Water quality analysis

PHASE:3 PROJECT

TEAM MEMBERS

D.SHREE THOVARTHENE

G.NILOFAR

E.PREETHI RAJALAKSHMI

K.MAHALAKSHMI

- PREPROCESSING THE DATASET
- PERFORMING ANALYSIS AND VISUALIZATION USING IBM COGNOS

Preprocessing the dataset:

Definition:

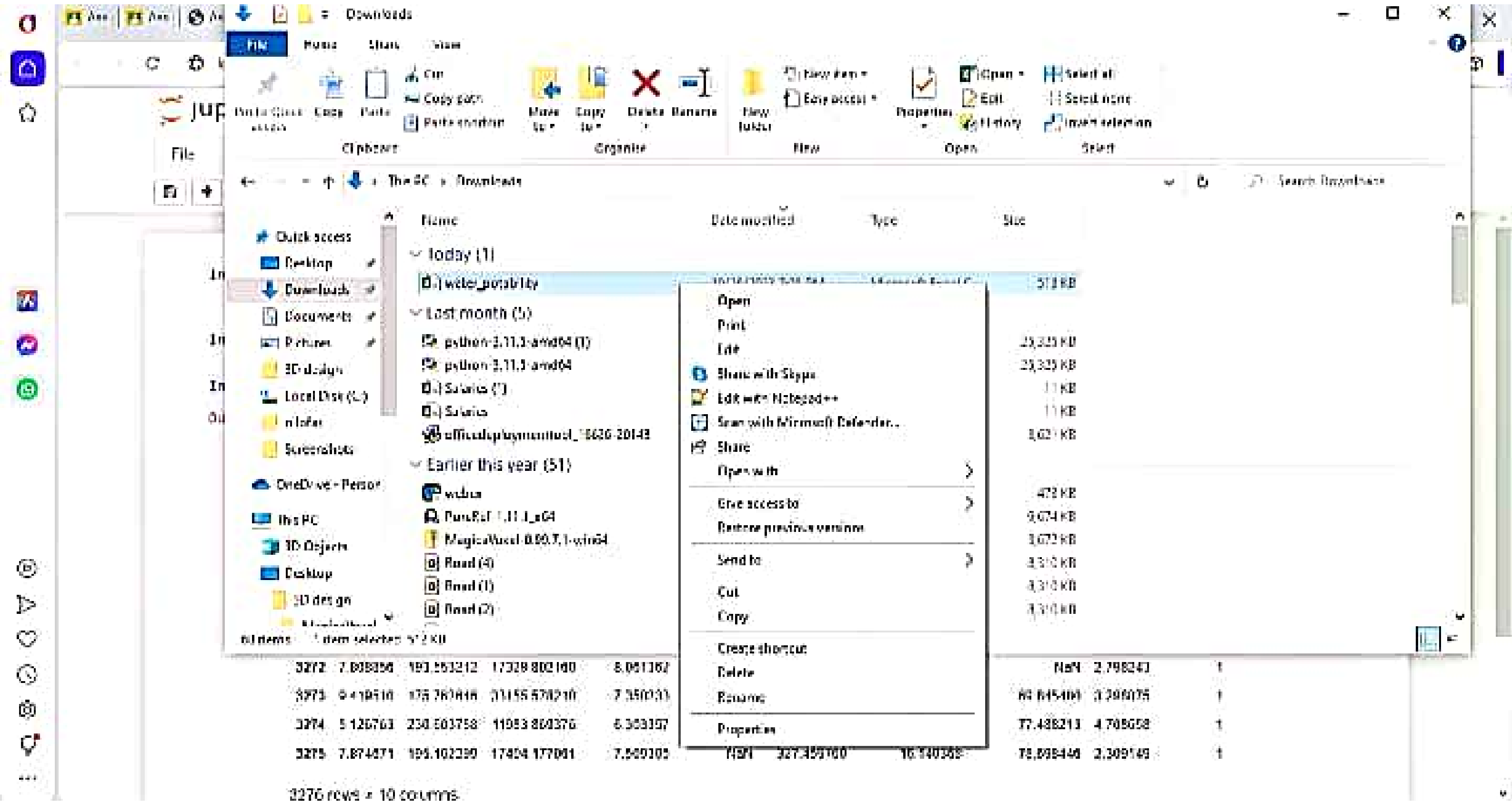
Preprocessing a dataset is a crucial step in data analysis and machine learning. It involves cleaning and transforming the data to make it suitable for analysis or model training.

Data cleaning:

- Data cleaning is an essential part of preprocessing your dataset to ensure its quality and reliability.

Handling Missing Data:

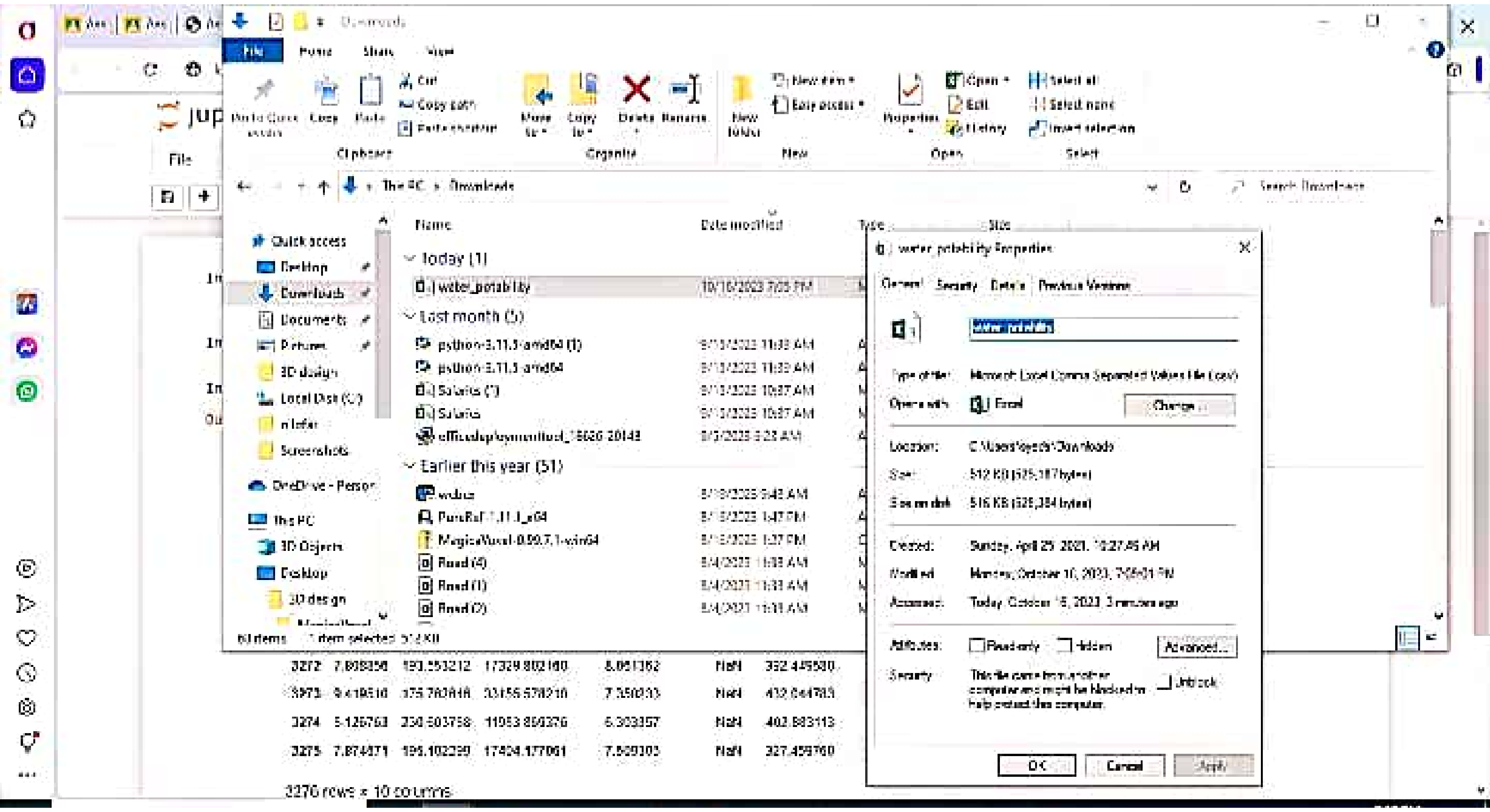
- Identify and analyze missing values in your dataset.
- Decide whether to impute missing values, remove rows or columns with missing data, or use other techniques like interpolation.



Name	Date modified	Type	Size
Today (1)			
water_potability	[illegible]	Microsoft Excel C	513 KB
Last month (5)			
python-3.11.3-amd64 (1)			25,325 KB
python-3.11.3-amd64			25,325 KB
Salaries (1)			11 KB
Salaries			11 KB
officedeploymenttool_16626-20148			3,621 KB
Earlier this year (51)			
webex			472 KB
PureRef-1.11.1_x64			9,674 KB
MagicaVoxel-0.99.7.1-win64			3,672 KB
Road (4)			4,310 KB
Road (1)			4,310 KB
Road (2)			4,310 KB

60 items   1 item selected 512 KB

```
3272  7.808856  193.553212  17329.802160  8.061362                                   NaN  2.798243  1
3273  9.419510  175.762646  33155.578218  7.350233                             69.845400  3.298875  1
3274  5.126763  230.603758  11983.869376  6.303357                             77.488213  4.708658  1
3275  7.874671  195.102299  17404.177061  7.509306  NaN  327.459760  16.140368   78.698446  2.309149  1

3276 rows × 10 columns
```



This PC > Downloads

Name	Date modified
Today (1)	
water_potability	10/10/2023 7:05 PM
Last month (5)	
python-3.11.5-amd64 (1)	9/13/2023 11:33 AM
python-3.11.5-amd64	9/13/2023 11:33 AM
Salaris (1)	9/13/2023 10:37 AM
Salaris	9/13/2023 10:37 AM
officedeploymenttool_16626-20148	9/3/2023 3:23 AM
Earlier this year (51)	
wdisc	8/19/2023 9:43 AM
PureRef-1.11.1_x64	8/15/2023 1:47 PM
MagicaVoxel-0.99.7.1-win64	8/15/2023 1:27 PM
Road (4)	8/4/2023 11:33 AM
Road (1)	8/4/2023 11:33 AM
Road (2)	8/4/2023 11:33 AM

60 items   1 item selected 512 KB

water_potability Properties

water_potability

Type of file: Microsoft Excel Comma Separated Values File (.csv)

Opens with: Excel

Location: C:\Users\syeda\Downloads

Size: 512 KB (525,187 bytes)

Size on disk: 516 KB (528,384 bytes)

Created: Sunday, April 25, 2021, 10:27:45 AM

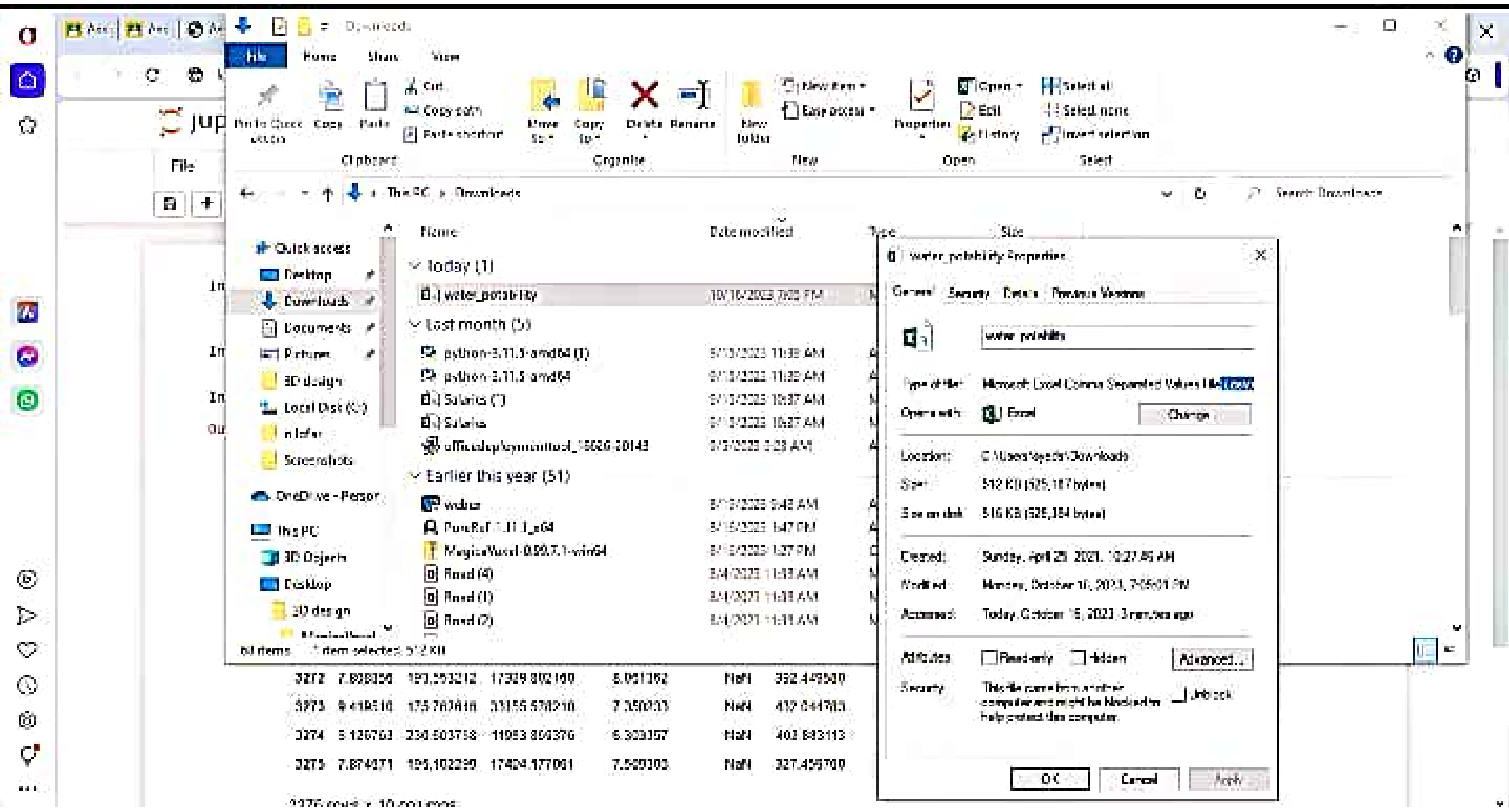
Modified: Monday, October 10, 2023, 7:05:01 PM

Accessed: Today, October 16, 2023, 3 minutes ago

Attributes: □ Read-only □ Hidden

Security: This file came from another computer and might be blocked to help protect this computer. □ Unblock

```
3272  7.808856  193.553212  17329.802160  8.061362  NaN  392.449580
3273  9.419510  175.762646  33155.578218  7.350233  NaN  432.044783
3274  5.126763  230.603758  11983.869376  6.303357  NaN  402.883113
3275  7.874671  195.102299  17404.177061  7.509306  NaN  327.459760

3276 rows × 10 columns
```



This PC > Downloads

Name | Date modified | Type | Size

Today (1)
water_potability 10/16/2023 7:02 PM

Last month (5)
python-3.11.5-amd64 (1) 9/13/2023 11:39 AM
python-3.11.5-amd64 9/13/2023 11:39 AM
Salaries (1) 9/13/2023 10:37 AM
Salaries 9/13/2023 10:37 AM
officedeploymenttool_16626-20148 9/3/2023 3:23 AM

Earlier this year (51)
vcdus 8/13/2023 9:48 AM
PureRef 1.11.1_x64 8/15/2023 1:47 PM
MagicaVoxel-0.99.7.1-win64 8/15/2023 1:27 PM
Road (4) 8/4/2023 11:33 AM
Road (1) 8/4/2023 11:33 AM
Road (2) 8/4/2023 11:33 AM

60 items 1 item selected 512 KB

water_potability Properties

General | Security | Details | Previous Versions

water_potability

Type of file: Microsoft Excel Comma Separated Values File (.csv)

Opens with: Excel Change...

Location: C:\Users\syeda\Downloads

Size: 512 KB (525,187 bytes)

Size on disk: 516 KB (528,384 bytes)

Created: Sunday, April 25, 2021, 10:27:45 AM

Modified: Monday, October 16, 2023, 7:05:01 PM

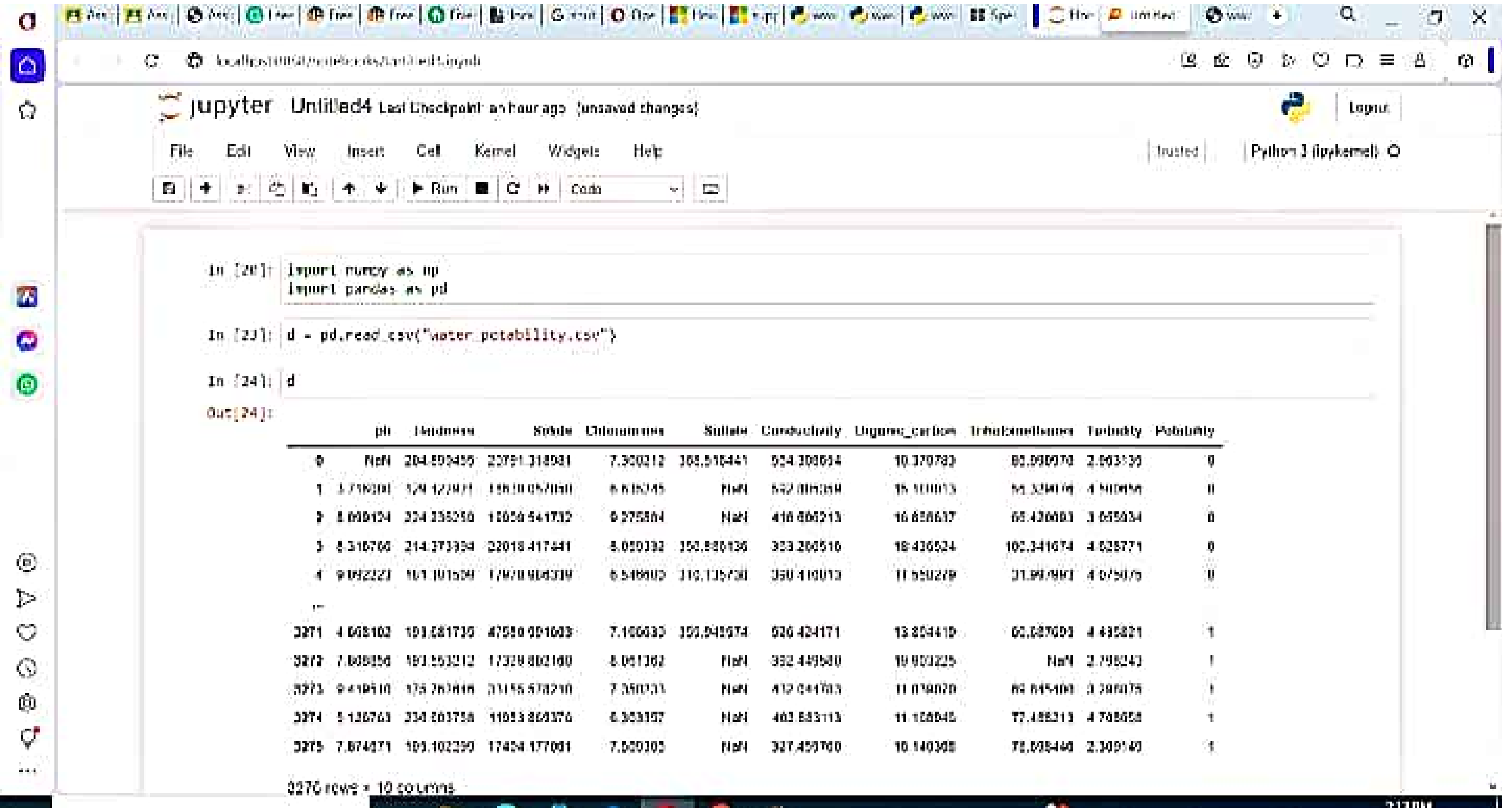
Accessed: Today, October 16, 2023, 3 minutes ago

Attributes: □ Read-only □ Hidden Advanced...

Security: This file came from another computer and might be blocked to help protect this computer. □ Unblock

OK Cancel Apply

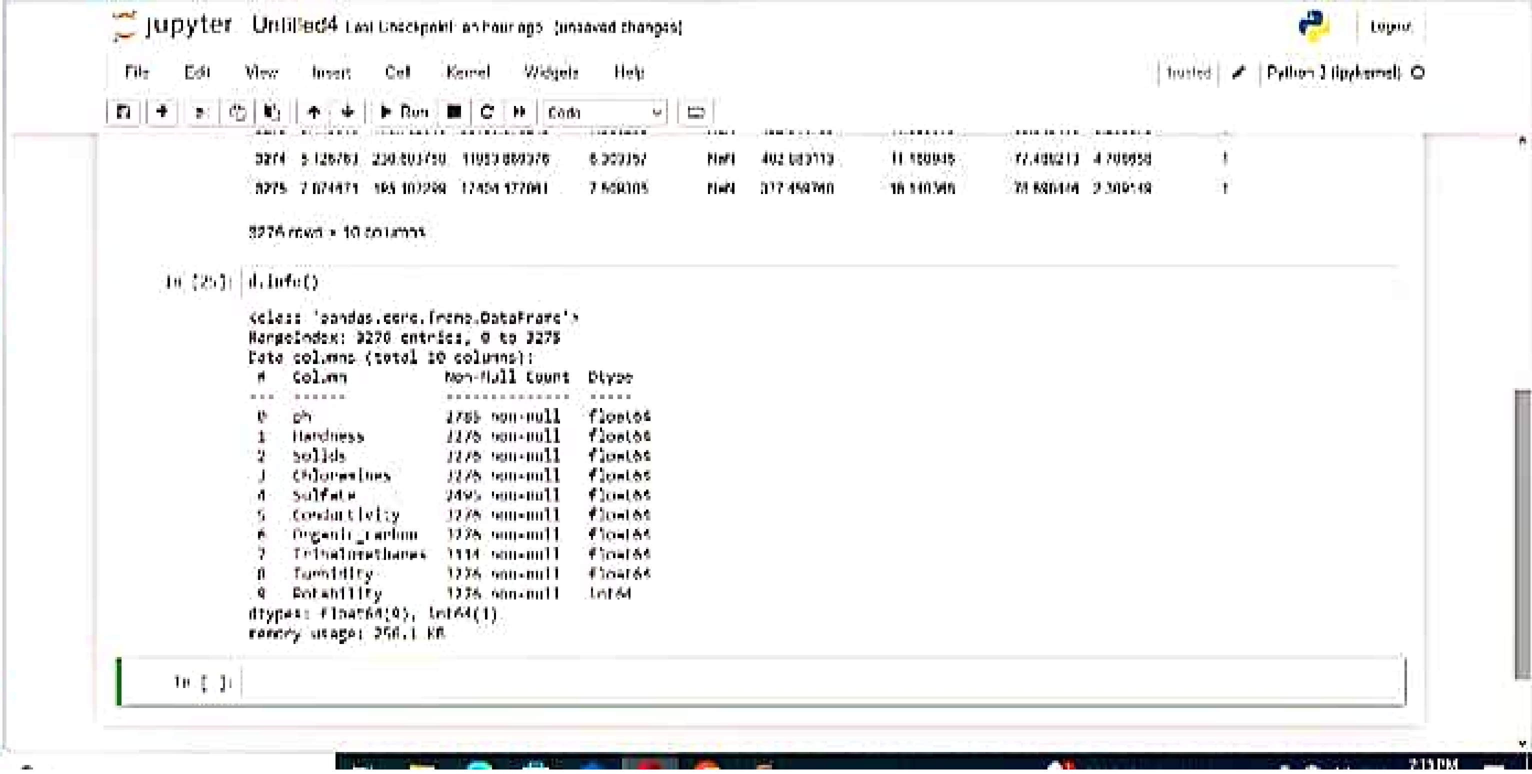
```
3272  7.808856  193.553212  17329.802160  8.061362  NaN  392.449580
3273  9.419510  175.762646  33155.578218  7.350233  NaN  432.044783
3274  5.126763  230.603758  11983.869376  6.303357  NaN  402.883113
3275  7.874671  195.102299  17404.177061  7.509306  NaN  327.459760

3276 rows × 10 columns
```



localhost:8888/notebooks/Untitled4.ipynb
jupyter Untitled4 Last Checkpoint: an hour ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [20]: import numpy as np
         import pandas as pd

In [23]: d = pd.read_csv("water_potability.csv")

In [24]: d
Out[24]:
```

```
            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
   2  8.099124  224.236259  19909.541732     9.275884         NaN    418.606213       16.868637        66.420093   3.055934           0
   3  8.316766  214.373394  22018.417441     8.059332  356.886136    363.266516       18.436524       100.341674   4.628771           0
   4  9.092223  181.101509  17978.986339     6.546600  310.135738    398.410813       11.558279        31.997993   4.075075           0
 ...
3271  4.668102  193.681735  47580.991603     7.166639  359.948574    526.424171       13.894419        66.687695   4.435821           1
3272  7.808856  193.553212  17329.802160     8.061362         NaN    392.449580       19.903225              NaN   2.798243           1
3273  9.419510  175.762646  33155.578218     7.350233         NaN    432.044783       11.039070        69.845400   3.298875           1
3274  5.126763  230.603758  11983.869376     6.303357         NaN    402.883113       11.168946        77.488213   4.708658           1
3275  7.874671  195.102299  17404.177061     7.509306         NaN    327.459760       16.140368        78.698446   2.309149           1

3276 rows × 10 columns
```



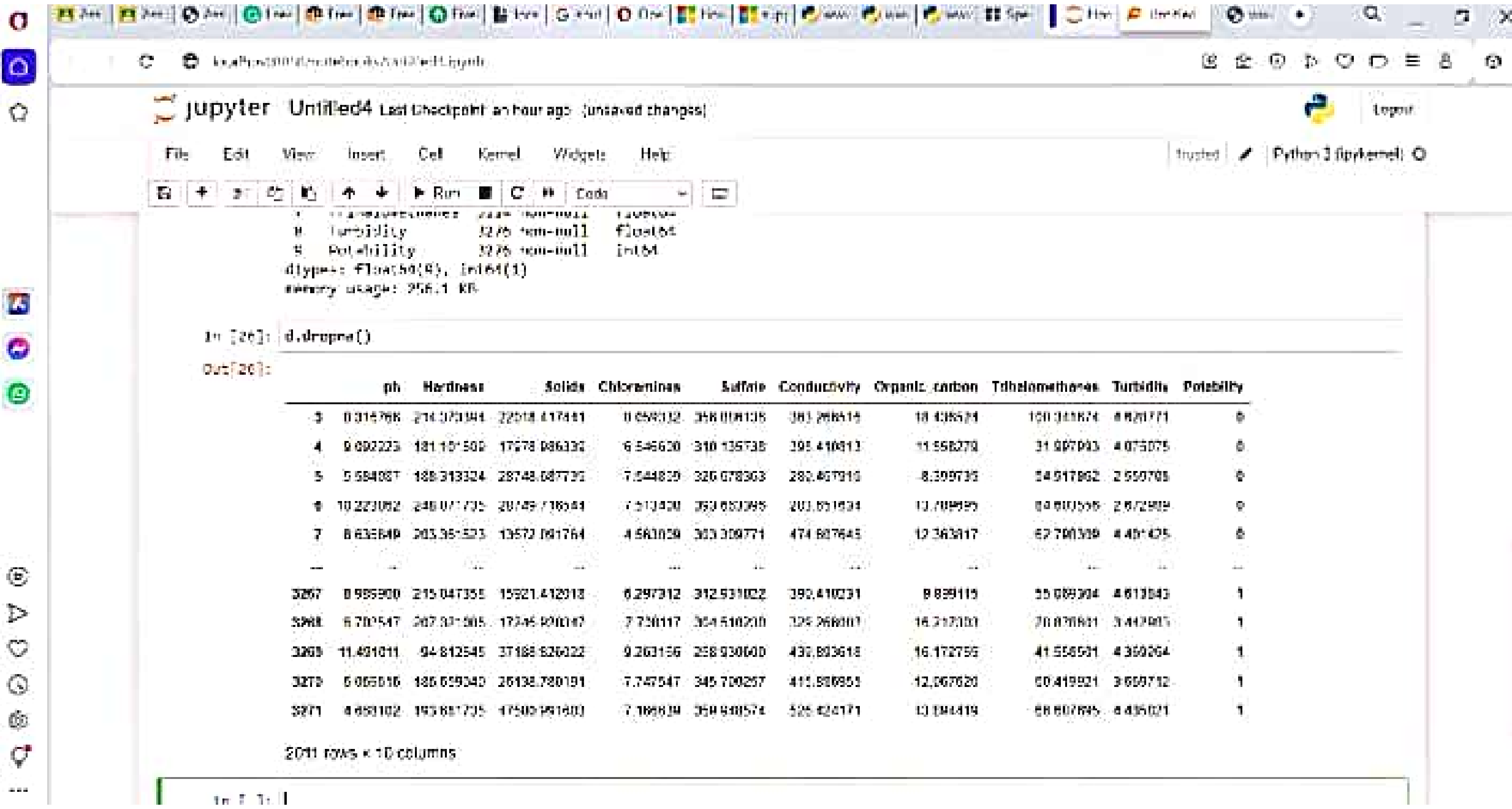
```
3274  5.126763  230.603758  11983.869376  6.303357  NaN  402.883113  11.168946  77.488213  4.708658  1
3275  7.874671  195.102299  17404.177061  7.509306  NaN  327.459760  16.140368  78.698446  2.309149  1

3276 rows × 10 columns
```

In [25]:

```
d.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 [ ]:



```
 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 [26]:

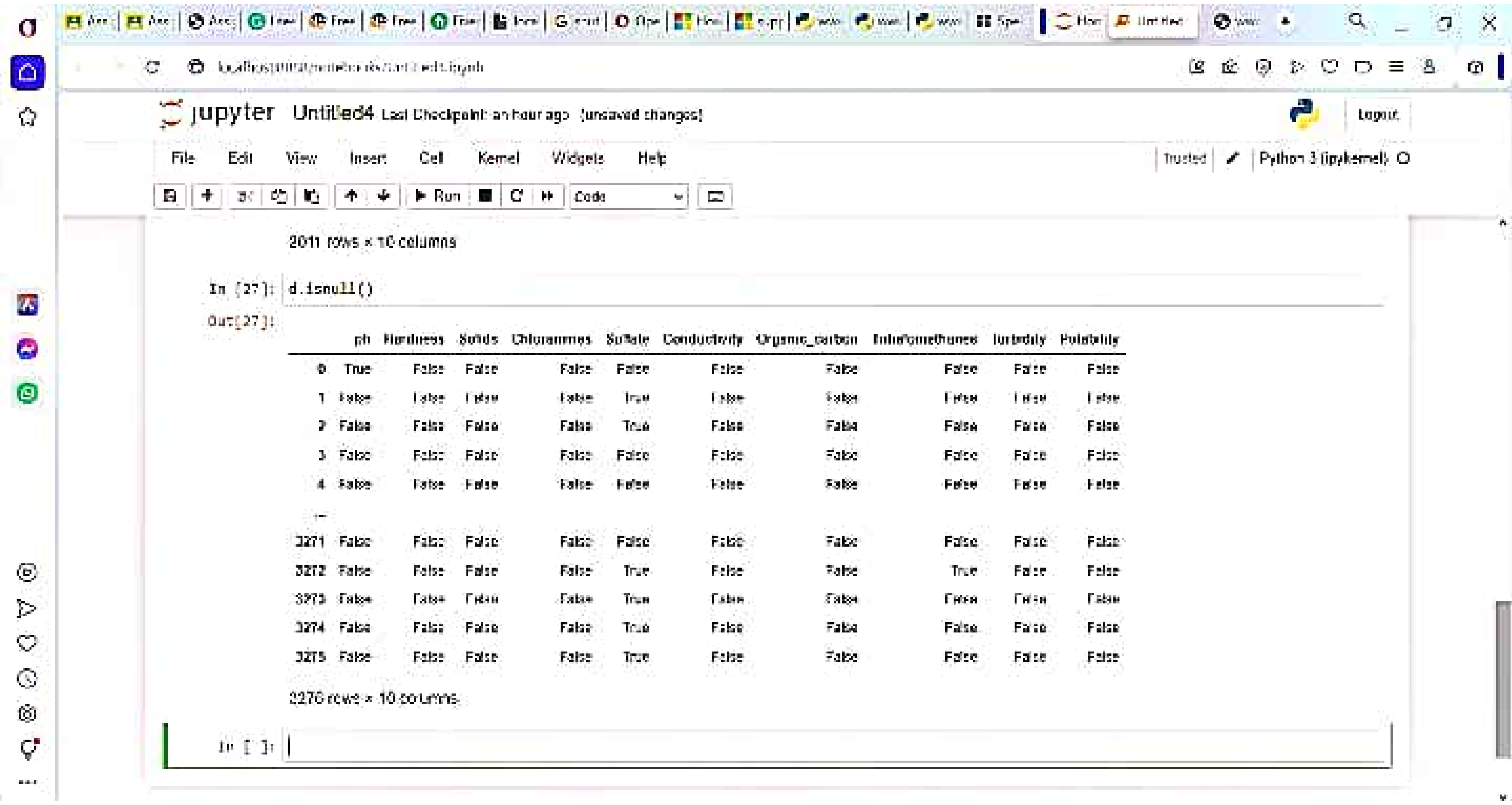
```
d.dropna()
```

Out[26]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
5	5.584087	188.313324	28748.687739	7.544869	326.678363	280.467916	8.399735	54.917862	2.559708	0
6	10.223862	248.071735	28749.716544	7.513408	393.663396	283.651634	13.789695	84.603556	2.672989	0
7	8.635849	203.361523	13672.091764	4.563009	303.309771	474.607645	12.363817	62.798309	4.401425	0
...	...	...	...	...	...	...	...	...	...	...
3267	8.989900	215.047358	15921.412018	6.297312	312.931022	390.410231	9.899115	55.069304	4.613843	1
3268	6.702547	207.321086	17246.920347	7.708117	304.510230	329.266002	16.217303	28.878601	3.442983	1
3269	11.491011	94.812545	37188.826022	9.263166	258.930600	439.893618	16.172755	41.558501	4.369264	1
3270	6.069616	186.659040	26138.780191	7.747547	345.700257	415.886955	12.067620	60.419921	3.669712	1
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1

2011 rows × 10 columns

In [ ]:



jupyter Untitled4 Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

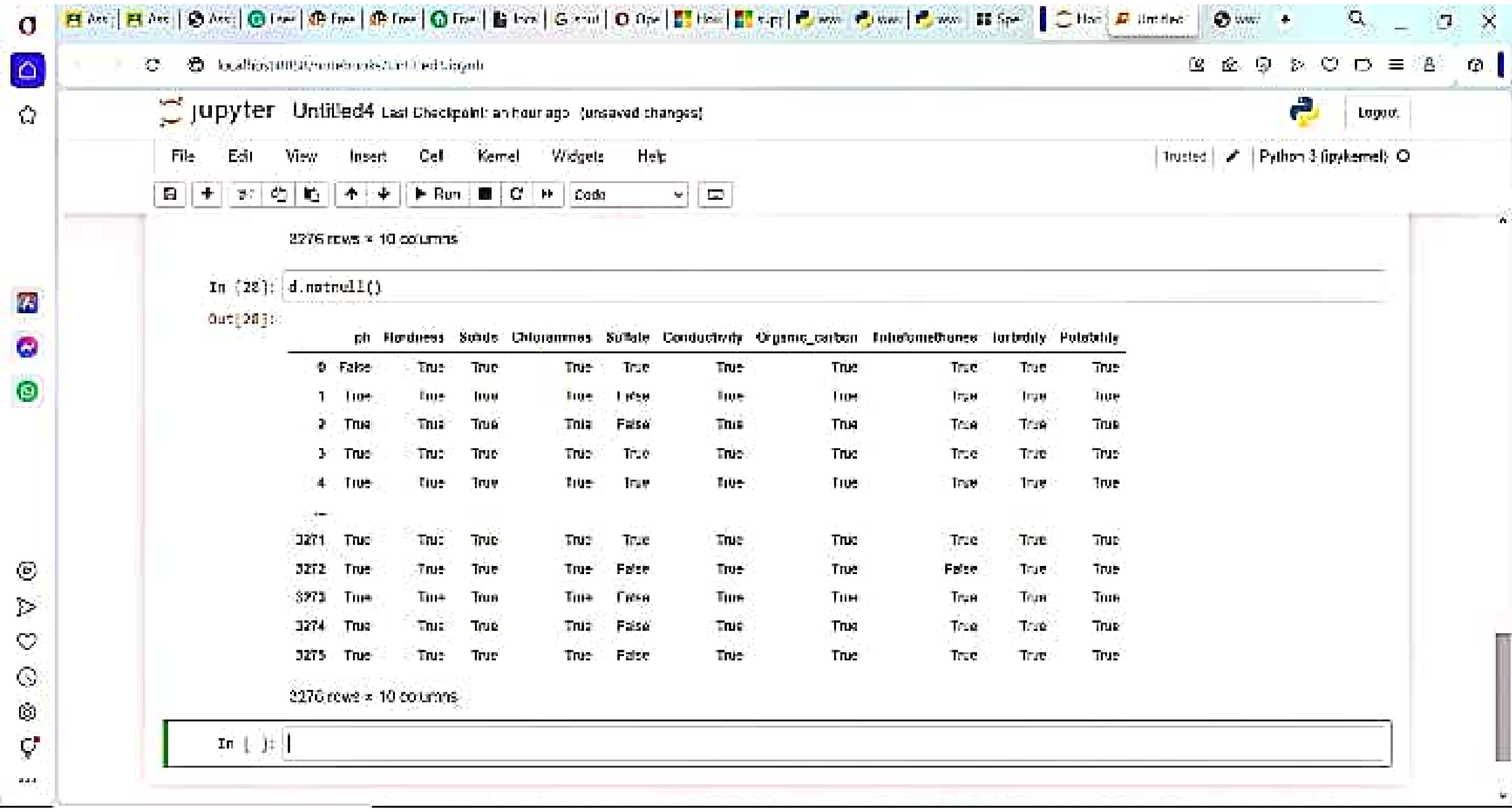
2011 rows × 10 columns

```
In [27]: d.isnull()
```

```
Out[27]:
         ph  Hardness  Solids  Chloramines  Sulfate  Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
   0   True     False   False        False    False         False           False            False      False       False
   1  False     False   False        False     True         False           False            False      False       False
   2  False     False   False        False     True         False           False            False      False       False
   3  False     False   False        False    False         False           False            False      False       False
   4  False     False   False        False    False         False           False            False      False       False
 ...
3271  False     False   False        False    False         False           False            False      False       False
3272  False     False   False        False     True         False           False             True      False       False
3273  False     False   False        False     True         False           False            False      False       False
3274  False     False   False        False     True         False           False            False      False       False
3275  False     False   False        False     True         False           False            False      False       False
```

3276 rows × 10 columns

```
In [ ]:
```



localhost:8888/notebooks/Untitled4.ipynb
jupyter
Untitled4 Last Checkpoint: an hour ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

3276 rows × 10 columns

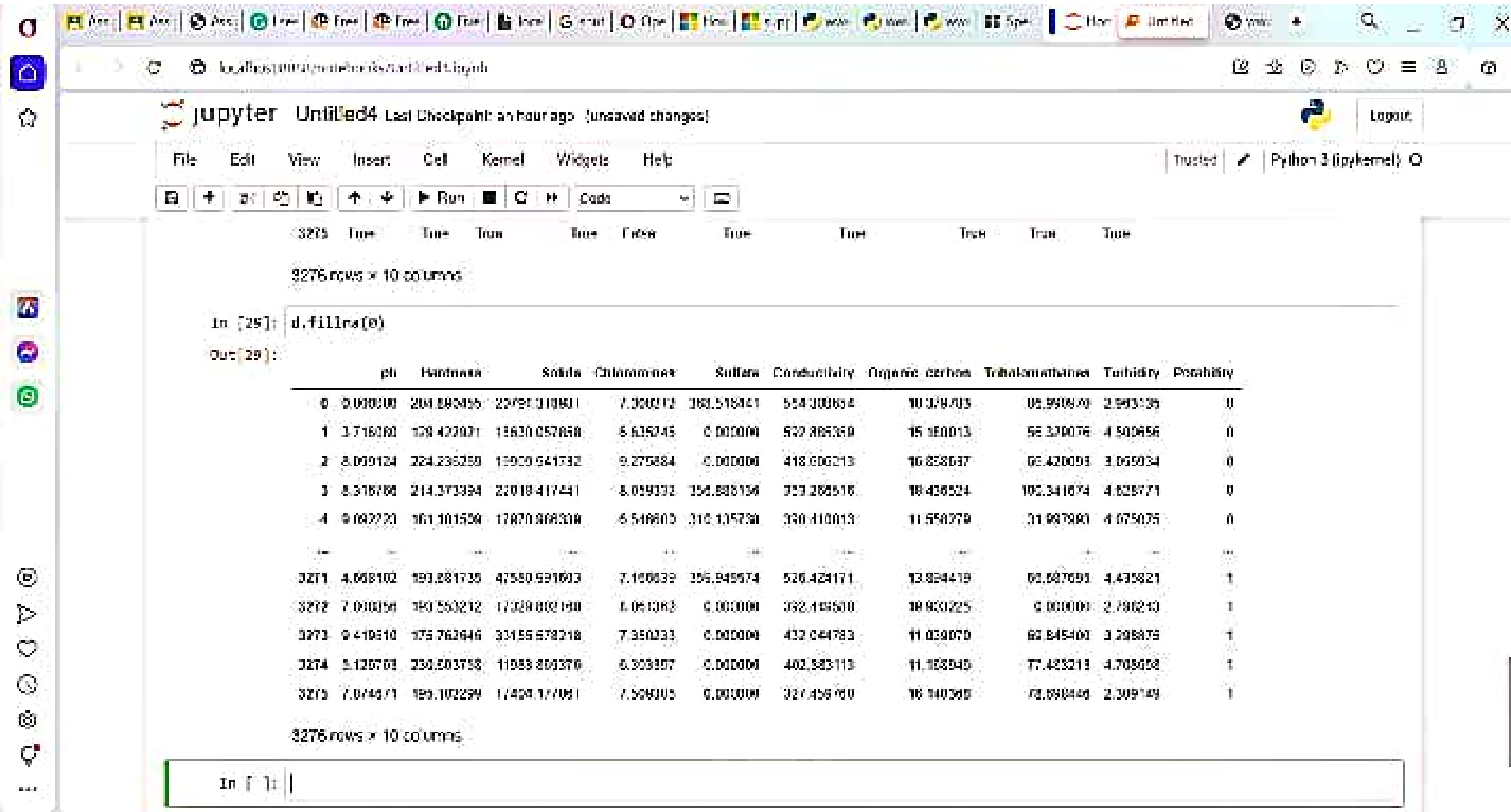
In [28]:
```
d.notnull()
```

Out[28]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	False	True	True	True	True	True	True	True	True	True
1	True	True	True	True	False	True	True	True	True	True
2	True	True	True	True	False	True	True	True	True	True
3	True	True	True	True	True	True	True	True	True	True
4	True	True	True	True	True	True	True	True	True	True
...										
3271	True	True	True	True	True	True	True	True	True	True
3272	True	True	True	True	False	True	True	False	True	True
3273	True	True	True	True	False	True	True	True	True	True
3274	True	True	True	True	False	True	True	True	True	True
3275	True	True	True	True	False	True	True	True	True	True

3276 rows × 10 columns

In [ ]:


localhost:8888/notebooks/Untitled4.ipynb
jupyter Untitled4 Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel) ○
Run
Code

```
3275  True  True  True  True  False  True  True  True  True

3276 rows × 10 columns
```

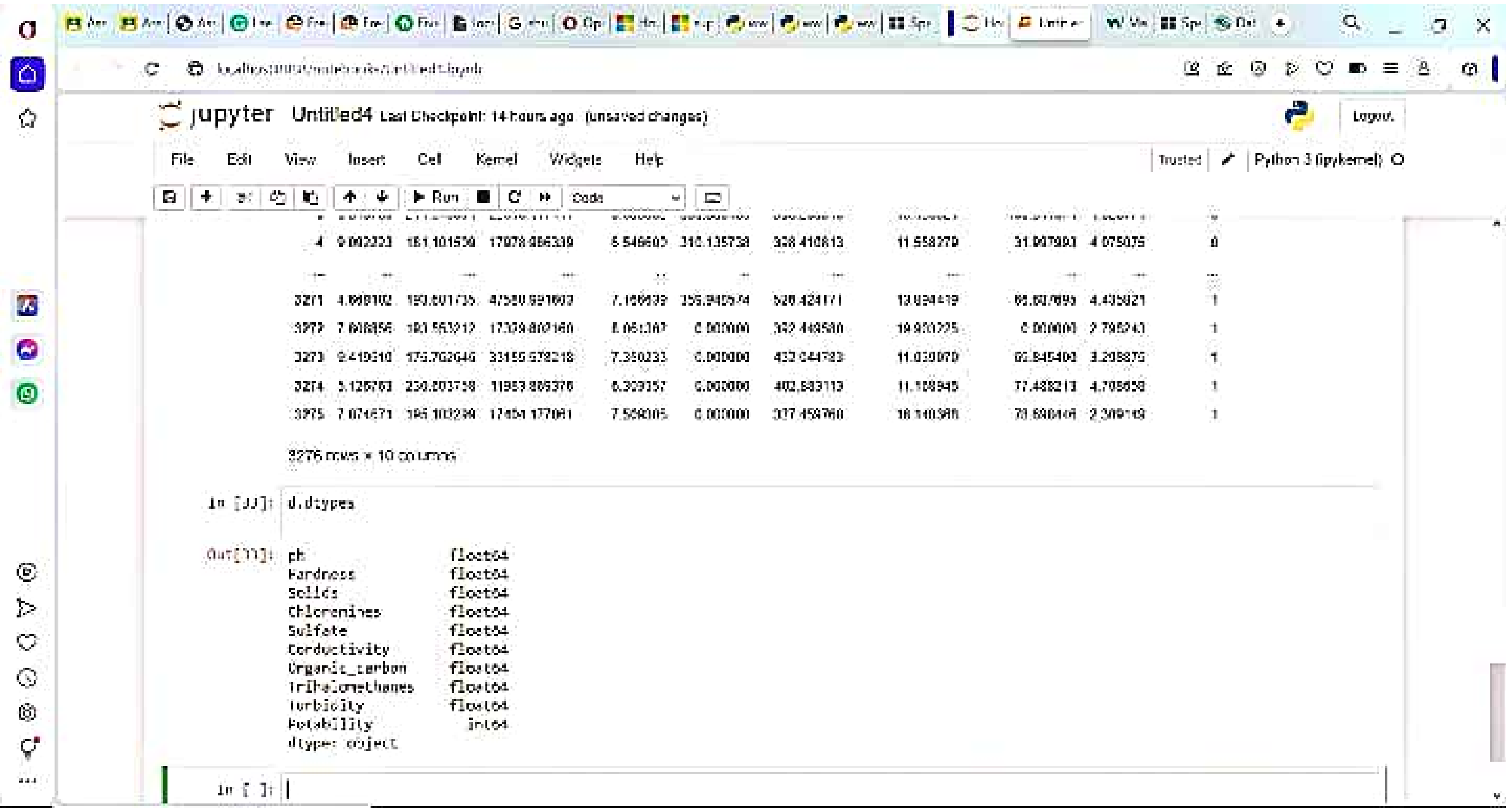
```
In [29]: d.fillna(0)
```

```
Out[29]:
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	0.000000	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	0.000000	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	0.000000	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	0.000000	392.449580	19.903225	0.000000	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	0.000000	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	0.000000	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	0.000000	327.459760	16.140368	78.698446	2.309149	1

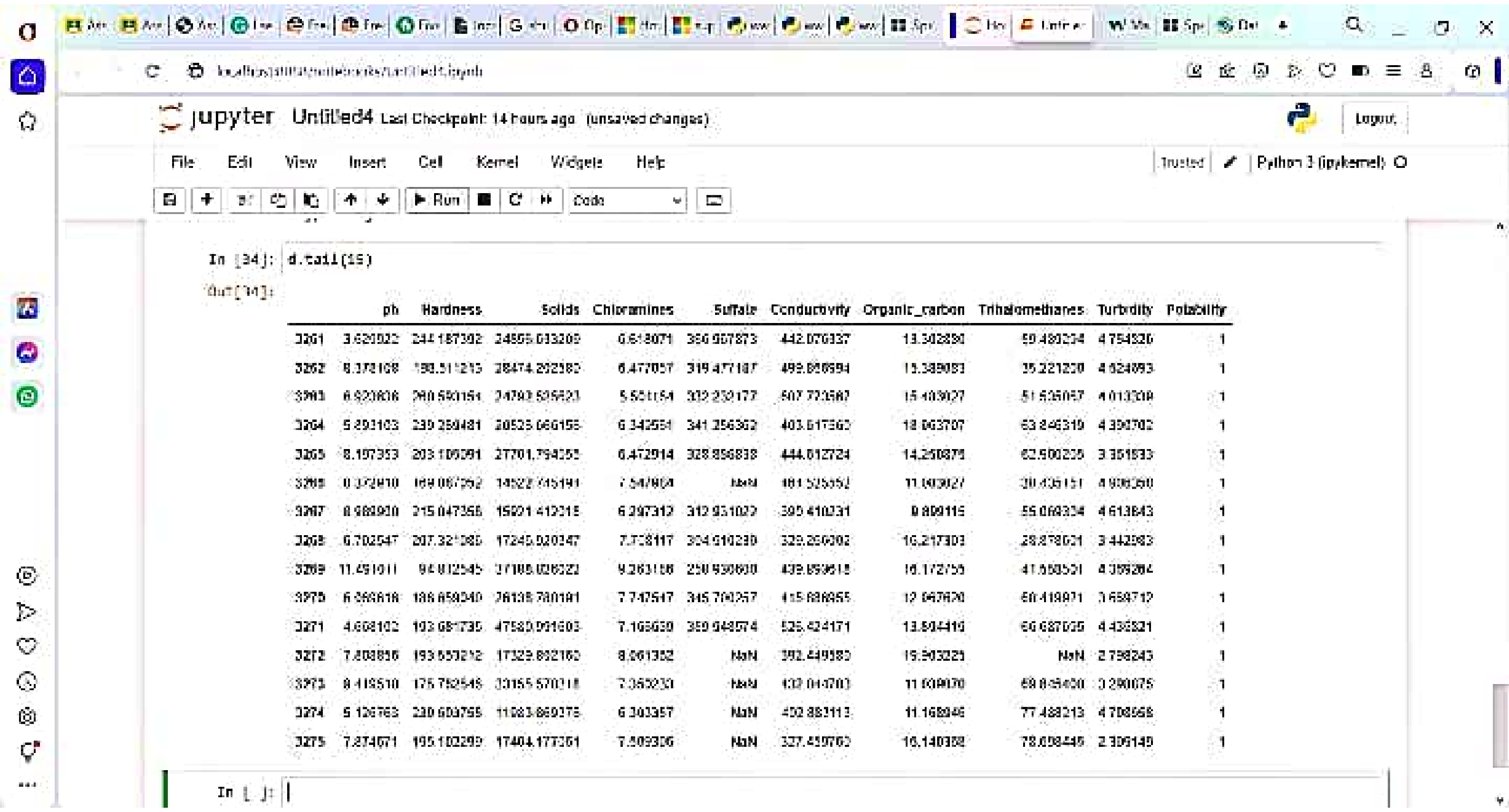
```
3276 rows × 10 columns
```

```
In [ ]: |
```


jupyter Untitled4 Last Checkpoint: 14 hours ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
   4  9.092223  181.101509  17978.986339   6.546600  310.135738  398.410813   11.558279   31.997993  4.075075   0
 ...       ...         ...           ...        ...         ...         ...         ...         ...       ...  ...
3271  4.668102  193.681735  47580.991603   7.166639  359.948574  526.424171   13.894419   66.687695  4.435821   1
3272  7.808856  193.553212  17329.802160   8.061362    0.000000  392.449580   19.903225    0.000000  2.798243   1
3273  9.419510  175.762646  33155.578218   7.350233    0.000000  432.044783   11.039070   69.845400  3.298875   1
3274  5.126763  230.603758  11983.869376   6.303357    0.000000  402.883113   11.168946   77.488213  4.708658   1
3275  7.874671  195.102299  17404.177061   7.509306    0.000000  327.459760   16.140380   78.698446  2.309149   1

3276 rows × 10 columns
```

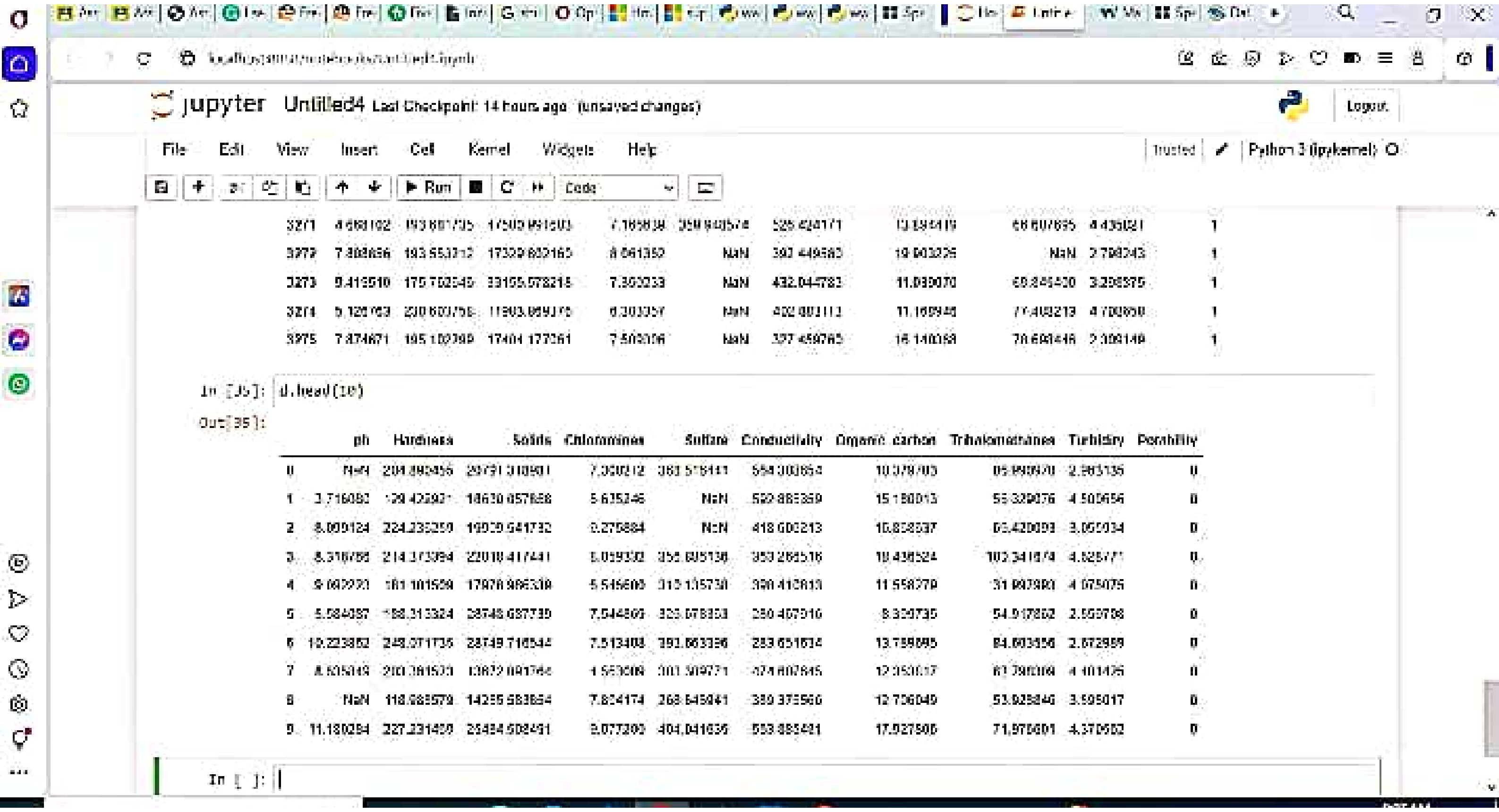
```
In [11]: d.dtypes

Out[11]: ph                 float64
         Hardness           float64
         Solids             float64
         Chloramines        float64
         Sulfate            float64
         Conductivity       float64
         Organic_carbon     float64
         Trihalomethanes    float64
         Turbidity          float64
         Potability           int64
         dtype: object

In [ ]: 
```



```
In [34]: d.tail(15)
```

```
Out[34]:
            ph    Hardness        Solids  Chloramines     Sulfate  Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
3261  3.620922  244.187392  24856.633209     6.618071  366.907873    442.076337       13.302880        59.489294   4.754526           1
3262  8.378108  198.511213  28474.202382     6.477057  319.477187    499.850594       15.389083        35.221220   4.624093           1
3263  6.920806  260.590154  24793.525823     5.501154  302.232177    507.773587       15.403027        51.535067   4.013308           1
3264  5.893103  239.269481  20526.666156     6.349561  341.256362    403.617560       18.963707        63.846319   4.390702           1
3265  8.197353  203.105091  27701.794055     6.472914  328.886838    444.612724       14.250875        62.906205   3.361833           1
3266  8.372910  169.087052  14622.745494     7.547984         NaN    464.525552       11.083027        38.435151   4.906358           1
3267  8.989900  215.047358  15921.412018     6.297312  312.931022    390.410231        9.899115        55.069304   4.613843           1
3268  6.702547  207.321086  17246.920347     7.708117  304.510230    329.266002       16.217303        28.878601   3.442983           1
3269 11.491011   94.812545  37188.826022     9.263166  258.930600    439.893618       16.172755        41.558501   4.369264           1
3270  6.069616  186.659040  26138.780191     7.747547  345.700257    415.886955       12.067620        60.419921   3.669712           1
3271  4.668102  193.681735  47580.991603     7.166639  359.948574    526.424171       13.894419        66.687695   4.435821           1
3272  7.808856  193.553212  17329.802160     8.061362         NaN    392.449580       19.903225              NaN   2.798243           1
3273  9.419510  175.762646  33155.578218     7.350233         NaN    432.044783       11.039070        69.845400   3.298875           1
3274  5.126763  230.603758  11983.869376     6.303357         NaN    402.883113       11.168946        77.488213   4.708658           1
3275  7.874671  195.102299  17404.177061     7.509306         NaN    327.459760       16.140368        78.698446   2.309149           1
```

```
In [ ]:
```
]]>


localhost:8888/notebooks/Untitled4.ipynb
jupyter
Untitled4 Last Checkpoint: 14 hours ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code

```
3271  4.668102  193.681735  47580.991603    7.166639  359.948574    526.424171    13.894419      66.687695  4.435821       1
3272  7.808856  193.553212  17329.802160    8.061362         NaN    392.449580    19.903225            NaN  2.798243       1
3273  9.419510  175.762646  33155.578218    7.350233         NaN    432.044783    11.039070      69.845400  3.298875       1
3274  5.126763  230.603758  11983.869376    6.303357         NaN    402.883113    11.168946      77.488213  4.708658       1
3275  7.874671  195.102299  17404.177061    7.509306         NaN    327.459760    16.140368      78.698446  2.309149       1
```

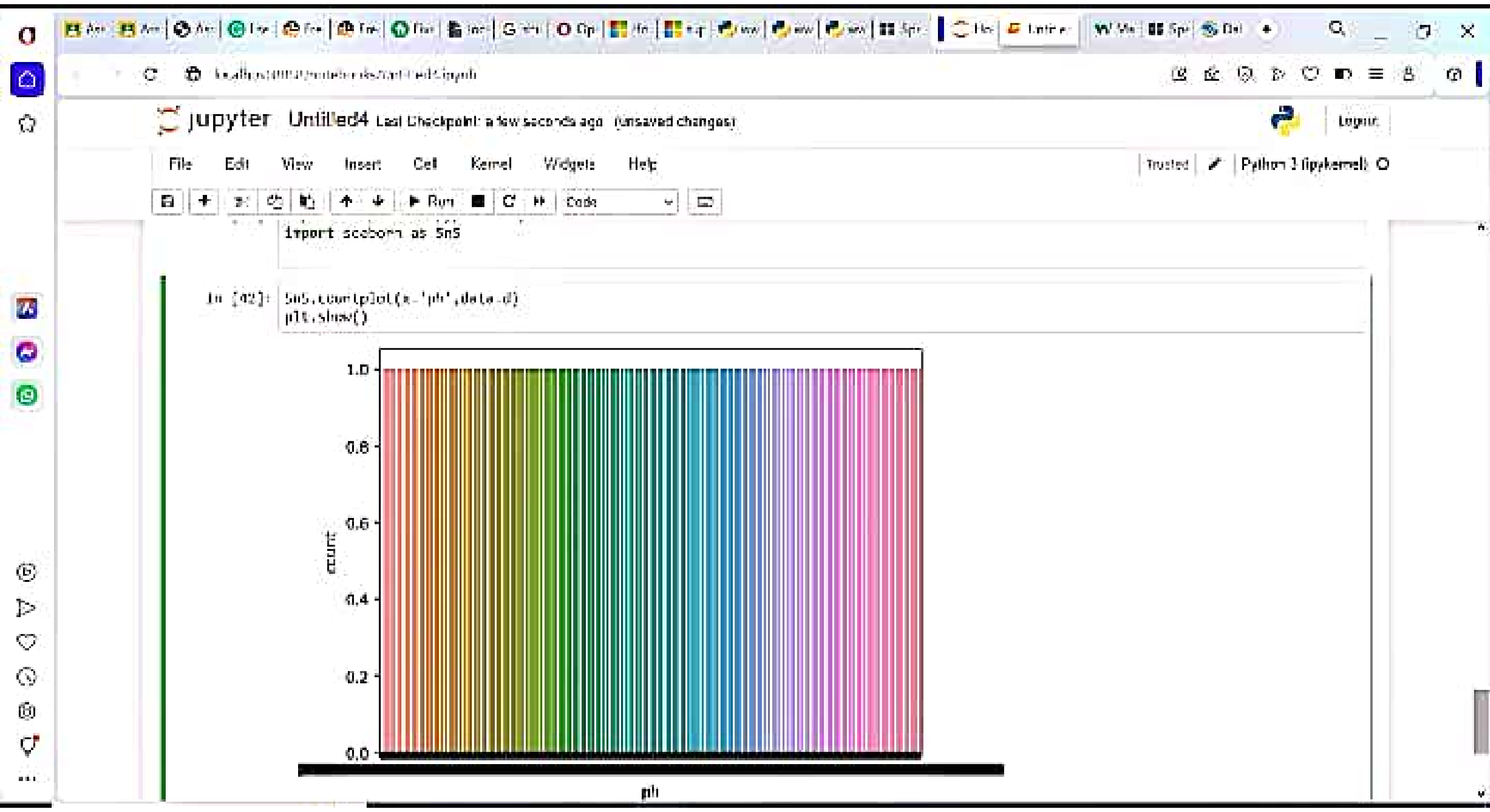
```
In [35]: d.head(10)
```

```
Out[35]:
          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
2   8.099124  224.236259  19909.541732     9.275884         NaN    418.606213       16.868637        66.420093   3.055934           0
3   8.316766  214.373394  22018.417441     8.059332  356.886136    363.266516       18.436524       100.341674   4.628771           0
4   9.092223  181.101509  17978.986339     6.546600  310.135738    398.410813       11.558279        31.997993   4.075075           0
5   5.584087  188.313324  28748.687739     7.544869  326.678363    280.467916        8.399735        54.917862   2.559708           0
6  10.223862  248.071735  28749.716544     7.513408  393.663396    283.651634       13.789695        84.603556   2.672989           0
7   8.635849  203.361523  13672.091764     4.563009  303.309771    474.607645       12.363817        62.798309   4.401425           0
8        NaN  118.988579  14285.583854     7.804174  268.646941    389.375566       12.706049        53.928846   3.595017           0
9  11.180284  227.231469  25484.508491     9.077200  404.041635    563.885481       17.927806        71.976601   4.370562           0
```

```
In [ ]:
```

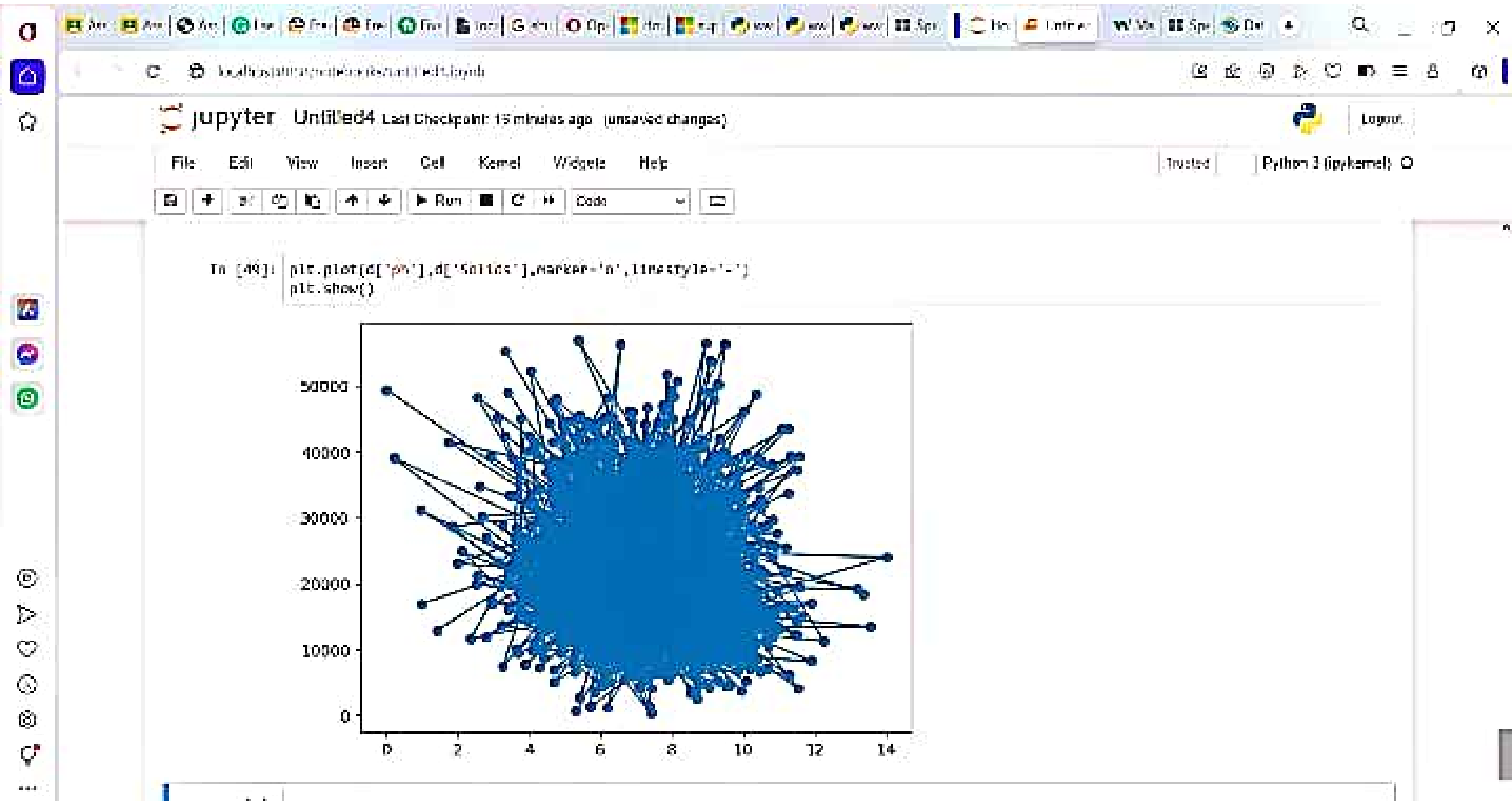
Performing analysis and Visualization:

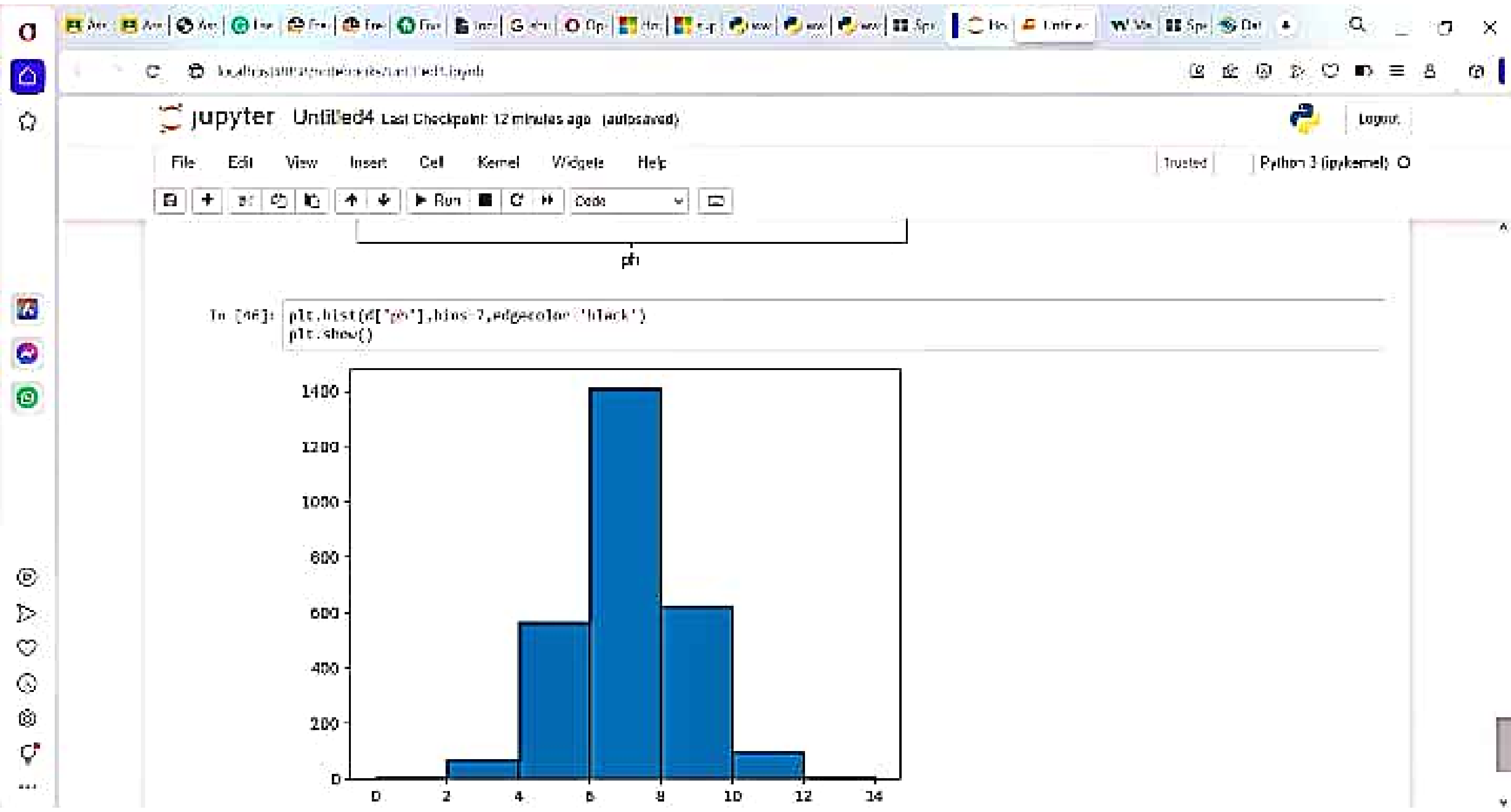
- Analyzing and visualizing a dataset is a common and crucial step in data analysis.
  - **Load the data**
  - **Explore the data**
    - Start by looking at the first few rows of your dataset to understand its structure and the type of data it contains.
    - Use functions like **head().**
  - **Data Cleaning**
    - This might involve imputing missing values, removing duplicates, or filtering out extreme outliers.
  - **Descriptive Statistics**
    - Calculate basic statistics like mean, median, standard deviation etc
  - **Data Visualization**
    - Histogram,Scatter plot,Line Charts ,Bar Charts,Box plot,etc



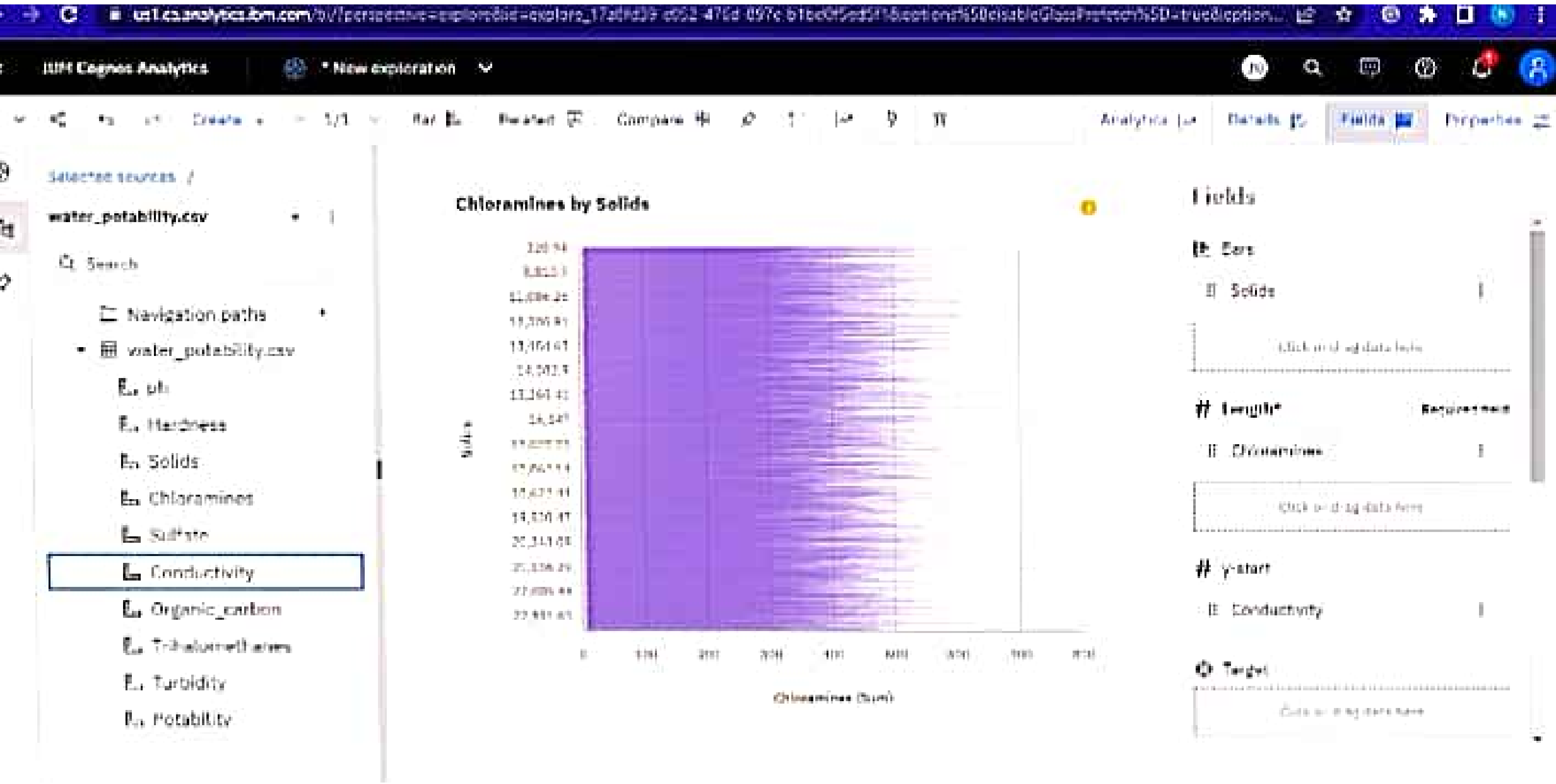
```
import seaborn as sns
```

```
In [42]: sns.countplot(x='ph',data=d)
         plt.show()
```





```
In [46]: plt.hist(d['ph'],bins=7,edgecolor='black')
         plt.show()
```



Bars
- Solids

Click or drag data here

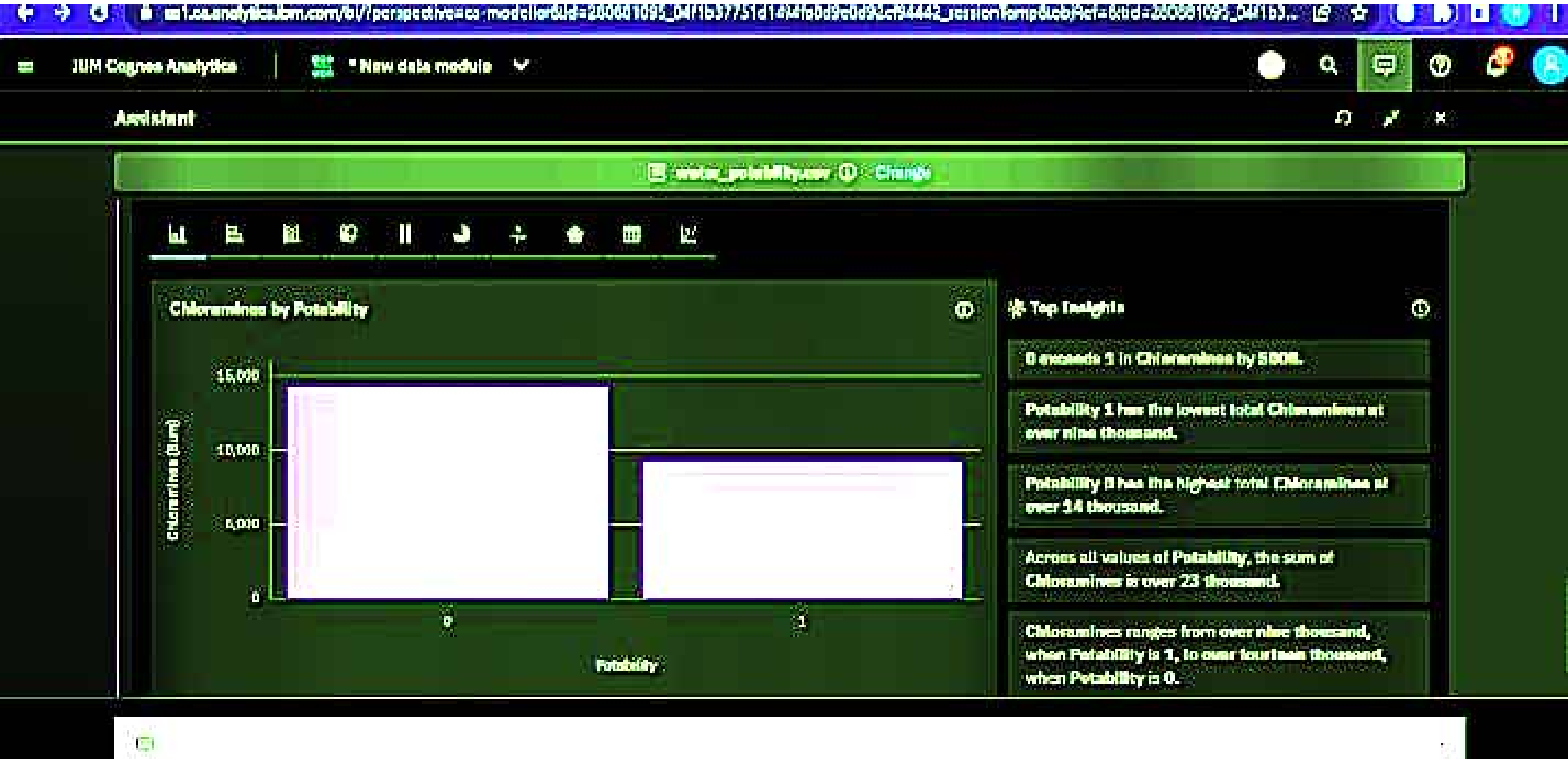
Length* Required field
- Chloramines

Click or drag data here

y-start
- Conductivity

Target

Click or drag data here



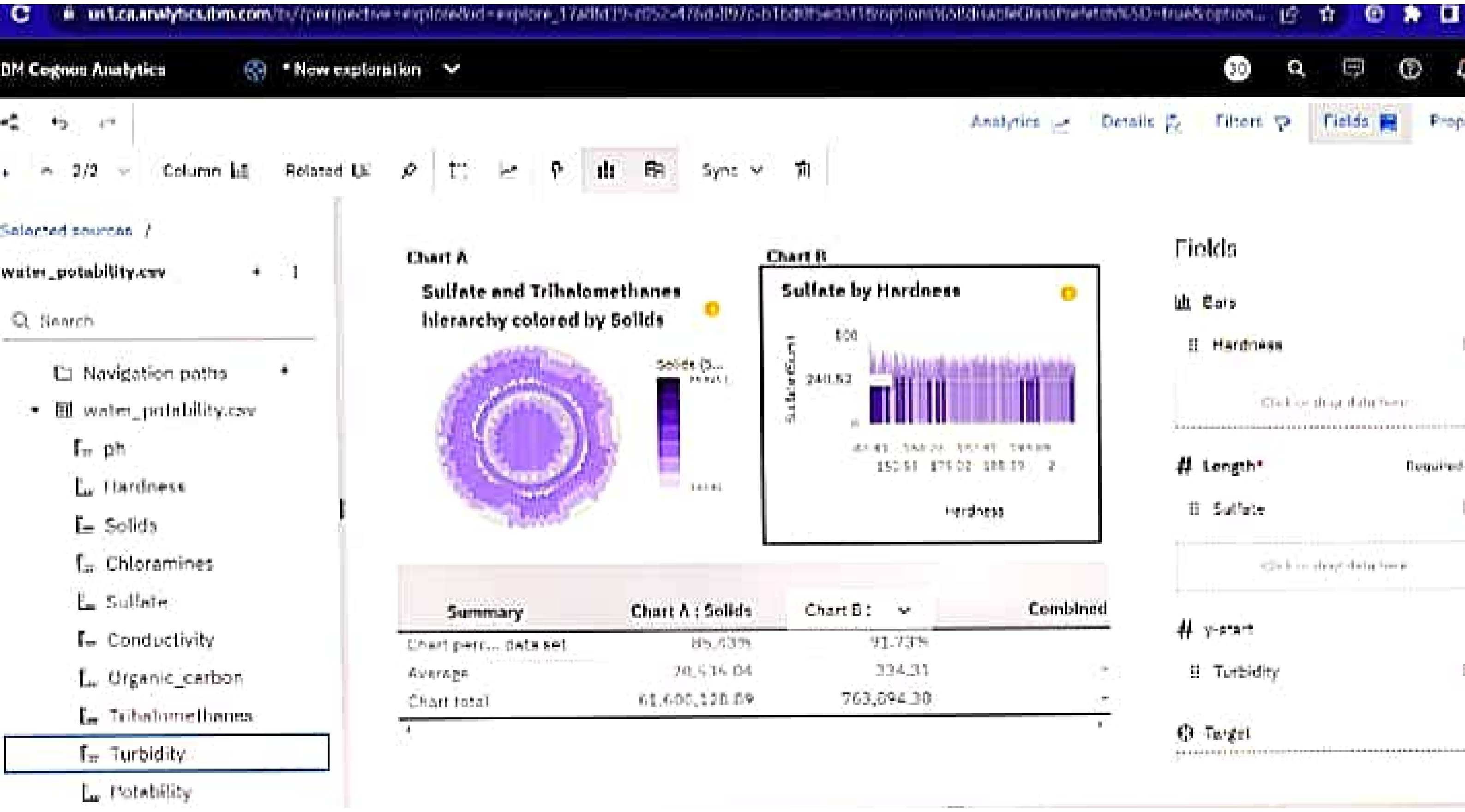
0 exceeds 1 in Chloramines by 5008.

Potability 1 has the lowest total Chloramines at over nine thousand.

Potability 0 has the highest total Chloramines at over 14 thousand.

Across all values of Potability, the sum of Chloramines is over 23 thousand.

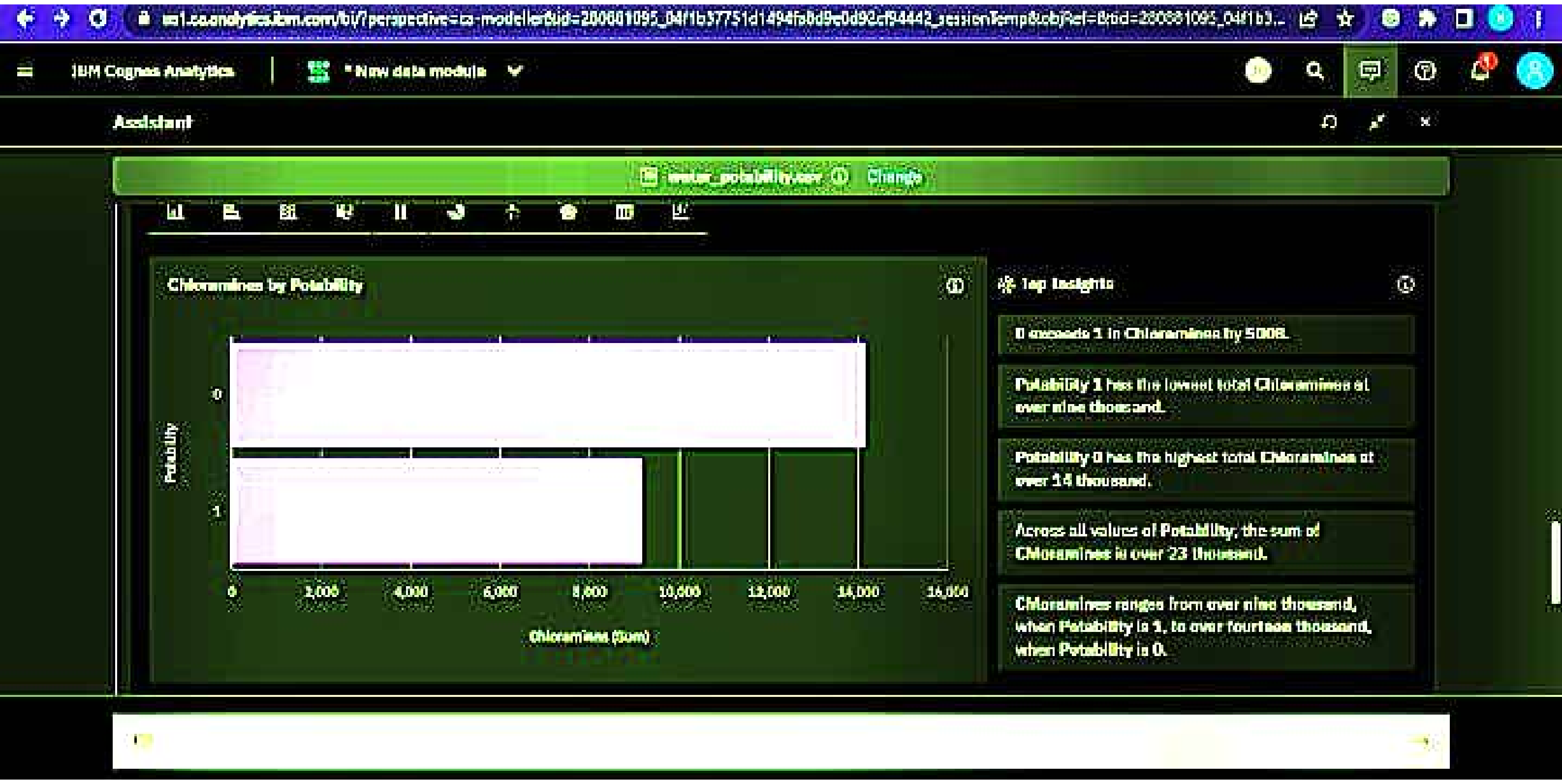
Chloramines ranges from over nine thousand, when Potability is 1, to over fourteen thousand, when Potability is 0.



Sulfate and Trihalomethanes hierarchy colored by Solids

Sulfate by Hardness

Summary	Chart A : Solids	Chart B :	Combined
Chart perc... data set	85.63%	91.73%	
Average	20,536.04	334.31	-
Chart total	61,600,128.09	763,094.30	-



us1.ca.analytics.ibm.com/bi/?perspective=ca-modeller&tid=280881095_04f1b37751d1494fa8d9e0d92cf94442_sessionTemp&objRef=&tid=280881095_04f1b3...
IBM Cognos Analytics
* New data module
Assistant
water_potability.csv
Change
Chloramines by Potability
Potability
0
1
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
Chloramines (Sum)
Top Insights

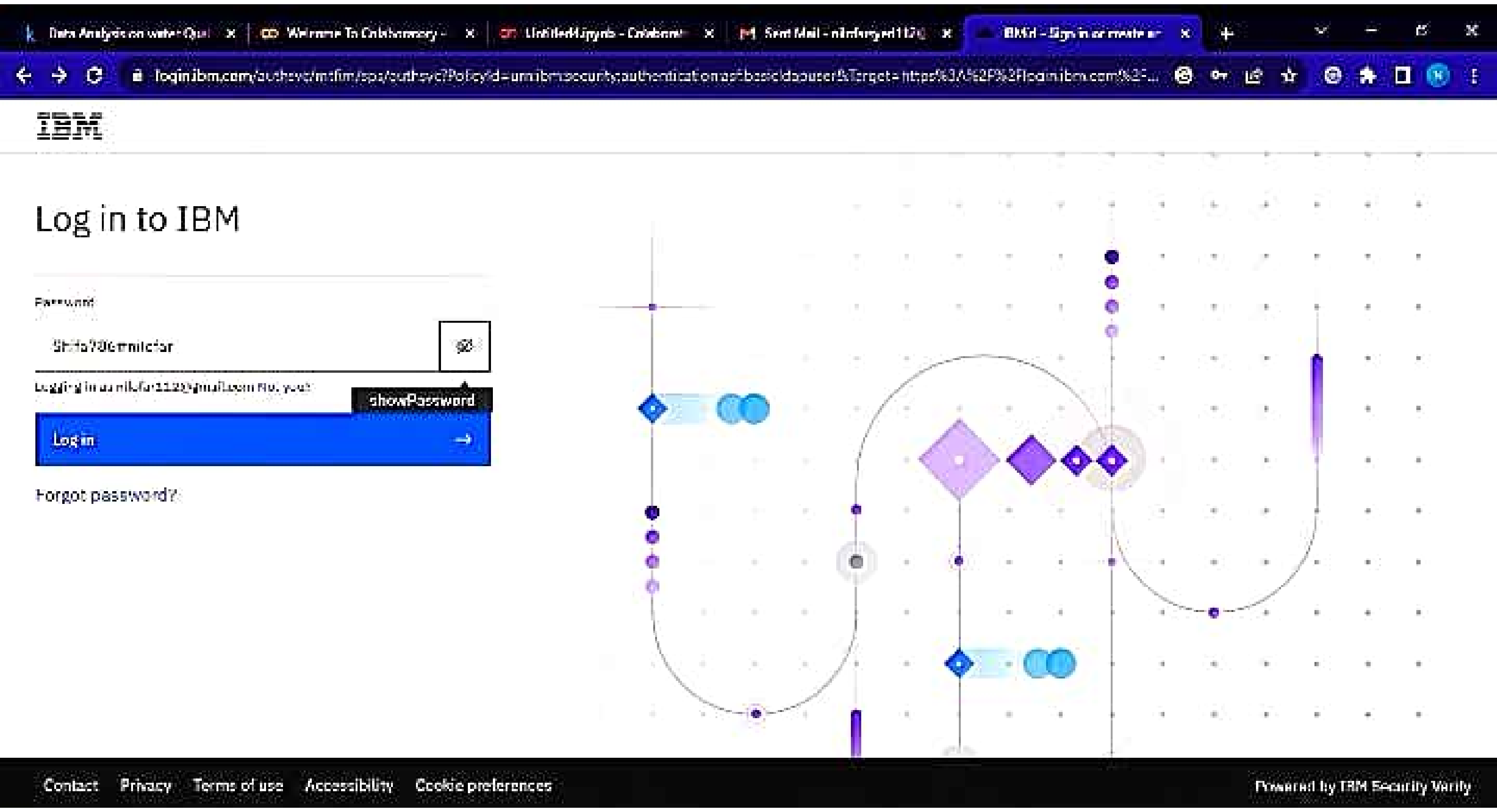
0 exceeds 1 in Chloramines by 5008.

Potability 1 has the lowest total Chloramines at over nine thousand.

Potability 0 has the highest total Chloramines at over 14 thousand.

Across all values of Potability, the sum of Chloramines is over 23 thousand.

Chloramines ranges from over nine thousand, when Potability is 1, to over fourteen thousand, when Potability is 0.



Data Analysis on water Qual
Welcome To Colaboratory -
Untitled4.ipynb - Colabora
Sent Mail - nilofaryed112@
IBMid - Sign in or create an
login.ibm.com/authsvc/mtfim/sps/authsvc?PolicyId=urn:ibm:security:authentication:asf:basicldapuser&Target=https%3A%2F%2Flogin.ibm.com%2F...
IBM

Log in to IBM

Password

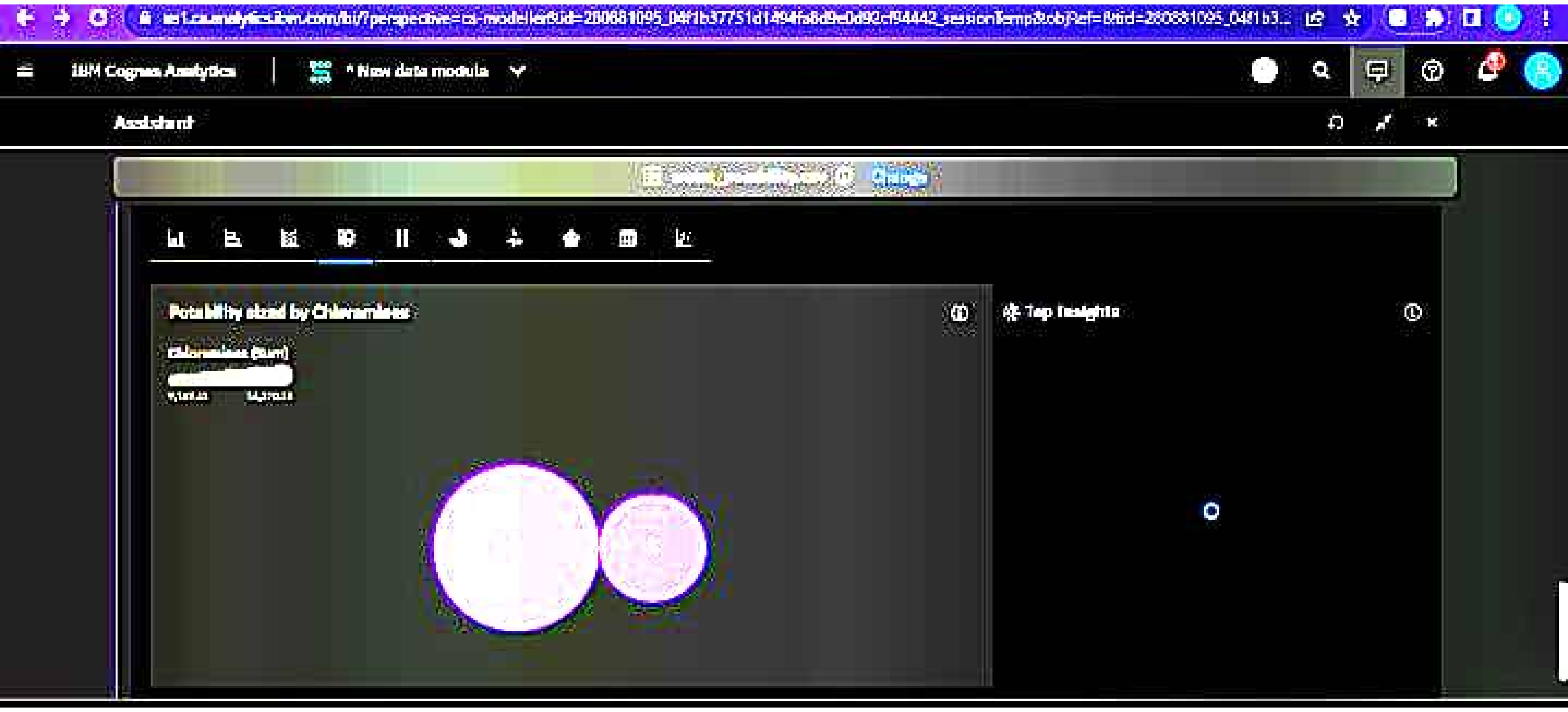
Shifa786nilofar

Logging in as nilofar112@gmail.com Not you?

Log in →

Forgot password?

Contact
Privacy
Terms of use
Accessibility
Cookie preferences
Powered by IBM Security Verify


us1.ca.analytics.ibm.com/bi/?perspective=ca-modeller&id=280881095_04f1b37751d1494fa8d9e0d92cf94442_sessionTemp&objRef=&tid=280881095_04f1b3...
IBM Cognos Analytics
* New data module
Assistant
Change
Potability sized by Chloramines
Chloramines (Sum)
Top Insights