Predicting Quarterly Sales for Customer Segments

A. Introduction and Problem Statement

Problem Statement: Predict Quarterly Sales for Customer Segments to Optimize Inventory and Marketing Budgets.

Dataset: A spreadsheet of company order shipments from 2013 to 2017.

Dataset Description: The dataset consists of thirteen (13) columns and eight hundred and twenty-five (825 rows). The first row indicates the shipping mode of the orders, the second row represents the customer segment and the third row denotes the dates.
The shipping modes are in four (4) categories - First Class, Same Day, Second Class and Standard Class.
The customer segments are three (3) - Consumer, Corporate and Home Office.

Column Key
FC = First Class
SD = Same Day
SC = Second Class
STC = Standard Class

B. Methods

Data Cleaning:
After the dataset was loaded and inspected, I cleaned the data. First, I renamed the columns as some displayed as 'Unknown' due to the spreadsheet formatting.
The first and second rows became redundant due to this and were dropped.
The column containing the dates was set as the index.
I also split the dataframe into 4 to better handle the visualisation and modelling of each shipping category. Empty rows in each of the 4 dataframes were removed.

Descriptive Statistics:
It was noted that the mean amount of order shipments was significantly different from the median values in each customer segment. This indicates the presence of extreme values (outliers) in this dataset. Figures 1 and 2 show the trend of standard class and same day shipments over time respectively, highlighting the extreme values.

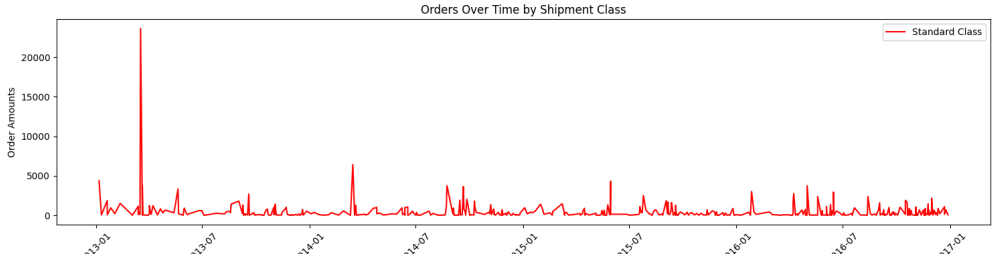

Fig 1: Line chart of the amount of order shipments over the period in standard class.

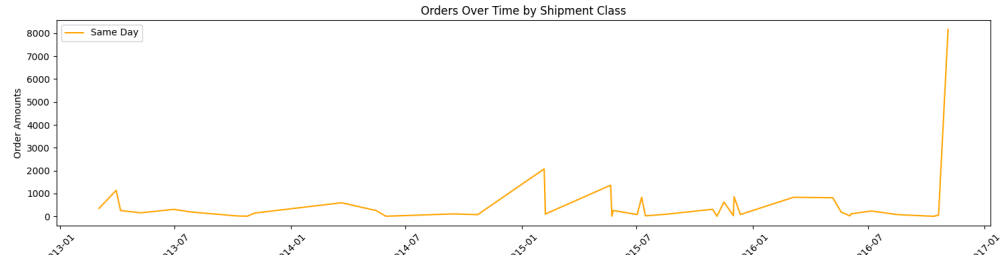

Fig 2: Line chart of the amount of order shipments over the period in the same day class.

Insights:

The customer segment with the most amount of order shipments was Home Office under standard class with #23661.228 on 18th March 2013. This is not consistent with the observed values in that segment.

The customer segment with the least amount of order shipments was Corporate in Standard Class with #1.167 on 16th December 2013.

It was noted that all four shipping modes were dominated by low-value orders. Figures 3, 4, 5 and 6 show a right-skewed distribution indicating that most order shipments cluster at the lower end of the sales range.

Across all four shipping categories, the consumer segments appeared the most frequently in the lower sales bins. This implies that consumers have the highest volume of small orders.

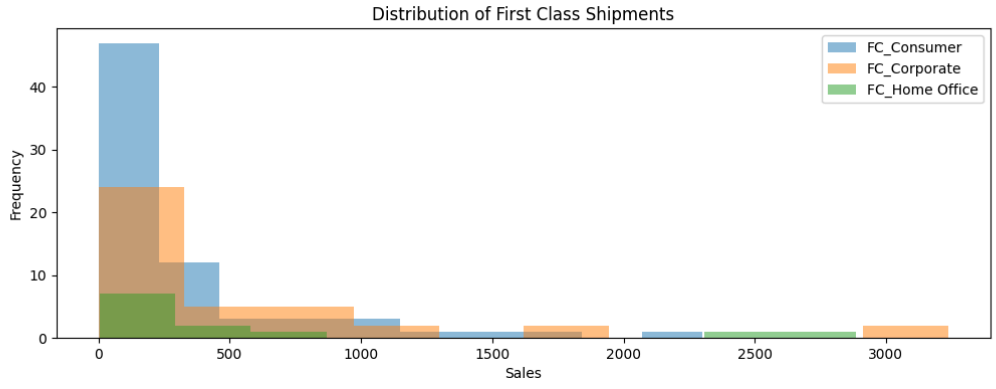

Fig 3: Histogram of first class shipments

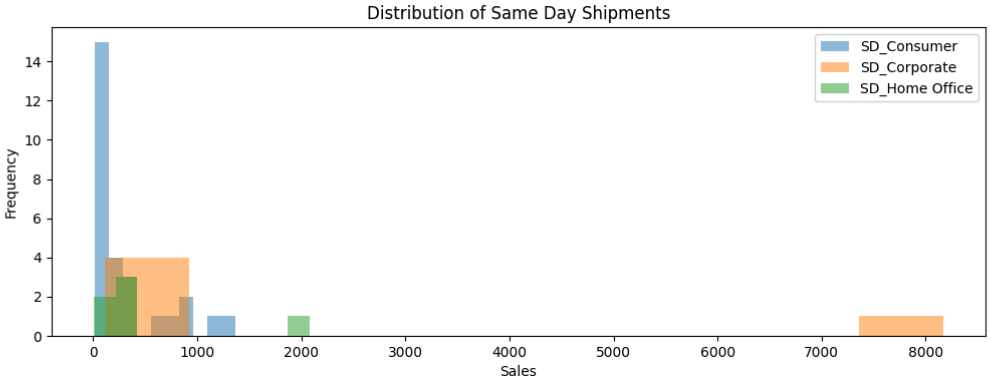

Fig 4: Histogram of same day shipments

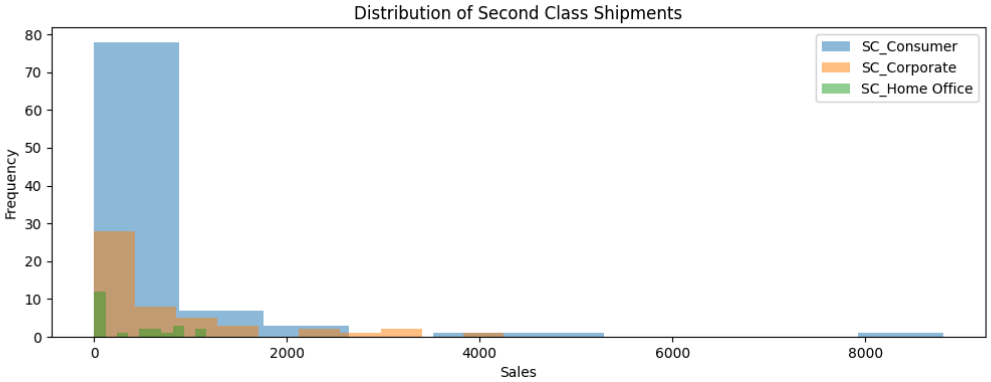

Fig 5: Histogram of second class shipments

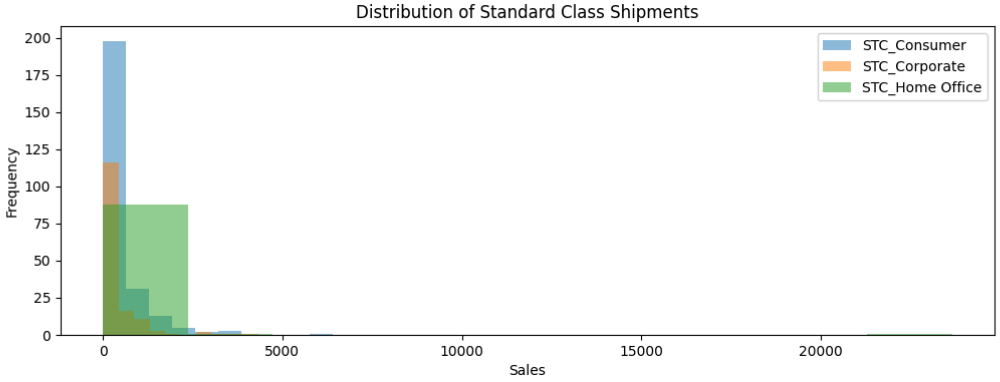

Fig 6: Histogram of standard class shipments

Business Implications:

1. Standard Class, and sometimes Second Class, handle shipments that can reach very high sales amounts. These outliers may require special handling such as larger storage space to ensure smooth fulfillment of orders.

2. Because the large volume of small orders is in the Consumer segments, specific marketing strategies like bulk deals are recommended to increase order size.

C. Model

Data Splitting: I used an 80-20% train-test split of the dataset.

The baseline model used was the median because of the presence of extreme outliers in the dataset.

The complex model used was linear regression. This model was chosen because of its simplicity and interpretability.

The linear regression model outperformed the median baseline according to the Root Mean Square Error (RMSE) performance metric. Table 1 summarises this.

Shipping Category	Median Baseline RMSE	Linear Regression RMSE
First Class	772.556	713.917
Same Day	2821.562	2730.833
Second Class	766.776	710.578
Standard Class	527.790	517.473

Table 1: RMSE values

Interpretation:
In every shipping category, the linear regression model achieved a lower RMSE compared to the median baseline model. This indicates that the linear regression model is better at capturing the underlying patterns and variability in the data than a simple median-based prediction.