I am Clare and let’s move into the explore data analysis. In this part, I will talk about how we visualize the raw data including the variety of plots such as …, and then take one of the glider name TwoRocks2014 as an instance, presenting how we perform the pre-processing by either deleting the invalid the data or the imputation method.

At the initial, we mapped the raw data and realize that the there are seven distinct gliders that spread across all around Australia with some of them located similarly, such as (RHS) TwoRocks2014 and 2013 that are both in Perth. Further, the diagrams such as heatmap also help to visualize the correlation between different features. As you can see from the left, the number in each grid represents the correlation coefficient and for PSAL here, DEPTH and TEMP are top two influential variables.

However, the raw data includes a significant amount of missing value. As the interpretation, white stripe represents the missing amount. For example, almost 100% of NTRA is missing while that amount on IRRA features, like IRRA443, are concentrated on the first 50% of raw data. Therefore, we decide to drop certain high missing percentage but less meaning variables like NTRA according to the dataset access from created by AODN 2016. Besides, due to the large amount of raw data and different missing distribution, it is better to split data by glider and analyze it one by one.

In that case, I will take TwoRocks2014 glider as an instance to assess feature one by one in different ways, including deleting and imputation. As for the dropping part, I will take the PSAL, which represents the seawater salinity, for example. It has five quality control types, and their corresponding interpretation are as shown below. I also plotted the PSAL time-series to investigate the variation of features. It shows that with time changes, the glider reached to the deepest at almost 200 meters on 21st Aug while the variation of PSAL is displayed in blue. In the following, I calculate the percentage on missing and bad data, showing that both amounts are very tiny when there is only 2 missing and the bad amount is only 0.7%. Therefore, simply deleting the value works perfectly in this situation.

However, not all bad or missing data can be straight deleted. For variable DOX2, after the similar analysis, a significant missing value at 8% that can’t be ignored. In that case, we will perform the KNN imputation that suggested by Sun in 2017 Chinese Tech Conference. This method gives a high accuracy at 89% that has been introduced by Raymon previously.

Thus, by similar methods, we analyzed all the 38 variables and then re-run the heatmap concentration, finding there is no white gap anymore.

In conclusion, although not all the seven gliders perform the same, we can visualize the value to diagnosis the basic features. Then to take either deleting or imputation as the filled methods.