Ass1Data is a dataset containing 300 columns with 44 variables. Having a quick at it, we can know columns 2 to 13 are factors among which only 'Priority', 'Price', 'Speed', 'Duration', and 'Temp' are ordered factors so they were converted into ordinal variables. 'ID' is a unique variable. The values in 'Date' column should be dates, so what l did was to convert this column into date type. The rest of the columns (1 and 15 to 44) are numeric. There isn't any missing data in the first 5 columns.

Categorical data

Under the categorical panel, there is a summary of the categorical variables. Categorical data has 300 observations of 11 variables with different numbers of levels. The numbers of observations of each level are quite different.

Mosaic plot displays the relationship of 'Author', 'Speed' and 'Agreed'. We can see that an 'yes' or 'no' answer does not have much effect on speed. But there are slightly more 'No' s to 'HH''s slow work. 'HH' has the most 'fast' outcomes, while 'KL' has the fewest 'fast' s and the most 'Medium's. 'KG' and 'XX' have almost the same outcomes no matter what 'speed' outcome is. There doesn't appear to be any zones of unusual rareness (red) of when these three variables intersect.

There's a boxplot showing each categorical variable against Y. My assignment shiny app allows me to choose the variable l would like to display. Take 'price' as an example, 'extravagant' is having the highest median while 'costly' has the lowest, which means 'price' may have effect on 'Y' value.

Numeric data

Glimpse shows the number of observations and variables and each observation, while summary, as the name indicates, has information like minimum, maximum, mean, median, quartiles and the count of missing data in each variable.

The box plot under numeric panel shows potential outliers for variables "sensor3", "sensor4", "sensor13", "sensor17', "sensor22", "sensor24" and "sensor27". All other numeric variables showed no outliers at the 1.5 IQR multiplier. These potential outliers are all high outliers. Without showing standardized, the outliers don't go away no matter how you change the IQR multiplier, which suggests that these variables are non-normal in distribution.

Correlogram gives us a basic understanding of the correlation of the data displayed, it's OLO ordered (using absolute of the correlation coefficient) and demonstrates the relationship between each variable. A Red shade means there's a negative relationship between the two variables, while blue means a positive relationship. The darker the color, the stronger the relationship. As can be seen in the plot, 'sensor' 1 and 'sensor 3' or 'sensor4' are negatively correlated, and the relationship is very weak. 'sensor3' and 'sensor4' have the strongest positive relationship. 'sensor1' and 'sensor5's within group absolute correlation are higher than the out-of-group correlation. The remaining variables had a patchwork of correlation but no obvious groupings.

As can be seen in the missingness plot, there's 3.6% of missing data in general. 'ID', 'Author', 'Date' and 'Priority' variables have no missing data, while 'sensor7' has 22% missing and the others have about 3% random distribution of missing values. Without clustering missingness, we can see there are no missing values after the first 290 observations. 'sensor18' and 'sensor20' have similar missing patterns. There are observations that are frequently missing, for example, the ones near 100^{th} are frequently missing as there appears to be a broken horizonal line there.

The ggpairs plot in which one numeric variable in the same data row is matched with another variable's value. Distributions of each variable can be seen in the plot, so are clustering and the correlation coefficients. 'Y' is approximately normal distribution. The correlation coefficients are displayed similarly in plot 6, we can see a strong relationship between 'sensor3' and 'sensor4', which validates what it shows in the correlogram plot.

Time series

The visualization also contains a time series plot of 'Y' in different times, since it's a unique numerical variable. At first glance, although it's quite hard to detect, the plot seems to follow a certain pattern and scatter around Y=22. We might then plot each year as a separate line in the same plot, which allows us to compare the year wise patterns side-by-side. Or we can group the data at seasonal intervals and see how the values are distributed within a given year or month and how it compares over time. If we want to build a time series model to predict Y, maybe differencing is also needed. Overall, this time series plot helps us get in the right direction of analysis.

Rising order chart

Rising order chart allows us to check continuity. As we can see, there are no unexpected discontinuities in variables 'Y', 'sensor1', 'sensor2' and 'sensor5', but there are big gaps (jumps) in the values "sensor3" and "sensor4". We might need to get back to how the data was collected and check if there's a mistake somewhere. In addition to this, when building a model, we need to try to make sure continuous variables are used in order to get a better model quality.

Homogeneity

As we can see in the last plot, which demonstrates Homogeneity of different numeric variables. 'Y', 'sensor1' and 'sensor2' are all scattered around 0, which means these three variables are similar. while 'sensor3' and 'sensor4' have similar distributions.