Course: DATA423-20S1

Assignment Report

Name: Zhen Huang ID: 74093323

Data Summary

Dataset of assignment 1 was load by readr::read_csv. There are 300 rows and 44 columns. We can see 12 character variables and 4(Price, Speed, Duration, Temp) of them could be ordinal factor variables and the others are unordered factor variables. We can transform ordered factors manually or automatically later in doing visualisation in Shiny. Besides, there are 31 numeric(double) type variables and one "Date" type variable.

Missing Data

The missing values chart showed that mostly all numeric variables (blue) had a similar (and random) distribution of missing values except "Y" has no missing data and "sensor7" has the most missing values (22%). The nominal variables had a similar condition, "ID", "Author" and "Priority" have no missing value. There is one Date type column with no missing value as well.

In all cases, the observations between 27 and 82 were not missing. There are 12 rows has just over 20% missing value. So the dataset is acceptable to do further exploration.

Novelty Variables

Numeric Data

The bar chart showed the ratio of continuity of variables. Columns are ascendingly sorted, filling different color in the column type. There are no weird numeric variables because their novelty of proportion (the number of unique values/number of rows) are close to 1. It is reasonable that the novelty of "sensor7" is just slightly below 0.8 due to its missing value (22%).

Categorical Data

The "ID" column is 100% unique, which suggested that it could be the observation identifier of the dataset. The other categorical data had a reasonable low ratio of continuity, which is normal.

The bar chart of novelty is not good enough to confirm the novelty of categorical data. So we use the mosaic plot to identify whether there is unusually rate categorical data or factors. There is no bright blue (too common) and no bright red (unusually rare) response found after I tried all the combination of categorical data in this dataset.

Data Overview

Boxplot

The box plots showed potential outliers for variables “sensor3”, “sensor4” and “sensor13”, "sensor17", "sensor22", "sensor24" and "sensor27". All other numeric variables showed no outliers at the 1.5 IQR multiplier. These potential outliers were all high outliers. The outliers do not go away even when the IQR multiplier reached 5.0, which suggest that these are unusual data which need to be reported.

Plot of each single numeric ariable

As a complement, the simple data point plot of selected numeric variable shown more detail of the potential outliers. Certain kinds of observations (rows) have unusually high values by sort in original observation order.

Corrgram

The "corrgram" implement a correlation chart. The graph can demonstrate 31 numeric variables as a whole and shown that four groups of variables in correlation method of "Pearson", which within each group the association was high and the out-of-group correlation was close to zero. The sorting way has no substantial impact on the variables in the group, regardless of whether the parameter is "OLO", "PCA", or "TRUE".

When we change to "spearman" correlation method, the outcome shown differently, there are three groups instead of four groups of sensors data, group1 contain "sensor1 - sensor 10 and Y", group2 contain "sensor11 - sensor 20" and group3 contain "sensor21 - sensor30". "Spearman" method can overcome the effect on the correlation coefficient from outliers. So we can report this finding as well and kept this kind of groups in the following steps.

Both methods have shown positive correlation within the numeric column groups, and there is no negative correlation in them.

GGpair show correlation of numeric data in groups

The ggpair plot clearly shown the distribution of each numeric data and its correlation with the other numeric data within a potential group. If we ignored the potential outliers in the plot, all the numeric are nearly normally distributed more or less. There are three groups in "Pearson" which shown positive linear associations. All seven variables with high outliers were put into one group, and we can hardly see the association cause those unusually high values. In contrast, those 7 variables are put into three groups, and we can see they have positive linear associations with the other variable within the group if we ignore the outlier.

GGpair show correlation of ordinal data

There are five columns of categorical data which are ordinal. I transfer the values into the number, so that ggpair can read numeric data and the plot showed that "Price" and "Priority", "Priority" and "Speed" both have a positive relationship with each other. The rest mutual relationship with this set of 5 variables is negative.

Discontinuty

Rising Value Chart

The rising value chart showed seven variables "sensor3", "sensor4" and "sensor13", "sensor17", "sensor22", "sensor24" and "sensor27" had a jump in their rising values. Sorting by original observations order, the values in about the last 50 observations were absent, which also means they are discontinued variables. "sensor7" had a small boost in values in the front part of the observations, due to the 22% missing values.

Multiple lines chart sorted by Date

The multiple lines chart showed the values of variables by sorting in "Date", we can also find similar situation demonstrated in the rising value chart. The outliers are unusually values.

Mixed Data Pattern

Tabplot

The data table, when sorted by the "ID" variable, showed one pattern of outliers. We can see those outlier related to certain kind of "ID" (Initial "D") and "Author" (value is "XX") and a certain period of Date (2006). This pattern suggested a potential problem in this dataset needed to be explained in context. Besides this, when sorted by the original order of dataset, we cannot see the other pattern.

Mixed Data in ggpairs

When categorical data combined with numeric data, we can see some other potential pattern. For example, when we checked data from sensor21 to sensor 30 with "Location", we can see that most of the outliers come from Boiler and Compressor.