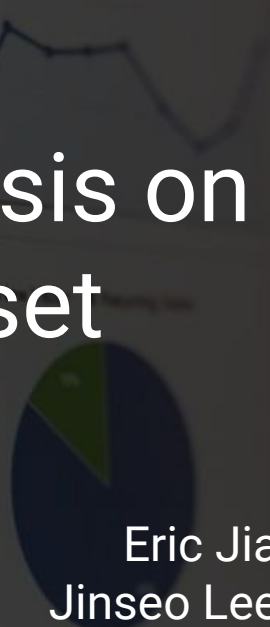


Exploratory Data Analysis on Water Quality Dataset



Eric Jia
Jinseo Lee
Dhanush Pucha

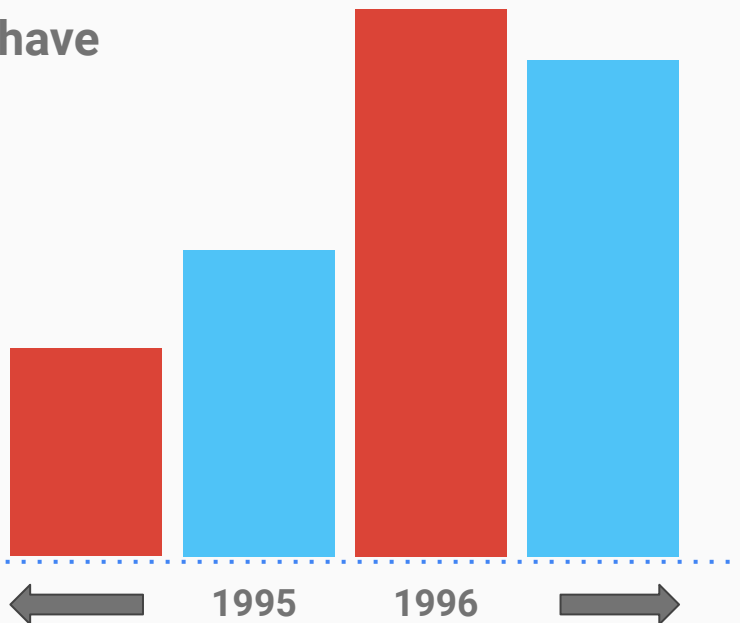




Observations of the Han River

Researchers from Japan and South Korea have been monitoring

- Salinity
- pH level
- Water temperature
- Water depth
- Air temperature



*The dataset included observations before and after 1995, as well as predictions based on data trends and scientific studies.

Initial Dataset:

	Site_Id	Unit_Id	Read_Date	Salinity (ppt)	Dissolved Oxygen (mg/L)	pH (standard units)	Secchi Depth (m)	Water Depth (m)	Water Temp (°C)	Air Temp-Celsius	Air Temp (°F)	Time (24:00)	Field_Tech	DateVerified	WhoVerified	AirTemp (C)
0	Bay	NaN	1/3/1994	1.3	11.7	7.3	0.40	0.40	5.9	8.0	46.40	11:00	NaN	NaN	NaN	8.000000
1	Bay	NaN	1/31/1994	1.5	12.0	7.4	0.20	0.35	3.0	2.6	36.68	11:30	NaN	NaN	NaN	2.600000
2	Bay	NaN	2/7/1994	1.0	10.5	7.2	0.25	0.60	5.9	7.6	45.68	9:45	NaN	NaN	NaN	7.600000
3	Bay	NaN	2/23/1994	1.0	10.1	7.4	0.35	0.50	10.0	2.7	36.86	NaN	NaN	NaN	NaN	2.700000
4	Bay	NaN	2/28/1994	1.0	12.6	7.2	0.20	0.40	1.6	0.0	32.00	10:30	NaN	NaN	NaN	0.000000
...
2366	Bay	NaN	10/11/2018	1.9	5.0	7.0	4.00	1.20	25.0	NaN	78.00	09:30	Sue Poe	11/13/2019	Christine Folks	25.555556
2367	Bay	NaN	10/24/2018	0.0	9.0	7.0	0.30	0.60	18.0	NaN	58.00	09:30	Sue Poe	11/13/2019	Christine Folks	14.444444
2368	Bay	NaN	10/28/2018	0.9	2.9	7.0	0.40	0.90	13.0	NaN	49.00	09:20	Sue Poe	11/13/2019	Christine Folks	9.444444
2369	Bay	NaN	11/7/2018	1.7	NaN	7.0	0.45	0.90	20.0	NaN	65.00	08:45	Sue Poe	11/13/2019	Christine Folks	18.333333
2370	Bay	NaN	12/11/2018	0.1	NaN	7.0	0.10	0.10	10.0	NaN	42.00	09:40	Sue Poe	11/13/2019	Christine Folks	5.555556

.drop() Function:

```
df.drop(columns = ['Site_Id', 'Unit_Id', 'Read_Date', 'Dissolved Oxygen (mg/L)', 'Secchi Depth (m)',  
                  'Air Temp-Celsius', 'Time (24:00)', 'Field_Tech', 'DateVerified', 'WhoVerified',  
                  'AirTemp (C)', 'Year'], inplace = True)
```

- The original Dataset had 2371 rows and 17 columns.
- pandas.drop() function was used to eliminate the unnecessary columns

.dropna() function

Gathering the 1st 400 rows

```
new_df = df.dropna()
```

```
new_df = new_df[:400]
```

Cleaned up Dataset:

	Salinity (ppt)	pH (standard units)	Water Depth (m)	Water Temp (?C)	Air Temp (?F)
0	1.3	7.3	0.40	5.9	46.40
1	1.5	7.4	0.35	3.0	36.68
2	1.0	7.2	0.60	5.9	45.68
3	1.0	7.4	0.50	10.0	36.86
4	1.0	7.2	0.40	1.6	32.00
...
482	1.0	7.9	0.65	19.5	75.00
483	1.2	8.7	0.45	23.0	71.00
485	1.5	8.4	0.64	26.0	73.70
486	1.5	8.5	0.85	26.0	83.00
487	1.0	8.6	0.80	27.0	77.00

*400 rows x 5 columns

- .dropna() function was used to remove all rows containing null values
- The first 400 rows is gathered and this is our new dataframe

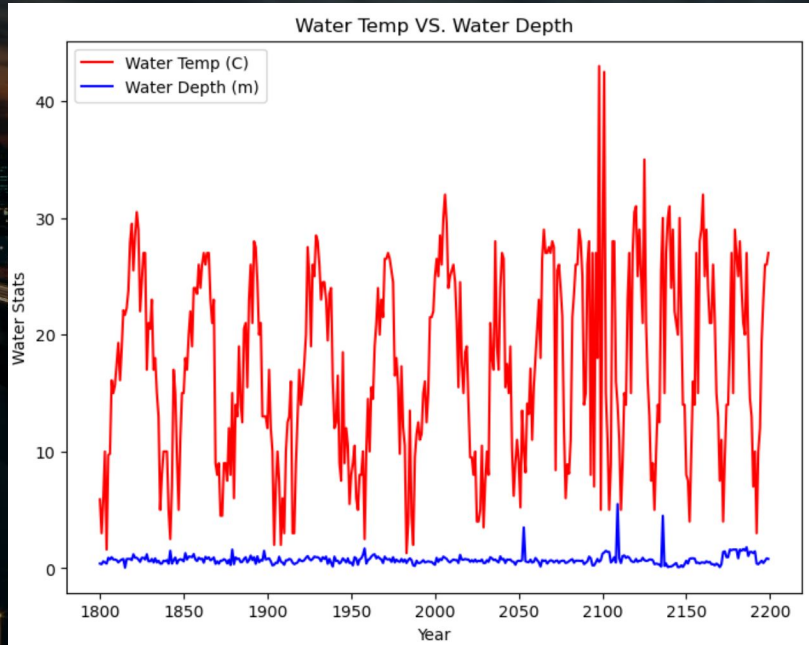
```
new_df[['Salinity (ppt)', 'pH (standard units)']].agg(['mean', 'median', 'var', 'std'])
```

	Salinity (ppt)	pH (standard units)
mean	2.154975	7.655500
median	1.900000	7.500000
var	2.130187	0.473015
std	1.459516	0.687761

An aggregate function was used to calculate the

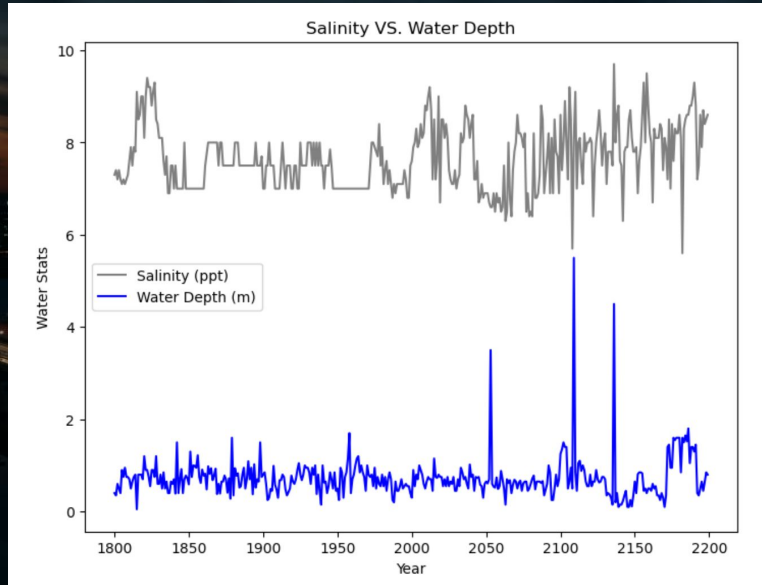
- Mean
- Median
- Variance
- Standard deviation (std)

Water Temp vs Water Depth



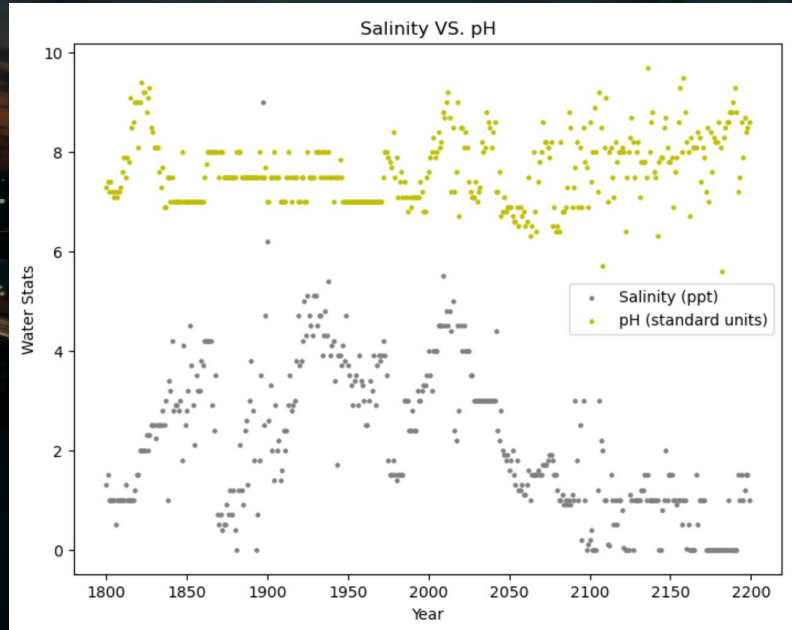
- Water temperature in the Han River (Korea) fluctuates significantly with seasonal changes every three months.
- Seasonal changes and weather patterns seem to have a greater influence on water temperature than water depth.
- The fluctuation in temperature suggests that water depth has little impact on the temperature of the Han River.

Salinity vs Water Depth



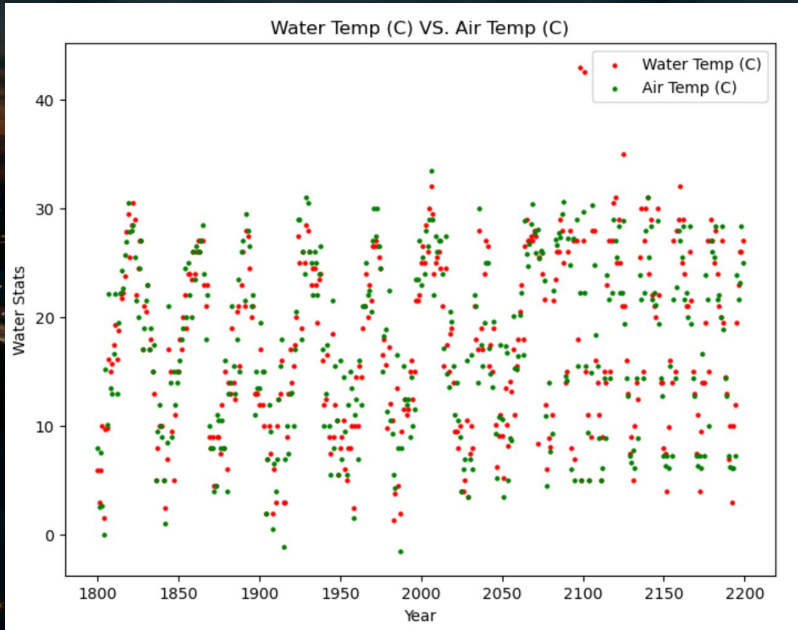
- A slightly consistent negative correlation was found between water depth and salinity in the river.
- As water depth increased, salinity tended to decrease, indicating a possible dilution effect with greater river flow.
- The correlation was not very strong, suggesting other factors could be influencing salinity levels in the river.

Salinity vs pH



- Water temperature in the Han River A weak correlation was found between salinity and pH in the Han River.
- Salinity levels do not seem to have a strong influence on the pH levels of the water.
- Other factors like pollution or natural sources could be more significant in determining pH levels in the river.

Water Temp vs Air Temp



- Strong positive correlation was found between water temperature and air temperature in the Han River.
- The temperature of the air is crucial in determining the temperature of the water in the river.
- Temperature of the water fluctuates due to heat transfer from the atmosphere, which is influenced by various environmental factors like weather patterns and seasonality.

Conclusion & Valuable Insights

- *Seasonal changes and weather patterns have a stronger influence on water temperature than water depth.*
- *A negative correlation exists between water depth and salinity levels, indicating a possible dilution effect from river flow.*
- *The weak correlation between salinity and pH levels suggests other factors, such as pollution or natural sources, may impact pH levels.*
- *A strong correlation was found between water temperature and air temperature, implying that air temperature significantly influences water temperature.*

**Further studies could provide additional insights into the complex relationships between these variables, informing water quality management and conservation efforts in the region.*