To begin with, we explore the number of unique gliders, totaling seven, and plot  
the location maps. The unique gliders named are shown in appendix such as TwoRocks20130215, Leeuwin20131017, AIMS20151127 and TwoRocks20140808.（第一段的文字需要修改的部分）

Result 4.1.1 （p7）: 放appendix  
  
We firstly visualized the distribution of gliders, acknowledging that they spread  
across all of Australia (figure 4.1.1). Further, three of them named TwoRocks20130215,  
TwoRocks20140808, and Leeuwin20131017 are all located around Perth, from figure 4.1.2 to figure 4.1.4.

Left:Figure 4.1.1 The distribution of gliders in Australia  
Right:Figure 4.1.2 The distribution of gliders of Leeuwin20131017

Left:Figure 4.1.3 The distribution of gliders of TwoRocks20140808  
Right:Figure 4.1.4 The distribution of gliders of TwoRocks20130215

Then we started to explore the raw dataset, plotting a heatmap of the correlation  
of different numerical features as described in figure 4.1.5 and dendrogram 4.1.6  
points out the similarity between each variable.

Figure 4.1.5 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 4.1.6 Dendrogram of Numerical Variables

The dendrogram 4.1.6 diagram indicates the attribute distance between each feature in terms of similarity. The key to interpret this 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 velocity on the eastward sea surface. However, the raw data includes  
a significant amount of missing values. According to figure 4.1.7, white vertical  
stripes represent the missing data. In that case, variables including NTRA, UCUR,  
VCUR and their related quality controls have a significant number of missing values while the missing data on IRRAD features, like IRRAD443, are concentrated on the first half of raw data. According to variable instance data from AODN[11], we first filter data with the valid range to generate the legal dataset. After selection, the total number of rows has decreased from 3,123,117 to 3,101,188.

Figure 4.1.7 Concentration of Missing Values in Raw Data

For variables consisting of more than 90% missing values, including UCUR, VCVR, UCUR GPS, VCVR GPS, and NTRA, we calculate the percentage of missing values against the type of glider and time slot. The results show that the missing  
values are distributed equally, spread over almost every glider and time. It can be concluded that it is missing at random for the above features.  
As shown in the IMOS project by the Australian National Facility for Ocean Gliders [15], 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 the frequency of missing data. NTRA, after communicating with the project supervisor, was determined to be a less relevant variable; therefore, we decided to delete the above features and their quality control counterparts.

The below result 4.1.9 is an instance of all glider types (platform code), seven in  
total, and the percentage of missing values in UCUR distributed over different gliders. （p9 这段删掉）  
Result 4.1.9 the platform genre and the missing VCUR given platform distribution percentage （这个图直接删掉）  
  
Due to the enormous size of the raw data and the different percentages of missing  
variables from each glider, it is better to analyze data from gliders located in different areas separately.  
  
Thus, we will explore the data sourced from Perth named TwoRocks20140808.  
Assessing each feature, we acknowledge that PSAL, which represents the seawater salinity, has five quality control types as defined in the raw dataset, 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 exists independent variable TIME, it is necessary to investigate the  
relationship between the variables and time. In such a case, we plot the diagram of PSAL against time and depth movement.  
Figure 4.1.8 timeseries on the change of PSAL and DEPTH  
It is clear from figure 4.1.8 that the glider TwoRocks2014 dove to it’s deepest  
point 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.  
We can observe that the PSAL variables is only missing two data points, a relatively minuscule amount. Therefore, it is a good idea to ignore these invalid rows.  
Similarly, we can also delete the bad quality corresponding values since the relative percentage of bad data is 0.7%.  
However, not all features can be straightforwardly 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 states in appendix（加这一句 P10原4.1.11 上面那一段）, the number of missing values is 55,056, a significant percentage of the data set with 8.08%. This moderate amount of missing data cannot easily be ignored; therefore, we continue to further imputation methods.

Result 4.1.11 number and percentage of bad and missing data for DOX2（放appendix）

Likewise, we can plot the time-series diagram to describe the variation of DOX2  
against the basic properties, including depth of gliders and seawater temperature.  
After comparing the figures from 4.1.9 to 4.1.10 and time-series of DOX2 against temperature in Appendix(p10最后一段), it can be seen that the data  
is missing during the gliders’ diving process. The gap in data is caused by the  
gliders’ movement as depth and temperature changes instead of specific or accidental reasons.  
Left: Figures 4.1.9 (把result 4.1.12 改成figure 4.1.9) timeseries on the missing DOX2 values against depth  
Right: Figures 4.1.10 timeseries on the good type DOX2 values against depth  
Left: timeseries on the missing DOX2 values against temperature  
Right: timeseries on the good type DOX2 values against temperature(放appendix,原4.1.14和4.1.15)

Concerning a conference paper about time series imputation methods [16], k-nearestneighbors (KNN) imputation, which is to match data points with its closest k neighbours 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 of dealing with all variety of missing data.  
Figure 4.1.11 partial correlation plot between each variable

The above figure 4.1.11(P12第一句), indicates the correlation values between every feature. High correlation with a coefficient between 0.5 and 1 can be considered further, the dark blue entries represents non or less correlated features. As the coefficient goes higher, the color of the entry turn light blue reflecting 0.5 correlated variables, orange correlation entries reflect between 0.6 and 0.7 coefficients and dark red entries reflect coefficients of 0.7 and above. Variables with coefficients of 0.5+ in addition to non-missing entries like PRES and DEPTH can be defined as predictable sets and can be used to evaluate the DOX2 missing values as a target set. First splitting the non-missing data into training and testing groups of 80% and 20%, we fit the model with the training set, predict the value of DOX2 and compare with their real values, which gives a significantly high accuracy fitted model at 89.2%. After acknowledging the good-fit model, it can be applied to the set of missing DOX2 values using the same predictable set, giving us the imputation result.  
By similar methods, either imputation or deletion, we analyzed all the variables  
that contained the missing values. We then re-run the heatmap concentration of  
missing value, figure 4.1.12, finding are no white gaps (which represents missing)  
anymore.  
Left: Figure 4.1.12 concentration of missing value after data cleaning of  
TwoRocks2014  
Right: Figure 4.1.13 concentration of missing value of raw Leeuwin20131017 data

However, not all seven gliders perform precisely the same due to recording differences. Take the glider named Leeuwin20131017 for example, the data related to IRRAD are 100% missing according to the figure 4.1.13.  
For reference, IRRAD is a measurement from the optical sensor of how much light from the sun penetrates the water. It is reasonable there are no records regarding the fields, so we deleted all four IRRAD variables and their corresponding quality controls totaling eight variables.  
After processing all the features, there are eight features different between the number in TwoRocks2014 at 47 and in Leeuwin at 39, as a result of the IRRAD variables.  
Overall, all the glider missions can be classified into the above two categories. One, after exploring and processing the raw data, the majority of the features are kept like twoRocks2014. And two, similar to the Leeuwin glider mission without all the essential IRRAD variables in the processed table. In that case, our modelling process will focus on the above two types.