**WATER QUALITY ANALYSIS**

**PHASE 4: DEVELOPMENT PART 2**

**Introduction:**

**We can define predictive models as quantitative mathematical projections that use statistical classifiers to determine the probability of a specific water quality event in the future.** **Water quality specialists use models for many purposes: Assessing water quality conditions and causes of impairment. Predicting how surface waters will respond to changes in their watersheds and the environment (e.g., future growth, climate change)**

**Data visualization:**

library(myrwaR) #package by Jeff that loads MyRWA's water data

library(dplyr) # for manipulating datasets

library(lubridate) #for manipulating date/times

library(ggplot2) #powerhouse for data visualization

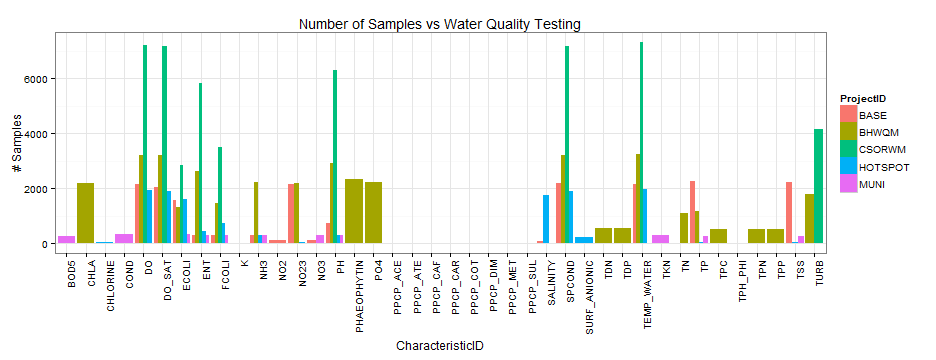
library(ggmap) # for map visualization theme\_set(theme\_bw()) #change the default theme to black/white opts\_chunk$set(tidy = FALSE)

**Number of Samples Across Different Kinds of Testings And Projects**

#wq is the data.frame pulled out from the water quality database

#CharacteristicID and ProjectID are corresponding IDs from the database

ggplot(wq, aes(CharacteristicID, fill=ProjectID)) + geom\_bar(position="dodge") + theme(axis.text.x=element\_text(angle=90, hjust=1)) + labs(title = "Number of Samples vs Water Quality Testing", y='# Samples')



Web-based WQI index can be used to identify pollution sources based on the graphical representation of water quality trend. The system is assessed by evaluating spatial variations water quality of a river in Peninsular Malaysia known as Langat River for 2005, 2010 and 2015.

## Organize Your Data:

## The first step is to organize and analyze the data so you can determine the results of your monitoring effort. We recommend using spreadsheet software such as Google Sheets (free) or Microsoft Excel (low cost, available on many public computers). For standardized data entry format, there are three options: long format, wide format, and tables (ex: by station, year, etc.). We recommend using a long format (see the images below). Simply put, long format often contains values that repeat in the first column, while wide format does not contain values that repeat in the first column. Using the long format example below, each row (horizontal) represents data that was collected at the same time and location. With our wide format example, each column (vertical) represents data that was collected at the same time and location.

## Chart Your Data:

## It is difficult to identify trends by just looking at the raw data, and frankly it can be overwhelming if you’ve collected a lot of information. This is especially important when working with large datasets. And, when working with data, it is of the utmost importance to objectively display the data (check your axis scales!). Utilizing the built in functionality of the selected spreadsheet program enables quick and easy ways to create charts and graphs. By visualizing data, it’s much easier to compare the data to water quality standards (i.e., dissolved oxygen, temperature) as well as identify outliers or transcription errors. And, if your data is set up in long format, programs like Google Sheets will even recommend helpful charts/graphs once you begin the process. We recommend looking at a variety of charts and graphs to really get a good sense of the information and trends.

## Communicate Your Data:

## It’s time to tell the world the results! And, honestly, it’s just perfectly fine if the results are not groundbreaking. Consistency and replication in science are very important. If the result is that the water is squeaky clean, then that is just awesome. If the result is that the water is not safe for swimming, well that is important to report and perhaps drive some action.  Utilizing digital tools such as blogs, website pages, social media platforms, etc are great ways to get your message out there. Just keep in mind that each platform might have a different audience, and therefore a different narrative! Another good way to communicate information is through a press release or discussion with a local reporter. Getting your data published in the media is a quick way to reach large audiences.

**prediction model:**

**Data-Pre processing:** The dataset included some null values. For handling such null values, the median method is used in this analysis. Furthermore, Min–Max scalar is used to scale the data, which makes the computation easier.

**Dimensionality-Reduction:** For dimensionality reduction, Principal Component Analysis (PCA) is applied to the dataset, which extracts the most dominant water quality parameters. PCA is a statistical analysis that reduced a dataset’s dimensionality that is influenced by the multi-correlated variables. Since PCA takes all the inter-correlated variables, it transformed them into a small number of non-correlated variables that described all the variances. The uncorrelated variables obtained from the PCA are known as principal components (PCs).

* **Data-Split:** The collected data are then divided into two sets after conducting PCA on the dataset: training and testing set with a proportion of 80 and 20 percent.

**Water quality parameter:**

**Water quality parameters are essential indicators used to evaluate the suitability and safety of water for various purposes. These parameters include temperature, pH, dissolved oxygen, turbidity, conductivity, and the presence of pollutants. They indicate the physical, chemical, and biological properties of water**

## Conclusion:

This paper demonstrated a method for predicting and classifying the water quality using machine learning algorithms. The water metrics, including PH, DO, SS, EC, Turbidity, Chloride, COD, TDS, and Alkalinity, were used in this study. For [data pre processing](https://www.sciencedirect.com/topics/computer-science/data-preprocessing), the median technique used to handle the null values and min–max scalar to scale the data. For the prediction purpose, we applied the principal component regression (PCR) method. After analyzing the performance of multiple PCR models, PCA with Support Vector Regression seems to be more effective with an accuracy of 95%. However, if the number of components reduced, then PCA with the Multiple Linear Regression model proved to be more effective. For the classification purpose, the Gradient Boosting classifier used to classify the water quality status. Besides, to check the performance of the model, the proposed model is compared with several state-of-art classifiers, including Ada-Boost Classifier, Support Vector Classifier, and Random Forest Classifier. Experimental results showed that the Gradient Boosting Classifier classified water quality status more efficiently. Despite the achievements outlined in this paper, some improvements are still possible, including we can collect more training samples to make the model more stable and more progress is possible on the prediction model. Those issues will be overcome in future research, perhaps by proper tuning of the PCR model and using deep neural network.