Exploratory data analysis of a given dataset

1 About the raw data

The raw data is from a CSV file without any introduction to the background.

2 The number of observations and variable types

This data frame has 300 observations of 44 variables. After we loaded all the data by read.csv() in R at the beginning, we found that there are only two data types: factor and numeric. The Date is a typical Datetime variable rather than a factor, and other variables like Priority, Price, Speed, Duration, and Temp with apparent orders should be ordinal variables. The types of these variables we mentioned need to be programmatically altered.

Additionally, the cardinality of ID is 300, which means the type of this variable should be string, but R treats ID as a factor (nominal variable). As it is not a predictor but only an identity, we just leave it as a factor. Besides, the variable Agreed has only two levels: Yes and No, it would be logical or nominal. We could convert it to a logical variable during the data preprocessing stage if it is necessary.

Table1 data types of variables

Data type		Count	Variable names
nominal	ordinal factors	5	Priority, Price, Speed, Duration, Temp
	not ordinal factors	7	(ID), Author, Location,(Agreed), State, Class, Surface
numeric		31	Y, sensor1~ sensor30
date		1	Date

3 Missing values

When all the variables are selected, we can find out that 3.6% of values are missing in general, and 96.4% of values exist. Firstly, there is no missing value in the first five variables(Y, ID, Author, Date, and Priority). By contrast, sensor7 has the most missing values, with a percentage of 22 %. While, for other variables, the proportions of missing values are not so high. Maybe we need to find out the reasons why sensor7 has more missing values than others.

Generally, missing values are scatted among the observations without obvious patterns. However, some slight patterns exist. For instance, if we focus on Surface and sensor2, we can discover than most of the missing values are concentrated at the top(around 10).

4 Novelties

Mosaic plots

Mosaic plots will identify unusually rare factors. I used a pick list of factor variables to tailor the plot, so there is not too much going on here for us to comprehend. If we pick Priority, Price, Speed, Duration, there is no zone in dark blue or dark red, which illustrates that there is no unusual rareness. We could also choose other factors to explore novelties.

Box plots

Box plots attempt to identify outliers of numeric data. Let select some variables (sesor1 to senser10). Based on a default IQR multiplier value of 1.5, the box plots show potential outliers for sensor3, sensor4, and sensor13. These potential outliers are extremely high, and they could not disappear even the IQR multiplier reaches 4. So, these three variables do not follow a normal distribution. Additionally, if outliers were hidden, we can also find that sensor7 is left-skewed, which means the mode of sensor7 is larger than its mean.

5 Correlation

Correlogram

The Pearson correlation matrix of numeric variables is illustrated in the correlogram (OLO ordered). It shows four blocks in dark blue, which means the numeric variables can be divided into four groups except for Y. Within the groups, correlation is pretty high, and the out-of-group correlation is much lower.

Table2 variable groups

Group	Variables
Group1	Sensor1, 2,5,6,7,8,9,10
Group2	Sensor3,4,13,17,22,24,27
Group3	Sensor11,12,14,15,16,18,19,20
Group4	Sensor21,23,25,26,28,29,30

Pairs plot

We can also use ggpairs to visualize the correlation between each pair of variables. We built four pairs-plots to show the correlation matrix respectively. The correlations in Group4 are all larger than 0.98. Besides, variables in Group1, Group3, and Group4 have linear relationships respectively. But the scatterplots look strange in Group2. In each scatterplot, the points are

clustered into two groups.

Mixed pairs

If we mixed nominal variables and numeric ones, the pairs plot could tell us more. For example, we choose Price, sensor1, sensor2, and sensor5, and we can find the number of observations with the costly price is the smallest. Sensor1, sensor2, and sensor5 are not normally distributed at the 'costly' level. Secondly, the price level can affect the correlation. In this case, the sensor1 and sensor2 have the strongest correlation (0.927) when the level of Price is "Costly", followed by "Cheap" (0.891) and "Extravagant" (0.89).

6 Continuity and homogeneity

Rising value chart

We can select some numeric variables to detect their continuity. For instance, by choosing sensor1, sensor2, and sensor 7, the rising value chart shows that sensor7 has a gap. If data are not scaled, we can find the value of sensor7 jumps from 10 to 20, so it is discontinuous. In the same way, sensor_i $i \in \{3,4,13,17,22,24,27\}$ are not continuous either. The values approximately between 200 and 1200 are absent.

Homogeneity

We select some numeric variables as default variables, and the right part(observation 270 to 300) of the plot looks significantly different to rest part.

7 Time Series

Y is selected to create a time-series object with a weekly frequency as there is one observation at each week. But the line fluctuates, and there is no special pattern here. Maybe the stationarity and seasonality worth our further exploration.