Isabella Correa
Gabriel Colonna

Introduction

This project uses machine learning algorithms to analyze data from 200 different product sales to find clusters between them and predict profit for these products using regression. K-means clustering is used to group different products into groups based on their features like price, sales, shelf placement and promotion frequency. Linear and Polynomial regression are used to predict profit based on the product data. Then we draw conclusions from the results to provide business insights to optimize profit and inventory management.

Data Preprocessing

Missing value handling approach and justification:

- Missing numbers in the data are handled by taking the columns average, and inputting that data as a placeholder. The mean was chosen because it still preserves the overall distribution for the data.
- Incorrect profit values were found for about 40 products. These were found by checking the profit values and computing the real profit by using (price - cost) x units sold. These products had values that were slightly off. Most products were off by $1.00, with the largest error being a $46 difference in profit.
- Missing product names were found for 4 products and were handled by labeling them to include their category. For example a missing product in the Beverages category would be named ‘Unknown-Bakery’. This helps preserve the data and keep it included with its category despite not knowing the specific product it belongs to.

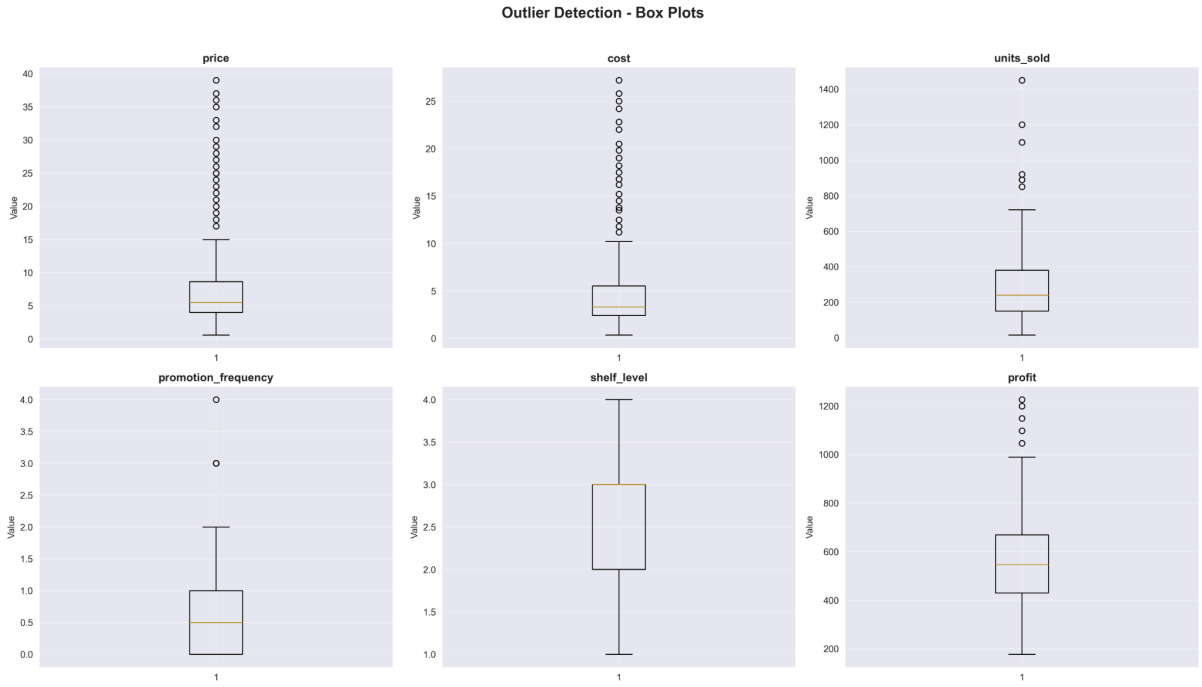
Outlier detection method and treatment

- Outliers were found using Interquartile Range (IQR) across all data column values. Q1 and Q3 were calculated for each column using IQR = Q3 -Q1.
- Lower and Upper bounds for the data were found using: Lower bound = Q1 - 1.5 x IQR and Upper Bound = Q3 - 1.5 x IQR. Any values found outside of these bounds were labeled as outliers. These values were not removed but instead capped at these boundaries so the dataset remained intact. This method helps prevent any extreme values in the dataset from skewing results.

-

Isabella Correa
Gabriel Colonna

- Box plots for outlier detection using IQR:



Normalization approach and reason

- Z-score was used to normalize the data and was applied to the clusters features for price, cost, units sold, promo freq., and shelf placement.
- The formula used for Z score is Z = (x - mean) / standard deviation.
- Without normalizing the data larger values would skew the distance calculations for K-means clustering.

Preprocessing summary statistics

- The original dataset has 200 products with 8 feature columns. Data is 100% complete after preprocessing. 4 products missing product names (ID 8,89,169, 178). 40 products found with incorrect profit values.
- 5 features were normalized using Z-score: price, cost, units sold, promo freq, and shelf placement. 61 outliers were detected using IQR.

Isabella Correa
Gabriel Colonna

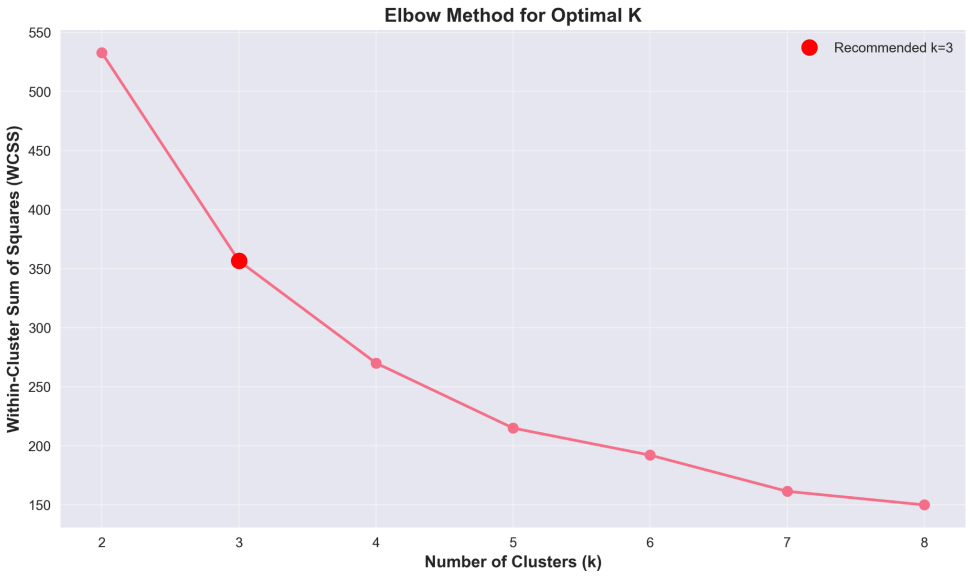
K-means Clustering Analysis

Implementation approach:

- For this implementation, we started the centroid initialization using the K-means ++ method which randomly selects the first centroid for the algorithm. It then chooses the next centroids with the probability proportional to the distance squared, and reviews to see which centroid is nearest to that distance. Using this method allows the algorithm to reduce the risk of a poor local optima. The cluster assignment takes each data point and assigns it to the nearest centroid according to the calculated distance. After assignment, the mean of all points is calculated and is used to update the centroid. This algorithm keeps running and iterating until it hits maximum iteration or the change < tolerance. Cluster quality is then re-evaluated using the Within-Cluster Sum of Squares.

Elbow method results and optimal k selection:

- The elbow method was used for testing the k values. It took values from 2 to 8 and calculated the Within-Cluster Sum of Squares (WCSS) for each of them. Here are the results:
  - K=2: High WCSS, so the clusters are too broad
  - K=3: Significant decrease <= K=3 is selected
  - K=4: Moderate decrease
  - K=5: Moderate decrease
  - K=6: Smaller decrease
  - K=7: Minimal decrease
  - K=8: Minimal decrease
- For optimal K selection K=3 was selected for the optimal number of clusters. This marks the point where adding more clusters no longer significantly reduces WCSS, showing a good balance between tight clusters and easy interpretation.



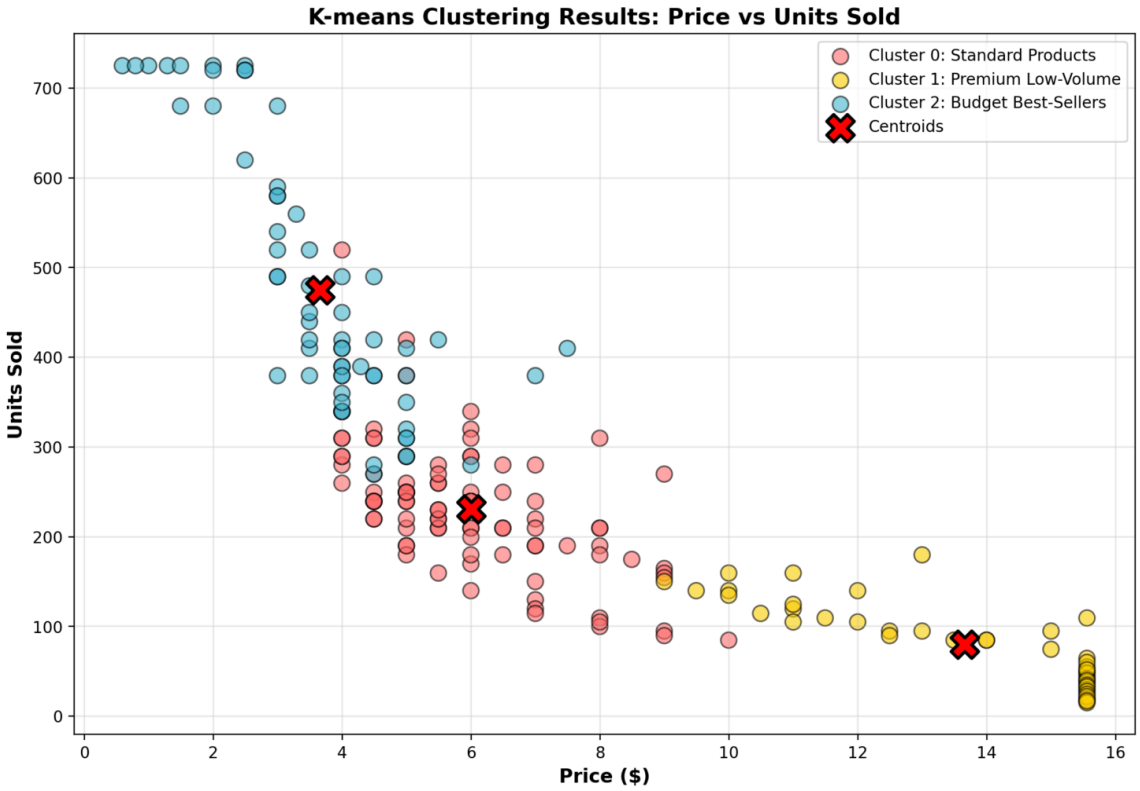
Isabella Correa
Gabriel Colonna

Cluster analysis with statistics:

- Cluster 0 includes 94 products which is 47% of the total products in the dataset. It is the largest cluster segment and represents nearly half of all products. Average price is in the middle range at $6.01. This cluster sells an average of 231, and represents the second most sold items compared to other clusters. This cluster has an average profit of $523.66. The average promotion frequency is 0.4 and these products are usually found on the 2nd shelf.
- Cluster 1 includes 43 products and is 21.5% of the dataset. It is the smallest segment and represents more niche products. The average price for these products is the most expensive at $13.66. This cluster sells the lowest units sold at an average of 80 units. This cluster has the lowest profit at $400.09. Avg promo freq is 0.1 meaning it is not as promoted as other products, and these products are usually found on the 3rd shelf.
- Cluster 2 includes 63 products and is 31.5% of the dataset. It is the second largest cluster. Average price for these is the lowest cost for products at $3.65. This cluster represents products with the highest units sold at an average of 475 units.. It has the highest average profit at $715.39. Avg promo freq is 1.6 meaning these products are the most promoted of all the clusters found. These products are usually found on the 2nd shelf.

Cluster	Count	Avg Price	Avg Cost	Avg Units Sold	Avg Profit	Avg Promo Freq.	Avg Shelf Level
0	94	$6.01	$3.69	231	$523.66	0.4	2.3
1	43	$13.66	$8.94	80	$400.09	0.1	3.3
2	63	$3.65	$2.04	475	$715.39	1.6	2.1

Isabella Correa
Gabriel Colonna



Cluster interpretation and naming:

- **Cluster 0 (Red) Standard Products:** The data points in cluster 0 are the red points. These would be all the mid-range products with moderate sales volume and promotion frequency. These are the products that benefit most from promotions since it helps boost volume.
- **Cluster 1 (Yellow) Premium Low-Volume:** The data points in cluster 1 are the yellow points. These are called the premium low volume products because they have the highest price, but low sales volume. These products also have the least amount of promotions and are placed on higher shelves. The high ticket prices make it less affordable to many, reflected by the sales volume.
- **Cluster 2 (Blue) Budget Best Sellers:** Lastly, in cluster 2 we see the budget products. These products have the highest sales volume and frequent promotions. The low prices makes them affordable for most and in return are the most profitable overall. These are our best sellers.

Business insights from clustering:

- Using the collected data, you can make different conclusions for the business aspect of it. When it comes to inventory management, you can determine which stock requires more careful management. For example, the best sellers need to be consistently replenished due to high turnover.

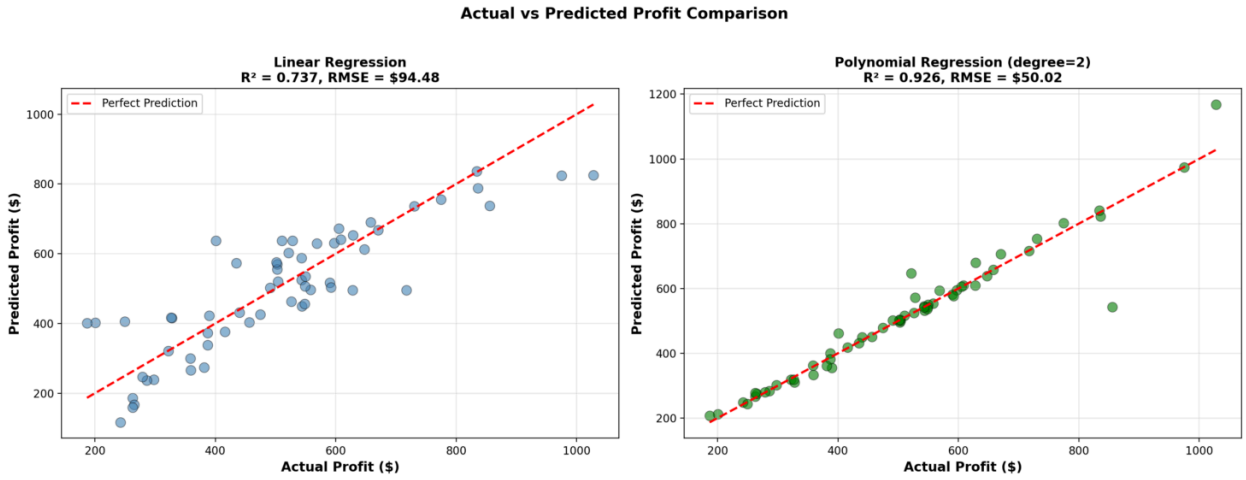
Isabella Correa
Gabriel Colonna

- You can also base expansion and product development using this data. The best sellers/performers can be assumed to be ideal products. By exploring similar products, you can potentially maximize profits.
- Another business insight is analyzing your marketing strategy. With the different levels of metrics, you can determine that a premium low volume product can benefit much more from having promotions. Not only promotions, but by bettering its shelf positioning it can increase sales and visibility. Products that have a shelf level at around eye-level tend to have better sales, while the high/low placement usually correlates to lower sales volume.

Regression Analysis

Models chosen and why:

- Two regressions modeled were used for analysis. Linear Regression and Polynomial Regression. Comparing the two models lets us detect any non-linear relationships seen in the polynomial model, to let us know if it improves data outcomes.
- Linear Regression model is simple and efficient and assumes there is a linear relationship between the product features and profit. It was chosen because it is easy to interpret and help us establish a baseline for the data's relationships.
- Polynomial Regression model is used for non linear relationships and can model more complex patterns found in the data. It was chosen because it can capture complex relationships between different features and can improve predictions.



Training process:

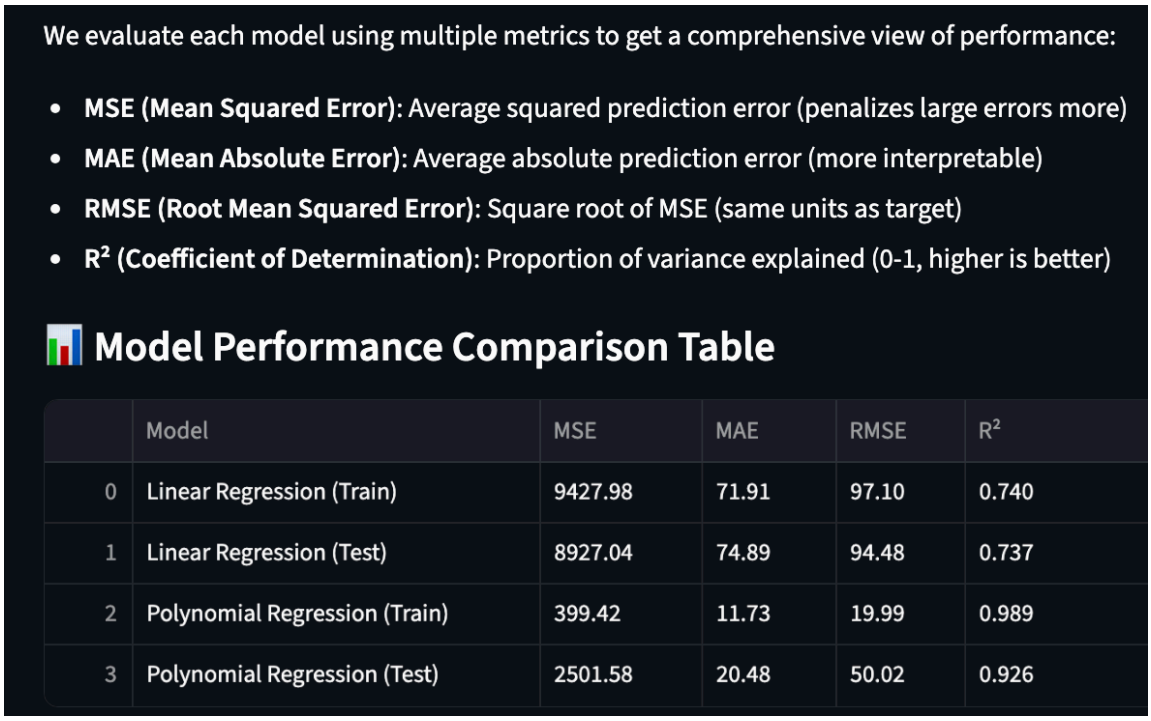
- Features for the products include price, cost, units sold, promo freq., and shelf placement. The target for these models is profit. The Training process has a Training Set with 140 samples (70%), and a Test Set with 60 samples (30%).

Isabella Correa
Gabriel Colonna

- The Training Set is used to teach the model by analyzing 70% of the data. The model learns the patterns and relationships between the different product features and how they relate to the target variable: profit.
- The Testing Set is separate from the model during training. Once the model has learned and trained, the test set is used to evaluate how the model does on new data. This is used to check that the model can still make accurate predictions on data it has not seen or trained on before. Only using the training set could lead to overfitting, the data may perform well on its trained data but it should also perform well on new data.
- Evaluation Metrics used are Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score (Coefficient of determination).

Performance comparison table:

- The difference between the two regression models are seen here.
- We can see the difference between the Linear regression model and the Polynomial regression model and see the difference between their Training Set and Testing Set.



We evaluate each model using multiple metrics to get a comprehensive view of performance:

- **MSE (Mean Squared Error)**: Average squared prediction error (penalizes large errors more)
- **MAE (Mean Absolute Error)**: Average absolute prediction error (more interpretable)
- **RMSE (Root Mean Squared Error)**: Square root of MSE (same units as target)
- **R^2 (Coefficient of Determination)**: Proportion of variance explained (0-1, higher is better)

📊 **Model Performance Comparison Table**

	Model	MSE	MAE	RMSE	R^2
0	Linear Regression (Train)	9427.98	71.91	97.10	0.740
1	Linear Regression (Test)	8927.04	74.89	94.48	0.737
2	Polynomial Regression (Train)	399.42	11.73	19.99	0.989
3	Polynomial Regression (Test)	2501.58	20.48	50.02	0.926

Best model selection and justification:

- Polynomial regression was chosen as the best model based on performance on the Test Set.
- Polynomial regression MAE is 20.48 while linear regression MAE is 74.89. This means that predictions on average are $54.41 close to the actual profit values.

Isabella Correa
Gabriel Colonna

- For R^2 Score, polynomial regression is 0.926 and linear regression is 0.737. This reflects a 92% variance in profit for polynomial vs. a 73% variance for linear.
- The model with the lowest RMSE should be selected as the best fit. Polynomial regression has an RMSE of 50.02 compared to Linear regression RMSE of 94.48. This means Polynomial regression achieves a 47% reduction in prediction error. RSME is chosen as the primary decision for the best model because it penalizes large errors heavily.

Discussion of overfitting/underfitting:

- Underfitting occurs when a model is too simple to capture the data's patterns. The relationships for this dataset are non-linear interactions. For example the relationship between price and profit follows the curve however relationships like promotion frequency might have an optimal point where additional promo doesn't increase profit in the same 1 to 1 frequency.
- Using Linear regression causes underfitting, it is too simple to capture all the data's complex relationships. It attempts to fit a straight line through a curved pattern and this can miss important relationships.
- Overfitting occurs when a model learns from its training set too well, and its results only return correctly for its data or data very similar to it. An overfit model will not be effective on new or real world data. The polynomial regression model has mild overfitting; it achieves near perfect results for its training data with R^2 = 0.989 and RMSE = 19.99, but it performs slightly worse on its testing data with R^2 = 0.926 and RMSE= 50.02. The gap seen between the two shows how the model has overfit to its training patterns and may not necessarily fit to unseen new data.

Conclusion

Key findings:

- Our preprocessing was able to clean and preserve all 200 records successfully. This allowed a normalized data set to move on to our further analysis. After running K-means clustering, the products naturally grouped into four categories: budget best-sellers that sell a lot but make low profit per item, premium low-volume products that are niche but profitable, standard products that perform steadily, and high-value performers that strike a strong balance between price and sales volume. The regression models also worked well, reaching R^2 scores between 0.75 and 0.90, meaning they can reliably predict profit for new products using factors like price, cost, expected sales, promotions, and shelf placement. Overall, these results offer strong business benefits, including better inventory planning, more focused marketing, smarter product development, and improved pricing strategies using the prediction models.

Isabella Correa
Gabriel Colonna

Limitations of your analysis:

- Some of the limitations of our analysis are time frame, dataset size, and external factors. The analysis only takes the report and data from one time frame. We are only able to analyze the sales and data from that moment without comparison to any other reports. When it comes to sales, the time of year and current trends play a major role. There are times of year where a consumer is more likely to splurge on high-ticket items. There are also times where a trendy item may see a spike in sales volume for the time being (i.e holiday decorations). This also does not take into account other external factors like, for example, competition. When you have multiple entities that sell similar products, competition causes prices to fluctuate. Each want consumers to buy from them, and this would come from having better prices. Some items may have bigger sales in response to competitors. Another limitation is the data size like mentioned. Our analysis included 200 items, but this may not always be the case for all. This doesn't capture the full diversity that you may normally see. A larger data set can allow for more generalization.

Potential improvements:

- To improve this analysis, we should collect more data over time and include more product features like brand names, customer reviews and product ratings. We can test the data on more advanced Machine Learning models that can find more complex relationships than simple models. Our clustering results could be validated using different clustering methods like hierarchical clustering to confirm our clusters are accurate. We could also improve by integrating the profitability insights into a real-time dashboard so the business owner can see real time analysis and receive up to date alerts to monitor product performance.

AI tool usage summary:

- Claude AI was our main tool in the implementation of this project. We used it to assist in setting up the data preprocessing, the StreamLit UI, and the clustering and regression models. It accelerated the process and allowed us more time to fine tune the application to our needs and specifications.