Data Cleansing

Conrad Thiele

2019-04-21

Data preparation plays pivotal role and in my opinion is inter-twined with data understaning as a foundation to ensuring that, any statistical analysis or algorithm, wrappers, and other assorted machine learning output accurately represent the derived sample. The output that forms the functional graph is made of up of careful tranforming, cleaning and examination of data.

In table 1.1 are two samples datasets downloaded from Queensland Government open access dataset and is comprised of wave sensory information from bouys off the Queensland coast. [1][2]. If you would like to download the data used in this tut-tworiel than you can do so via:

* [Wave Data mooloolaba](./data/mooloolaba_2018-01-01t00_00-2018-10-31t23_30.csv)
* [Wave Data Caloundra](./data/caloundra-pob_2018-01-01t00_00-2018-10-31t23_30.csv)

First things first, import the dataset. In this case, I will be using RStudio, however there are many software packages for performing all the task showning in the tworiel.[RStudio-Vers]

mooloolaba.waves <- read.csv(file = "./data/mooloolaba\_2018-01-01t00\_00-2018-10-31t23\_30.csv", header = T)  
caloundra.waves <- read.csv(file = "./data/caloundra-pob\_2018-01-01t00\_00-2018-10-31t23\_30.csv", header = T)

With the data loaded, it is often good to take a quick look for missing values (NA), not a number (NaN), incorrect column headings, and dates/time formats are imported correctly. This can be down with the function **head()** and **tail()** which, show a subset portion of the first and last rows.

The field names and descriptors are as follows:

* **Hs** - Significant wave height, an average of the highest third of the waves in a record (26.6 minute recording period).
* **Hmax** - The maximum wave height in the record.
* **Tz** - The zero upcrossing wave period.
* **Tp** - The peak energy wave period.
* **Dir\_Tp TRUE** - Direction (related to true north) from which the peak period. Field name Peak.Direction in CSV file.
* **SST** - Approximation of sea surface temperature

head(mooloolaba.waves)

## Date.Time Hs Hmax Tp Tz Peak.Direction SST  
## 1 1/01/2018 0:00 -99.900 -99.90 -99.900 -99.900 -99.9 -99.9  
## 2 1/01/2018 0:30 0.513 0.81 10.315 4.748 -99.9 -99.9  
## 3 1/01/2018 1:00 0.566 0.93 10.778 5.003 92.0 26.4  
## 4 1/01/2018 1:30 0.557 0.85 9.984 4.990 91.0 26.4  
## 5 1/01/2018 2:00 0.569 0.96 9.277 5.214 81.0 26.4  
## 6 1/01/2018 2:30 0.571 0.88 9.901 5.121 80.0 26.4

tail(mooloolaba.waves)

## Date.Time Hs Hmax Tp Tz Peak.Direction SST  
## 14587 31/10/2018 21:00 0.802 1.24 12.155 5.121 123 23.60  
## 14588 31/10/2018 21:30 0.771 1.34 12.111 4.856 122 23.60  
## 14589 31/10/2018 22:00 0.787 1.25 12.130 5.160 129 23.55  
## 14590 31/10/2018 22:30 0.784 1.32 11.817 5.214 120 23.55  
## 14591 31/10/2018 23:00 0.777 1.20 12.269 5.382 129 23.55  
## 14592 31/10/2018 23:30 0.816 1.18 11.968 5.261 125 23.55

*Table 1.1 Head and tail of dataset. Dir\_Tp TRUE = Peak.Direction*

As you can see there are rows in *Table 1.1* with the value -99 that does not appear to match up with the field and descriptions. There are two good reason to remove these values, firstly, they are some kind of sensory data issue which, a form of sampling error and secondly, this will skew the mean and median values for each column as seen in *Table 1.2* which, is close to five number Summary with mean values.

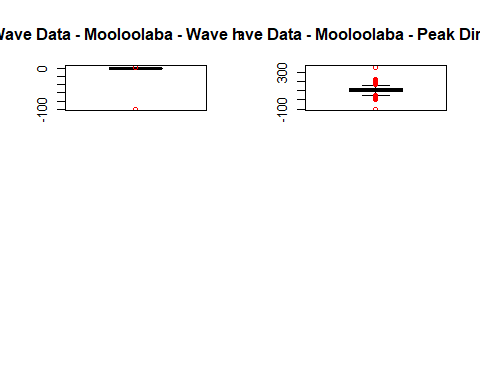
summary(mooloolaba.waves)

## Date.Time Hs Hmax   
## 1/01/2018 0:00 : 1 Min. :-99.9000 Min. :-99.900   
## 1/01/2018 0:30 : 1 1st Qu.: 0.8037 1st Qu.: 1.350   
## 1/01/2018 1:00 : 1 Median : 1.1120 Median : 1.870   
## 1/01/2018 1:30 : 1 Mean : 1.0747 Mean : 1.929   
## 1/01/2018 10:00: 1 3rd Qu.: 1.6080 3rd Qu.: 2.700   
## 1/01/2018 10:30: 1 Max. : 4.2570 Max. : 7.262   
## (Other) :14586   
## Tp Tz Peak.Direction SST   
## Min. :-99.900 Min. :-99.900 Min. :-99.9 Min. :-99.90   
## 1st Qu.: 7.526 1st Qu.: 5.033 1st Qu.: 91.0 1st Qu.: 21.00   
## Median : 9.132 Median : 5.567 Median :105.0 Median : 22.95   
## Mean : 9.072 Mean : 5.513 Mean :102.5 Mean : 23.09   
## 3rd Qu.: 10.901 3rd Qu.: 6.256 3rd Qu.:119.0 3rd Qu.: 25.95   
## Max. : 21.121 Max. : 10.146 Max. :358.0 Max. : 28.65   
##

*Table 1.2 Adapted Five number summary with mean values*

Another way to determine outliers is via boxplot as it does a true five number summary. However as each columns range of values (min - max) are completely different then, it is best to do a boxplot for each field except the data.time column. Figure 1.1 from left to right shows a boxplot of the columns selected; wave height (Hs), and Peak.Direction.

par(mfrow=c(2,2))  
boxplot(mooloolaba.waves$Hs,  
 data=mooloolaba.waves,  
 outcol="red",  
 main="Wave Data - Mooloolaba - Wave height")   
  
boxplot(mooloolaba.waves$Peak.Direction,  
 data=mooloolaba.waves,  
 outcol="red",  
 main="Wave Data - Mooloolaba - Peak Direction")

 *Figure 1.1 Wave height and Peak Direction Boxplot* As you can see, the box portion of the plot which represents the range between the first and third quartile range is hard to interpret. This often happens when the distance between outliers is extreme compared to the distance between quartile ranges.

As such, it is extremely difficult to determine anything of real value from this plot other than, it definitely has outliers and they possibly need to be removed. However, you should always ask yourself these questions before removing outliers:

* What is abnormal and therefore, what does your team/indivdual consider to be normal?
* What should we do with the outliers?
* Is this in fact a new opportunity.
* A beginning of a new trend.
* Is this something this person knows and is it possible other will learn from this outlier behaviour and start a new trend.
* Or is this a sign of a massive changes in circumstance.[3]

An outlier(s) are “data point(s) that are an abnormal distance from other values in a dataset.”, and as such, abnormal will need to be defined. [3] Outliers are often found using a five number summary that can be then shown as a boxplot.

Now that the outlier -99.9 has been removed the boxplot, seen in **Figure 1.2**, appears more uniformed even though there are now more outliers. Yet the outlier that now exist are clearly normal behaviour which can be explained.

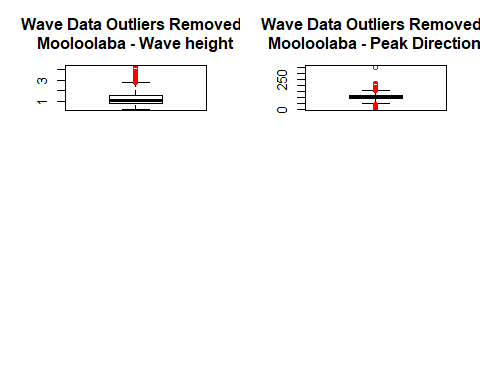
Examining the first plot on the left, Wave height, the average wave height for Mooloolaba is 1 metre, still during storms/cyclones, one would expect the wave height could reach between the ranges of 3-4 metres.

In turn, looking at the Peak Direction plot, second graph on the right, the average direction is ~100 degree, yet, as there are 360 degrees in a circle, then the outliers ~350-0 degrees are completely feasible and should be kept.

In both instances, the abnormal values are crucial to explaining the whole picture and as such, it would be detrimental for the analyst to exclude them and would produce what is know as an **error due to bias**.

mooloo.RM.outlier <- mooloolaba.waves[  
 !(apply(mooloolaba.waves[,2:6], 1, function(y) any(abs(y + 99.9) < 1e-9))),]  
  
# any(abs(y + 99.9) < 1e-9) )  
# dplyr::near(y, -99.9))

par(mfrow=c(2,2))  
boxplot(mooloo.RM.outlier$Hs,  
 data=mooloo.RM.outlier,  
 outcol="red",  
 main="Wave Data Outliers Removed:\nMooloolaba - Wave height")   
  
boxplot(mooloo.RM.outlier$Peak.Direction,  
 data=mooloo.RM.outlier,  
 outcol="red",  
 main="Wave Data Outliers Removed:\nMooloolaba - Peak Direction")

 *Figure 1.2 Wave height and Peak direction for Mooloolaba outliers removed*

To sum up, missing values are not always labled as Not Availabe (NA) as is the case with the sensor data reading -99.9. Removing incorrect values, such as these, can make a dramatic improvement in understanding the basics of the data at hand and hopefully now the reader will have a better understand of the need to examine outlier value and question whether or not to remove them.

### References

[1] “Coastal Data System - Waves (Mooloolaba) - Datasets | Data | Queensland Government.” [Online]. Available: <https://data.qld.gov.au/dataset/coastal-data-system-waves-mooloolaba>. [Accessed: 12-Dec-2018]

[2] “Coastal Data System - Waves (Caloundra) - Datasets | Data | Queensland Government.” [Online]. Available: <https://data.qld.gov.au/dataset/coastal-data-system-waves-caloundra>. [Accessed: 12-Dec-2018]

[3] E. Davila, “Outliers.” [Online]. Available: <https://www.lynda.com/Business-Skills-tutorials/Outliers/427473/531429-4.html>. [Accessed: 29-Apr-2019]