FH MTM Data Project Group 1

Select a Dataset

source: https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data/data

Data Analysis

Classification methode

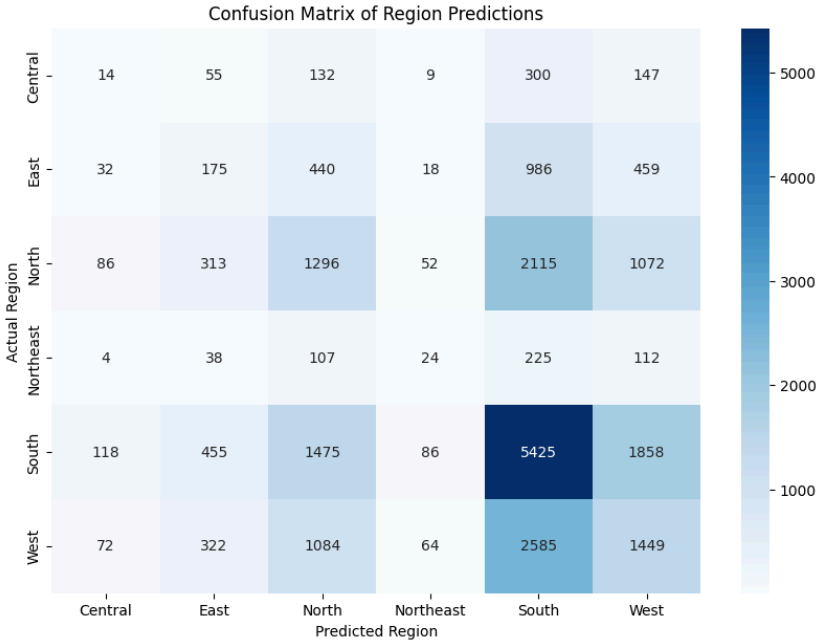
Accuracy: 0.36127391828995004

Classification Report:

	precision	recall	f1-score	support
0	0.04	0.02	0.03	657
1	0.13	0.08	0.10	2110
2	0.29	0.26	0.27	4934
3	0.09	0.05	0.06	510
4	0.47	0.58	0.52	9417
5	0.28	0.26	0.27	5578
accuracy			0.36	23204
marco avg	0.22	0.21	0.21	23204
weighted avg	0.33	0.36	0.34	23204

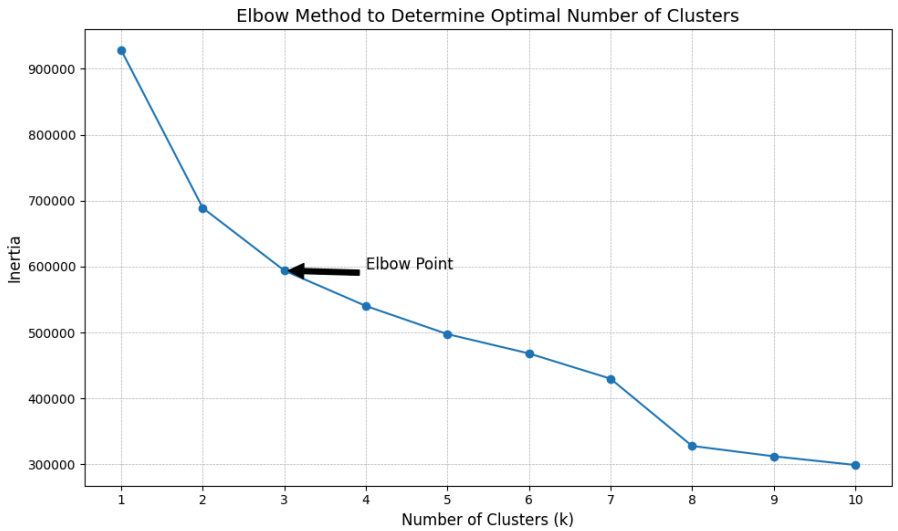
The performance results of the Random Forest classifier in predicting regions indicate several areas of concern. The model achieved an overall accuracy of 36.13%, suggesting that it correctly predicts the region for only about one-third of the instances in the test set. Class 4 had the highest performance, with a precision of 0.47, recall of 0.58, and an F1-score of 0.52, indicating that the model is relatively better at predicting this region. However, the performance for other classes is significantly lower. The macro average scores and weighted average scores further underscore the model's overall inadequacy in predicting the regions accurately.

Confusion Matrix



The model shows the highest prediction accuracy for the "South" region, with 5425 correct predictions. Significant misclassifications are observed between the "South" and "West" regions, as well as between the "North" and other regions. The "Northeast" region is particularly challenging for the model, with only 24 correct predictions. This confusion matrix highlights the need for improvements in model training, feature selection, or data preprocessing to reduce misclassification rates and improve overall prediction accuracy.

Elbow Point



Before the Elbow: As the number of clusters increases, the inertia, which measures the internal consistency of clusters, decreases rapidly. This indicates that each added cluster

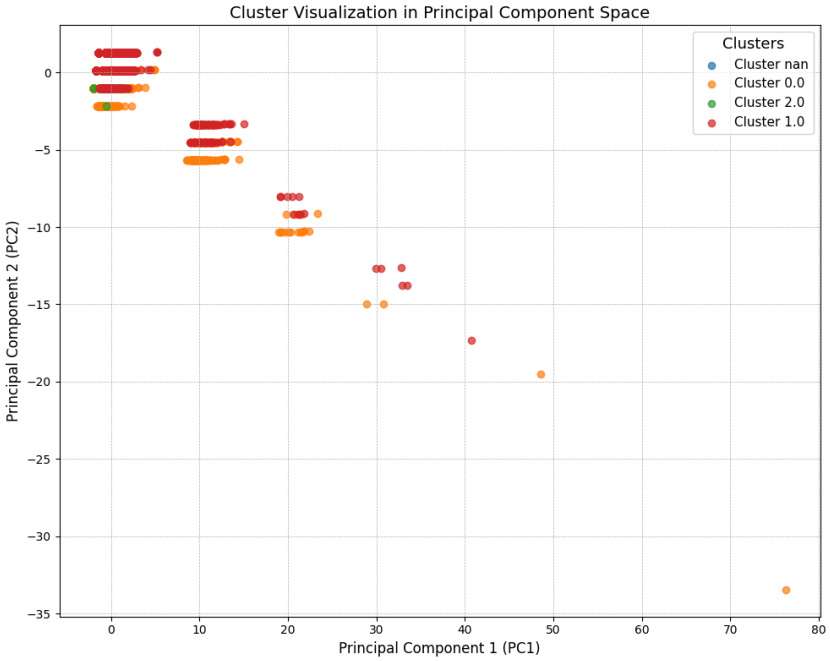
significantly improves the clustering by reducing the variance within each cluster. The sharp decline in inertia highlights the substantial benefits of adding more clusters in this initial phase.

At the Elbow: This is the point on the plot where a noticeable bend or elbow forms. It signifies a transition where the addition of more clusters starts to yield diminishing returns. In other words, increasing the number of clusters beyond this point does not significantly enhance the clustering quality. The elbow marks an optimal balance between the number of clusters and the within-cluster variance reduction.

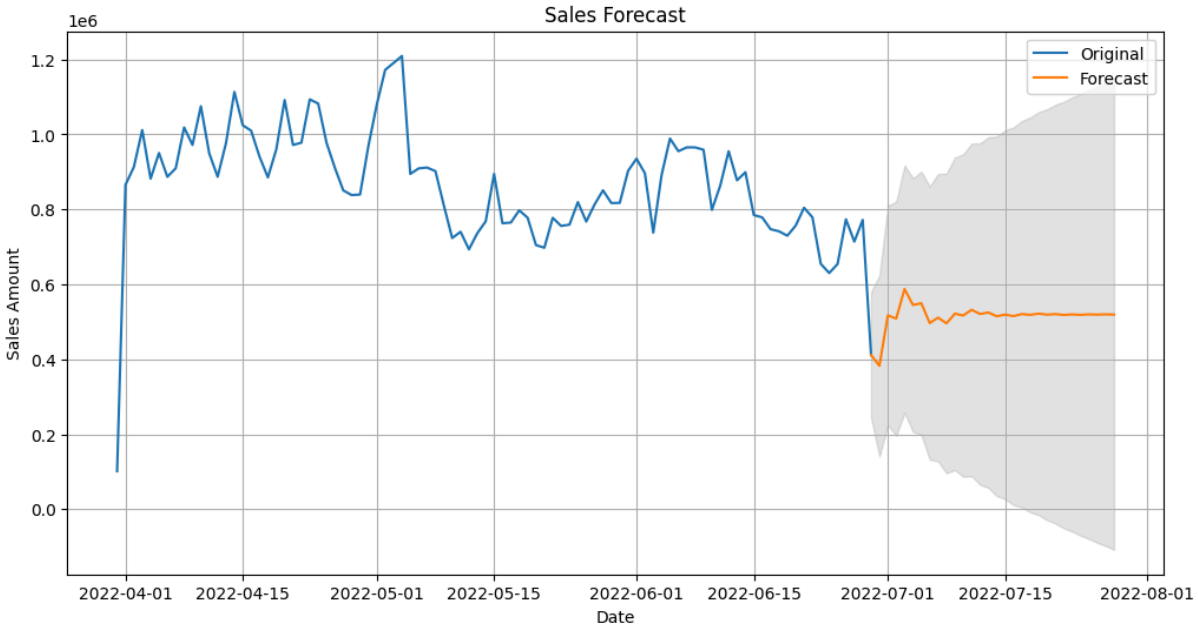
After the Elbow: Beyond this critical point, the inertia continues to decrease but at a much slower rate. This slower decline suggests that additional clusters contribute minimally to reducing the within-cluster variance. Essentially, adding more clusters past the elbow does not substantially improve the clustering outcome, indicating that the optimal number of clusters has been reached around the elbow point.

Interpretation of Clustering Results

The analysis of the clusters reveals distinct order characteristics: Cluster 0 represents lower-priced orders with an average amount of $407.65 and single-item orders, Cluster 1 includes higher-priced orders averaging $1157.05, and Cluster 2 consists of mid-priced orders averaging $692.45. All clusters show a consistent pattern of orders placed mid-week (Wednesdays) and around mid-May. The findings suggest that order timing and price per item are key differentiators among the clusters, providing valuable insights for targeted marketing, inventory management, and customer service strategies.



Time Series Conclusion



The time series analysis reveals distinct seasonal and trend components in the sales data. There is a noticeable upward trend in sales from early April to mid-May, after which sales begin to decline. This decline continues until the end of June, marked by fluctuations and an overall downward trend. The sales data demonstrates considerable volatility, with multiple peaks and troughs observed throughout the period. These sharp increases and decreases in daily sales suggest the influence of possible promotions, seasonal effects, or other

external factors impacting sales. Additionally, the data contains some outliers, particularly noticeable at the start of April and the end of June, where there are sharp drops in sales. From mid-May to the end of June, the overall level of sales declines, and the peaks become less pronounced compared to earlier periods. This period indicates a trend towards reduced stability in sales performance. Further analysis is recommended to better understand these trends and the external factors influencing sales volatility and fluctuations.