Firstly, I loaded the dataset and made histograms and boxplots for numerical variables like “Sales”, “Profit”, “Quantity” and “Discount” to detect outliers and skewness. Secondly, I changed the original “Order Date” and “Ship Date” columns to datetime format, and used label encoding to transform categorical columns into numeric values and store the mappings. Next, I cleaned the data by handling missing values, removing or normalizing outliers, through sampling to solve the class imbalance problem, and used RandomForestClassifier to evaluate the important features for the target label and remove insignificant features to improve the model efficiency. Then, I made bar charts to show the top 10 most frequent values for each categorical column, and used stacked bar plots to analyse categorical column change over time (monthly, quarterly and yearly). Lastly, I did a correlation heat map to analyse all numerical variables to extract highly correlated variable pairs.

According to the histograms and boxplots for numerical variables, “Sales” is right-skewed, with most values from $0 to $500, and some extreme values above $4000. “Profit” centres around 0, but some losses below -$1000. Chairs produced the highest sales revenue, with Tables is second. “Quantity” mostly ranges from 2-6 units, but some values exceed 12. “Discount” mostly ranges from 0-0.3 but spikes up to 0.7. Those are outliers. According to the correlation heat map, “Customer Name” is almost the same as “Customer ID” (corr = 0.995), “Sub-Category” is highly correlated with “Product ID” (corr = 0.90). “Discount” and “Profit” are negative correlations (corr = 0.57), which means discount significantly affects profit.