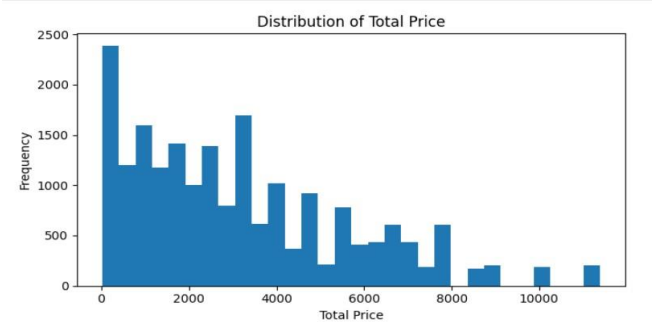
ElectroMart is an online electronics store that heavily invests in customer promotions and retention strategies but lacks the understanding of what truly influences customer spending or value. The purpose of the project was to reveal the factors (behavioral or demographic) that truly drive customer spending, how those specific factors help to predict high-value, low-value customers, and also uncover distinct customer segments that provide more targeted marketing strategies.

The **first step** was data preparation. I converted the data's date column (which was in a mess) into a date-time format to enable segregation based on month and quarters, which helped me to create seasons (Spring, Summer, Winter, and Fall). This was important to be able to explore seasons associated with the most spending. I also classified each customer into high versus low spenders based on the median of the total price. As you would see in the EDA stage, the total price distribution was right-skewed, so using the mean would have been quite misleading. I also converted the gender column into binary mode.

The **second step** was the EDA process, which helped me to investigate the dataset to summarize its main characteristics, gain insights, and prepare the data for modeling. Like I mentioned previously, below is a visualization of the total price, which is right-skewed.



I also explored EDA for variables like quantity, add-on total, product type, shipping type, payment method, etc.

The **next step** after EDA was to conduct hypothesis testing to help me make objective decisions regarding certain claims. A few of these hypothesis tests included whether buying add-ons related to the likelihood of being a repeat customer (*outcome:* there was no significant difference in total spending between customers who purchased add-ons and those who did not), whether spending differs between product categories (*outcome:* there was a statistically significant relationship as customers spend significantly more on smartphones, smartwatches, laptops, and tablets), or whether certain products sell better in certain seasons (*outcome:* The overall interaction model was highly significant ($p \approx 8.19e218$. This showed that the effect of product category on spending depended on the season. For instance, Smartphones and Smartwatches show increased spending in Spring and Summer, and Tablets spiked in Summer).

Over here, I also explored the demographic variables and realized none of them had any significant relationship with total price, and so these were eliminated from the analysis moving forward. I concluded that for ElectroMart, only behavioral variables like Product Type, Shipping Type, Payment Method, Season, Quantity, etc were useful.

In **step four**, I explored predictive modeling approaches, including linear regression, logistic regression, and decision tree analysis. In **linear regression**, the goal was to predict total price using the behavioral features identified. I split the data into 75% and 25% for training and testing, respectively. The training MSE was 2.59M, whereas the test MSE was 2.66M. Because some of the prediction errors ranged from 1000 to 2000, squaring them resulted in high MSE values, but this did not necessarily mean that the model was bad. The difference between the test and training MSE is not too huge and therefore suggests

minimal overfitting. I also explored the Lasso and tuned Lasso models as a way of regularizing the initial model, but found that they did not improve the model's accuracy. This means the initial predictors were informative and not excessively noisy.

From **the logistic regression**, the goal was to classify customers as high and low value.

	actual_highvalue	predicted_highvalue
0	1	1
1	0	0
2	0	0
3	1	1
4	1	1
5	1	1
6	0	0
7	0	0
8	0	0
9	0	1

Here is a sample of the prediction results from the logistic regression. The model resulted in an accuracy score of 81.84% demonstrating that behavioral features provide strong predictive ability to distinguish high-value from low-value customers.

	Predicted lowvalue	Predicted highvalue
Actual lowvalue	2098	472
Actual highvalue	436	1994

Also, the confusion matrix shows lower misclassifications for high and low-value customers.

For the **decision tree analysis**, the objective again was to predict high and low-value customers. Since it is able to capture non-linear variations, I decided to explore this as well and compare it to the logistic regression model, which follows a more linear approach. The accuracy score was 87.88%, which confirms that the decision tree better captured the non-linear nuances in the data.

	Predicted lowvalue	Predicted highvalue
Actual lowvalue	2348	222
Actual highvalue	384	2046

Also, the confusion matrix shows even lower misclassifications for high and low-value customers.

In **step 5**, the goal was to create clusters/customer segments based on the identified behavioral features analyzed through the predictive modeling. I used the elbow method and a conservative approach to settle on clusters (k =3). I then assigned each customer in the original dataset a cluster value and performed a cluster summary that provided interesting insights. Cluster 0, named **Low-Spend Efficient Buyers,** had customers with the lowest spending and the lowest quantity. Cluster 1, named **Add-On Lovers / Accessory-Heavy Buyers,** had a higher total spending compared to cluster 0 but lower than cluster 2. Most importantly, they were those with the highest add-on spending. Cluster 2, named **Bulk/High Quantity & High Spend Buyers,** were those with the highest total spending and highest average quantity spending per customer.

I also compared these clusters with the initial high versus low value assignments I made to each customer at the data preparation stage, and realized that there was a strong alignment.

HighValue segment	0	1
0	80.448598	19.551402
1	37.441574	62.558426
2	27.597977	72.402023

As shown, Cluster 0 shows **80% of the customers are low-value customers**. Custer 1 shows **62% are High Value Customers (driven by add-ons)**, and Cluster 2 shows **72% are High Value Customers (driven by quantity and premium goods).**

In my last step, I provided strategic recommendations for ElectroMart based on my analysis. Some of these recommendations included switching marketing segmentation

from demographics to customer behaviors (e.g., product type, quantity, ratings) to accurately identify high-value customers. Also, allocating premium marketing offers, exclusive deals, and early product launches specifically to Cluster 2 and high-value prediction customers to maximize ROI. Targeting Cluster 1 (heavy add-on spenders) with accessory bundles, cross-sell recommendations, and checkout prompts to increase attachment rates. Furthermore, using "if-then" Decision Tree rules to design automated, relevant email flows (upgrades, accessory recommendations, service recovery). Lastly, introducing tiered benefits based on actual high-value behaviors (premium products, addons, high ratings), not just loyalty program membership, focusing perks on Clusters 1 and 2.

In conclusion, this integrated analysis helped to provide ElectroMart with the necessary insights to transition from broad demographic marketing to precise, behavior-driven strategies.