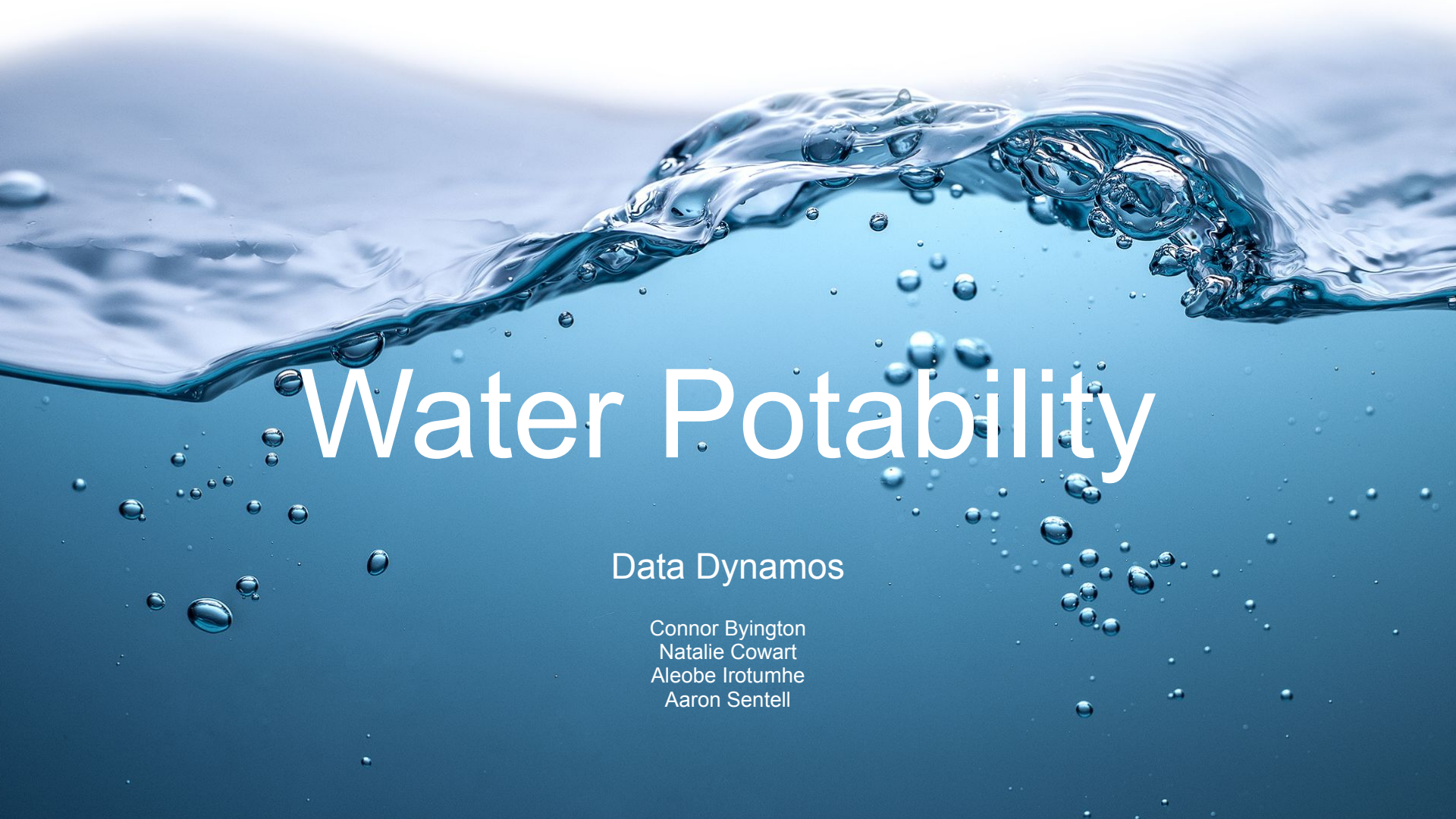


Water Potability

Data Dynamos

Connor Byington
Natalie Cowart
Aleobe Irotumhe
Aaron Sentell

Project Importance

🤷‍♂️ *Why are we here?*
- Potable water is essential for maintaining biological processes.

🤷‍♂️ *Why should I care?*
- Less than 1% of water on Earth is potable, most living things require potable water.

🤷‍♂️ *What should we do about it?*
- Continue to analyze water features to maintain water supplies for industrial and civil applications.

Initial Research Questions

🤷‍♀️ Do high sulfate levels indicate a high chance of non-potability?

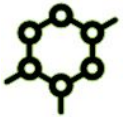
🤷‍♀️ Does water hardness have an inverse or direct relationship with turbidity?

🤷‍♀️ What is the greatest indicator of water non-potability?

Water Evaluation

Chemistry Indicator	Description	EPA Acceptable Range and/or Maximum
pH		6.5 - 8.5
Hardness		180-210 ppm
Solids (Total Dissolved Solids-TDS)		500 ppm
Chloramines		4 ppm
Sulfates		400 ppm

Water Evaluation Cont'd

Chemistry Indicator	Description	EPA Acceptable Range and/or Maximum
Conductivity		N/A (WHO recommended value is below 400 µS/cm)
Total Organic Carbon (TOC)		< 4 ppm in source water
Trihalomethanes (THMs)		80 ppm
Turbidity		N/A (WHO recommended value is below 5.00 NTU)
Potability		N/A (Potable water = 1, Non-potable water = 0)

Initial ETL

- Imported Dependencies
  - matplotlib for basic visualizations and graph labels
  - pandas for data cleansing and basic statistical summary
  - numpy for multi-dimensional arrays
  - seaborn for advanced statistical visualizations
- Read in .csv with raw data from Kaggle to a DataFrame using pandas

In [1]:
```
# Dependencies and Setup
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns

# Study data files
readin = "../dataset/water_potability.csv"

# create a dataframe
OG_df = pd.read_csv(readin)
OG_df.head()
```

Data Transformation

- Drop rows that contain N/A values for any water chemistry component
- Divide solids column by 100 (scale only goes up to 2,000)

```
# Drop rows w/ blank values
df2 = OG_df.dropna(axis=0,how='any')
df2.head(10)
```

```
# Fix weird solids column error (all values multiplied by 100)
df2['Solids'] = df2['Solids'].div(100)
df2.head()
```

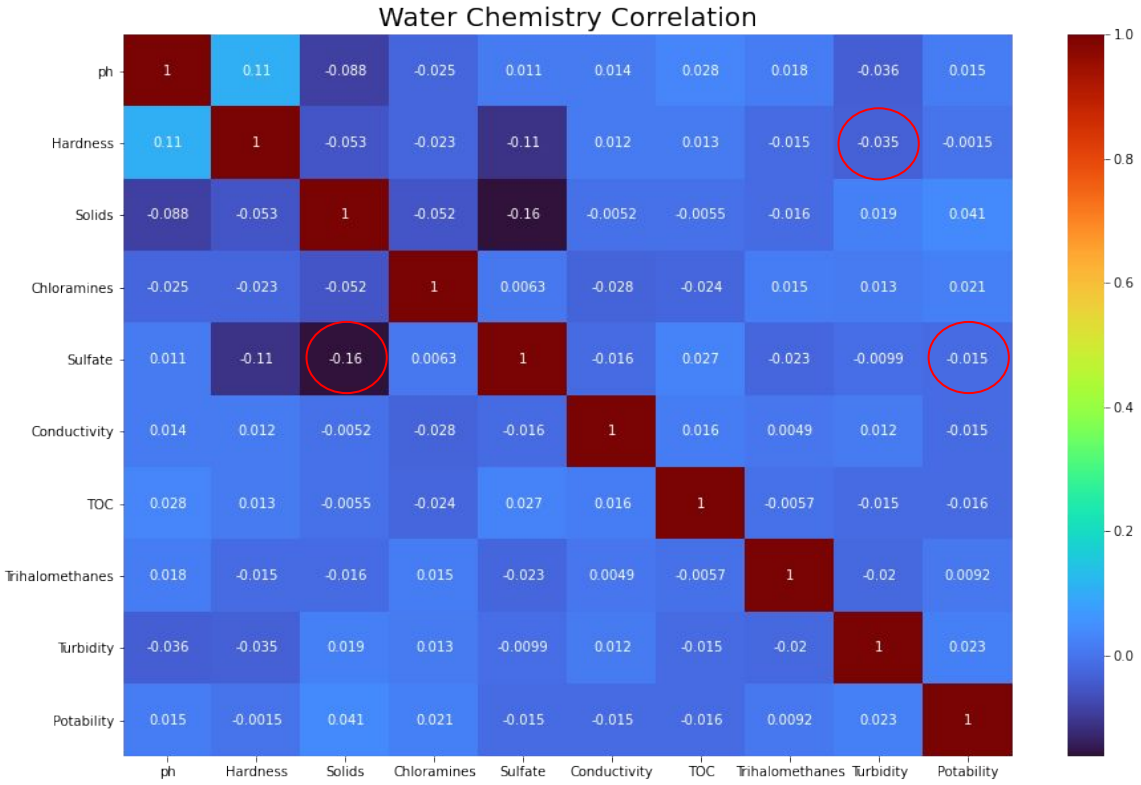
Before

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0

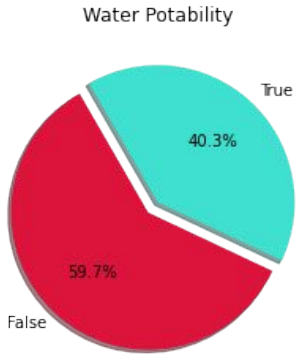
After

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
3	8.316766	214.373394	220.184174	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	179.789863	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

What Correlation?

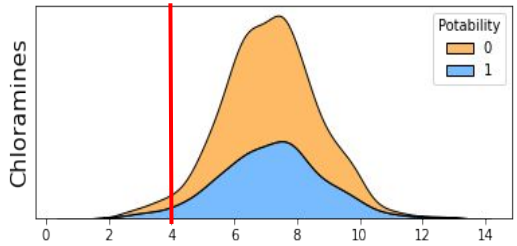


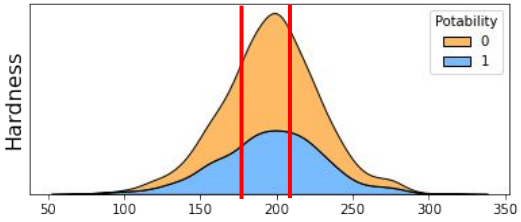
- Weak/no correlation between any of the water chemistry components and potability

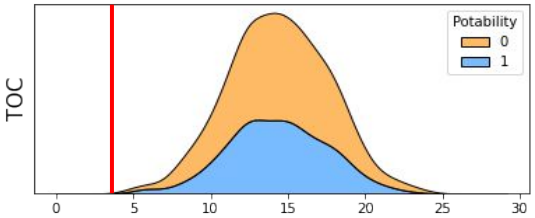


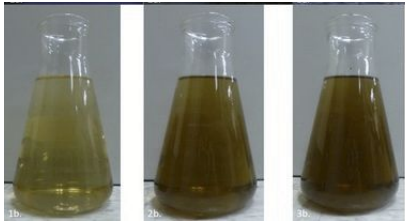
Distribution Analysis

- Distributions picked by feature
- *0: Non-Potable, 1: Potable*
- Chloramines- data outside acceptable range
- Hardness- data outside acceptable range (180-210 ppm)
- TOC- Well outside of acceptable range (< 2-4 ppm)
- Whether water represented to be potable is actually safe for consumption









Are Any Samples Safe for Consumption?

- Original data frame was filtered by EPA and WHO water feature parameters
- Contrasting water standards nation to nation
- Garbage in, garbage out
- Cultures adapting to different water conditions

In [19]:
```
narrowdf = OG_df.loc[(OG_df["ph"] <= 8.5) & (OG_df["ph"] >= 6.5) & \
                     (OG_df["Hardness"] >= 180) & (OG_df["Hardness"] <= 210) & \
                     (OG_df["Solids"]<= 500) & \
                     (OG_df["Chloramines"] <= 4) & \
                     (OG_df["Sulfate"] <= 400) & \
                     (OG_df["Conductivity"] <= 400) & \
                     (OG_df["Organic_carbon"] <= 4) & \
                     (OG_df["Trihalomethanes"] <= 80) & \
                     (OG_df["Turbidity"] <= 5)]
narrowdf.head(10)
```

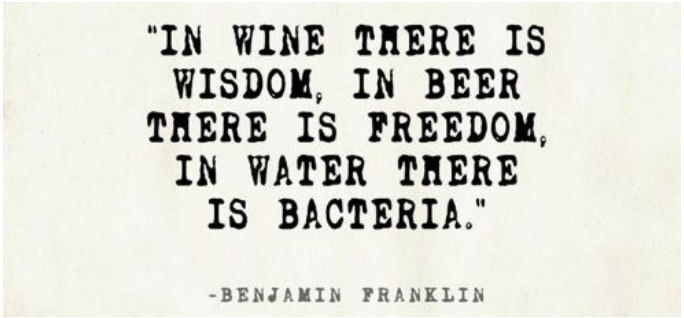
Out[19]:

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability

In [ ]:

What Did We Learn?

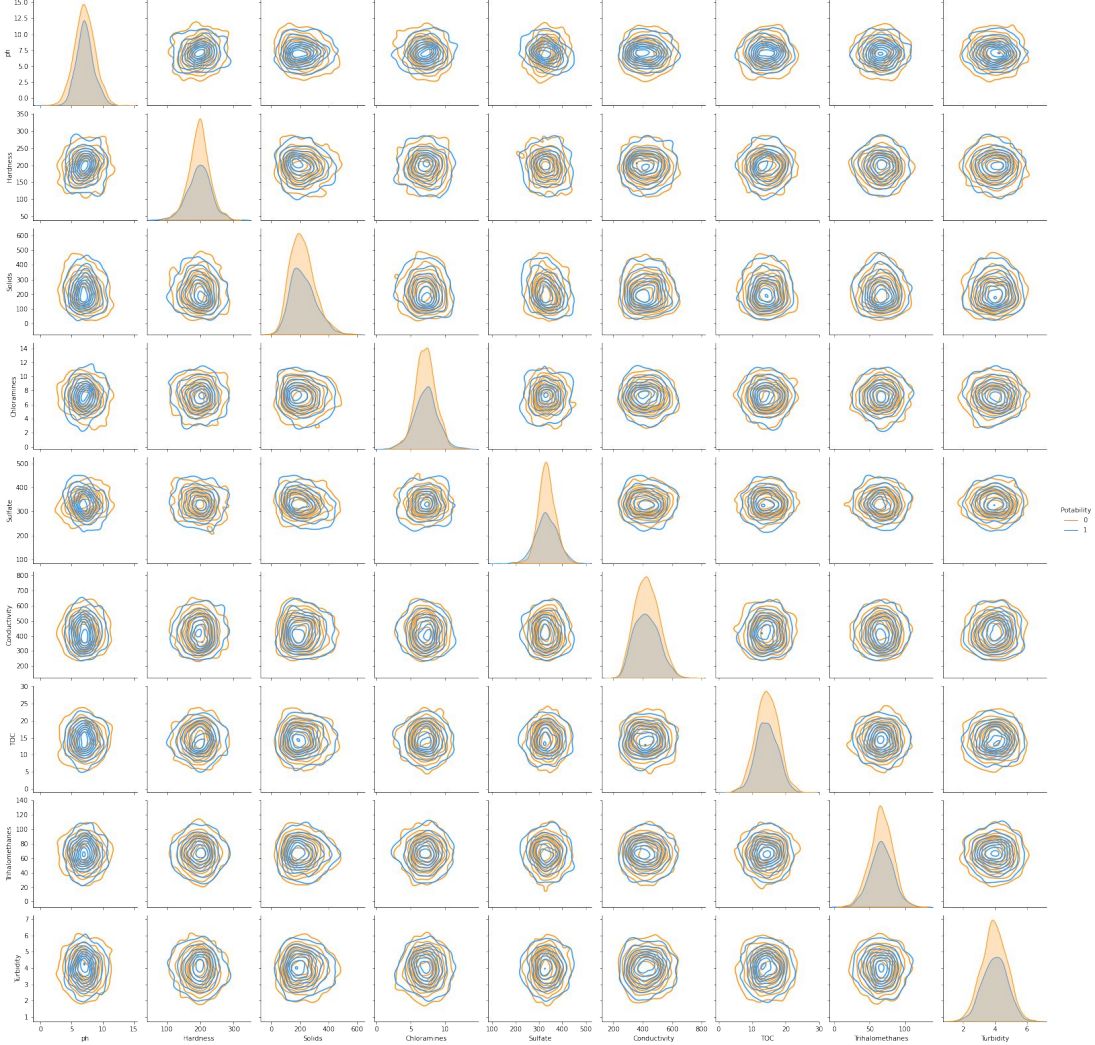
- Appearance of data reputability ≠ actual reputability
- Use accepted benchmarks / relevant standards to verify quality of data and assertions
- A bad data set can teach you as many things as a good data set
- The greatest indicator of water potability is likely Fecal Coliform *not* the features included in our dataset



Sources

- Acceptable drinking water parameters: https://epa.gov
- Backup standards for non-EPA regulated water components: https://who.int
- Potable Water Dataset: https://www.kaggle.com/adityakadiwal/water-potability
- Slide Aesthetic: https://www.revivedwater.eu/
- Benjamin Franklin Quote: https://me.me/i/in-wine-there-is-wisdomi-in-beer-there-is-treedomi-15067068
- Fecal Coliform Primary Indicator of Potability: https://www.water-research.net/index.php/water-testing/bacteria-testing/fecal-coliform-bacteria
- Hardness: https://www.realtor.com/advice/home-improvement/what-is-hard-water/
- TOC: https://science.nd.edu/undergraduate/minors/sustainability/capstone-projects/2014/elser/

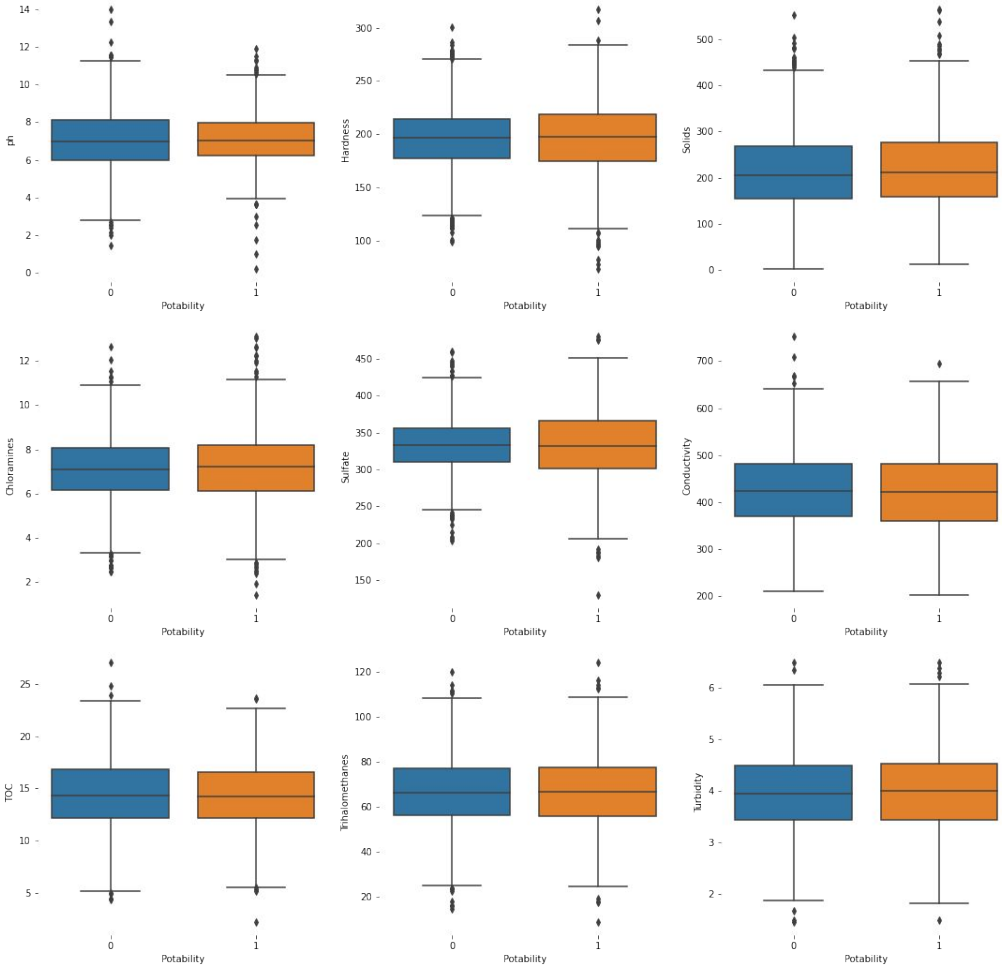
Kernel Density Estimation

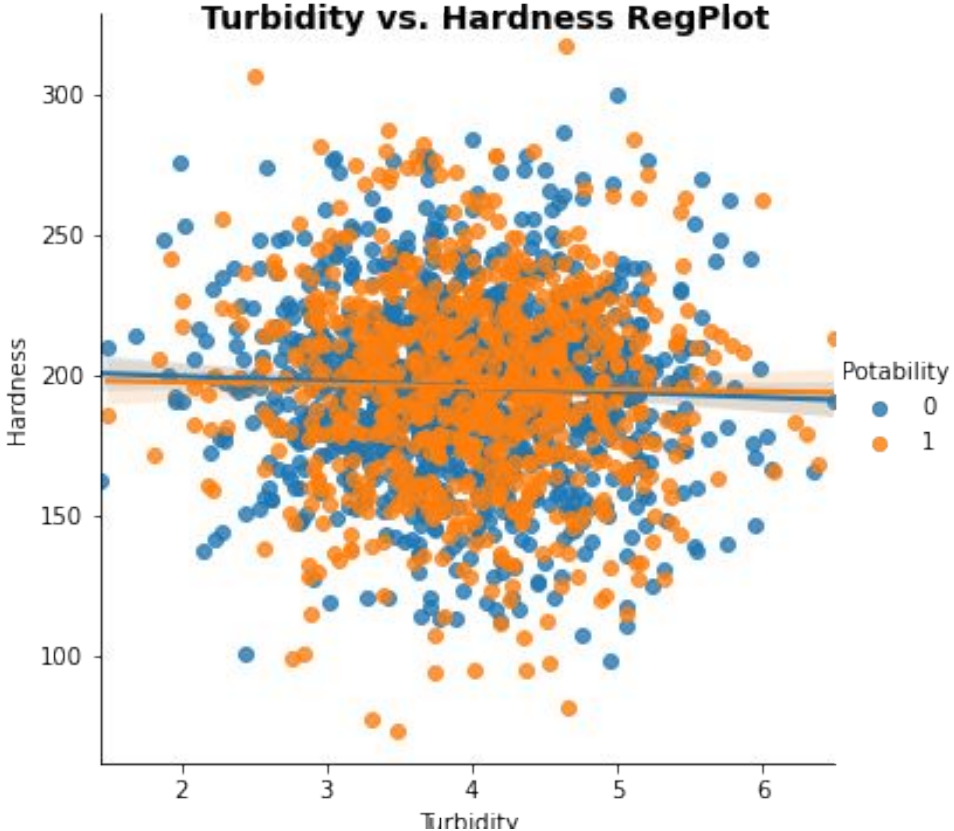


Correlation of Potability:

	Features	Correlation
1	Hardness	0.001505
7	Trihalomethanes	0.009244
0	ph	0.014530
4	Sulfate	0.015303
5	Conductivity	0.015496
6	Organic_carbon	0.015567
3	Chloramines	0.020784
8	Turbidity	0.022682
2	Solids	0.040674

Feature Boxplot by Potability Feature




Turbidity vs. Hardness RegPlot
Hardness
300
250
200
150
100
2
3
4
5
6
Turbidity
Potability
0
1