In [13]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Exploring Data

In [5]:
```
data = pd.read_csv('/content/water_potability.csv')
data
```

Out[5]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.
...	...	...	...	...	...	...	
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.
3273	9.419510	175.762646	33155.578218	7.350233	NaN	432.044783	11.
3274	5.126763	230.603758	11983.869376	6.303357	NaN	402.883113	11.
3275	7.874671	195.102299	17404.177061	7.509306	NaN	327.459760	16.

3276 rows × 10 columns

In [21]:
```
data.head()
```

Out[21]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carb
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558

In [7]:
```
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
```

In [15]:
```
data.describe().T
```

Out[15]:

	count	mean	std	min	25%	50%
ph	2785.0	7.080795	1.594320	0.000000	6.093092	7.036752
Hardness	3276.0	196.369496	32.879761	47.432000	176.850538	196.967627
Solids	3276.0	22014.092526	8768.570828	320.942611	15666.690297	20927.833607
Chloramines	3276.0	7.122277	1.583085	0.352000	6.127421	7.130299
Sulfate	2495.0	333.775777	41.416840	129.000000	307.699498	333.073546
Conductivity	3276.0	426.205111	80.824064	181.483754	365.734414	421.884968
Organic_carbon	3276.0	14.284970	3.308162	2.200000	12.065801	14.218338
Trihalomethanes	3114.0	66.396293	16.175008	0.738000	55.844536	66.622485
Turbidity	3276.0	3.966786	0.780382	1.450000	3.439711	3.955028
Potability	3276.0	0.390110	0.487849	0.000000	0.000000	0.000000

Pre-processing

In [9]:
```
#percentage of missing values in the dataset
data.isna().mean() * 100
```

Out[9]:
```
ph                14.987790
Hardness           0.000000
Solids             0.000000
Chloramines        0.000000
Sulfate           23.840049
Conductivity       0.000000
Organic_carbon     0.000000
Trihalomethanes    4.945055
Turbidity          0.000000
Potability         0.000000
dtype: float64
```

In [17]:
```
#checks missing values in the dataset and gives the sum
data.isnull().sum()
```

Out[17]:
```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

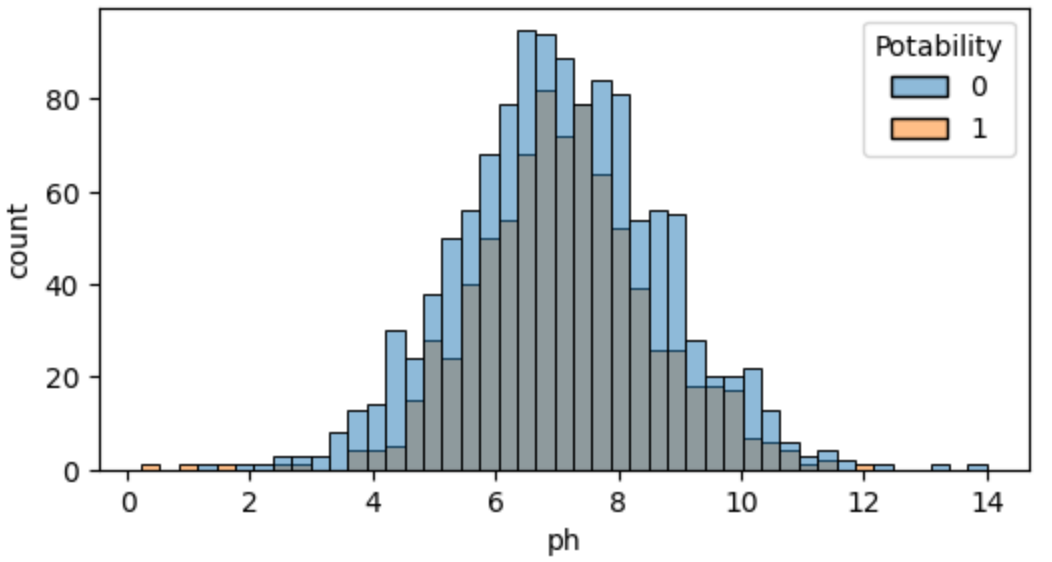
In [22]:
```
#eliminates missing values
data = data.dropna()
data.isnull().sum()
```

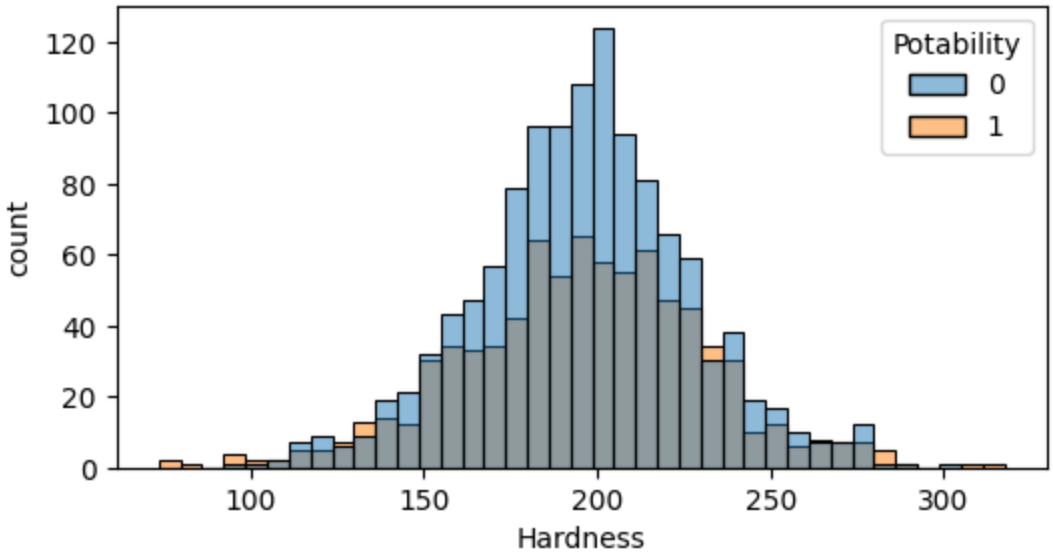
Out[22]:
```
ph                0
Hardness          0
Solids            0
Chloramines       0
Sulfate           0
Conductivity      0
Organic_carbon    0
Trihalomethanes   0
Turbidity         0
Potability        0
dtype: int64
```

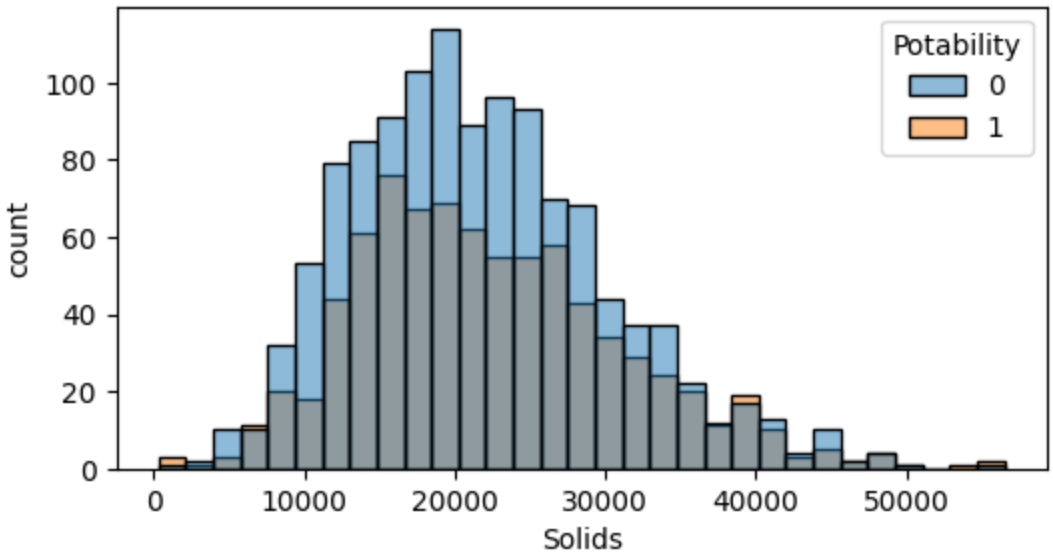
Plots

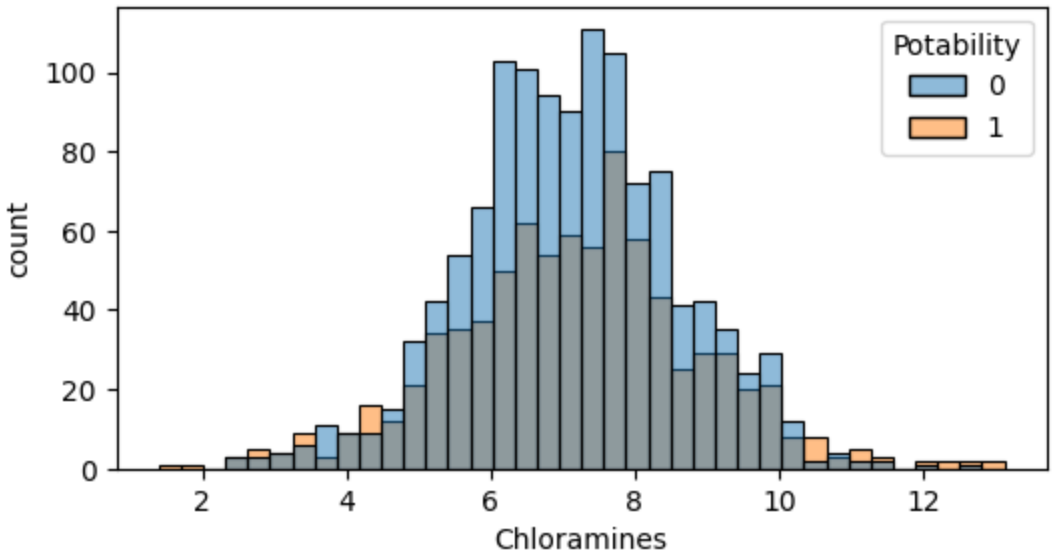
In [23]:
```
def histplot(var): #define a histogram plot that containts a var argument, that var
    plt.figure(figsize = (6,3))
    sns.histplot(data = data, x = data[var], hue = data.Potability)
    plt.xlabel(var)
    plt.ylabel("count")
    plt.show()

numeric_vars = ["ph","Hardness","Solids","Chloramines","Sulfate","Conductivity","Or
for n in numeric_vars: #loops every numeric_vars in the array to create a histogram
```

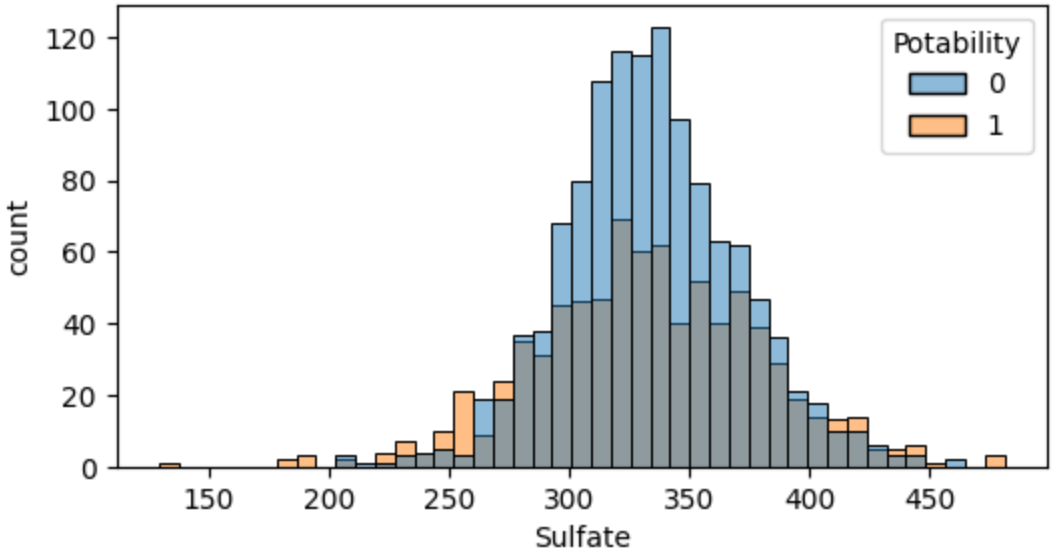
```
#generates every histoplot on every numeric_vars that affect potability
```

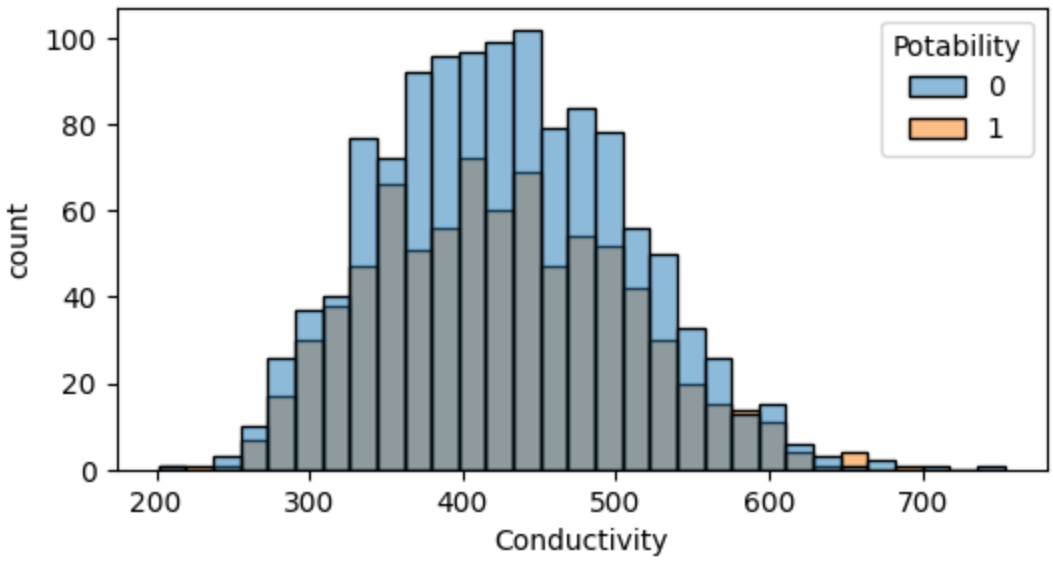


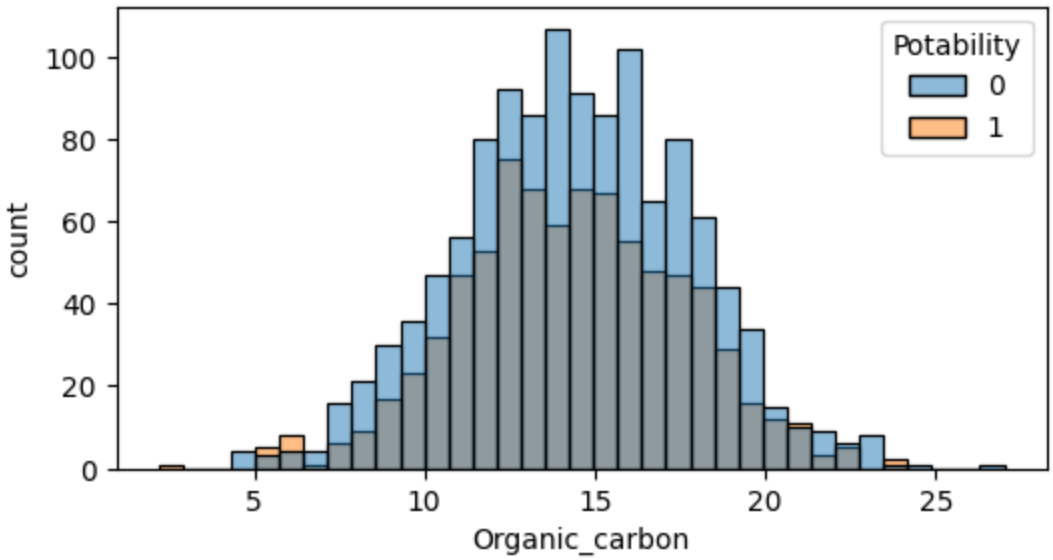


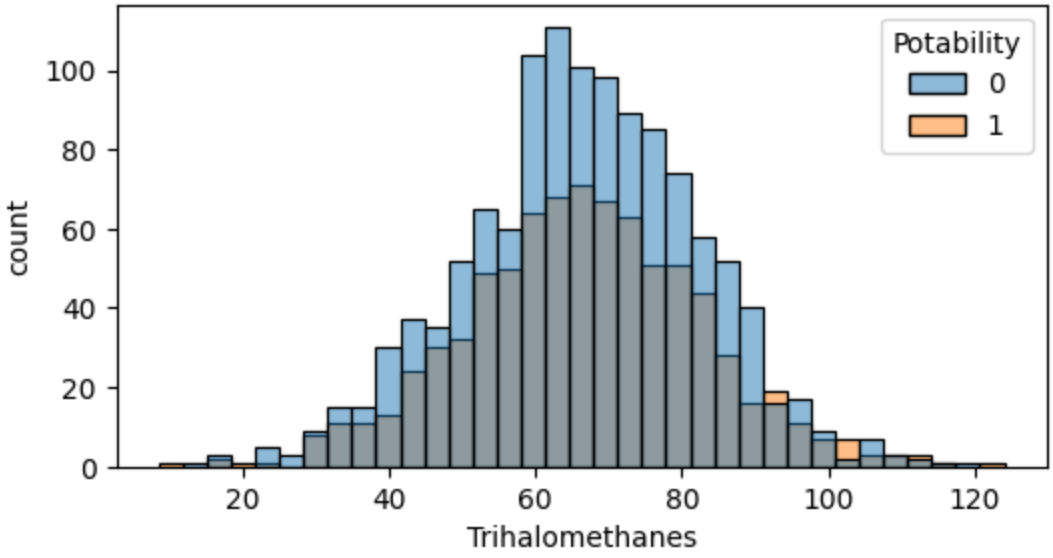


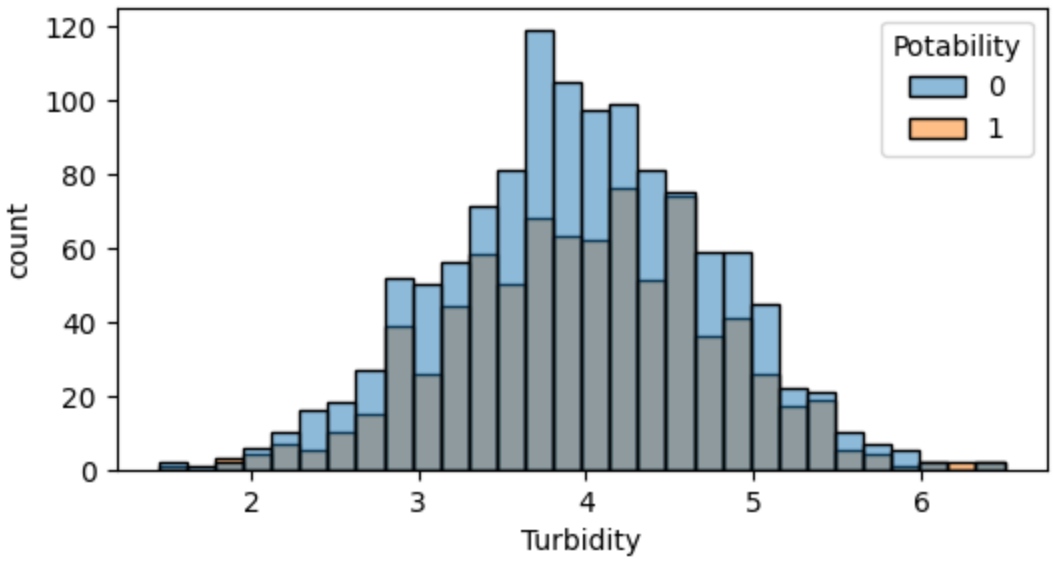

Potability
0
1
count
Chloramines
0
20
40
60
80
100
2
4
6
8
10
12


Potability
0
1
count
Sulfate
0
20
40
60
80
100
120
150
200
250
300
350
400
450


Potability
0
1
count
Conductivity
0
20
40
60
80
100
200
300
400
500
600
700







Correlation

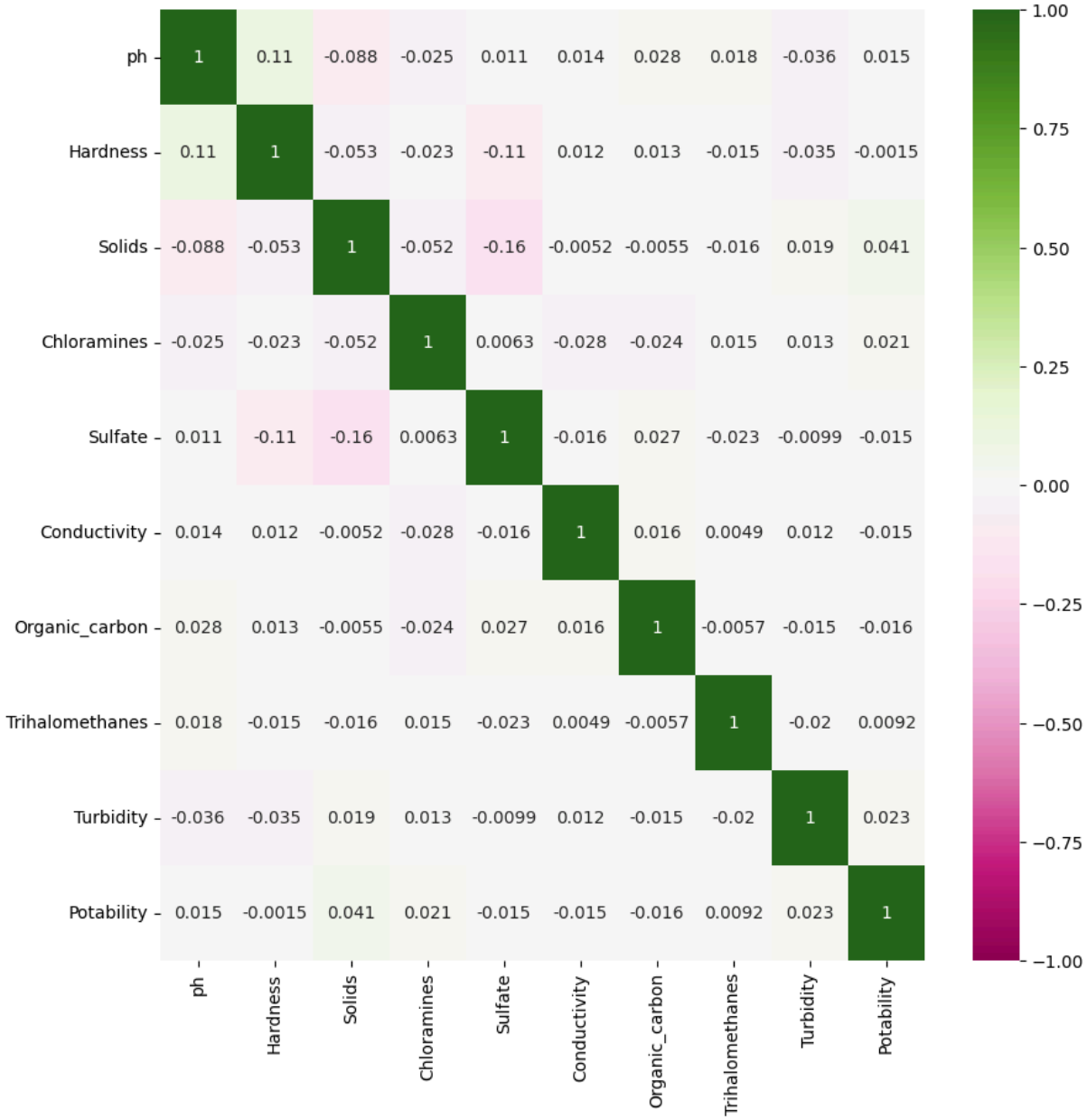
In [25]:
```
corr = data.corr()
corr
```

Out[25]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Org
ph	1.000000	0.108948	-0.087615	-0.024768	0.010524	0.014128	
Hardness	0.108948	1.000000	-0.053269	-0.022685	-0.108521	0.011731	
Solids	-0.087615	-0.053269	1.000000	-0.051789	-0.162769	-0.005198	
Chloramines	-0.024768	-0.022685	-0.051789	1.000000	0.006254	-0.028277	
Sulfate	0.010524	-0.108521	-0.162769	0.006254	1.000000	-0.016192	
Conductivity	0.014128	0.011731	-0.005198	-0.028277	-0.016192	1.000000	
Organic_carbon	0.028375	0.013224	-0.005484	-0.023808	0.026776	0.015647	
Trihalomethanes	0.018278	-0.015400	-0.015668	0.014990	-0.023347	0.004888	
Turbidity	-0.035849	-0.034831	0.019409	0.013137	-0.009934	0.012495	
Potability	0.014530	-0.001505	0.040674	0.020784	-0.015303	-0.015496	

In [55]:
```
plt.figure(figsize = (10,10))
sns.heatmap(corr, cmap='PiYG', vmin=-1, vmax=1, annot=True,)
```

Out[55]: <Axes: >


ph 1 0.11 -0.088 -0.025 0.011 0.014 0.028 0.018 -0.036 0.015
Hardness 0.11 1 -0.053 -0.023 -0.11 0.012 0.013 -0.015 -0.035 -0.0015
Solids -0.088 -0.053 1 -0.052 -0.16 -0.0052 -0.0055 -0.016 0.019 0.041
Chloramines -0.025 -0.023 -0.052 1 0.0063 -0.028 -0.024 0.015 0.013 0.021
Sulfate 0.011 -0.11 -0.16 0.0063 1 -0.016 0.027 -0.023 -0.0099 -0.015
Conductivity 0.014 0.012 -0.0052 -0.028 -0.016 1 0.016 0.0049 0.012 -0.015
Organic_carbon 0.028 0.013 -0.0055 -0.024 0.027 0.016 1 -0.0057 -0.015 -0.016
Trihalomethanes 0.018 -0.015 -0.016 0.015 -0.023 0.0049 -0.0057 1 -0.02 0.0092
Turbidity -0.036 -0.035 0.019 0.013 -0.0099 0.012 -0.015 -0.02 1 0.023
Potability 0.015 -0.0015 0.041 0.021 -0.015 -0.015 -0.016 0.0092 0.023 1
ph
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity
Potability
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00