IS-733 DATA MINING

Homework-1

Submitted By:-

Shashank Puppala

UB01976

1.Create a dataset profile table that gives an overview of the dataset.

1 Ans:- The dataset consists of data related to impurities present in the water, based on which water quality can be measured and analyzed. The dataset contain 2370 rows with 17 different attributes describing the data. By using the data profiling library, we are able to create the dataset profile table as follows;

a.) Total number of instances/rows - **2370**

b.) Total number of features/columns - **15**

c.) Statistics about each column are as shown in the figure below of the dataset profile; (*Continued on the next slide…*)

Dataset Profile

Pandas Profiling Report

Overview Variables Interactions Correlations Missing values Sample

Overview Alerts 26 Reproduction

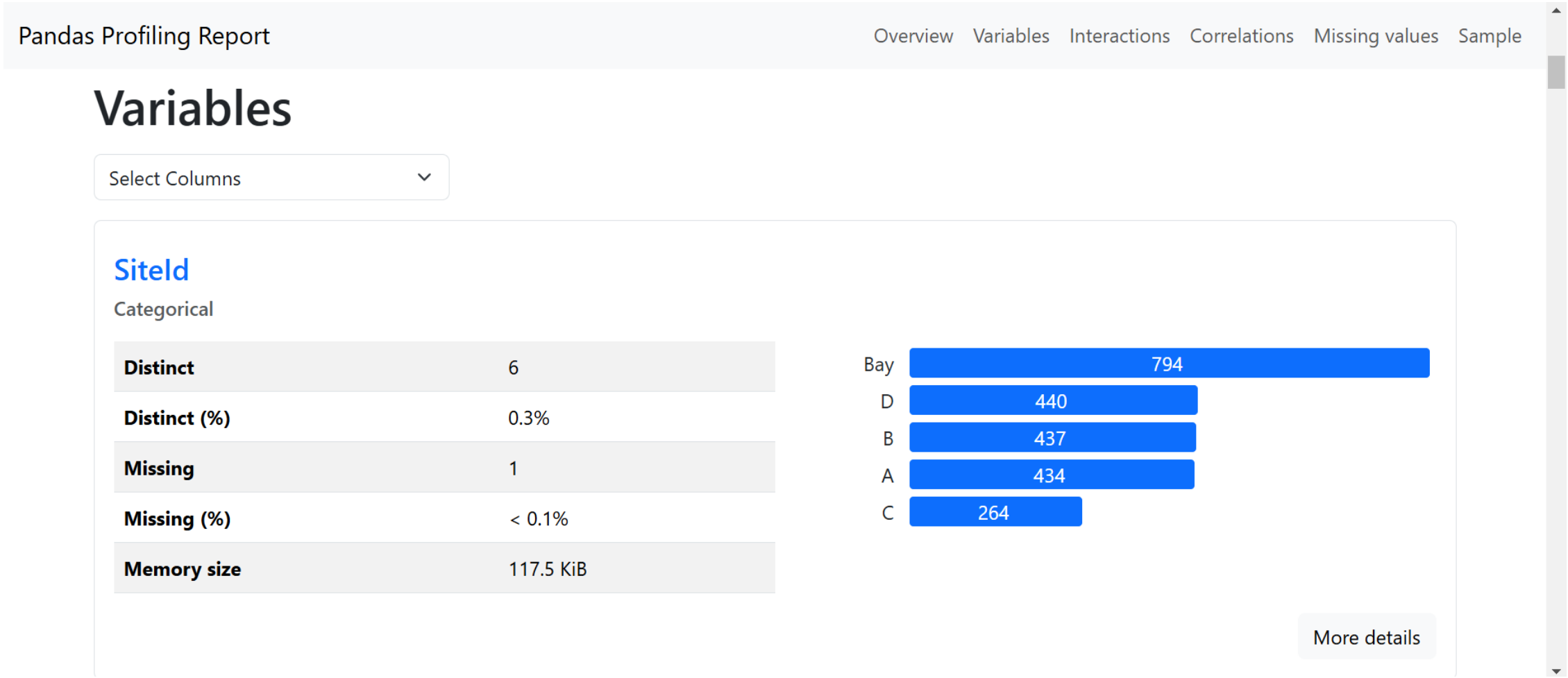
Dataset statistics

Number of variables	15
Number of observations	2371
Missing cells	9981
Missing cells (%)	28.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	903.1 KiB
Average record size in memory	390.0 B

Variable types

Categorical	5
DateTime	1
Numeric	8
Text	1

The statistics for individual attributes are displayed as follows in dataset profile, giving information regarding whether it is numeric or discrete (Categorial), or temporal ;



For all the attributes, it calculates the number of distinct values, missing values, and what percentage are missing

Pandas Profiling Report

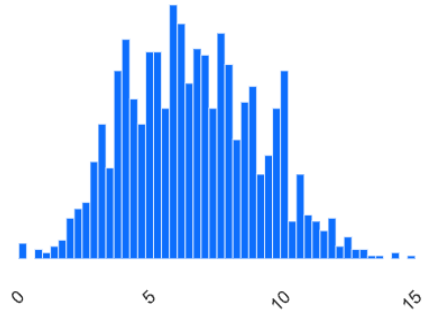
Overview Variables Interactions Correlations Missing values Sample

Dissolved Oxygen milligram per lit

Real number (ℝ)

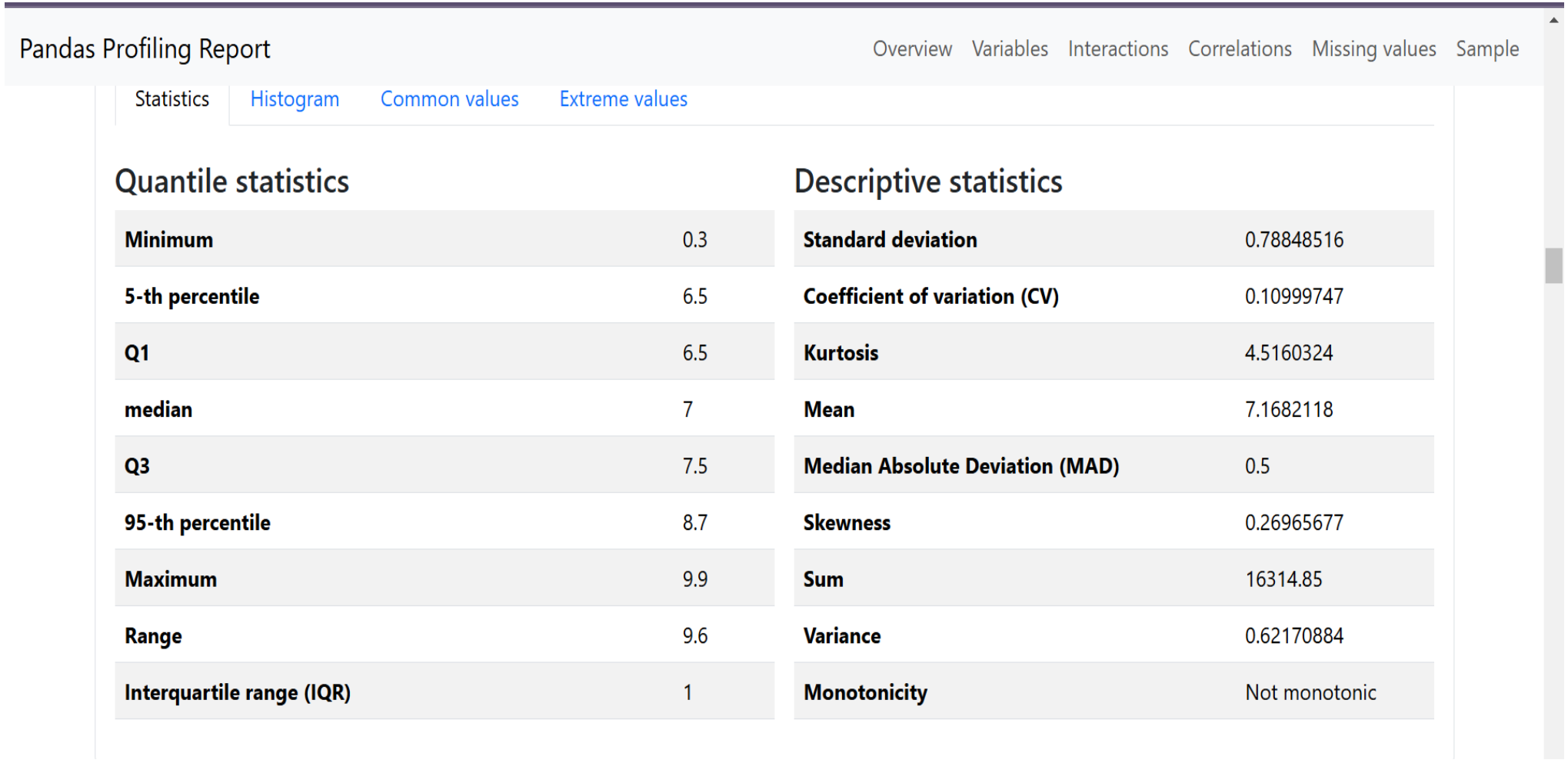
High correlation Missing

Distinct	154	**Minimum**	0
Distinct (%)	10.1%	**Maximum**	15.1
Missing	851	**Zeros**	4
Missing (%)	35.9%	**Zeros (%)**	0.2%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6.6462632	**Memory size**	18.7 KiB



More details

For the numeric attributes, the values of mean, median, and standard deviation are also calculated by using the profiling package (for example, below one shows statistics of pH attribute);



Pandas Profiling Report

Overview Variables Interactions Correlations Missing values Sample

Statistics Histogram Common values Extreme values

Quantile statistics

Minimum	0.3
5-th percentile	6.5
Q1	6.5
median	7
Q3	7.5
95-th percentile	8.7
Maximum	9.9
Range	9.6
Interquartile range (IQR)	1

Descriptive statistics

Standard deviation	0.78848516
Coefficient of variation (CV)	0.10999747
Kurtosis	4.5160324
Mean	7.1682118
Median Absolute Deviation (MAD)	0.5
Skewness	0.26965677
Sum	16314.85
Variance	0.62170884
Monotonicity	Not monotonic

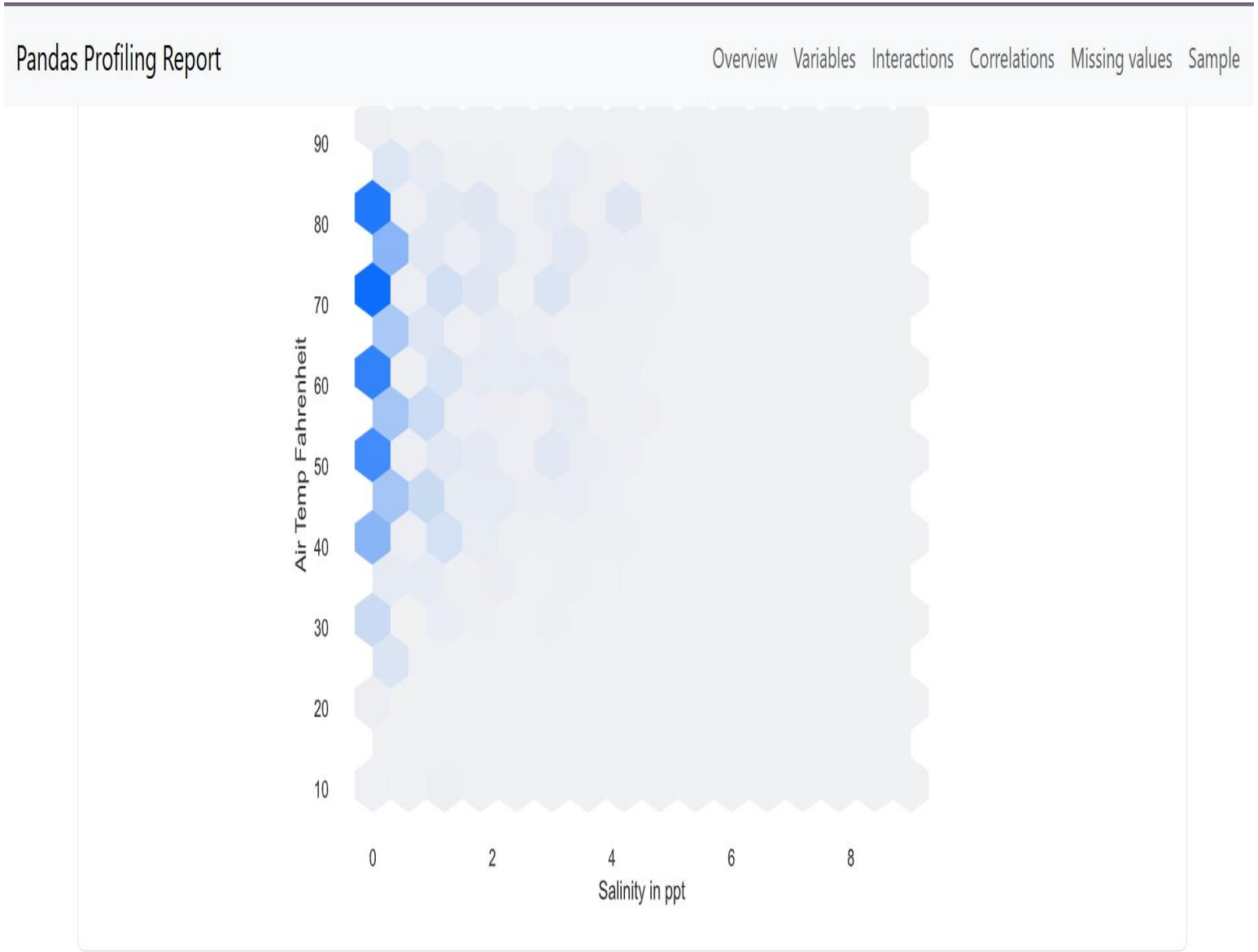
For the discrete attributes, we calculate the number of unique values per attribute, and top three attributes with largest count;

```
Total Number of Unique Values per Attribute:
SiteId: 6 unique values
UnitId: 2 unique values
ReadDate: 801 unique values
Time 24 clcok: 90 unique values
Field Technician: 14 unique values
DateVerified: 44 unique values
WhoVerified: 12 unique values

Top Three Attribute Values with the Largest Count:

SiteId:
SiteId
Bay    794
D      440
B      437
Name: count, dtype: int64
```

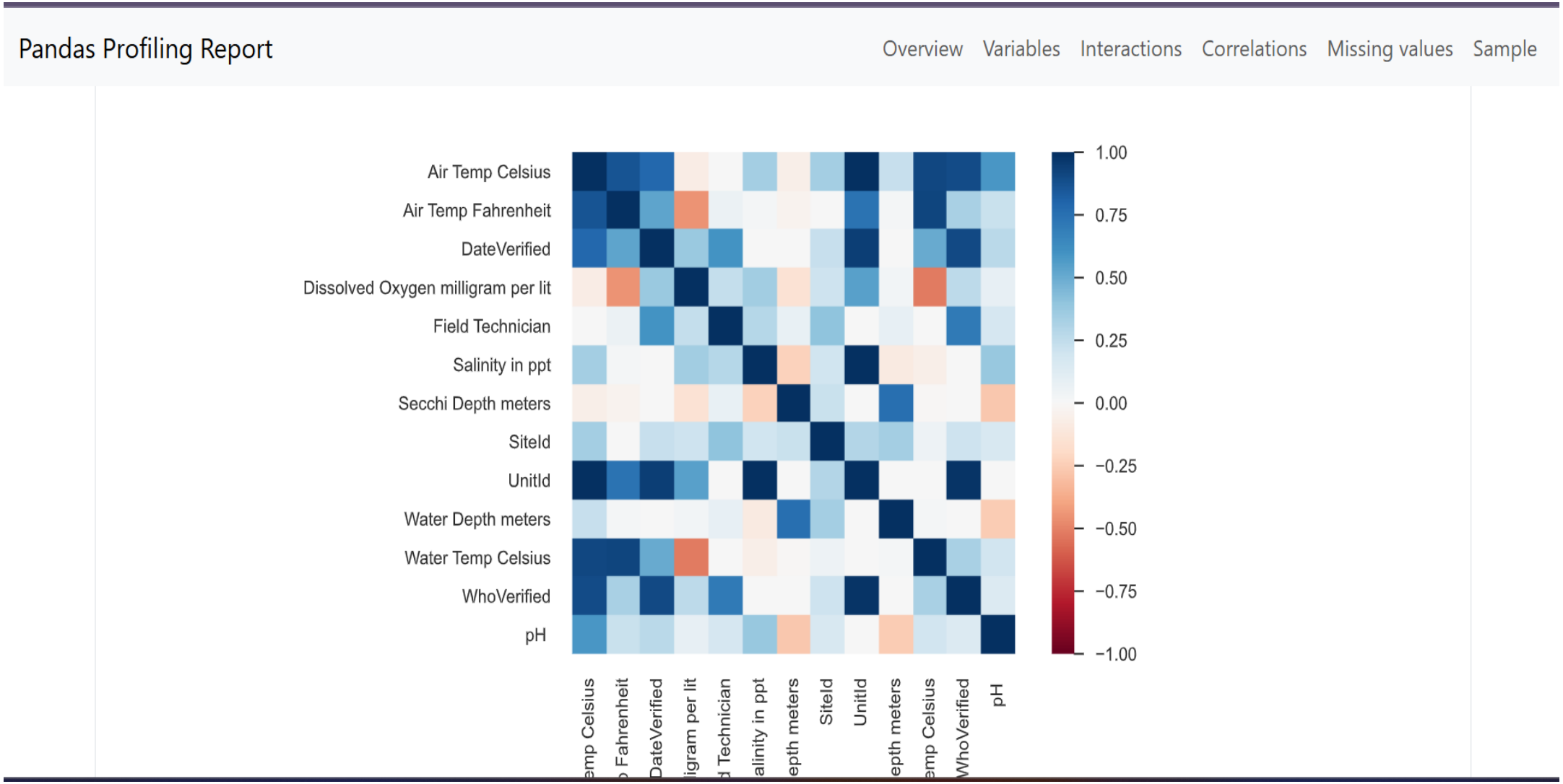
For each attribute, number of missing values in each column and what is the percentage of missing values is also calculated;

- We even have the interaction plots and correlation plots, to understand the behavior of one attribute with respect to other. This will allow to understand how the attributes are affecting each other.
- We can display the sample data as well, i.e., the first and last few rows, to let user check and relate it to the data as well.

- Interactions Plot-



Correlations Plot -



Sample data displayed in the Dataset Profiling:-

Pandas Profiling Report

Overview Variables Interactions Correlations Missing values Sample

Sample

First rows Last rows

	SiteId	UnitId	ReadDate	Salinity in ppt	Dissolved Oxygen milligram per lit	pH	Secchi Depth meters	Water Depth meters	Water T
0	Bay	NaN	01-03-1994	1.3	11.7	7.3	0.40	0.40	5.9
1	Bay	NaN	1/31/1994	1.5	12.0	7.4	0.20	0.35	3.0
2	Bay	NaN	02-07-1994	1.0	10.5	7.2	0.25	0.60	5.9
3	Bay	NaN	2/23/1994	1.0	10.1	7.4	0.35	0.50	10.0
4	Bay	NaN	2/28/1994	1.0	12.6	7.2	0.20	0.40	1.6
5	Bay	NaN	03-07-1994	1.0	9.9	7.1	0.20	0.90	9.7
6	Bay	NaN	3/14/1994	0.5	10.4	7.2	0.25	0.75	9.8

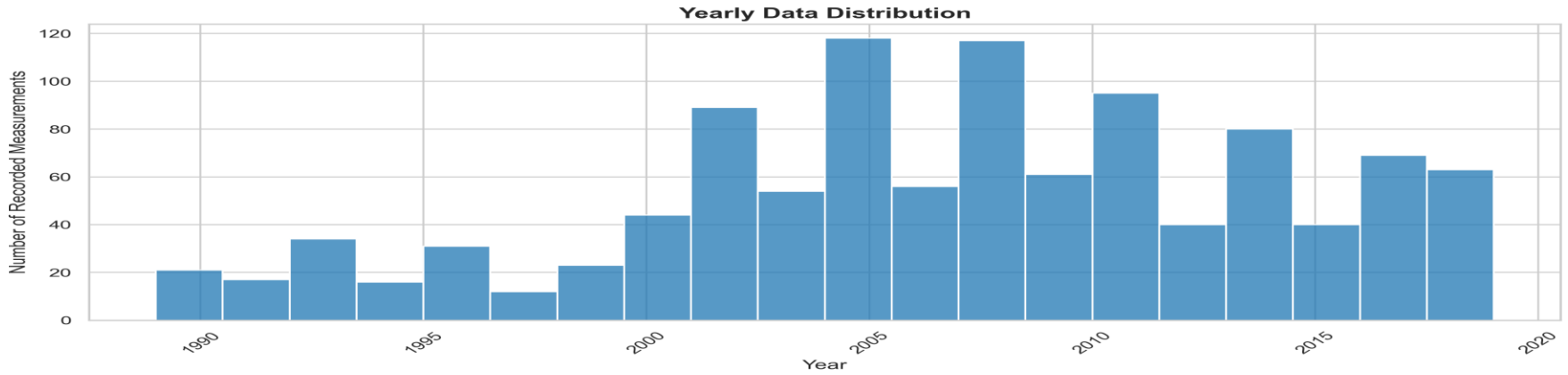
2. Generate a series of plots to describe the temporal pattern (year-to-year, monthly, and day-of-week) or other aggregate patterns.

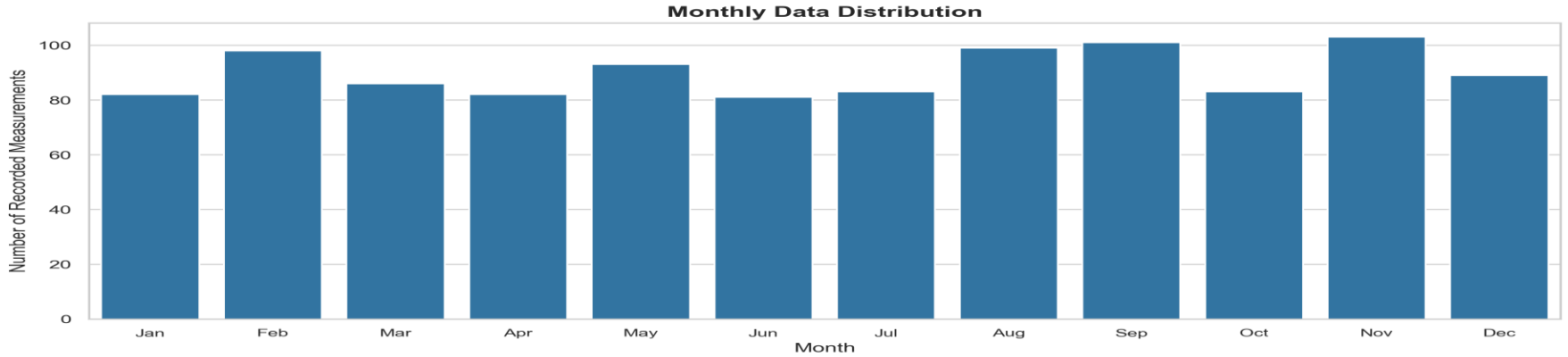
2 Ans:- Temporal patterns generally refers to sequences of events, or trends over time. They give insights into how the data behaves in a time-dependent context. As, out dataset consists of data of water metrics over years, we can draw various plots over years, months, and day of weeks.

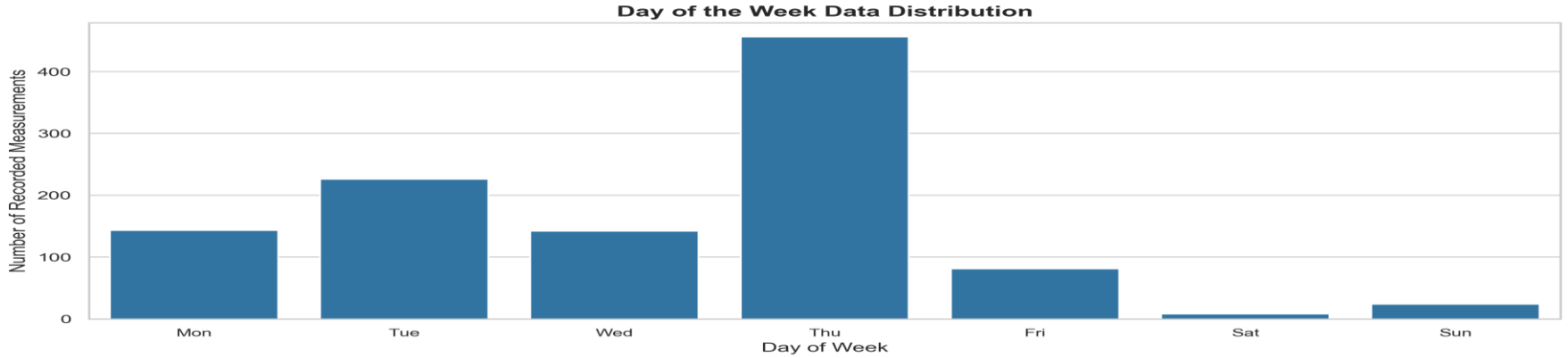
The temporal patterns are as shown in the below diagram;

(*Continued on the next slide..*)

TEMPORAL PATTERNS OF DATA







Analysis from the Temporal patterns:-

- From the yearly pattern, there is no much historical data before 1995, but the data is collected in a scheduled pattern from the years about 2000. And the maximum data is collected between the years 2005 to 2010.
- In case on the monthly trends, we could see that data is collected in more numbers during the starting of the year in January month and during the end of the year in November.
- Coming to the case of days of the week, the data collected on Thursday is highest compared to all the other days of week, and very low on weekends.

3. Generate a plot describing the distribution of your data, think of what machine learning problem could be around.

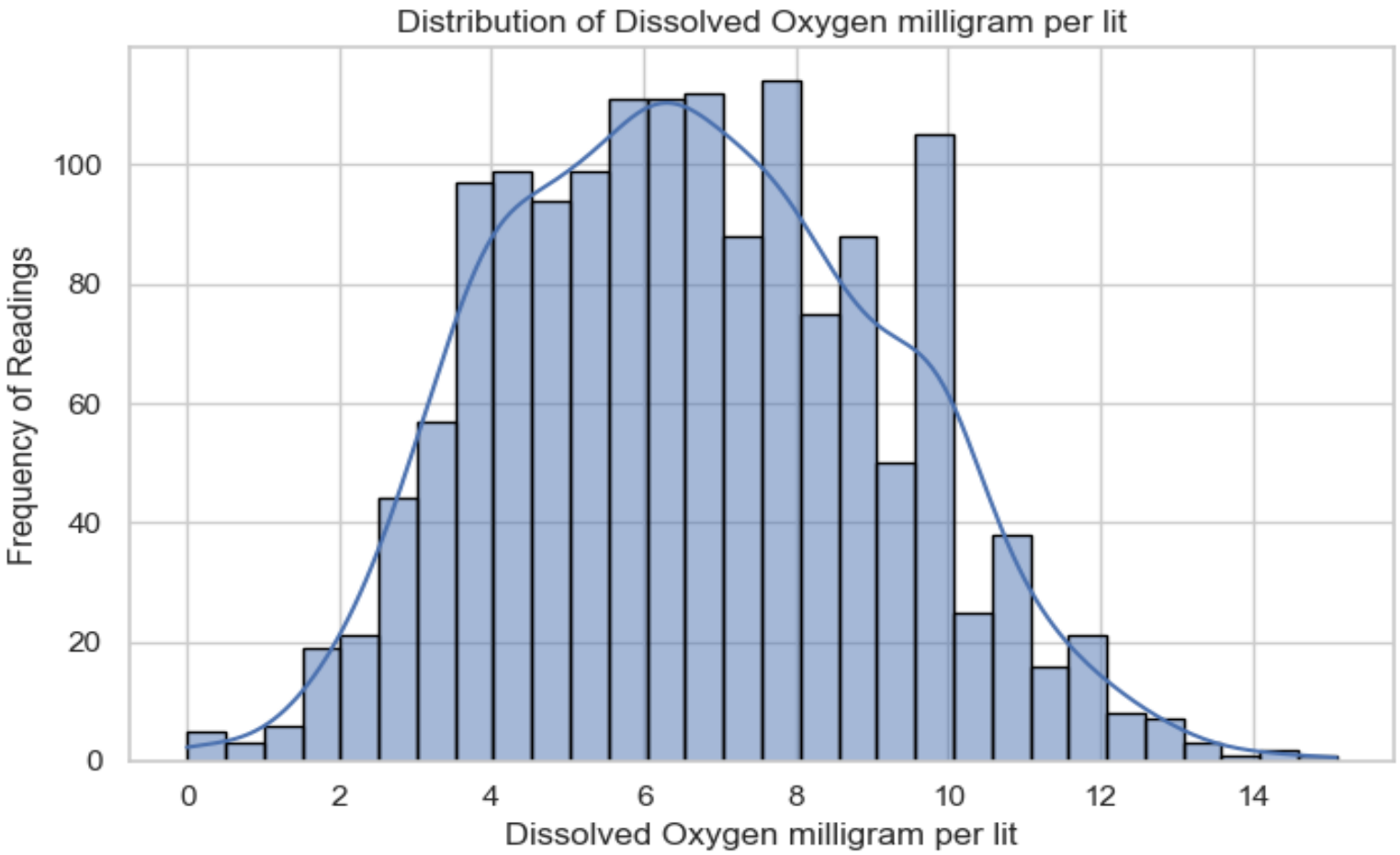
3 Ans:- The histogram plots below give the distribution of various water quality metrics. On the y-axis the frequency of the readings is taken and on the x-axis the water quality metric. Based on the distribution of data and after analyzing the data, one machine learning problem would be **Analyzing whether the water quality is good or bad**.

Using the metrics, we can draw insights whether the water is good or bad and can conclude whether it can be used or not.

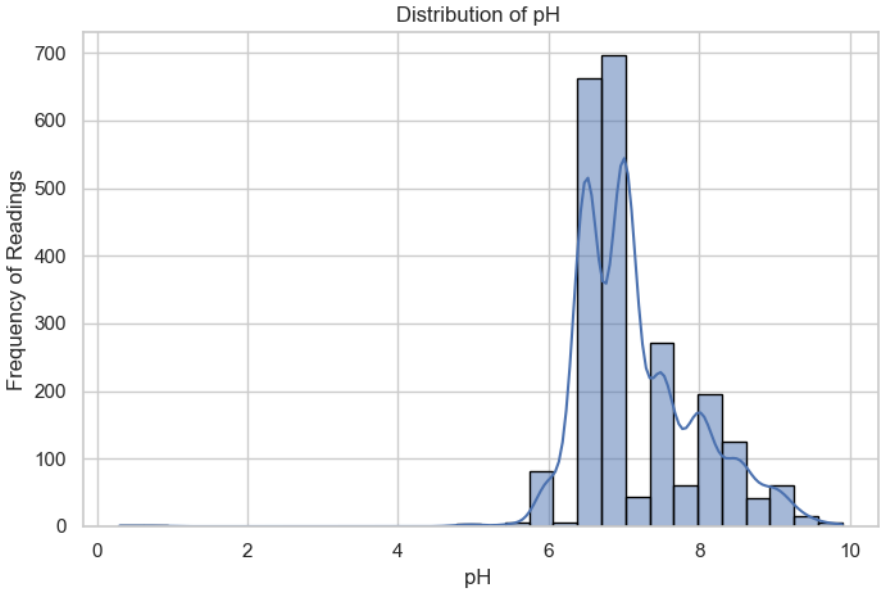
(*Continued on next slide* ..)

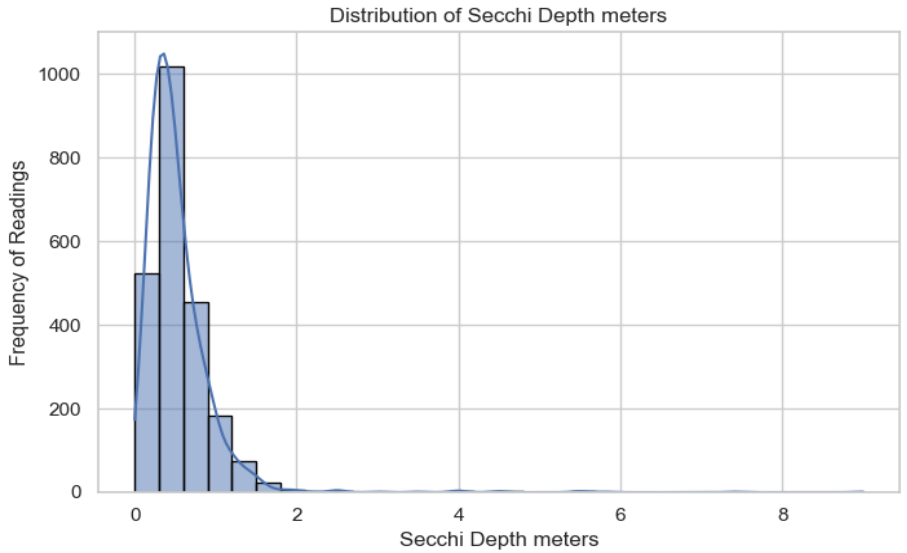
Distribution of Data for various Metrics:-

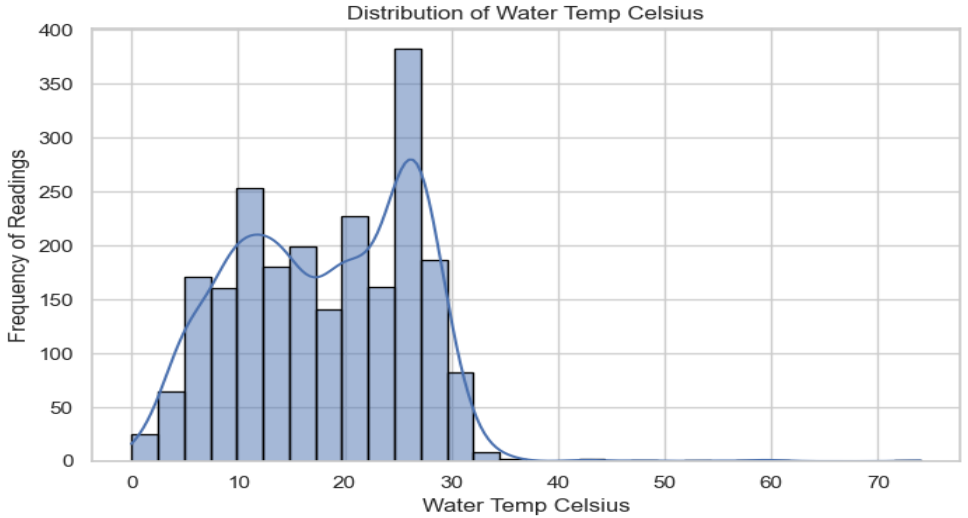
(Below is the plot for Dissolved oxygen)



Similarly, the plots for pH, Secchi Depth, Water temperature are as follows;







The basic description and information regarding the metrics is as follows;

```
RangeIndex: 2371 entries, 0 to 2370
Data columns (total 15 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   SiteId                                2370 non-null   object
 1   UnitId                                32 non-null     object
 2   ReadDate                              2366 non-null   object
 3   Salinity in ppt                       2241 non-null   float64
 4   Dissolved Oxygen milligram per lit    1520 non-null   float64
 5   pH                                    2276 non-null   float64
 6   Secchi Depth meters                   2298 non-null   float64
 7   Water Depth meters                    2300 non-null   float64
 8   Water Temp Celsius                    2250 non-null   float64
 9   Air Temp Celsius                      85 non-null     float64
 10  Air Temp Fahrenheit                   2300 non-null   float64
 11  Time 24 clcok                         2308 non-null   object
 12  Field Technician                      2332 non-null   object
 13  DateVerified                          453 non-null    object
 14  WhoVerified                           453 non-null    object
dtypes: float64(8), object(7)
memory usage: 278.0+ KB
None
       Salinity in ppt  Dissolved Oxygen milligram per lit           pH   \
count       2241.000000                        1520.000000  2276.000000
mean           0.717068                           6.646263     7.168212
std            1.230819                           2.506608     0.788485
min            0.000000                           0.000000     0.300000
25%            0.000000                           4.800000     6.500000
50%            0.000000                           6.500000     7.000000
75%            1.000000                           8.500000     7.500000
max            9.000000                          15.100000     9.900000

       Secchi Depth meters  Water Depth meters  Water Temp Celsius  \
count          2298.000000         2300.000000         2250.000000
mean              0.524898            0.762559           18.062138
std               0.473663            0.621140            8.298246
min               0.000000            0.010000            0.000000
25%               0.300000            0.400000           11.000000
50%               0.400000            0.650000           19.000000
75%               0.650000            0.950000           25.000000
max               9.000000           12.000000           74.000000

       Air Temp Celsius  Air Temp Fahrenheit
count         85.000000          2300.000000
mean          16.437647            62.051637
std           11.754138            15.492236
min            0.000000            10.500000
25%            9.000000            49.000000
50%           15.000000            63.000000
75%           21.700000            75.000000
max           74.000000            92.300000
```

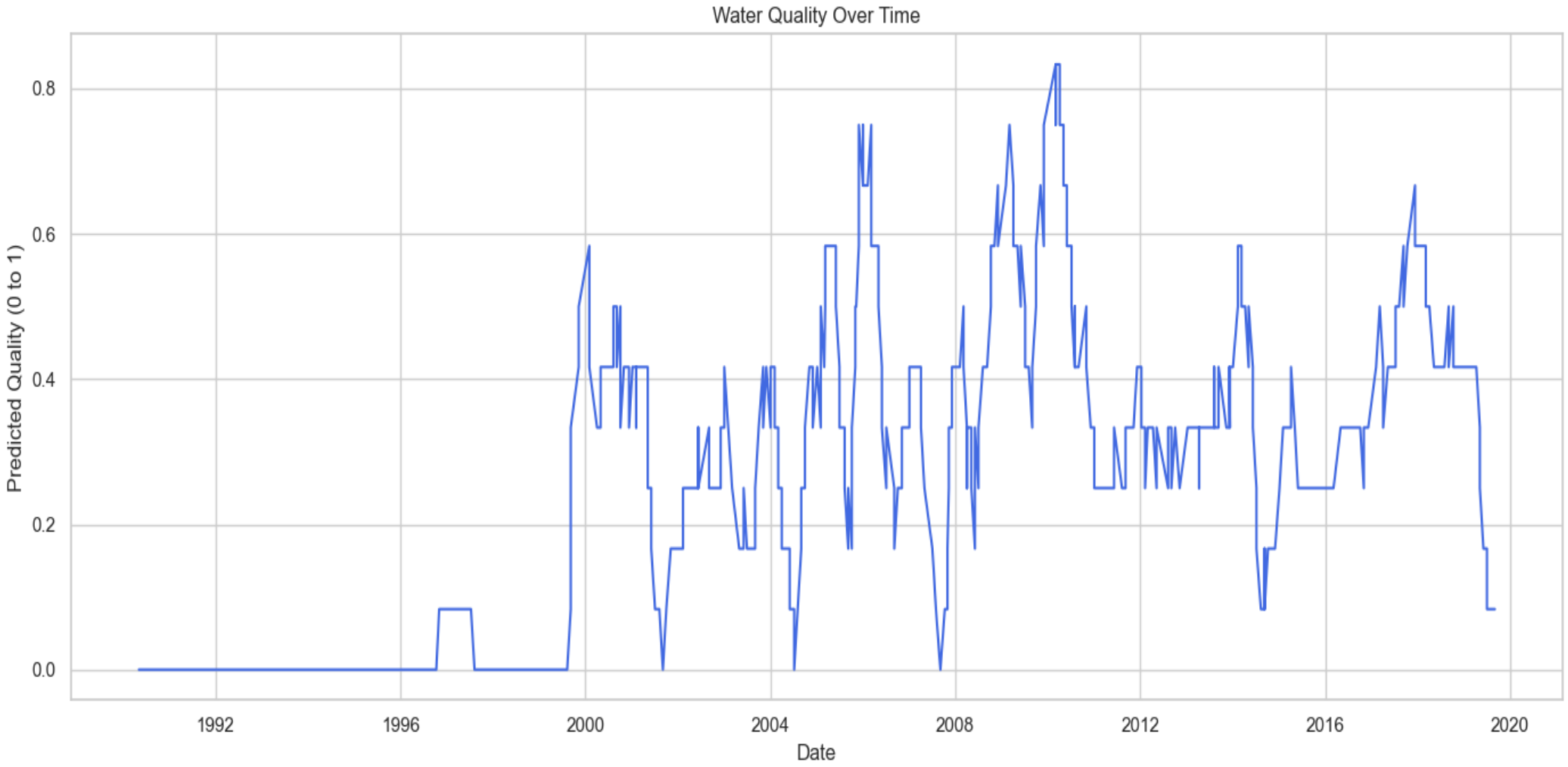
4. Generate a series of plots to illustrate to support your story and make your points clear.

4 Ans:- To show how the water quality is predicted, whether it is good or bad, we can use ML techniques like Random forest to analyze the data and train the model (We are using 80% of the data from the dataset to train the model, and 20% of the data to test the model). Once the model, is trained on the data, it will be able to determine whether water quality is good or bad.

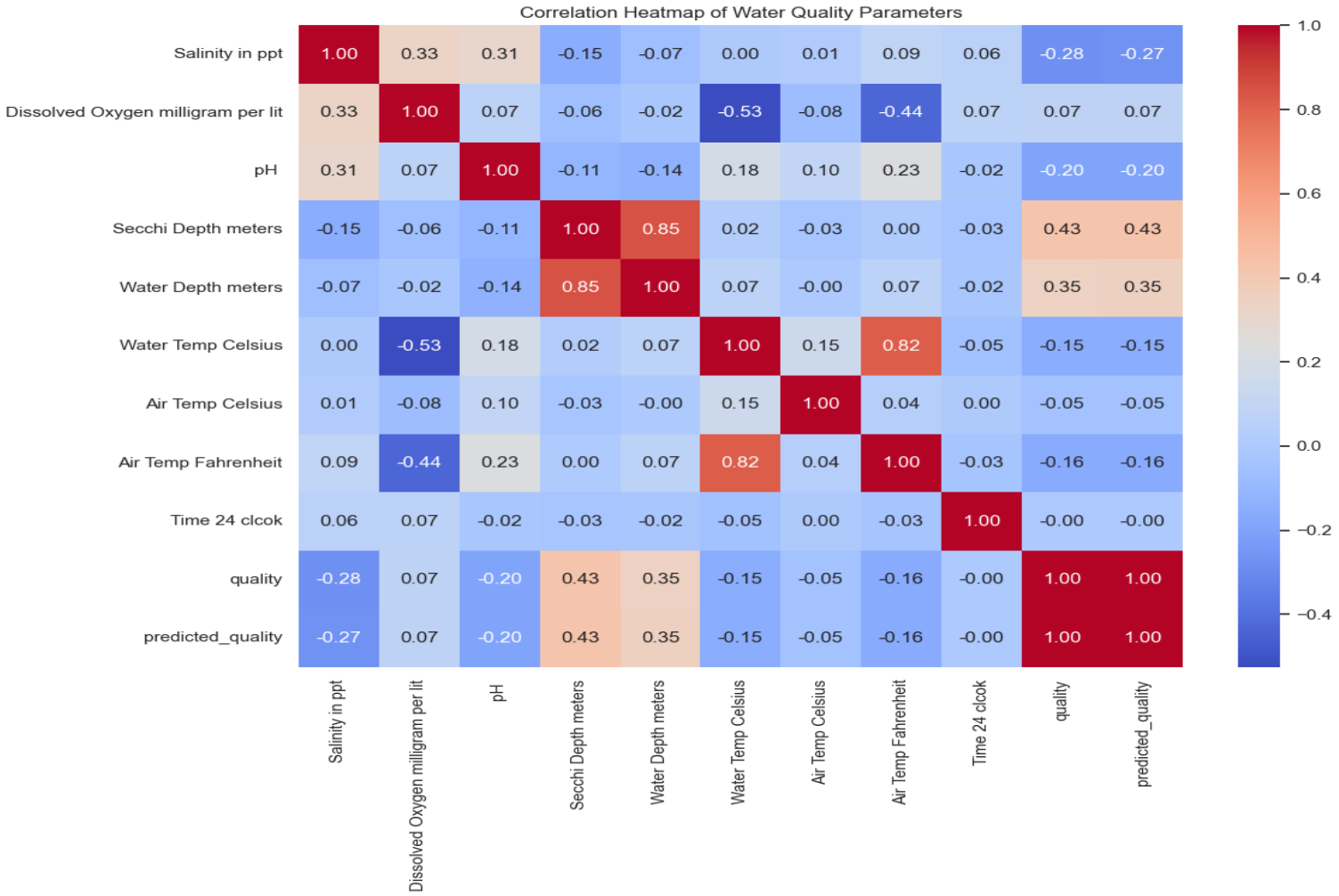
By using the line plot and heat map we understand the water quality. Please find the plots on the next slides to understand the pattern in the data;

(*Continued on next slides..*)

Line Plot of Water Quality



Water Quality Correlation Heat Map



Insights from Line plot and Heatmap:-

- The line plot clearly gives the statistics over the years on how the water quality is changing. Based on the metrics, water quality is measured (i.e. **0 - indicates Bad or Low Quality**, **1 – indicates Good or High Quality**). During the years, between **2008 and 2012** the water quality improved a lot and was good till **2011**, later it got reduced and gradually increased by **2018**.
- The correlation Heatmap, gives the information regarding which **parameters must be increased**, and **which must be decreased**. The **one's with negative values must be reduced** to maximum extent to increase the water quality and **one's with positive must be increased** to maintain or increase the quality of water.
- From the heatmap plot, **pH, salinity in water, water temp., air temperature must be reduced**, and **Dissolved oxygen, Secchi depth, and water depth must be increased** to increase water quality.

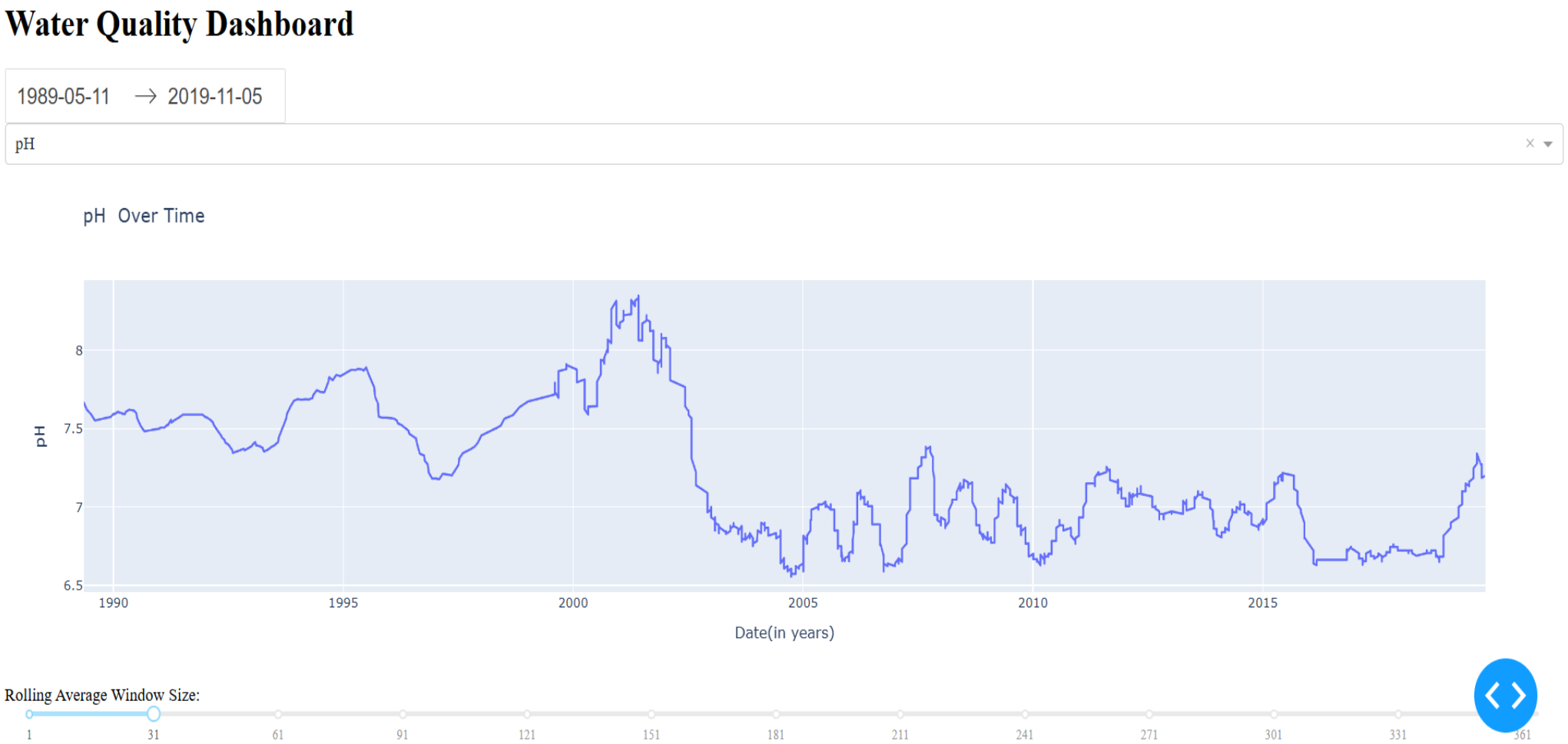
5. Design a dashboard that allows users to explore the data pattern. You may get inspiration from tasks 2-4, but feel free to add insights.

5 Ans:- A Dashboard is created with all the water quality metrics spread over years span, allowing users to explore different patterns of water quality over years. Provided are the options to zoom in and out, download the plot in case if required, reset the axes, Auto scale the plot.

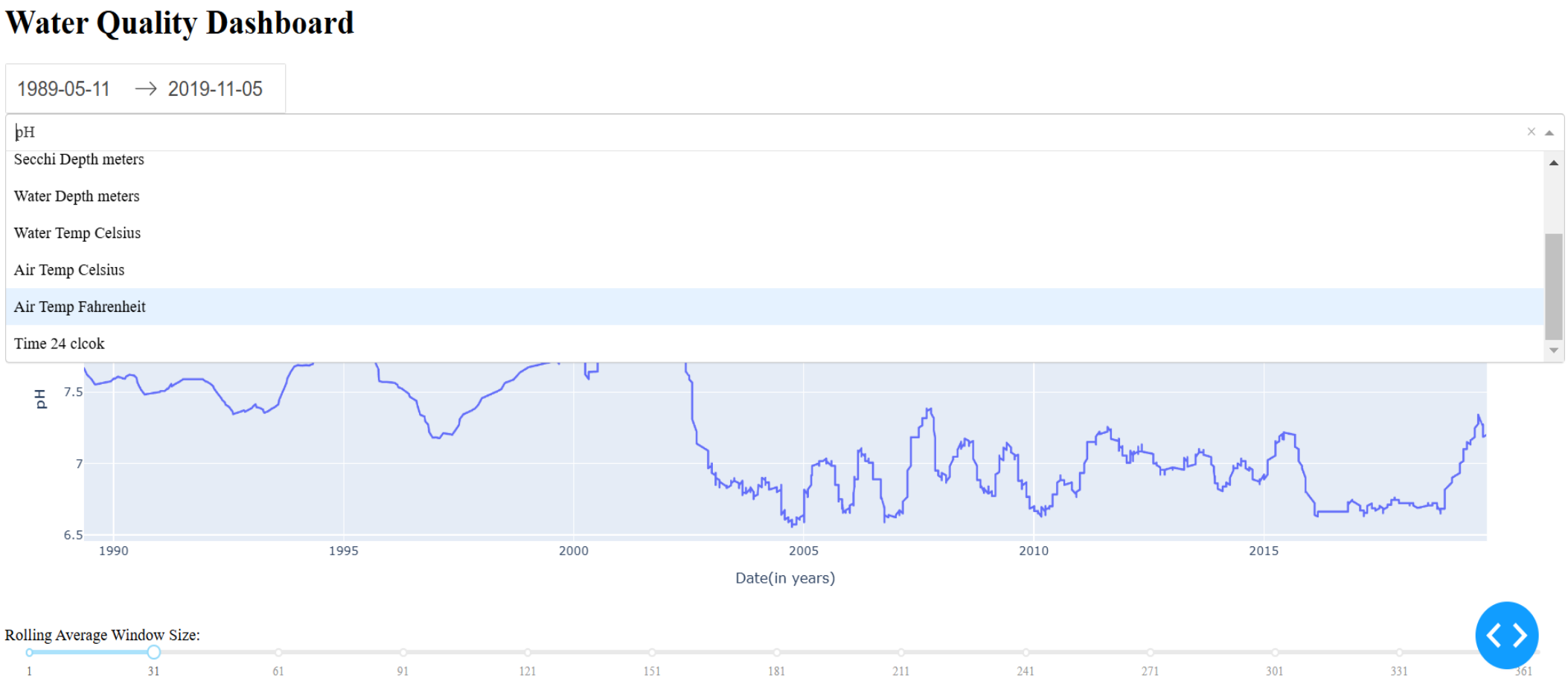
There are options to even highlight the plot to have a closer look at the value at that time instance. This way we can compare overall span with one particular date as required. Below is the dashboard generated;

(*Continued on next slide..*)

Dashboard showing Data Insights



The Dashboard provides dropdown as well to select various attributes and check their trends over years;



There is even a scaler as "**Rolling average window size**", which is used to smooth out fluctuations during the small intervals and reduce the noise. This helps to adjust size according the requirement, recent changes(smaller window) and overall (larger window).

Using of the Dashboard :-

- To view the dashboard, it is necessary for the dash application to run and generate the plot.
- So, need to execute code related to Question-5 to get the server running and dash application running, to generate the dashboard.
- The Dashboard is found at the following address:-

http://127.0.0.1:8050/

(*Please run the code before clicking on the link to view the dashboard*)

Conclusion

- The water quality is determined over years, and over the years between **2008 to 2011 it has greatly improved**, and after that it gradually decreased and maintained steady fluctuations at **0.6-0.8** range (i.e., where water quality is Average) over years of **2018 to 2020**.
- The dashboard gives clear understanding on how the attributes are changing over the period of time. Especially Salanity, pH, and Dissolved oxygen which are very much important to determine the water quality.
- One more important point that is understood from the plots is, there is need to increase the dissolved oxygen in water to make it good and deccrease the pH, and salanity of the water.

The Corresponding python code is uploaded to GitHub. Please find it in the below link:-
https://github.com/UB01976/is7332025/tree/main/data-mining-project-repo/hw1/hw1_UB01976

The folder consists of the Dataset, Python codes file to generate the plots, Dataset Profile report, and jupyter notebook.

References:-

- Dataset - https://catalog.data.gov/dataset/water-quality-data-41c5e

- Dataset Landing page - https://iris.fws.gov/APPS/ServCat/Reference/Profile/117348