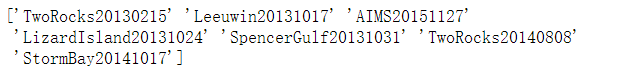
Chapter 4

Exploratory Data Analysis

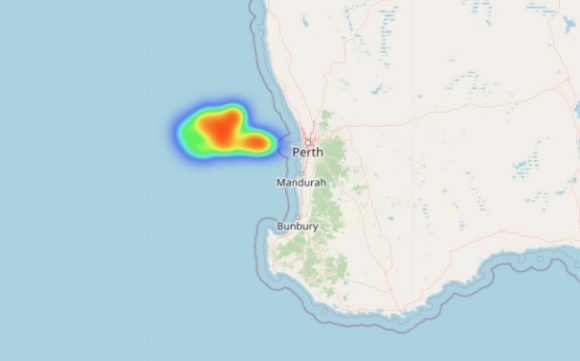
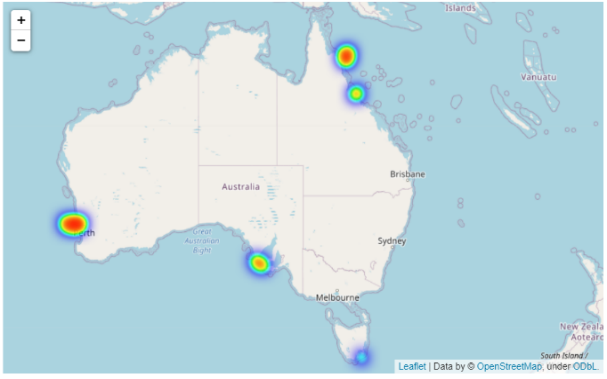
4.1 Using Python

See <https://github.com/RaymonQ/Ocean-Project-6/tree/master/Data_Proprocessing/Python> and corresponding codes in the appendix for more details.

To begin with, we explore the number of unique gliders, which are seven in total, and plot the location maps. The unique gliders named are shown in Result 4.1.1.



Result 4.1.1 The gliders distribution in Australia

We firstly visualized the distribution of gliders, acknowledging that they spread across all of Australia (figure 4.1.2). Further, three of them named TwoRocks20130215, TwoRocks20140808, and Leeuwin20131017 are all located around Perth, from figure 4.1.3 to figure 4.1.5.

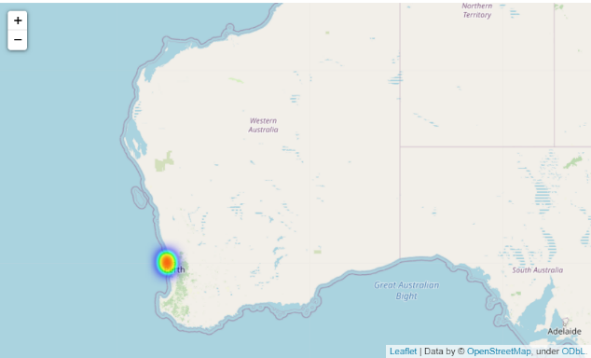
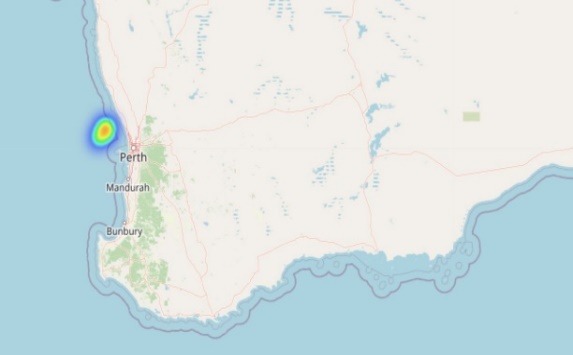
Figure 4.1.2 The gliders distribution in Australia Figure 4.1.3 The gliders distribution of Leeuwin20131017

Figure 4.1.5 The gliders distribution of TwoRocks20130215

Figure 4.1.4 The gliders distribution of TwoRocks20140808

Then we started to explore the raw dataset, plotting the heatmap on the correlation of different numerical features as described in figure 4.1.6 and dendrogram 4.1.7 points out the similarity between every two variables.

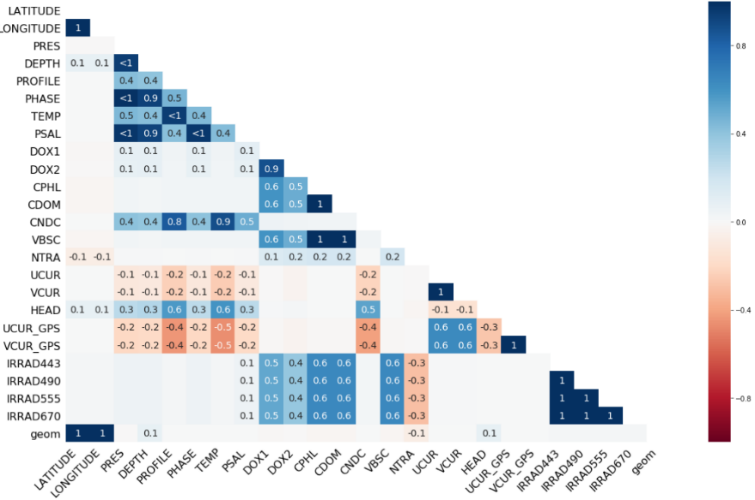


Figure 4.1.6 Correlation between Numerical variables

The above heatmap shows the quantitative correlation between the numerical features. As the coefficient goes higher, block color changes from red to blue. For example, DOX1 has 0.1 relationships with DEPTH, PROFILE, TEMP and VBSC, 0.9 with DOX2, 0.6 with CPHL, CDOM, VBSC and 0.5 with all the IRRAD related variables. It can be concluded that DOX1 has a high correlation with DOX2 and CPHL however, features like DEPTH has little impact on it.

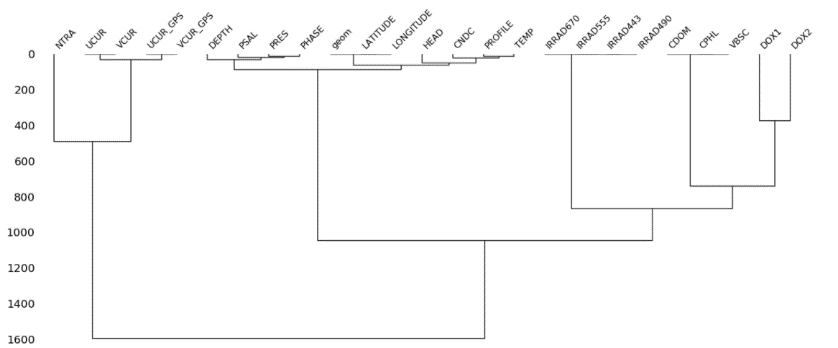
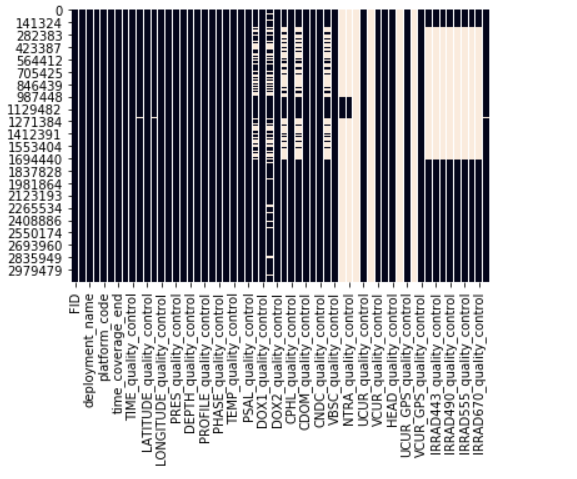


Figure 4.1.7 Dendrogram of Numerical Variables

The dendrogram 4.1.7 diagram indicates the attribute distance between each feature in terms of similarity. The key to interpret is to focus on the height at which any two objects are joined. In the example above, we can see that UCUR and UCUR\_GPS are similar. UCUR records the eastward seawater velocity, and the latter is the value of velocity on the eastward surface.

However, the raw data includes a significant amount of missing value. According to figure 4.1.8, white vertical stripe represents the missing data. In that case, variables include NTRA, UCUR, VEVR and their related quality controls have a significant number of missing value while the missing data on IRRA features,

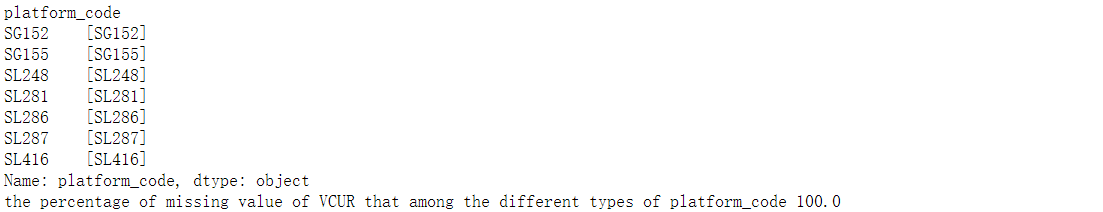
Figure 4.1.8 Concentration of Missing Values in Raw Data

like IRRA443, are concentrated on the half first of raw data.

According to variable instance data form (AODN 2016)【2】, we first filter data with the valid range to generate the legal dataset. After the selection, the total number of rows has decreased from 3,123,117 to 3,101,188.

For variables that contained more than 90% of missing values, including UCUR, VCVR, UCUR\_GPS, VCVR\_GPS, and NTRA, we calculate the percentage of missing values against the type of gliders and time slots. While the result shows that the missing value are distributed equally, that spreads almost every glider and time. It can be concluded that it is a random missing situation for the above features.

As shown in the IMOS project by Australian National Facility for Ocean Gliders (ANFOG 2017)【1】, UCUR and VCUR are the zonal and meridional components of the depth-integrated current velocity. They can only be calculated when the glider is close to the surface, which explains so many gaps. NTRA, after communicating with the supervisor, she agreed that it is not a much meaningful variable. Therefore, we decide to delete the above features and their related quality control.

The below result 4.1.9 is an instance of all glider types (platform\_code), seven in total, and the percentage of missing value in UCUR that distributes in different gliders.

Result 4.1.9 the platform genre and

the missing VCUR given platform distribution percentage

Due to the enormous raw data and the different percentages of missing variables in each glider, it is better to analyze data by gliders located in various areas separately.

Thus, we will explore the data of which is located in Perth named TwoRocks20140808.

Assessing features by features, we acknowledge that PSAL, which represents the seawater salinity, has five quality control types as defined in the raw table, including 0 (No QC performed), 1(Good data), 4 (Bad data), 3(Bad data that are potentially corrected) and 9 (Missing Data).

As there also has independent variable TIME, it is necessary to investigate the relationship between feature value changes with time series. In that case, we plot the diagram of PSAL against time and depth movement.

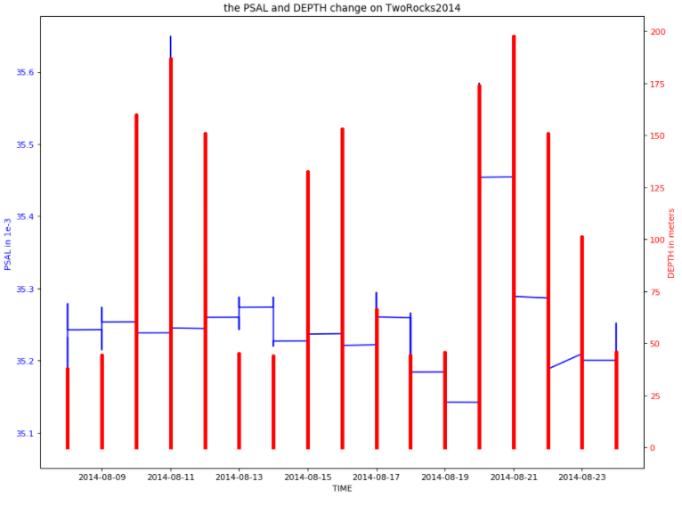
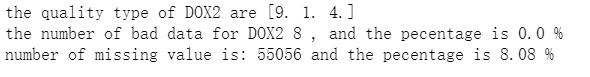


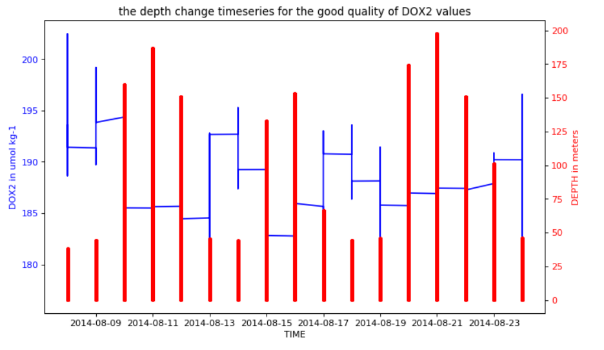
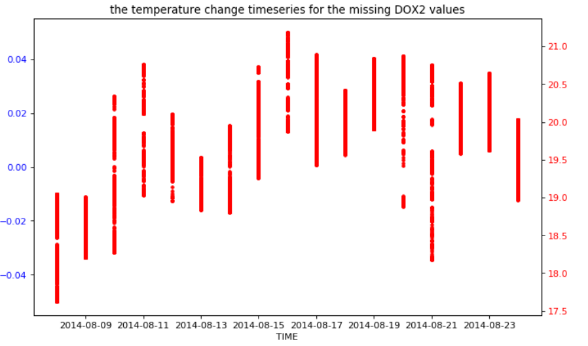
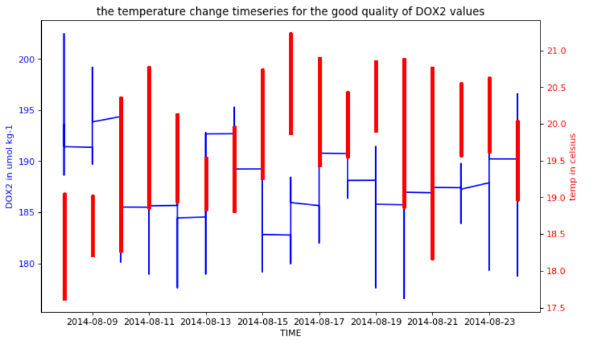
Figure 4.1.10 timeseries on the change of PSAL and DEPTH

It is manifest from the 4.1.10 that the glider TwoRocks2014 went to the deepest under the sea at almost 200 meters on 21st Aug 2014, indicated by the right vertical line in red. Seawater salinity, however, distributed evenly between 35.15 to 35.45 1e-3 in August.

It is also be calculated the number of missing values in PSAL is only two, which is a too tiny amount. Therefore, it is a good idea to ignore those invalid rows. Similarly, it can also delete the bad quality corresponding value since the percentage of bad data given good one at 0.7% is still small.

However, not all the features can be straight deleted on the missing rows. Take the following variable DOX2, which is the value of moles of oxygen per unit mass in seawater for example in result 4.1.11, the number of missing values is 55,056 at a significant percentage of all data set with 8.08%. The moderate missing size cannot be easily ignored; therefore, we would do further imputation methods.

Result 4.1.11 number and percentage on bad and missing data for DOX2

Likewise, we can plot the time-series diagram to describe the variation of DOX2 again the basic properties, including depth of gliders goes and temperature of the seawater. After comparing the figures from 4.1.12 to 4.1.15, it can be realized that the data is missing on the gliders' diving process. The gap is led by the movement as depth and temperature changes instead of specific or accidental reasons.

Result 4.1.15 timeseries on the good type DOX2 values

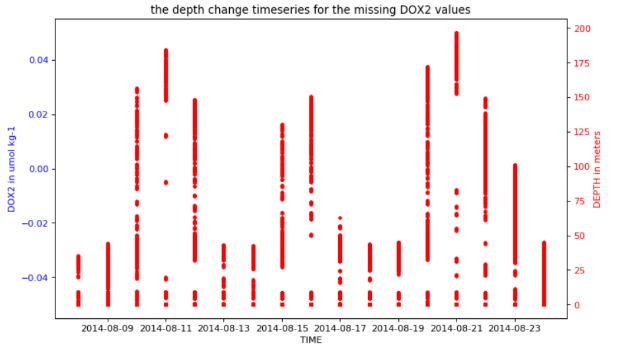
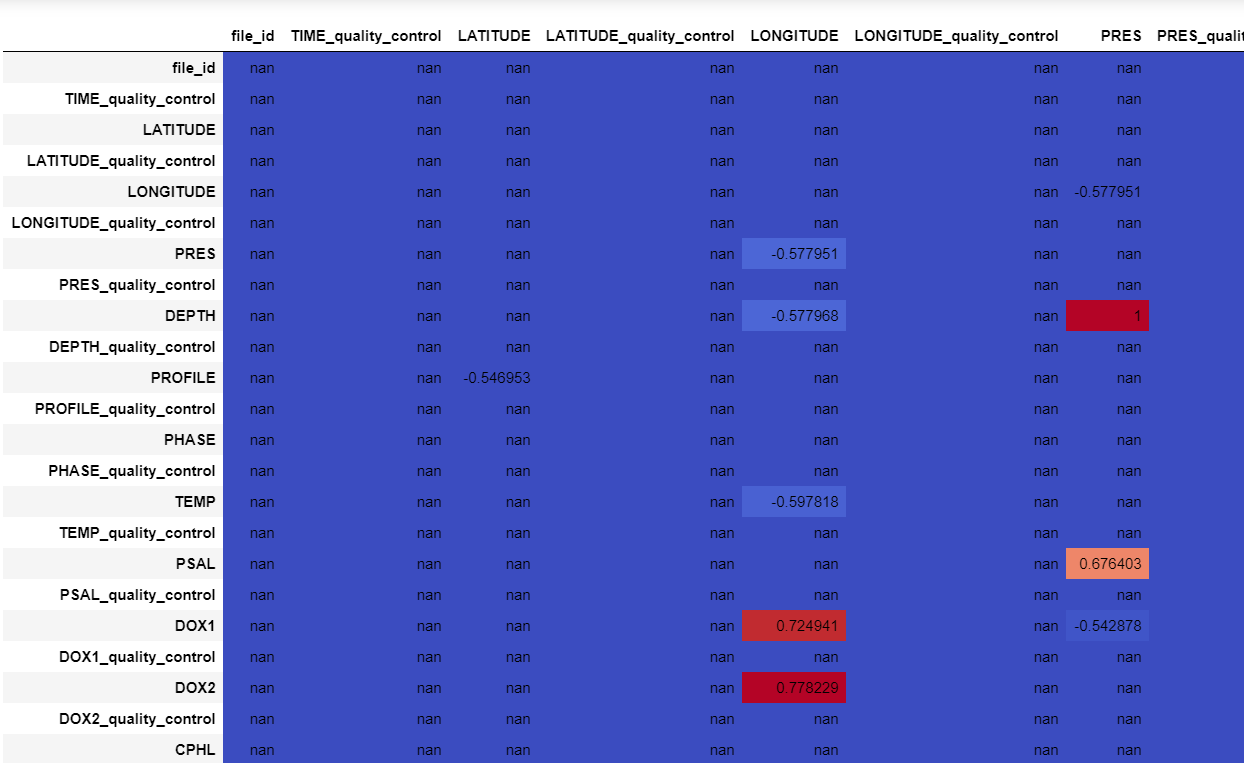
Against temperature

Result 4.1.14 timeseries on the missing DOX2 values

Against temperature

Result 4.1.13 timeseries on the good type DOX2 values against depth

Result 4.1.12 timeseries on the missing DOX2 values against depth

Concerning conference paper about time series imputation methods (Sun, Bin; Ma, Liyao 2017)【3】, k-nearest-neighbors (KNN) imputation which is to match points with its closest k in a multi-dimensional space and replace the non-value obtained from related cases in the whole set of records is an effective and accurate way to dealing with all variety of missing data.

Result 4.1.16 partial correlation plot between each variable

As exhibited in the above figure 4.1.16, it indicates the correlation value between every feature. While the high correlation with a coefficient between 0.5 and 1 will be considered, the dark blue part represents non or less correlated features. As the coefficient goes broader, the color on the plot is in baby blue meaning 0.5 relation variables, orange color for the correlation between 0.6 and 0.7 and dark red for the high-related features with the coefficient 0.7 and above.

Therefore, the coefficient of 0.5+ non-missing features like PRES and DEPTH is defined as the predictable sets can be used to evaluate the DOX2 missing value as the target set. By splitting the non-missing data into train and test group into 80% and 20% firstly, we fit the model with the training set, predicting the value of DOX2 and comparing with their real number, which gives a significant high accurate of fitted model at 89.2%. After acknowledging the good-fit model, it can be fit to predict the missing DOX2 value with the same predictable set, that gives the imputation result.

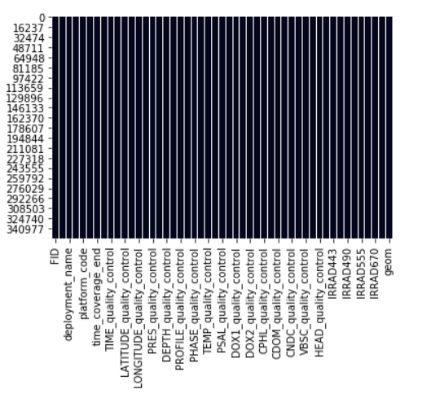
By similar methods, either imputation or deleting, we analyzed all the variables that contained the missing values. Then we re-run the heatmap concentration figure 4.1.17 of missing value, finding there is no white gap (which represents missing) anymore.

Figure 4.1.17 concentration of missing value after data cleaning of TwoRocks2014

## However, not all seven gliders perform precisely the same due to the record difference. Take the glider named Leeuwin20131017 for example, the data related to IRRAD are 100% missing according to the figure 4.1.18.

## 

Figure 4.1.18 concentration of missing value of raw Leeuwin20131017 data

## By further reference, IRRAD is a measurement of the optical sensor of how much light from the sun penetrates in the water. It is reasonable there are no records regarding the fields, so we deleted all four IRRAD variables and their corresponding quality controls which are eight variables in total.

## 

Result 4.1.18 the number of variables comparison after preprocessing of twoRock2014 and Leeuwin

## After processing all the features, it shows in result 4.1.18 that there are eight features difference between the number of twoRocks2014 at 47 and the Leeuwin at 39, which is resulted by the IRRAD correlated variables.

Overall, all the gliders can be classified into the above two categories. One is that after exploring and processing the raw data, the majority of the features are kept like twoRocks2014. And the other is similar to the Leeuwin glider performance without all the essential IRRAD variables in the processed table. In that case, our modelling step will focus on the above two types.

Reference：

1.

ANFOG, Australian National Facility for Ocean Glider. 2017. *Integrated Marine Observing System.* http://imos.org.au/emii.html; http://imos.org.au/anfog.html.

2. AODN, Australian Ocean Data Network. 2016. *OPeNDAP Dataset Access Form.* November. <http://thredds.aodn.org.au/thredds/dodsC/IMOS/ANFOG/slocum_glider/Forster20170911/IMOS_ANFOG_BCEOPSTUV_20170911T071056Z_SL287_FV01_timeseries_END-20171002T010328Z.nc.html>

3. <https://www.researchgate.net/publication/320087317_An_Improved_k-Nearest_Neighbours_Method_for_Traffic_Time_Series_Imputation>

Sun, Bin; Ma, Liyao. 2017. "An Improved k-Nearest Neighbours Method for Traffic Time Series Imputation." *2017 Chinese Automation Congress (CAC).*