From 2016/11/11 13:54:56 to 2016/11/26 22:49:23, there are 4752 rows, meaning 4752 tonnage data.

Sort tonnage data. Range is from 1 to 206.

I suggest we should delete the extreme ones, since they seem to be outliers. Statistics on Tonnage data.

|  |
| --- |
| **Statistics** |
| Tonnage |
| N | Valid | 4751 |
| Missing | 0 |
| Mean | 19.0422 |  |
| Median | 18.9500 |  |
| Mode | 18.90 |  |
| Std. Deviation | 4.02115 |  |
| Minimum | 1.00 |  |
| Maximum | 206.00 |  |
| Percentiles | 20 | 17.6000 |
| 25 | 17.8500 |
| 40 | 18.5900 |
| 50 | 18.9500 |
| 60 | 19.3500 |
| 75 | 20.0500 |
| 80 | 20.3000 |

99.7% data should fall in the following range.

(Mean – 3\*std, mean + 3\*std) = (6.9788, 31.1057)

Thus, data outside of this range are deleted.

Below 6.9788, there are 2 tonnage data.

Above 31.1057, there are 11 tonnage data.

As a result, 4737 tonnage data are left.

100 worst cases, 100 middle cases, 100 best cases are selected.

(1-100), (2321-2420), (4640-4739)

I merged detailed machine info for all 300 time-period into one file. Since for different time-period, there are different rows for each period, I sampled 100 rows for each time-period.

Thus, the data is 300\*100 rows.

However, there are some errors in the data. For example, the start time and end time are the same. After deleting these error data, I got 201 good time-period.

These 201 rows can be categorized into three groups according to tonnage: low, middle, and high. Like this

I suggest that, you may look into the three groups of data. And try to find some difference between them. If we can find some distinctive difference, we can make very good presentation.