Water Quality

Data Analytics

Agenda

- A water technician came with a request to predict the drinkability of 3 batches of treated water.
- He also came with a report of past water measurements and their drinkability results.
- My task is to give him a breakdown of which measurements will contribute most to drinkability, and also to give him a percentage of success in identifying which batch of water which is not drinkable.

- Dataset source:
- https://www.kaggle.com/adityakadiwal/water-potability?select=water_potability.csv

I need to know if my treated water is drinkable or not….

My past measurements:

	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
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681736	47580.99160	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.80216	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.57822	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.86938	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.17706	7.509306	NaN	327.459761	16.140368	78.698446	2.309149	1

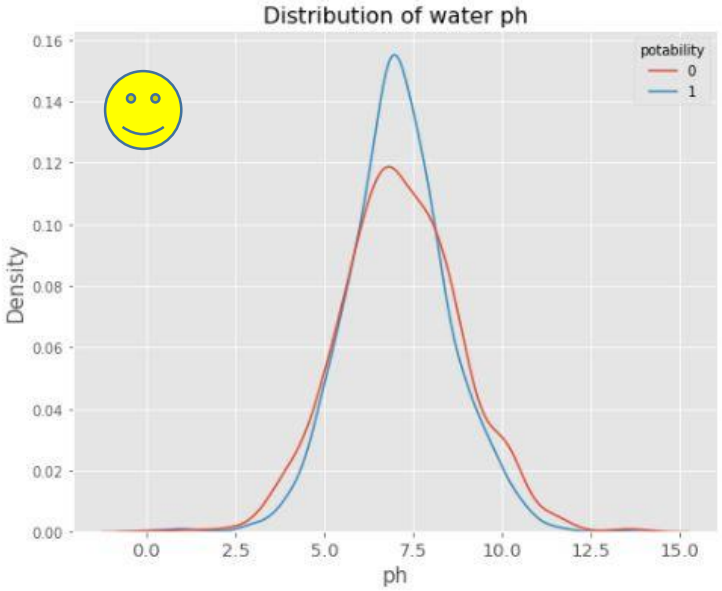
3276 rows × 10 columns

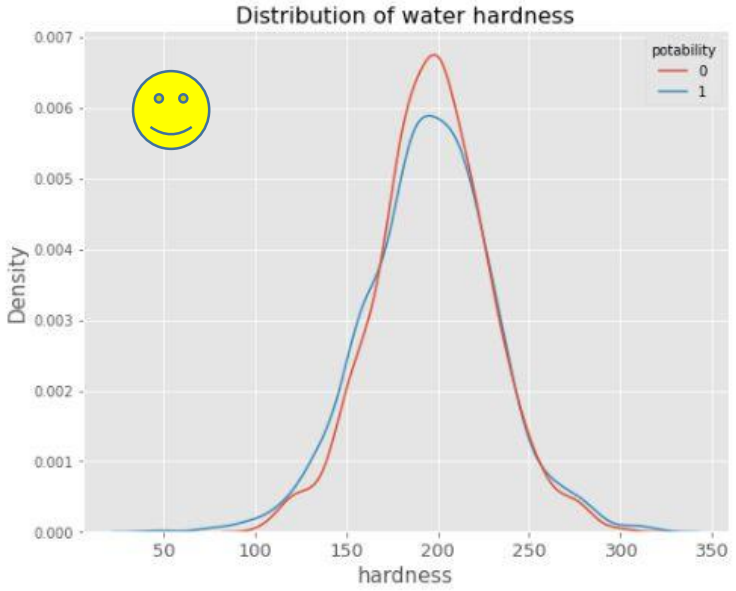
My current batch:

	batch	production_date	ph	hardness	solids	chloramines	sulfate	conductivity	organic_carbon	trihalomethanes	turbidity
1	batch_1	25/7/2021	8.316766	214.373394	22018.41744	8.059332	356.886136	363.266516	18.436525	100.341674	4.628771
2	batch_2	26/7/2021	4.668102	193.681736	47580.99160	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821
3	batch_3	27/7/2021	9.092223	181.101509	17978.98634	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075

- Base on your predictions and probability, we will decide whether to release it for consumption or not to release.
- I will also like to visualize how my measurements will look like between those which are drinkable (aka ‘potable’) and those which are not.

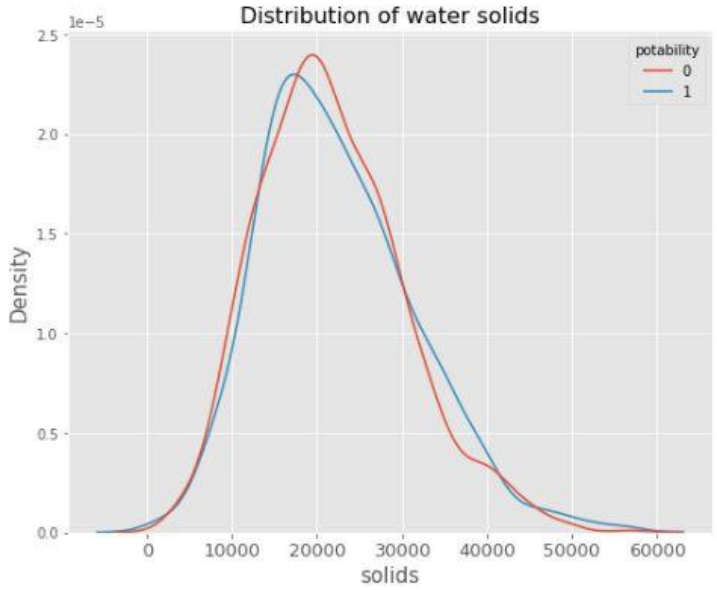
The graphs represent the characteristics of each property

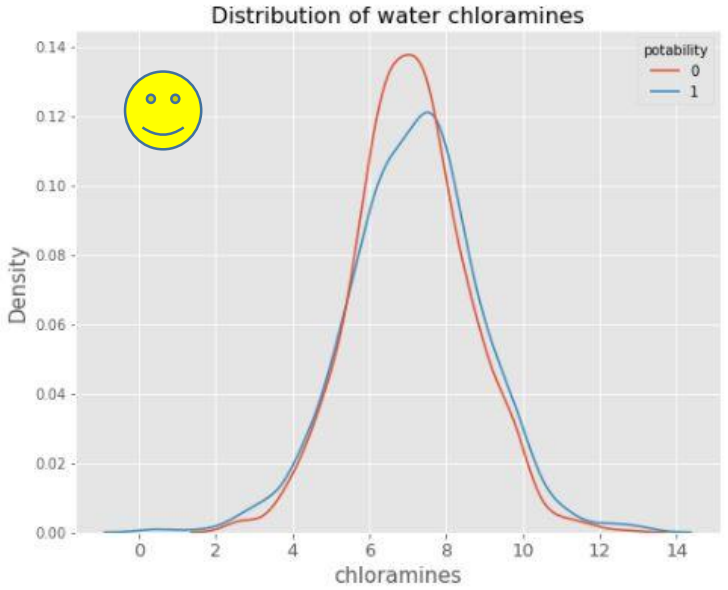


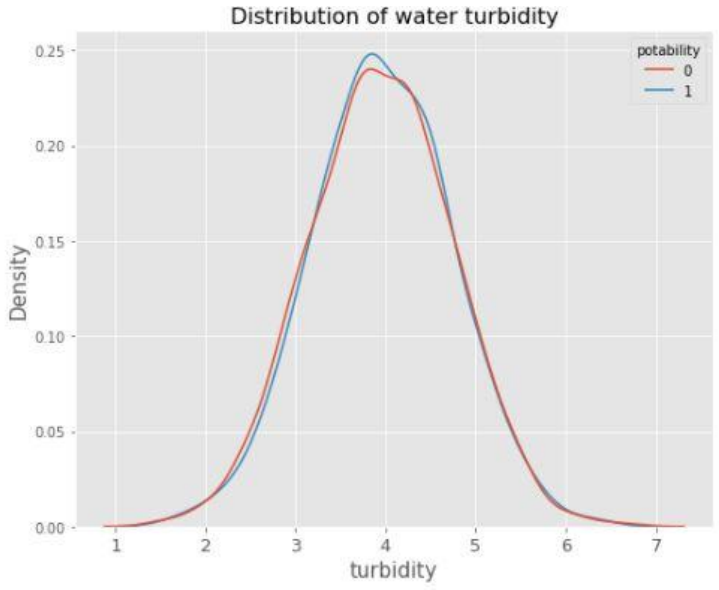


- Blue represents drinkable
- Red is not drinkable

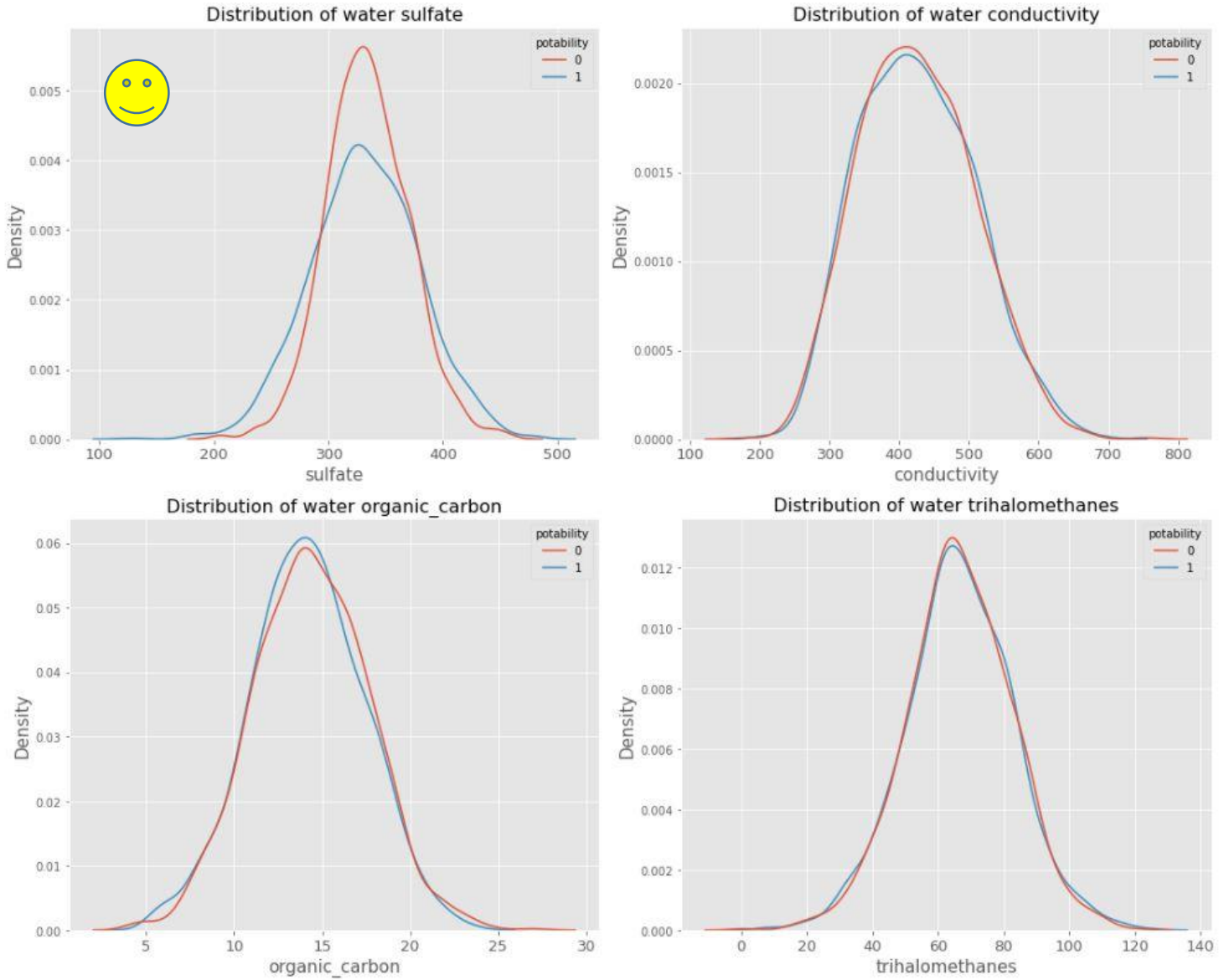
- The difference is more obvious for ph, hardness and chloramines







Out of the 9 properties, 4 of them relates more to drinkability



- Blue represents drinkable
- Red is not drinkable

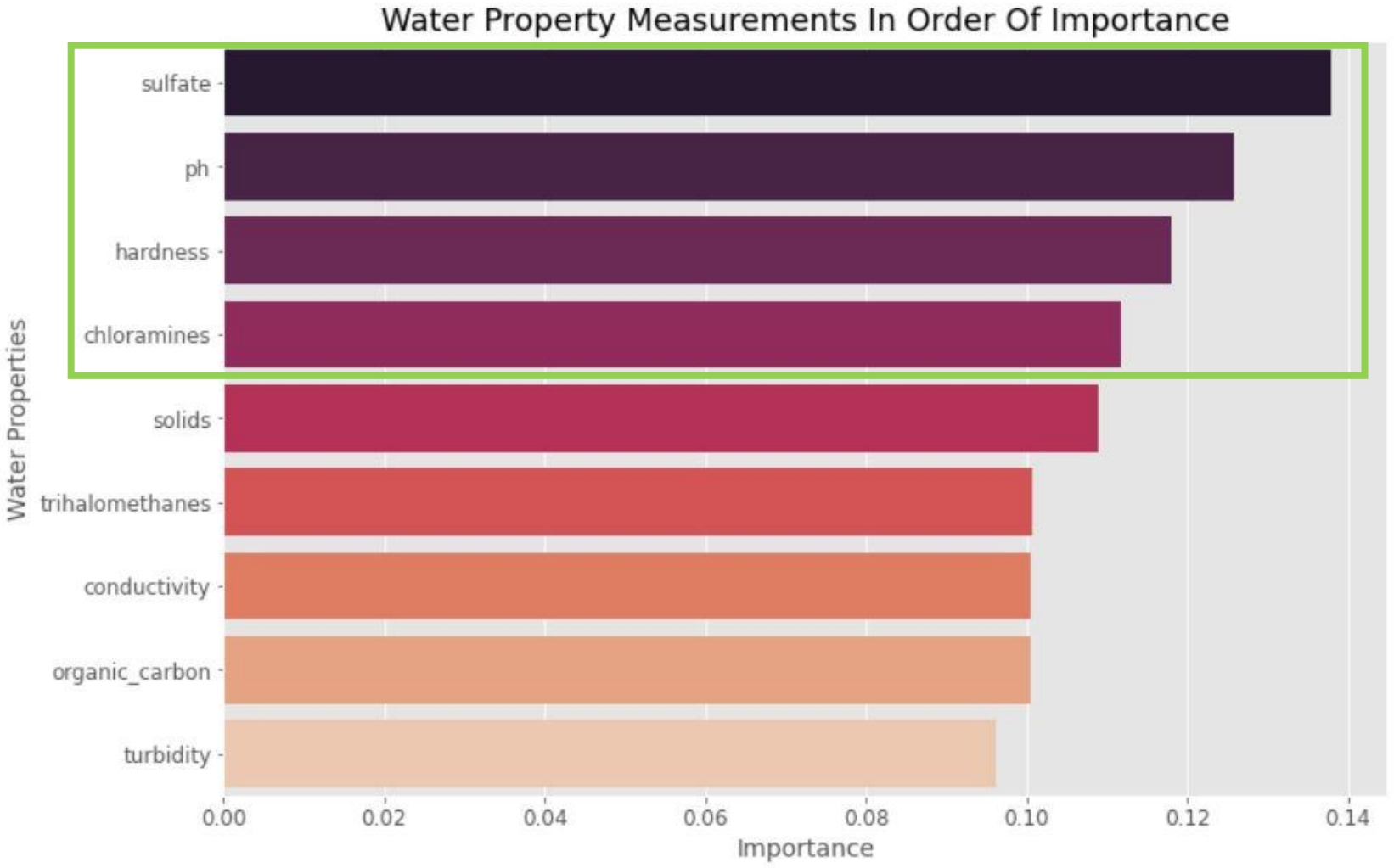
- The difference is more obvious for sulfate.

- Therefore the 4 main properties which contributed to your water's drinkability are:

- Ph
- Hardness
- Chloramines
- Sulfate

- May I know which one contributes the most?

Out of the 9 properties, 4 of them relates more to drinkability



- The first 4 properties relates the most to your water's drinkability. (green box)
- This means that if you can deal with the first 4 well enough, this will contribute the most to your final outcome.
- That sounds logical, everything is about priority and yielding the maximum results out of it!
- Can we do something prediction now?

Sure! Here are the results.

In [31]:
```
sample = pd.read_csv('water_samples.csv')
sample.index = sample.index + 1
```

In [32]:
```
sample
```

Out[32]:

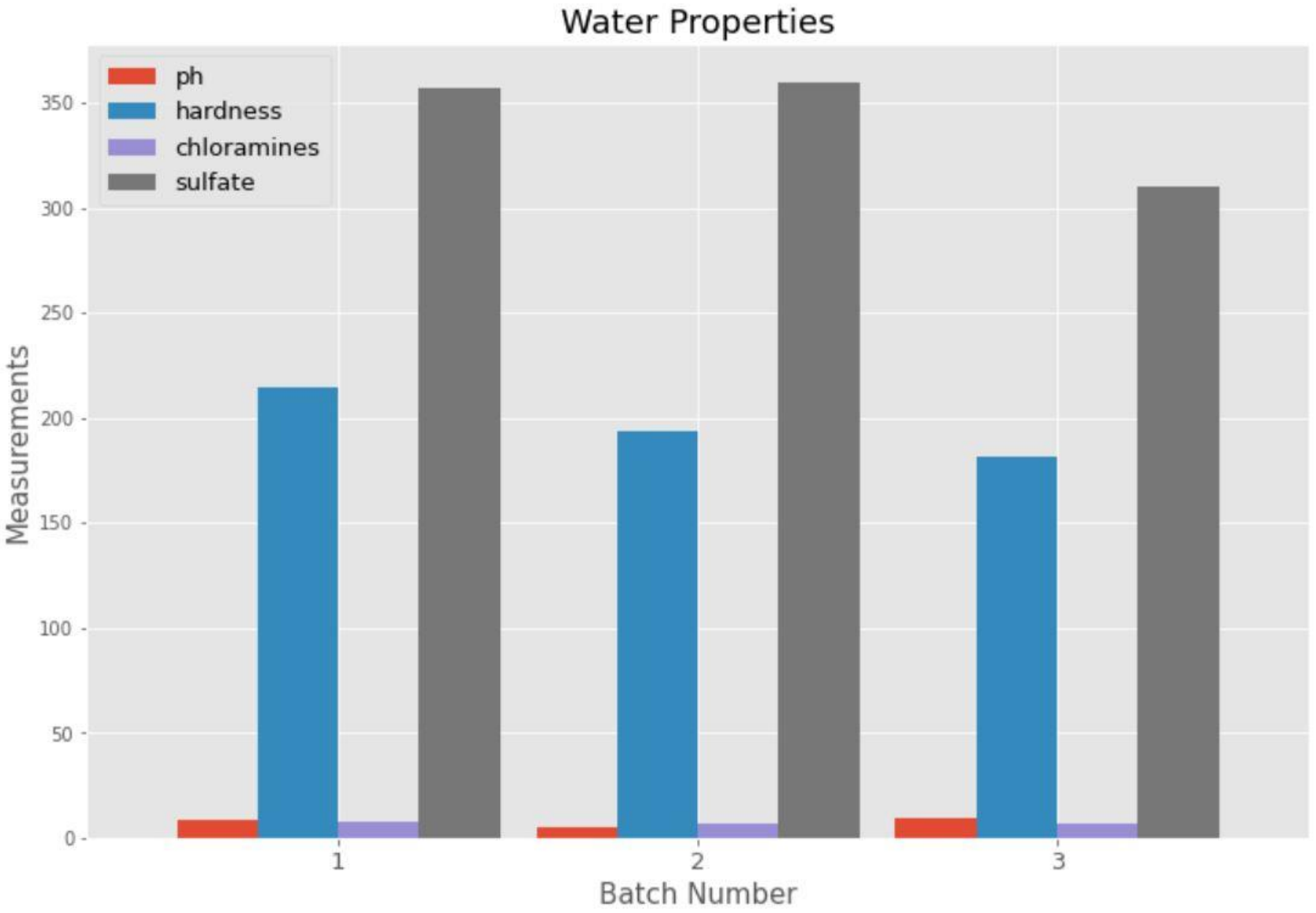
	batch	production_date	ph	hardness	solids	chloramines	sulfate	conductivity	organic_carbon	trihalomethanes	turbidity
1	batch_1	25/7/2021	8.316766	214.373394	22018.41744	8.059332	356.886136	363.266516	18.436525	100.341674	4.628771
2	batch_2	26/7/2021	4.668102	193.681736	47580.99160	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821
3	batch_3	27/7/2021	9.092223	181.101509	17978.98634	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075

In [33]:
```
s1 = sample[['ph','hardness','chloramines','sulfate']]
knn.predict(s1)
```

Out[33]:
```
array([0, 1, 0], dtype=int64)
```

- Batch 2 (red box) is drinkable, whereas batch 1 & 3 are not drinkable. ('0' means undrinkable)
- I see, that's not good.
- Can I visualize all the 4 properties and comparing between the 3 batches as well? This will give me an idea how these will look like.

Yes. We can do that too!



- As you can see, batch 2 is drinkable.
- For batch 1, Sulfate is not an issue as it is similar to batch 2. However Hardness, Chloramines and ph is the problem.
- Regarding batch 3, chloramines is not the problem, but the rest are.

- Yes! Achieving the right balance is the key. Now we know which property to take priority on.

- As our goal is to capture water which are not drinkable, may I know what are our chances of getting our guesses correct?

I hope this report answers your question…

```
In [30]:  1 print('Classification report: \n',report)
```

```
Classification report:
                precision    recall  f1-score   support

           0       0.80      0.85      0.83      1998
           1       0.74      0.67      0.71      1278

    accuracy                           0.78      3276
   macro avg       0.77      0.76      0.77      3276
weighted avg       0.78      0.78      0.78      3276
```

- Of all the undrinkable water (class 0), we have an 85% chance of getting our guess correct. (red box)
- We also have an overall accuracy score of 78%
- Hope that answers your question!