EDA COURSE PROJECT

Description of Dataset

The dataset, “Water Quality “, which I selected is related with drinking water potability. Drinking water potability is an international crucial issue for human rights and health quality. Importance of water quality leads us to sustain some standards which is detected by WHO (World Health Organization).

First of all, pH value is acid-base balance evaluation metric. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. Similar with pH value, the dataset involves 8 more features which are hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes and turbidity.

Hardness is defined as the capacity of water to precipitate soup caused by Calcium and Magnesium.

Total dissolved solids (calcium, potassium, sodium and so on), TDS, has a desirable limit for 500 mg/l and maximum limit is 1000 mg/l.

Chloramines are the major disinfectants used in public water system. Chloramine level up to 4 mg/l are considered safe in drinking water.

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

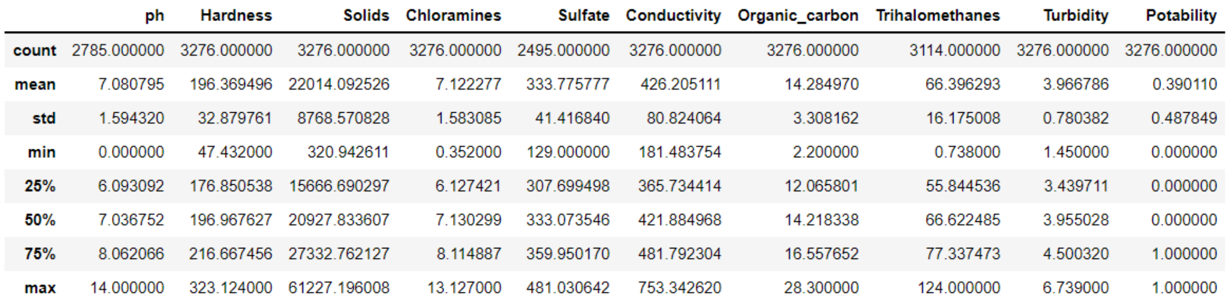
Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water. According to WHO standards, conductivity value should not exceeded 400 μS/cm.

Organic carbon is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

Trihalomethanes (THM) are chemicals which may be found in water treated with chlorine. THM levels up to 80 ppm is considered safe in drinking water.

The turbidity of water depends on the quantity of solid matter present in the suspended state. The average turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

Finally, target is specified as “potability” in binary format. According to those 9 features, water’s quality is expressed as potable (1) and non-potable (0).



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

Now, I can see first insight about features and target, above. Features' ranges, means, standard deviations and quartiles can be observed.

Initial Plan for Data Exploration

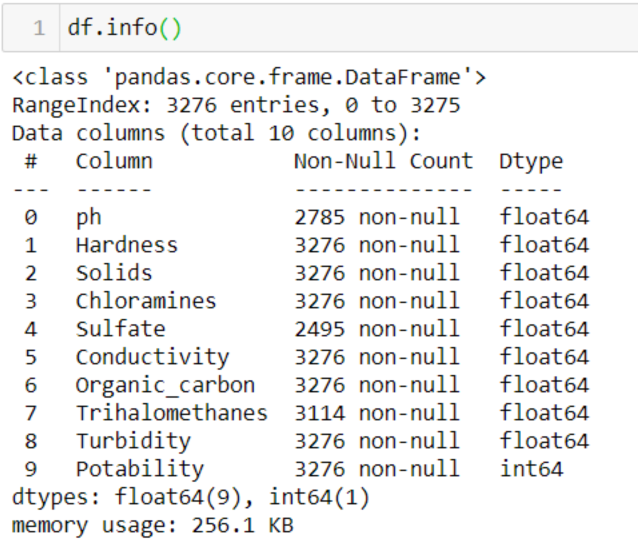
In the beginning, I aim to observe that how many null cells dataset has, then clear the unnecessary part of them. Similarly, missing values and outliers are also an issue that should be handled. I will pick an appropriate method according to dataset's attribute.

Afterwards, I take a closer look to relations relation between features and target, their weights on each other by visual items, then we do some feature engineering to get clearer dataset in order to use it for after courses.

Finally, hypothesis testing will be last step to end up the report.

Dataset Overview

Water Potability dataset includes 3276 rows (entries) and 10 columns (features and target). As table illustrates, there are some null values that should be handled and the majority comes from "ph" feature.

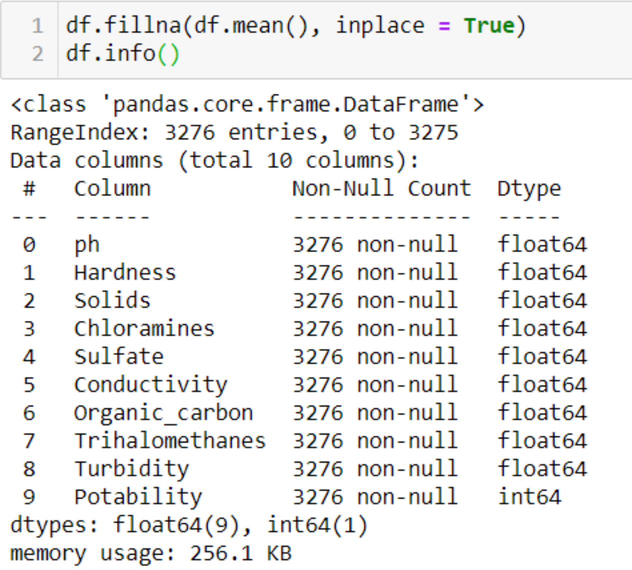


```
df.info()
```

```
<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
```

Data Cleaning

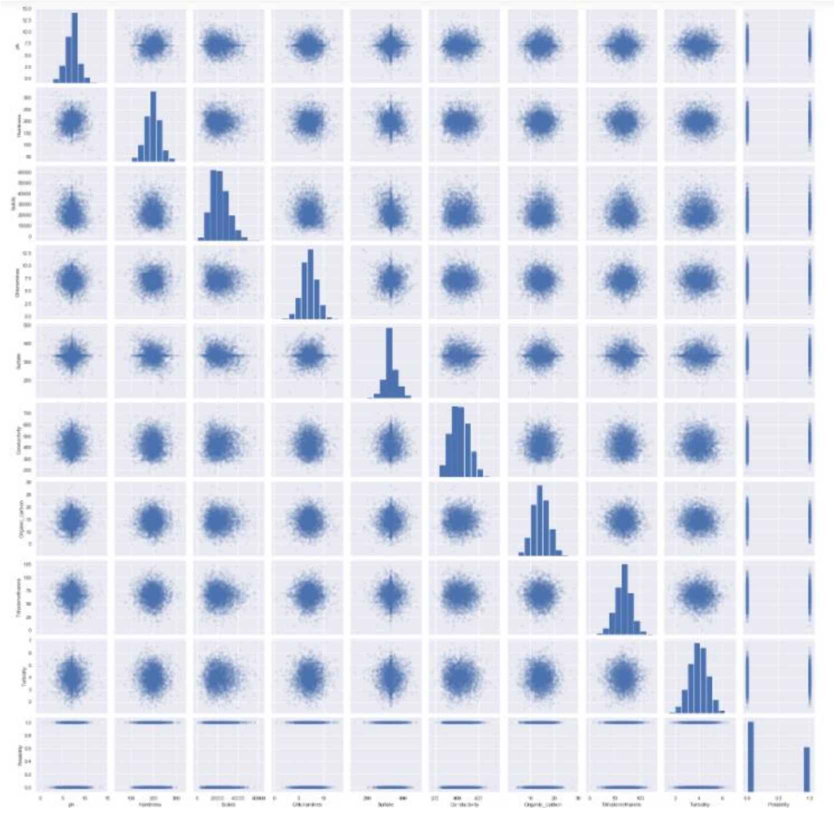
For a beginning of cleaning, I tried to mask null values but skewness of features are very low, there should be another way to solve missing value problem. Therefore, I put each columns' mean instead of the null value.



```
df.fillna(df.mean(), inplace = True)
df.info()
```

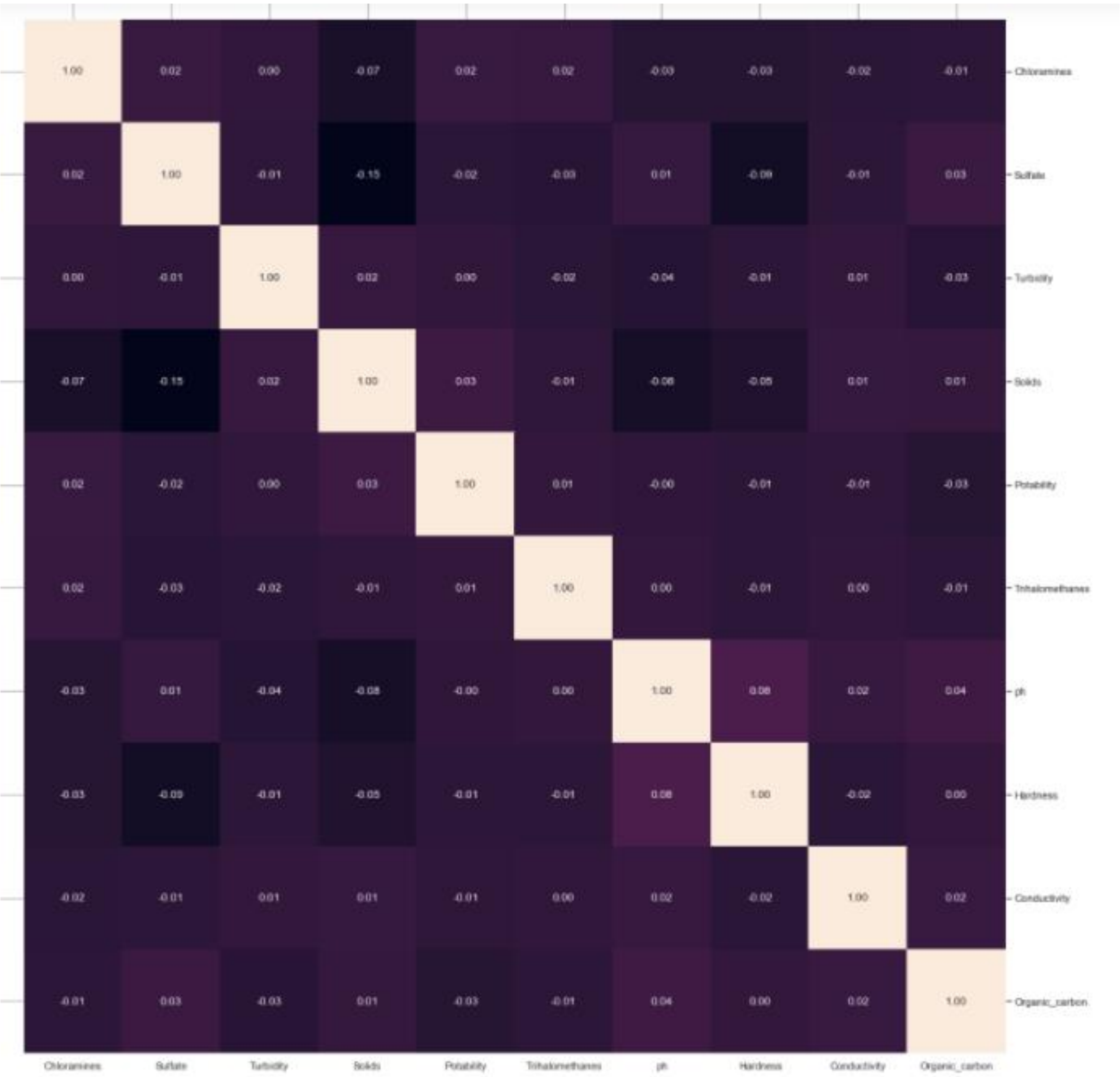
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   ph               3276 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          3276 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3276 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
```

Pair Plot of Features



As we can observe that, 9 of the features have majority of the median values. Moreover, their distributions are fine to move forward.

On the other hand, I want to use correlation matrix and its heat map to visualize features' relationships in a different way. In the heat map, when correlation value is closer to 1, it means that there is a strong tie each other, if correlation value is closer to 0, vice versa.

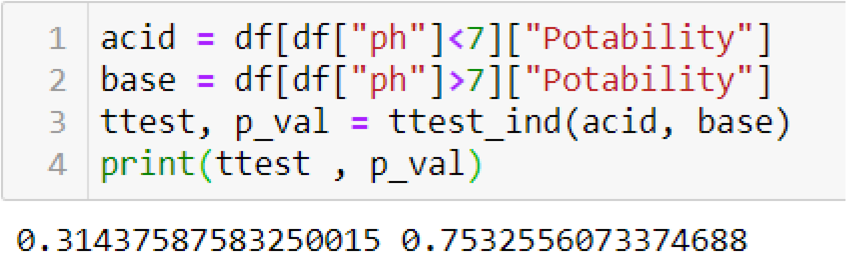


	Chloramines	Sulfate	Turbidity	Solids	Potability	Trihalomethanes	ph	Hardness	Conductivity	Organic_carbon
Chloramines	1.00	0.02	0.00	-0.07	0.02	0.02	-0.03	-0.03	-0.02	-0.01
Sulfate	0.02	1.00	-0.01	-0.15	-0.02	-0.03	0.01	-0.09	-0.01	0.03
Turbidity	0.00	-0.01	1.00	0.02	0.00	-0.02	-0.04	-0.01	0.01	-0.03
Solids	-0.07	-0.15	0.02	1.00	0.03	-0.01	-0.08	-0.05	0.01	0.01
Potability	0.02	-0.02	0.00	0.03	1.00	0.01	-0.00	-0.01	-0.01	-0.03
Trihalomethanes	0.02	-0.03	-0.02	-0.01	0.01	1.00	0.00	-0.01	0.00	-0.01
ph	-0.03	0.01	-0.04	-0.08	-0.00	0.00	1.00	0.08	0.02	0.04
Hardness	-0.03	-0.09	-0.01	-0.05	-0.01	-0.01	0.08	1.00	-0.02	0.00
Conductivity	-0.02	-0.01	0.01	0.01	-0.01	0.00	0.02	-0.02	1.00	0.02
Organic_carbon	-0.01	0.03	-0.03	0.01	-0.03	-0.01	0.04	0.00	0.02	1.00

The heat map shows that the features have no similarity or any relationship between each other. Therefore, feature reduction is not a good option to develop data analysis, because reduction will cause misleading condition.

Hypothesis Testing

1) The null hypothesis is that whether water is acidic or basic, has no effect on potability.

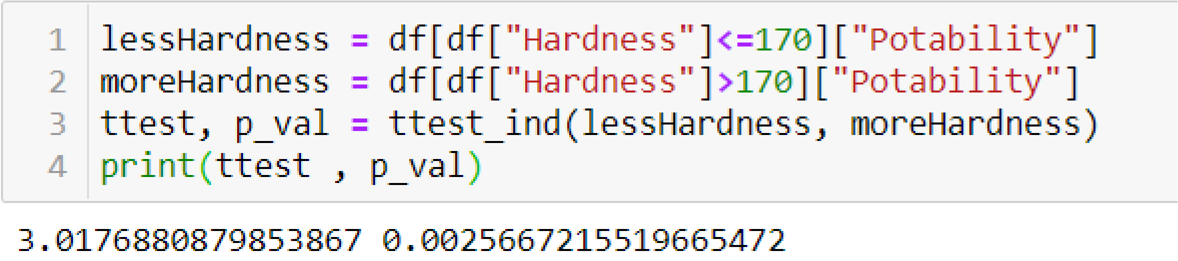


```
acid = df[df["ph"]<7]["Potability"]
base = df[df["ph"]>7]["Potability"]
ttest, p_val = ttest_ind(acid, base)
print(ttest , p_val)
```

```
0.31437587583250015 0.7532556073374688
```

P-value is greater than 0.05, therefore it claims that there is no relationship between being acid or base, and potability. Hence we can accept the null hypothesis.

2) The null hypothesis is that a water whose hardness is greater than 170, has no effect on potability.

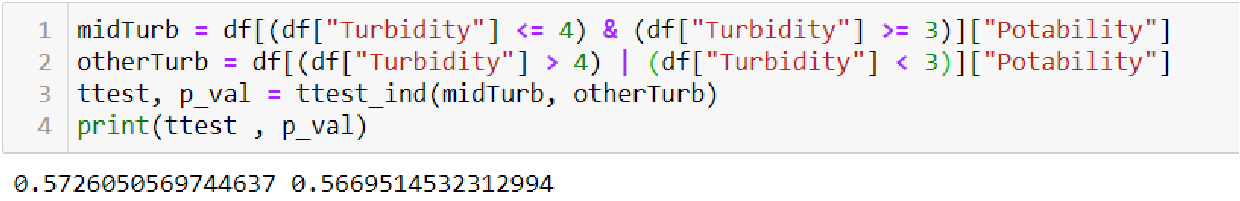


```
lessHardness = df[df["Hardness"]<=170]["Potability"]
moreHardness = df[df["Hardness"]>170]["Potability"]
ttest, p_val = ttest_ind(lessHardness, moreHardness)
print(ttest , p_val)
```

```
3.0176880879853867 0.0025667215519665472
```

P-value is less than 0.05, therefore it claims that there is a relationship between the hardness value that higher than 170, and potability. Hence we can reject the null hypothesis.

3) The null hypothesis is that a water whose turbidity is in between 3 and 4, has no effect on potability.



```
midTurb = df[(df["Turbidity"] <= 4) & (df["Turbidity"] >= 3)]["Potability"]
otherTurb = df[(df["Turbidity"] > 4) | (df["Turbidity"] < 3)]["Potability"]
ttest, p_val = ttest_ind(midTurb, otherTurb)
print(ttest , p_val)
```

```
0.5726050569744637 0.5669514532312994
```

P-value is greater than 0.05, therefore it claims that there is no relationship between staying 3-4 turbidity boundaries or not, and potability. Hence we can accept the null hypothesis.

Comments

Water quality dataset gain clear attributes after done with the data engineering.

Results shows that dataset is ready for machine learning model. Features have no relationship between each other. Null values are replaced with their each columns' average values.

As a suggestion, I might prefer to use Logistic Regression Model for dataset. Because, the potability is a binary value and the features behave more potable when they stay in detected boundaries.