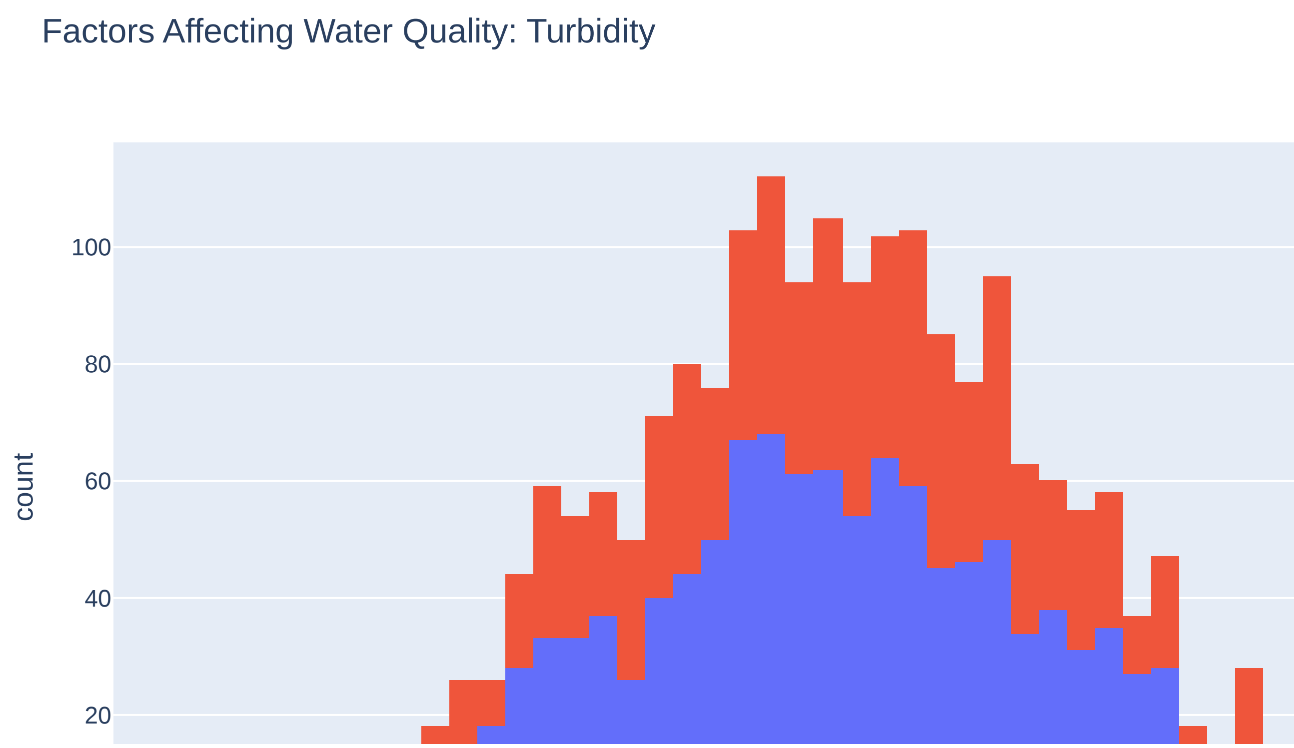
In [19]:
```
figure = px.histogram(data, x = "Trihalomethanes",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Trihalomethanes")
figure.show()
```



The graph above illustrates the distribution of trihalomethanes (THMs) in the water dataset. THMs are compounds present in water treated with chlorine. Drinking water with THM levels below 80 milligrams is deemed safe.

Let's now proceed to examine the next factor impacting drinking water quality within the dataset:

In [18]:
```
figure = px.histogram(data, x = "Turbidity",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Turbidity")
figure.show()
```



The above figure depicts the distribution of turbidity in water. Turbidity is determined by the quantity of suspended solids in the water. Water with a turbidity level below 5 milligrams is considered suitable for drinking.

Water Quality Prediction Model using Python

In the preceding section, we've investigated the various factors influencing water quality. Our next endeavor involves training a machine learning model for water quality analysis using Python. To accomplish this, we'll utilize the PyCaret library. If you're new to PyCaret, fret not, as it can be swiftly installed on your system via the pip command:

```
pip install pycaret
```

Before proceeding with model training, let's examine the correlation between all features and the 'Potability' column in the dataset:

In [14]:
```
correlation = data.corr(numeric_only=True)
correlation["ph"].sort_values(ascending=False)
```

Loading [MathJax]/extensions/Safe.js

Out[14]:
```
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
```

Now below is how you can see which machine learning algorithm is best for this dataset by using the PyCaret library in Python:

In [15]:
```python
from pycaret.classification import *
clf = setup(data, target="Potability",verbose=False,session_id=786)
compare_models()
```

	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	0.0480
rf	Random Forest Classifier	0.6780	0.6844	0.4040	0.6696	0.5024	0.2854	0.3063	0.0790
qda	Quadratic Discriminant Analysis	0.6745	0.7091	0.3866	0.6795	0.4879	0.2746	0.3013	0.0070
gbc	Gradient Boosting Classifier	0.6532	0.6558	0.3564	0.6297	0.4517	0.2257	0.2473	0.0770
lightgbm	Light Gradient Boosting Machine	0.6432	0.6658	0.4869	0.5719	0.5232	0.2416	0.2453	0.0880
xgboost	Extreme Gradient Boosting	0.6333	0.6677	0.4729	0.5540	0.5074	0.2193	0.2224	0.0400
nb	Naive Bayes	0.6212	0.6280	0.2506	0.5728	0.3474	0.1344	0.1581	0.0060
lr	Logistic Regression	0.6020	0.5093	0.0318	0.6467	0.0600	0.0220	0.0657	0.6160
ridge	Ridge Classifier	0.5984	0.5188	0.0282	0.6267	0.0534	0.0137	0.0499	0.0080
lda	Linear Discriminant Analysis	0.5970	0.5189	0.0299	0.5867	0.0564	0.0115	0.0421	0.0070
dummy	Dummy Classifier	0.5970	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0060
dt	Decision Tree Classifier	0.5956	0.5784	0.4902	0.4981	0.4927	0.1570	0.1576	0.0080
ada	Ada Boost Classifier	0.5949	0.5823	0.3087	0.4993	0.3796	0.1034	0.1109	0.0310
knn	K Neighbors Classifier	0.5423	0.5226	0.3262	0.4122	0.3625	0.0145	0.0145	0.3180
svm	SVM - Linear Kernel	0.4989	0.4713	0.4982	0.2008	0.2863	-0.0014	-0.0104	0.0070

Out[15]:

▾ ExtraTreesClassifier

```
ExtraTreesClassifier(bootstrap=False, ccp_alpha=0.0, class_weight=None,
                     criterion='gini', max_depth=None, max_features='sqrt',
                     max_leaf_nodes=None, max_samples=None,
                     min_impurity_decrease=0.0, min_samples_leaf=1,
                     min_samples_split=2, min_weight_fraction_leaf=0.0,
                     monotonic_cst=None, n_estimators=100, n_jobs=-1,
                     oob_score=False, random_state=786, verbose=0,
                     warm_start=False)
```

Based on the preceding outcome, it appears that the extraTrees classification algorithm is optimal for

ne learning model for water quality analysis.

Let's proceed to train the model and evaluate its predictions.

In [17]:
```
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

Out[17]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethane
3	8.316766	214.373398	22018.417969	8.059333	356.886139	363.266510	18.436525	100.34167
4	9.092223	181.101517	17978.986328	6.546600	310.135742	398.410828	11.558279	31.99799
5	5.584086	188.313324	28748.687500	7.544869	326.678375	280.467926	8.399734	54.91786
6	10.223862	248.071732	28749.716797	7.513409	393.663391	283.651642	13.789696	84.60355
7	8.635849	203.361526	13672.091797	4.563009	303.309784	474.607635	12.363816	62.79830

Summary

Access to safe drinking water stands as a fundamental requirement for all individuals. Legally, the provision of drinking water is recognized as a basic human right. Given the multitude of factors influencing water quality, it remains a prominent research domain within the field of machine learning.