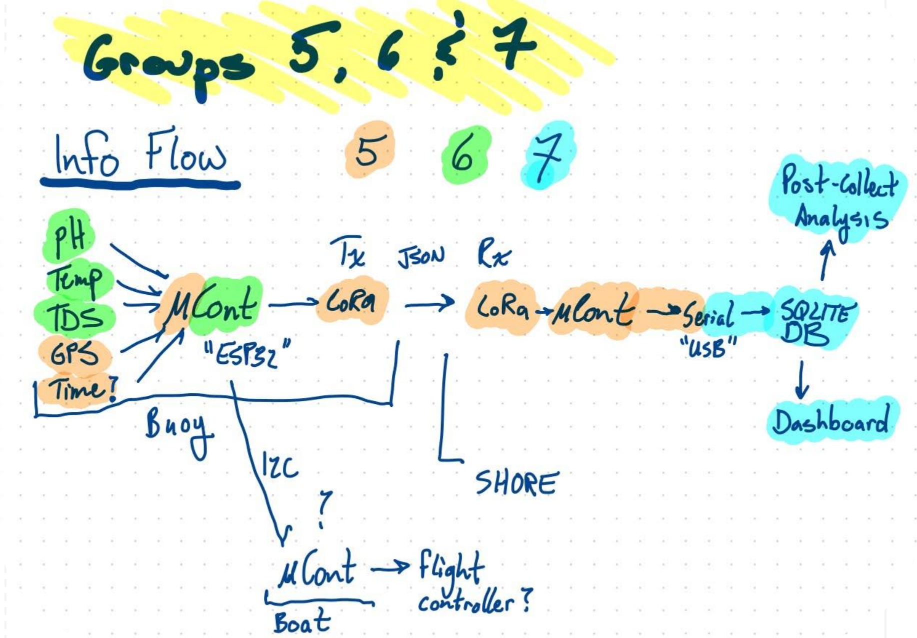
Group 7: Data Analysis & Visualization

Tanmayi Dasari, Sebastian Hazlett, Claire Lee, Dorothy Zhang, Alex Miller

1 Background/Scope

Overall Goal

We program the server to receive LoRa communication, process the received Lake Miramar buoy sensor data, analyze it and make conclusions & cool models.



	A: ...ctivity	B: Responsible	C: 18	D: 19	E: 20	F: 21	G: 22	H: 25	I: 26	J: 27	K: 28	29	1
	7.1 Data Storage												
	7.1.1 Create the Database	Whole Team											
4	7.1.1.1 Determine database software to use - SQLite	Whole Team											
5	7.1.1.2 Map the data tables and types	Whole Team											
6	7.1.1.3 Create the tables	Whole Team											
7	7.1.1.4 Populate and test with dummy data	Claire											
8	**7.1.2 Data Transfer**												
9		Team 7 and 6											
10	7.1.2.1 Set the transfer data format DONE	Alex											
11	7.1.2.2 Read data from serial connection DONE	Sebastian											
12	7.1.2.3 Verify data formatting DONE	Sebastian											
13	7.1.2.4 Log the processes DONE	Sebastian											
14	7.1.2.5 Write data to the database DONE	Sebastian Alex											
15	**7.2 Live Data Feed**												
16	7.2.1 Determine libaries to use												
17	7.2.2 Setup development environment	Whole Team											
18	7.2.3 Mock up the dashboard interface	Dorothy Tanmayi											
19	7.2.4 Read data from the database	Dorothy Tanmayi											
20	7.2.5 Code visualizations	Dorothy Tanmayi											
21	7.2.6 Code automatic refreshing	Dorothy Tanmayi											
22	**7.3 Testing**												
23	7.3.1 Merge with telemetry and data acquistion	Whole Team											
24	7.3.2 Test all processes in lab	Whole Team											
25	7.3.3 Iterate designs/code based on test results	Whole Team											
26	7.3.4 Test all processes in pool	Whole Team											
27	7.3.4 Iterate designs/code based on test results	Whole Team											
28	**7.4 Collect Data from Miramar Lake**												
29	**7.5 Data Analysis**												
30	7.5.1 Determine locations to survey	Teams 6 and 7											
31	7.5.2 Determine libaries to use	Claire Alex											
32	7.5.3 Setup development environment	Claire Alex											
33	7.5.4 Source real data from past surveys	Claire Alex											
34	7.5.5 Read data from database	Claire Alex											
35	7.5.6 Exploratory Data Analysis	Claire Alex											
36	7.5.6.1 Correlations	Claire Alex											
37	7.5.6.2 Histograms, box-whisker charts, summary statistics	Claire Alex											
38	7.5.6.3 Heatmaps based on location	Claire Alex											
39	7.5.7 Modeling	Whole Team											
	7.5.7.1 K-means to group data points based on location	Claire Alex											
	7.5.7.2 Ordinary least squares and ANOVA	Claire Alex											
	...5.8 Analysis and Conclusions	Claire Alex											
	...ster												

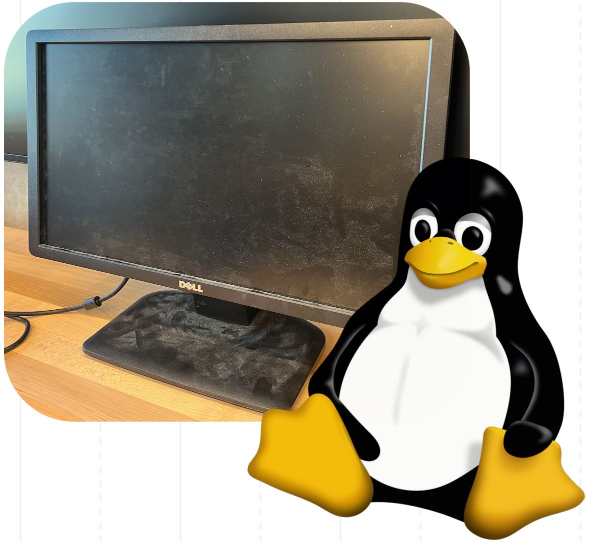
Gantt Chart

2 Procedure/Challenges

Hardware

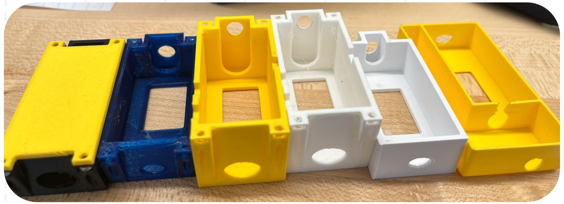
Base station (July 15-19)

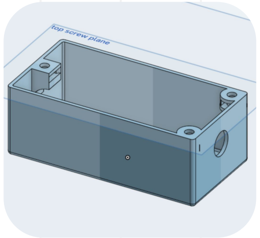
Set up base station server with Linux

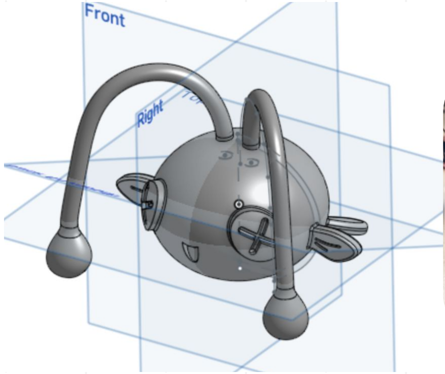


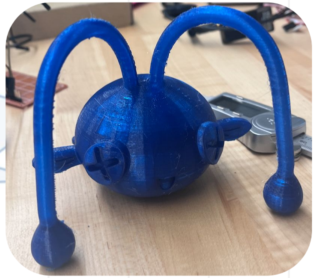
CAD/3D Printing (July 21-22)

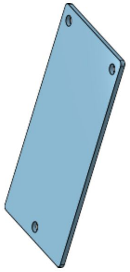
esp32 box & chinchou receiver case





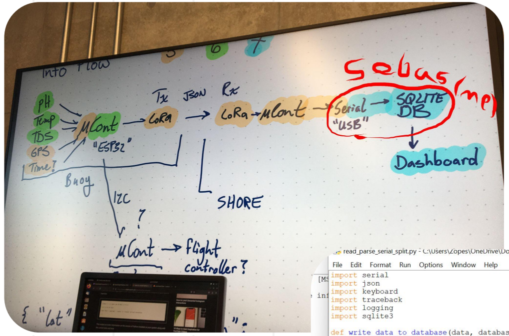


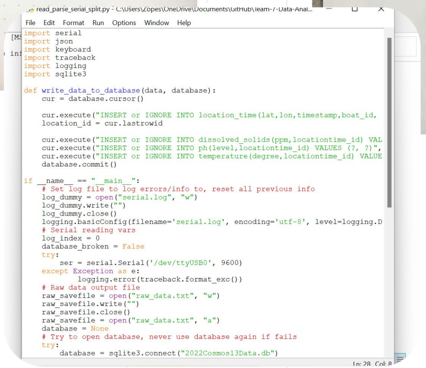




LoRa → Server Groundwork (July 18-26)

My task: take sensor data in serial from LoRa microcontroller into server (July 18)

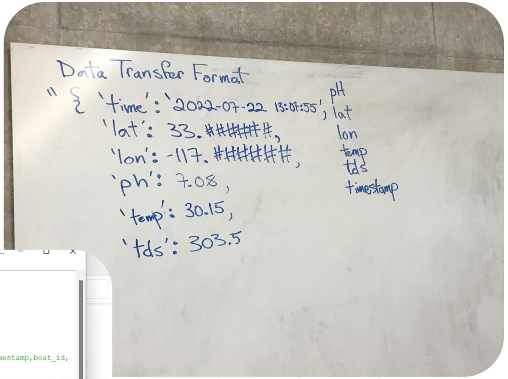




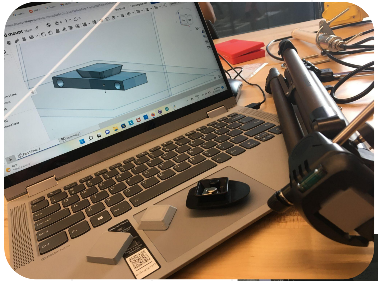
```
read_parse_serial_split.py - C:\Users\Zopes\OneDrive\Documents\GitHub\Team-7-Data-Anal...
File Edit Format Run Options Window Help
import serial
import json
import keyboard
import traceback
import logging
import sqlite3

def write_data_to_database(data, database):
    cur = database.cursor()

    cur.execute("INSERT or IGNORE INTO location_time(lat,lon,timestamp,boat_id,
    location_id = cur.lastrowid

    cur.execute("INSERT or IGNORE INTO dissolved_solids(ppm,locationtime_id) VAL
    cur.execute("INSERT or IGNORE INTO ph(level,locationtime_id) VALUES (?, ?)",
    cur.execute("INSERT or IGNORE INTO temperature(degree,locationtime_id) VALUE
    database.commit()

if __name__ == "__main__":
    # Set log file to log errors/info to, reset all previous info
    log_dummy = open("serial.log", "w")
    log_dummy.write("")
    log_dummy.close()
    logging.basicConfig(filename='serial.log', encoding='utf-8', level=logging.D
    # Serial reading vars
    log_index = 0
    database_broken = False
    try:
        ser = serial.Serial('/dev/ttyUSB0', 9600)
    except Exception as e:
            logging.error(traceback.format_exc())
    # Raw data output file
    raw_savefile = open("raw_data.txt", "w")
    raw_savefile.write("")
    raw_savefile.close()
    raw_savefile = open("raw_data.txt", "a")
    database = None
    # Try to open database, never use database again if fails
    try:
        database = sqlite3.connect("2022Cosmos13Data.db")
Ln: 28 Col: 8
```

LoRa received data -> database code (July 18-28)

Data transfer format (July 18)



Rapid prototyping to make the radio-to-stand connector by 3D print, with the radio connecting to LoRa and the server (July 20-21)

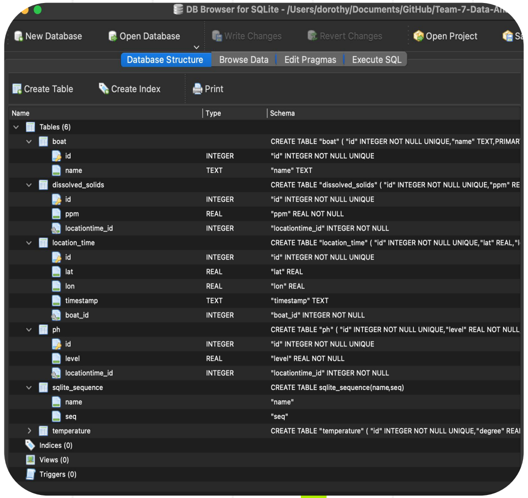


Testing the radio (July 26)

Database → Dashboard

Database

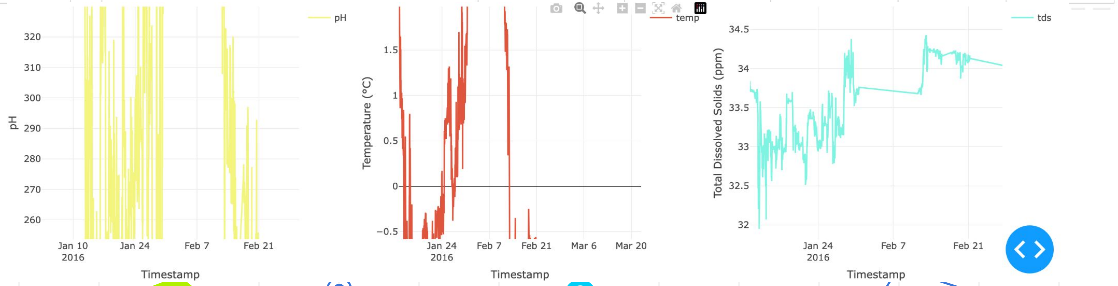
Loaded the dummy sensor data to database using DB Browser for SQLite



Dashboard

Used Pandas to access the SQL database and create Plotly formatted dataframe. Used Plotly Dash library to create datatable, scatterplot map, and line graphs

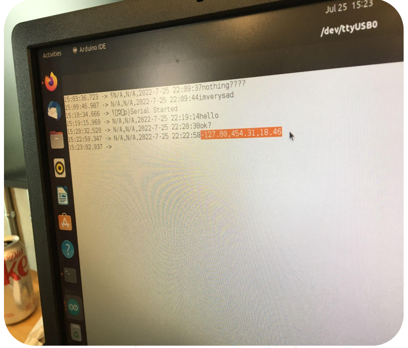
id	boat	timestamp	lat	lon	ph	temp	tds
10004	1	2022-07-25 10:05:00	1	1	1	1	1
10003	1	2022-07-25 09:34:00	0	0	0	0	0
10002	1	2016-02-21 05:58:59	-63.937	-57.68	227.96	-0.74	34.13
10001	1	2016-02-21 05:56:15	-63.936	-57.682	227.48	-0.75	34.13
10000	1	2016-02-21 05:53:30	-63.933	-57.685	226.99	-0.74	34.13
9999	1	2016-02-21 05:50:46	-63.931	-57.687	226.21	-0.72	34.13
9998	1	2016-02-21 05:48:01	-63.93	-57.689	225.13	-0.74	34.13
9997	1	2016-02-21 05:45:16	-63.929	-57.691	223.92	-0.73	34.13
9996	1	2016-02-21 05:42:32	-63.931	-57.693	223.57	-0.72	34.14
9995	1	2016-02-21 05:39:47	-63.933	-57.694	224.87	-0.72	34.14



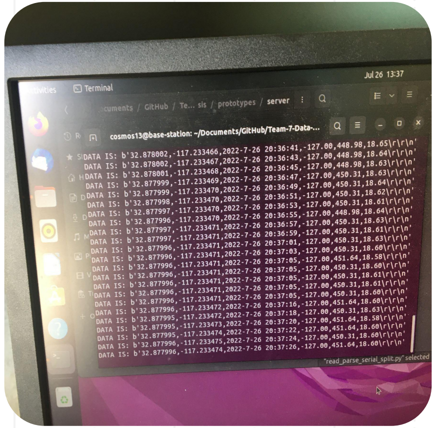

Timestamp
Timestamp
Timestamp
Long Beach
Mission Viejo
Temecula
Palm Springs

Buoy-Server LoRa Pipeline In Action (July 25-29)

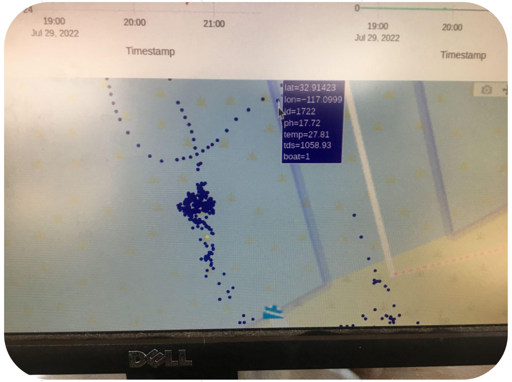
LoRa bad output because of a miswiring in the sending buoy (July 25)



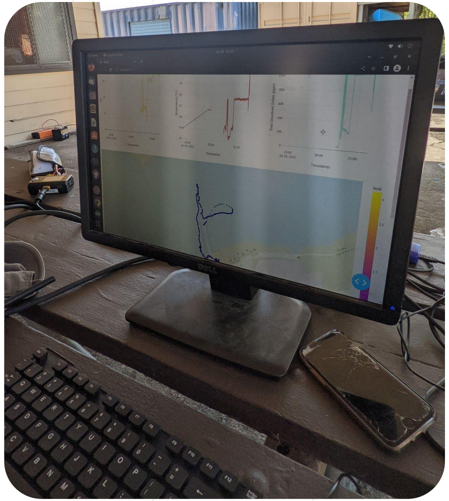
Radio communication read from LoRa serial into command line (July 26)



Visualizing the movement of our buoy+boat on Lake Miramar on server via GPS on Friday test day (July 29)



Basestation with live dashboard (July 29)



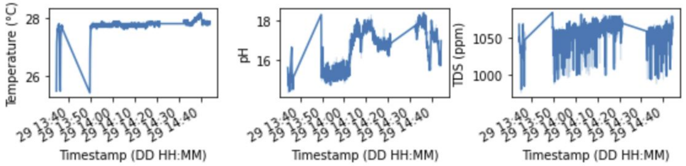
Findings & Conclusions

3

Data Analysis

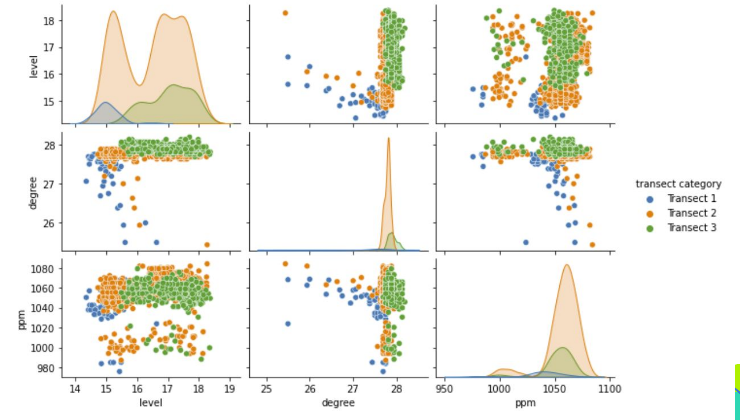
Exploratory Data Analysis

- Line Plot of Measurements vs Timestamp



- Pair Plot of Measurements

Dip in level vs level
Ppm vs ppm seems like it could be normal



- Path of cleaned data

Path of the cleaned data

Linear Regression

Dep. Variable:	level	R-squared:	0.038
Model:	OLS	Adj. R-squared:	0.037
Method:	Least Squares	F-statistic:	47.79
Date:	Wed, 03 Aug 2022	Prob (F-statistic):	7.61e-12
Time:	21:00:49	Log-Likelihood:	-1726.5
No. Observations:	1224	AIC:	3457.
Df Residuals:	1222	BIC:	3467.
Df Model:	1		
Covariance Type:	nonrobust		

OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.7465	1.773	-2.678	0.008	-8.224	-1.269
ppm	0.0116	0.002	6.913	0.000	0.008	0.015

Omnibus:	581.153	Durbin-Watson:	0.196
Prob(Omnibus):	0.000	Jarque-Bera (JB):	67.878
Skew:	-0.102	Prob(JB):	1.82e-15
Kurtosis:	1.865	Cond. No.	6.59e+04

- Correlation Plot of Measurements

Highest correlation is 0.2

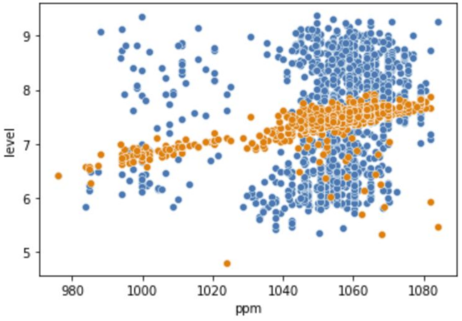
	level	degree	ppm
level	1.00	0.20	0.19
degree	0.20	1.00	-0.03
ppm	0.19	-0.03	1.00

Data Analysis

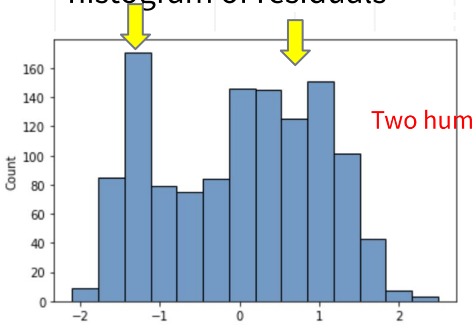
Both the Q-Q Plot and the normality test suggest that the residuals are not pulled from a normal distribution, and the normality assumption may not hold.

Evaluating Goodness of Fit -- Checking Model Assumptions

- level ~ degree + ppm



- histogram of residuals

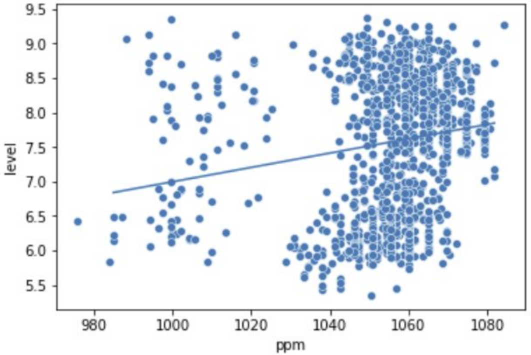


Two humps

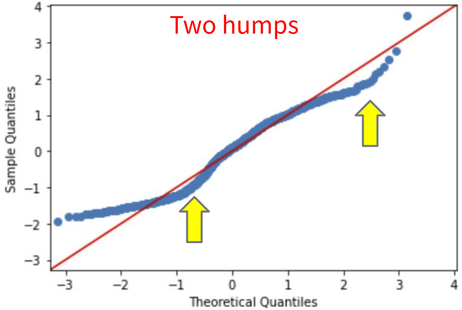
- normality test

```
Z-score statistic 581.1532781197341
p-value 6.370434963499156e-127
```

- predictions

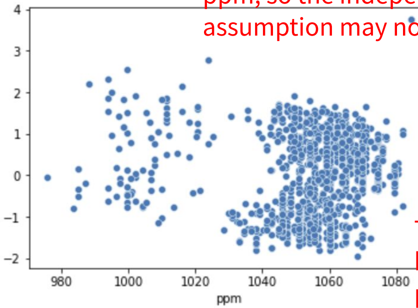


- qq plot



- level ~ ppm

There may be some clustering around 1000 ppm and 1060 ppm, so the independence assumption may not hold.



The constant variance assumption appears to hold because there appears to be the same variance at low and high ppm values.

The Mean Zero Assumption appears to hold because there are an equal amount of data points above and below the line at similar distances.

Data Analysis

Cleaned Data

- linear regression

Dep. Variable:	level	R-squared:	0.155
Model:	OLS	Adj. R-squared:	0.153
Method:	Least Squares	F-statistic:	101.2
Date:	Wed, 03 Aug 2022	Prob (F-statistic):	4.36e-41
Time:	11:47:14	Log-Likelihood:	-1481.3
No. Observations:	1110	AIC:	2969.
Df Residuals:	1107	BIC:	2984.
Df Model:	2		
Covariance Type:	nonrobust		

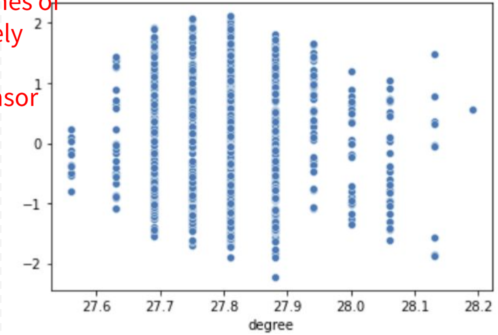
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-100.0278	8.941	-11.187	0.000	-117.572	-82.484
ppm	0.0322	0.003	11.324	0.000	0.027	0.038
degree	2.6433	0.301	8.774	0.000	2.052	3.234

Omnibus:	165.959	Durbin-Watson:	0.286
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.558
Skew:	0.068	Prob(JB):	1.56e-09
Kurtosis:	2.073	Cond. No.	3.43e+05

The R-squared value is now 0.155, which is higher than before.

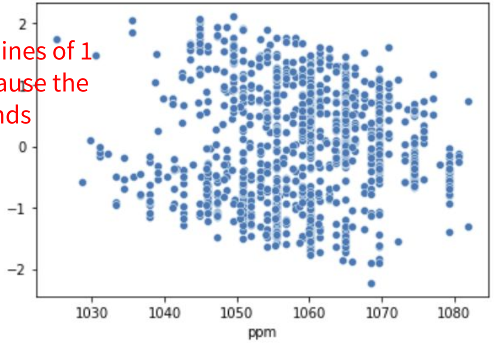
- degree residuals

Group around lines of 0.05 degrees likely because the temperature sensor rounds

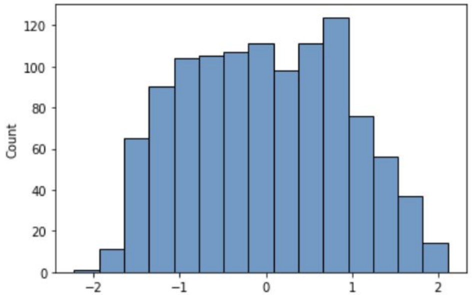


- ppm residuals

Group around lines of 1 ppm likely because the tds sensor rounds



- distribution of residuals



The two humps are no longer visible, but the distribution still does not look normal. It is too flat, but better fitted than before.

	level	degree	ppm
level	1.00	0.24	0.31
degree	0.24	1.00	-0.01
ppm	0.31	-0.01	1.00

Future Opportunities

4

Possible Improvements

Dashboard

- Make completely live (only updated when reloaded)
- Make the dashboard accessible

Analysis

- Find a better way to adjust the pH
- Prevent the rounding of sensor values
- Train with multiple groups instead of just one with 80% of the data

Receiver

- Remove database shutdown on bad data
- Continuously write logs, don’t reset them.
- Find LoRa serial port automatically
- Neater cmd line output

Hacking for Oceans: Data Analysis & Visualization

Tanmayi Dasari, Sebastian Hazlett, Claire Lee, Dorothy Zhang, Alex Miller

Purpose

To program the final server communication and explore the collected data in meaningful ways by processing, visualizing & analyzing the data, and by making conclusions & informative models.

Project Goals

The goals of the project are to finish the chain of data communication and visualize/analyze that collected data. To do this, we first receive serial data from LoRa to the server, then create map plots, datatables, and line graphs for data visualization, and then do exploratory data analysis to analyze the data.

Procedures

1. Set up base station by installing Linux on server and 3D printing hardware protective cases
2. Send sensor data from LoRa microcontroller to populate the server SQLite database
3. Data visualization - use Pandas to access the SQL database and create Plotly formatted dataframe. Used Plotly Dash library to create datatable, scatterplot map, and line graphs
4. Data analysis - use exploratory data analysis, K-means clustering, linear regression, ANOVA, goodness of fit to analyze data

Dashboard

Fig 1: Plotly dashboard with datatable, scatterplot map, and line graphs

Acknowledgements

Professor Jack Silberman
J. Michael Tritchler
Ivan Ferrier
Dallas Rodriguez
Melody Gill
Devanshi Jain

Source Code

Abstract

Oceans cover 70% of the surface of the Earth. It is important to monitor and understand these large bodies. Cluster 13 is designing and testing an autonomous vessel to remotely track and record water quality. Our group was responsible for capturing and storing pH, temperature, total dissolved solids, and location in real time into an onshore database. Then we programmed a web-based dashboard that included a map plot, line graphs, and a datatable. The dashboard updated as new data arrived. We used exploratory data analysis, and linear regression to visualize our results such that we could find patterns and derive conclusions about the water in Lake Miramar. We found that none of the measurements had an extremely high correlation with each other. When developed a model to predict the pH based on temperature and total dissolved solids. We evaluated the model for Goodness of Fit. Some of the assumptions may not hold. Consequently, we cleaned the data for the to remove anomalous records. The R-squared value of the new model was more than three times higher than before, but the model assumptions did not improve.

Challenges

Before cleaning the data, pH levels ranged from 16 to 18 on a 14 point scale and some values included jumbled letters and symbols for when the arduino was unable to properly relay information. After testing the lake water in the lab, the Data Acquisition team was able to find that the pH should be around 8. Consequently, we subtracted 9 from each pH data point to try and normalize our data around the 8 pH range. We also got rid of any faulty data points and converted the timestamp from a string to datetime and the measurements we took to float64. We also made sure to convert the timezone from the default GMT to PDT.

EDA:

Fig 2: pairplot comparing the three measurements against each other

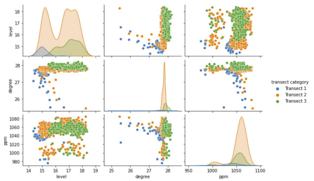


Fig 3: pairplot comparing the three measurements against each other

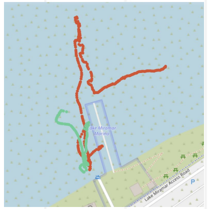


Fig 4: correlation plot has a highest value of 0.20 for level vs degree

	level	degree	ppm
level	1.00	0.20	0.19
degree	0.20	1.00	-0.03
ppm	0.19	-0.03	1.00

Results

We cleaned the data to get rid of faulty data, converted the timestamps from a string to datetime, our measurements to floats, and also converted the timestamps to reflect our PDT time. Using the pH from the Data Acquisition Team, we subtracted 9 from the pH data points to more accurately normalize the pH around the correct pH of 8. We then filtered to only include the datapoint from when the buoy was in the water by only including pH values above 4. We then created and analyzed the data using pairplots, correlation plots, and mapbox. To analyze the goodness of fit, we created a linear regression model, qq plot, and normality test. We found that the Independence Assumption and Normality Assumption didn't appear to hold, but the Constant of Variance assumption and Mean Zero Assumption did. We cleaned the data to ignore the values in this smaller hump to see if that model would be better and ran the test again. The R-squared value went from around 4 to 15 percent, so using the dataset that was cleaned more may be a better model. In the end, however, it seems that none of our models were ideal for predicting the pH based on temperature and total dissolved solids

Fig 5: line of best fit for the pH based on temperature and TDS

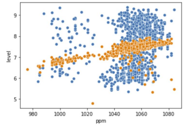


Fig 6: qq plot has two humps

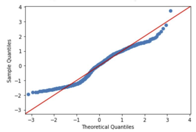


Fig 7: OLS shows R-squared value of 0.038

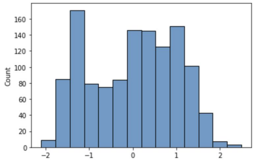
Dep. Variable:	level	R-squared:	0.038
Model:	OLS	Adj. R-squared:	0.037
Method:	Least Squares	F-statistic:	47.79
Date:	Wed, 03 Aug 2022	Prob (F-statistic):	7.61e-12
Time:	21:00:49	Log-Likelihood:	-1726.5
No. Observations:	1224	AIC:	3457.
Df Residuals:	1222	BIC:	3467.
Df Model:	1		
Covariance Type:	nonrobust		

OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.7465	1.773	-2.678	0.008	-8.224	-1.269
ppm	0.0116	0.002	6.913	0.000	0.008	0.015

Omnibus:	581.153	Durbin-Watson:	0.196
Prob(Omnibus):	0.000	Jarque-Bera (JB):	67.878
Skew:	-0.102	Prob(JB):	1.82e-15
Kurtosis:	1.865	Cond. No.	6.59e+04

Fig 8: Histogram of residuals shows two humps. It does not mass the normality assumption.

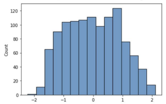


Dep. Variable:	level	R-squared:	0.155
Model:	OLS	Adj. R-squared:	0.153
Method:	Least Squares	F-statistic:	101.2
Date:	Wed, 03 Aug 2022	Prob (F-statistic):	4.36e-41
Time:	11:47:14	Log-Likelihood:	-1481.3
No. Observations:	1110	AIC:	2969.
Df Residuals:	1107	BIC:	2984.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-100.0278	8.941	-11.187	0.000	-117.572	-82.484
ppm	0.0322	0.003	11.324	0.000	0.027	0.038
degree	2.6433	0.301	8.774	0.000	2.052	3.234

Omnibus:	165.959	Durbin-Watson:	0.286
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.558
Skew:	0.068	Prob(JB):	1.56e-09
Kurtosis:	2.073	Cond. No.	3.43e+05

new distribution of residuals



THANK YOU!

Special thanks to Professor Jack Silberman, J. Michael Tritchler, Ivan Ferrier, Dallas Rodriguez, Melody Gill, and Devanshi Jain