This project first read the 'X store sales' dataset and gained an initial understanding of the variable types and meanings based on the first ten rows. Next, the ‘describe()’ function was used to automatically calculate the mean and standard deviation, revealing that Sales and Profit exhibit significant variability. Then, box plots and histograms were created for the four numerical features (‘Sales’, ‘Quantity‘, ‘Discount’, and ‘Profit’) to examine data distribution and outliers. However, although ‘Row ID’ and ‘Postal Code’ are numerical variables, they have no significant impact on predictions, so they were converted to character type. ‘Order Date’ and ‘Ship Date’ were converted to date format and split into features such as year, month, day, quarter, and whether it was a weekend for modeling purposes. Subsequently, one-hot encoding was performed to generate a purely numerical training set.

In data visualization, we gradually enlarged the time granularity and found that the target variable showed a significant growth trend in the fourth quarter of each year, with an overall steady upward trend year by year. In addition, there is a weak linear relationship between the target variable and ‘Profit’.

During the data preprocessing stage, we first checked for missing values, then identified and removed outliers using the ‘3×IQR’ method. It is interesting to note that there was no imbalance in the regression task. In the preliminary feature selection, we used Lasso regression to filter out a group of variables that were most predictive of sales.

Furthermore, we plotted a correlation matrix and heat map to identify the correlations between the variables selected by Lasso regression. From the plots, we found that ‘order_quarter’ and ‘ship_month’ were highly correlated. Since the former had a higher weight in the previous Lasso regression, we ultimately retained it as one of the predictive features. There were no significant correlations among the other selected variables, which could also be used as predictive features.