Name: Nathaniel Rodriguez
Student ID: 6272810

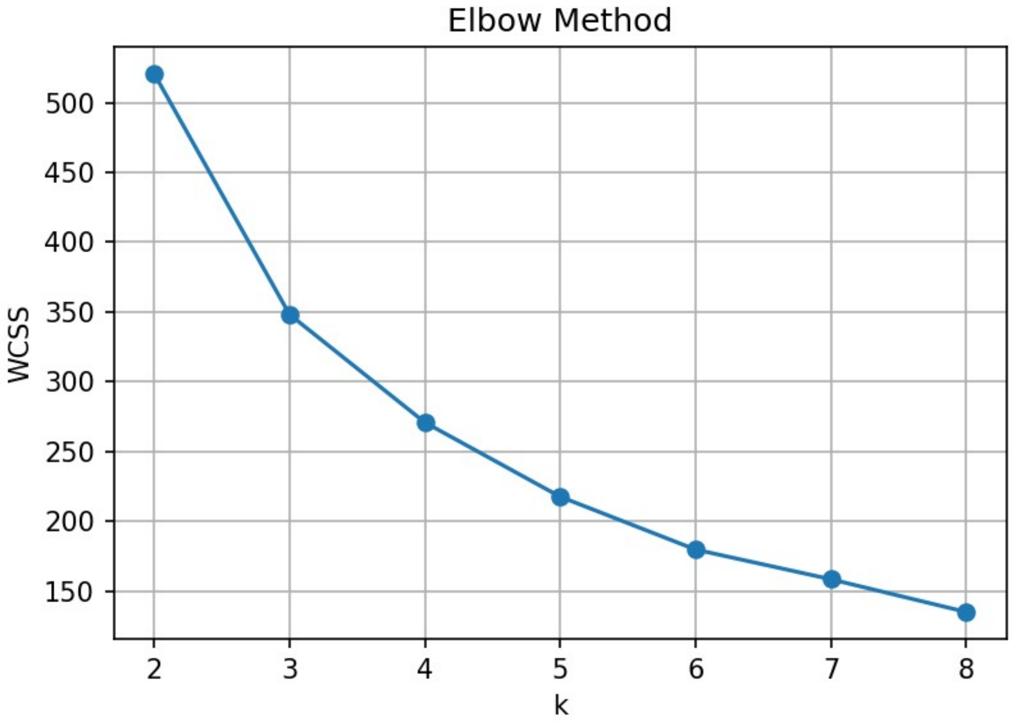
Introduction

This project analyzes product sales data using two main techniques: K-means clustering and regression modeling. The goal is to understand product groups, predict units sold, and extract business insights from the dataset.

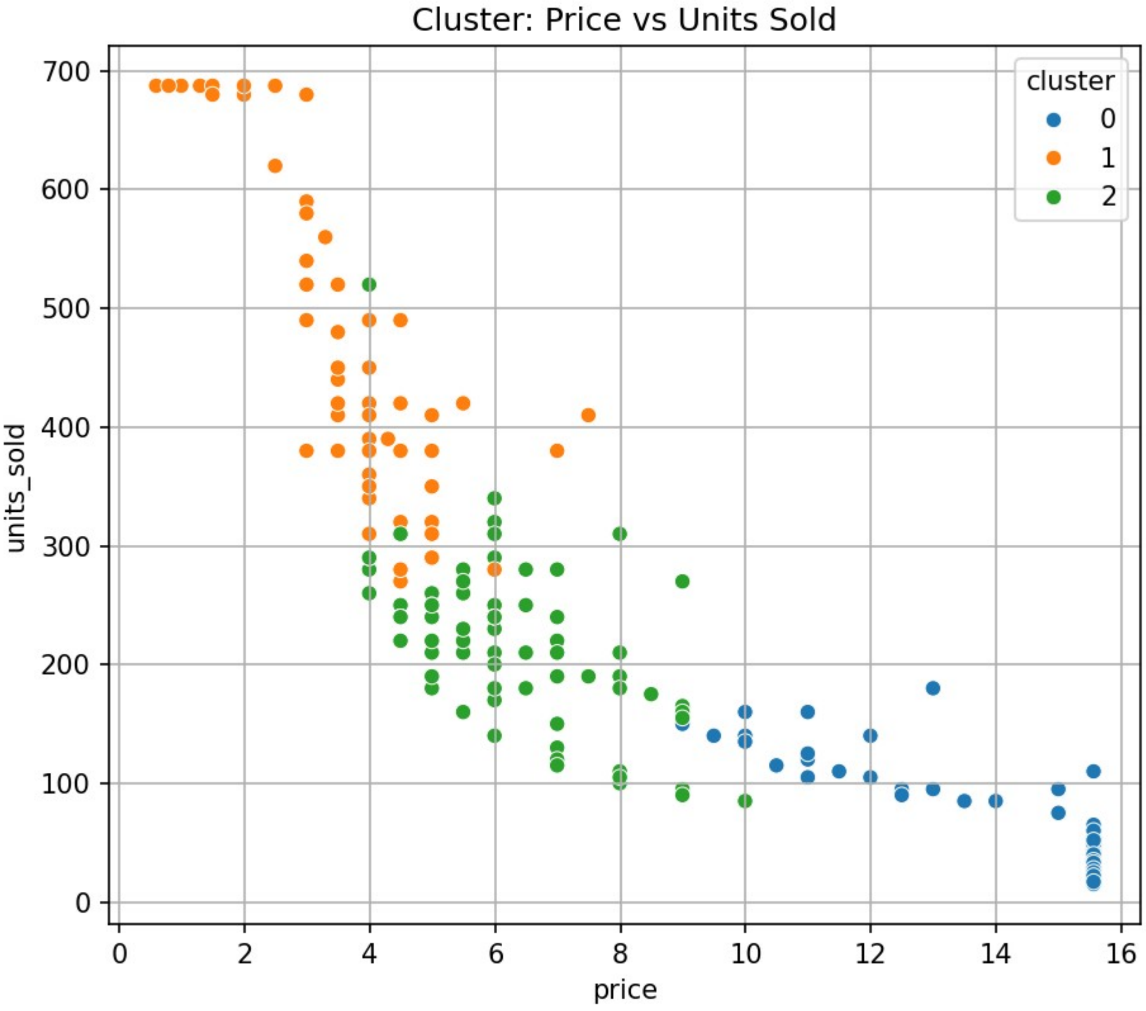
Data Preprocessing

We handled missing values by checking for gaps and filling or removing them depending on how important the field was. Outliers were detected using visual checks and statistical thresholds and removed when they distorted the analysis. Normalization (standard scaling) was applied so that all features were on similar scales, which helps clustering and regression models perform better. Summary statistics were reviewed to confirm the data was clean and consistent.

K-means Clustering Analysis

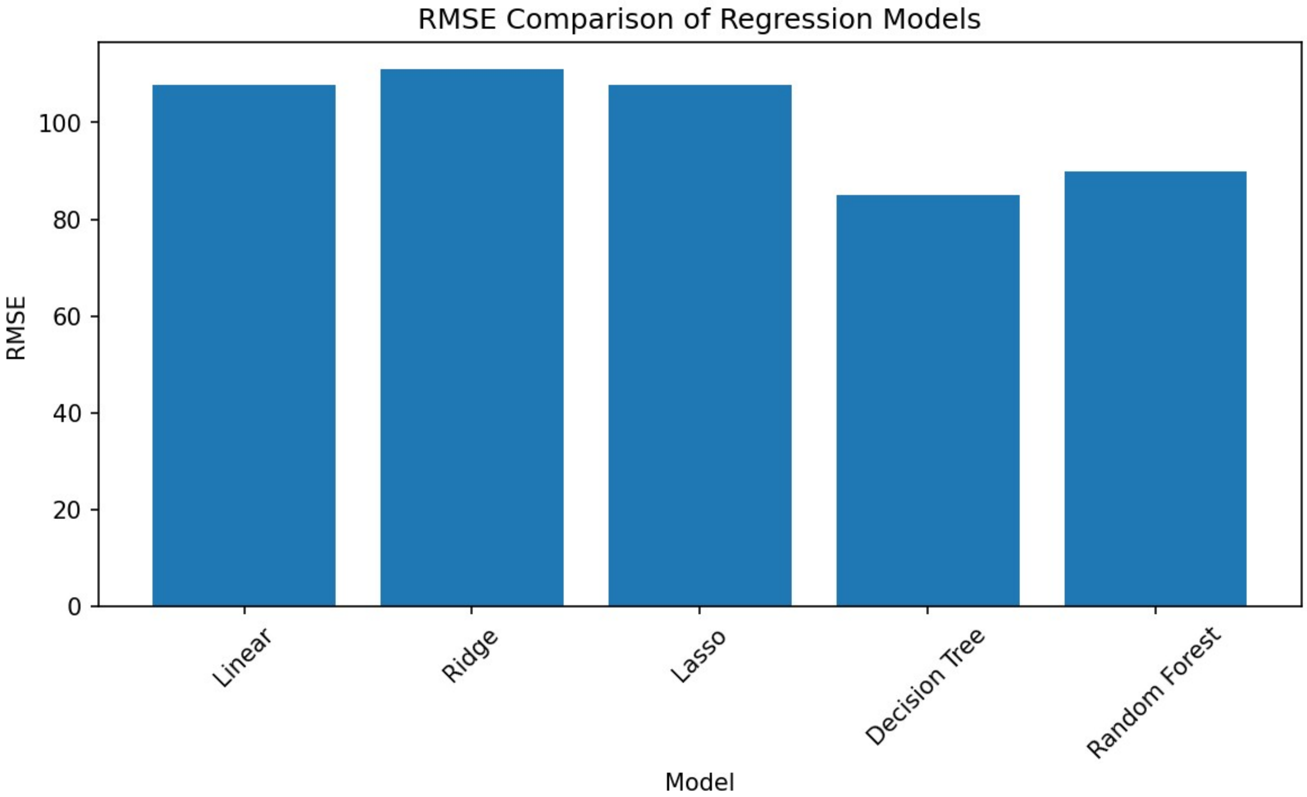


K-means was used to group products based on their attributes. The elbow method was applied by calculating inertia across different k values, and the point where improvements slowed down was chosen as the optimal k.

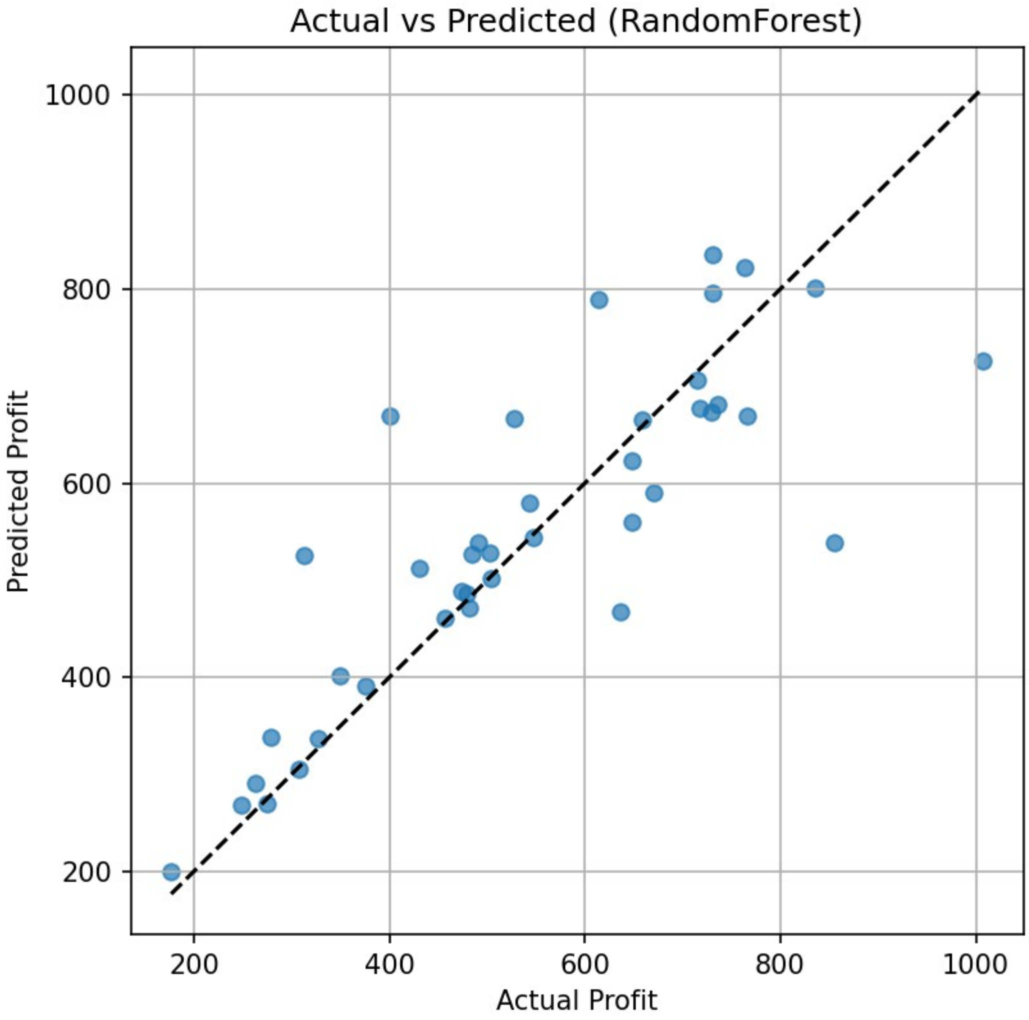


Cluster statistics were generated to show the average cost, price, units sold, and promotion frequency for each cluster. Each cluster was interpreted and given a practical meaning based on its characteristics, helping identify product types and performance patterns.

Regression Analysis



We tested multiple models including Linear Regression, Random Forest Regression, and Polynomial Regression. The training process involved splitting data into train/test sets and evaluating each model's accuracy. A performance comparison table helped determine which model predicted units sold the best.



The best model was selected based on R², MAE, and MSE values. We also discussed overfitting by checking how well models generalized to unseen data.

Conclusion

The main findings were identifying product clusters and determining which regression model predicts performance most accurately. Limitations include model simplicity and potential improvements such as trying more algorithms or expanding the dataset.

AI Tool Usage Summary

AI tools were used to help write code.