

Strategic Thinking - Capstone Project

US Candy Distributor:

Sales Prediction and Demand Forecast

**Aline Menezes Da Silva**

**18/05/2025**

**https://github.com/CCT-Dublin/ca1-capstone-project-proposal-Aline-Menezes-Silva**

Table of Contents

[1. Introduction 3](#_Toc196854499)

[2. Objectives 4](#_Toc196854500)

[3. Problem Definition 4](#_Toc196854501)

[4. Scope and Management 5](#_Toc196854502)

[5. Data Sources 5](#_Toc196854503)

[6. Ethical Considerations 6](#_Toc196854504)

[7. Data Understanding 7](#_Toc196854505)

[8. Technology/Methodology 14](#_Toc196854506)

[8.1 Libraries 14](#_Toc196854507)

[8.2 Machine Learning 14](#_Toc196854508)

[8.3 Time Series 14](#_Toc196854509)

[8.5 Hypothesis Test 15](#_Toc196854510)

[8.6 Hyperparameter Tuning and Cross Validation 16](#_Toc196854511)

[9. Results and Discussion 17](#_Toc196854512)

[9.1 Business Questions 17](#_Toc196854513)

[9.2 Hypothesis Testes 22](#_Toc196854514)

[9.3 ML Regressor Models 24](#_Toc196854515)

[9.4 Hyperparameter Tuning for tree-based models 26](#_Toc196854516)

[9.5 Feature Importance 29](#_Toc196854517)

[10. Time Series Analysis 30](#_Toc196854518)

[11. Conclusion 34](#_Toc196854519)

[12. References 35](#_Toc196854520)

Table of Figures

[Figure 1 - Project Scoop and Management. 6](#_Toc198854703)

[Figure 2 - Candy Dataframe .info 8](#_Toc198854704)

[Figure 3 - Dataset Assessment before preprocessing. 9](#_Toc198854705)

[Figure 4 - Missing values Pattern. 10](#_Toc198854706)

[Figure 5 – Numerical features outliers, before applying Log transformation 11](#_Toc198854707)

[Figure 6 - Outliers treated after applying log transformation 11](#_Toc198854708)

[Figure 7 - Annual Candy Sales 2021-2024 13](#_Toc198854709)

[Figure 8 - Pie chart - Shipping Method Popularity 14](#_Toc198854710)

[Figure 9 - Statistical Analysis Of Categorical Features 14](#_Toc198854711)

[Figure 10 - Comparing Sales and Profits Between Different Countries 17](#_Toc198854712)

[Figure 11 - Sales volume and profitability margins variation across states 18](#_Toc198854713)

[Figure 12 - Lowest Selling Products 19](#_Toc198854714)

[Figure 13 - Revenue by Factories in the last four years 20](#_Toc198854715)

[Figure 14 - Factory Performance Across Countries and Factories 21](#_Toc198854716)

[Figure 15 - Hypothesis Test 1. Profit Distribution by Shipping Method 23](#_Toc198854717)

[Figure 16 - Hypothesis Test 2. Profit by Unit Cost 24](#_Toc198854718)

[Figure 17 - ML Models Performance 26](#_Toc198854719)

[Figure 18 - ML Models Performance dataframe 27](#_Toc198854720)

[Figure 19 \_ Models Performance After Applying Hyperparameter Tunning 28](#_Toc198854721)

[Figure 20 - Models performance- Hyperparameter Tunning Dataframe 29](#_Toc198854722)

[Figure 21 - Feature Importance Using Gradient Boosting 29](#_Toc198854723)

[Figure 22 - Candy Demand trends over time 30](#_Toc198854724)

[Figure 23 - Candy Annual Growth 2021-2024 31](#_Toc198854725)

[Figure 24 - Seasonal Differentiation - ACF and PACF 32](#_Toc198854726)

[Figure 25 - Interpreting SARIMA Parameters 32](#_Toc198854727)

[Figure 26 - Demand Forecasting - SARIMA Model 33](#_Toc198854728)

[Figure 27 - The best parameters found using Grid Search 34](#_Toc198854729)

[Figure 28 - Demand forecasting after applying Grid Search 34](#_Toc198854730)

[Figure 29 - SARIMA Performance 35](#_Toc198854731)

[Figure 30 - Final Future Demand Forecast 35](#_Toc198854732)

# Introduction

Sweets have always been a popular treat enjoyed by people of all ages, and the demand for them keeps growing. This presents great opportunities for both new businesses and established brands. However, succeeding in this busy market requires a clear and simple plan. A good strategy focuses on knowing what customers want, keeping up with new trends, and making sure products reach shops and customers smoothly.

The sweets and confectionery industry is changing fast. People want exciting new flavours and healthier options. While chocolate remains a favourite, especially as a gift, other treats like gummy sweets and sugar-free products are becoming more popular, especially among health-conscious buyers. Seasonal occasions such as Christmas, Easter, and Halloween also play a key role in boosting sales, as sweets are often bought for celebrations.

*In recent years, the US confectionery market has been characterised by a steady shift toward healthier and premium offerings, as well as increased demand for nostalgic and seasonal products. (A Deep Dive into the Confectionery Market in the US, 2025)*

In today’s fast-paced and competitive market, it is vital for distributors to plan carefully and make smart decisions. One of the best ways to do this is by using data to understand what customers want and predict future sales. This report examines sales data from a candy distributor in the United States. By using Machine Learning and modern forecasting techniques, the aim is to discover patterns that can help the company manage stock more effectively and increase profits. These insights will help the business make better decisions, reduce waste, and stay ahead of competitors.

# Objectives

The main goal of this analysis is to understand how the USA Candy business is performing in terms of sales and profits, and to find ways to improve. It will look closely at how well different products are selling, how demand changes over time and in different regions, and which items are making the most or least profit. By spotting trends and patterns, the report will help identify what’s working well and where things could be better.

It will also highlight the best and worst-selling products so the business can make better decisions about how much stock to keep and whether prices need adjusting. The report will look at how efficient the factories and shipping processes are, aiming to find cheaper and faster ways to get products to customers.

Another part of the analysis will test important business ideas, for example, whether the way products are shipped affects how much profit is made, or if the cost of making a product has a big impact on how much money the business earns from it.

To help with future planning, the analysis will also use forecasting tools to predict how sales might change over the next year. A time series model will be used to estimate future demand for sweets, showing when sales are likely to be high or low. This will help the company get ready for busy times, avoid having too much or too little stock, and keep costs under control.

# Problem Definition

A US-based sweet distributor is finding it difficult to boost sales and profits, mainly because it's unclear which sales patterns to follow and how best to price its products. Without reliable predictions of customer demand, the company often orders too much stock, or not enough. This can lead to wasted money or missed sales opportunities.

In addition, shipping methods may not be cost-effective, and product prices might not be set in a way that brings in the best returns. The company also risks missing out on key opportunities for growth by not knowing which regions, products or customer types are the most successful.

To stay competitive in a challenging market, the distributor needs a clearer picture of what’s working and what’s not. This means using data to understand which items sell best, where they sell best, and which customer groups are the most profitable.

This analysis will explore the company's past sales data to answer practical questions, such as:

* Which shipping options are too expensive?
* Are higher-priced items actually more profitable?
* When do sales tend to rise or fall, and when should promotions be used?

*“Having the right stock at the right time means businesses adequately manage their supply chains and customer service operations. All this improves efficiency and profitability.”* (Schwarz, 2024)

# Scope and Management

This project was carried out over two academic semesters and followed a clear, step-by-step plan. The aim was to explore the topic in detail while making improvements along the way. To manage the work effectively, the project was split into several key stages. Each stage focused on a different part of the analysis, such as research, data collection, and interpretation.

After completing each stage, the progress was reviewed, and any necessary changes were made before moving forward. This helped to make sure that the final results were reliable and well thought out. By regularly checking and refining the work, the project stayed on track and adapted to new findings as they came up.

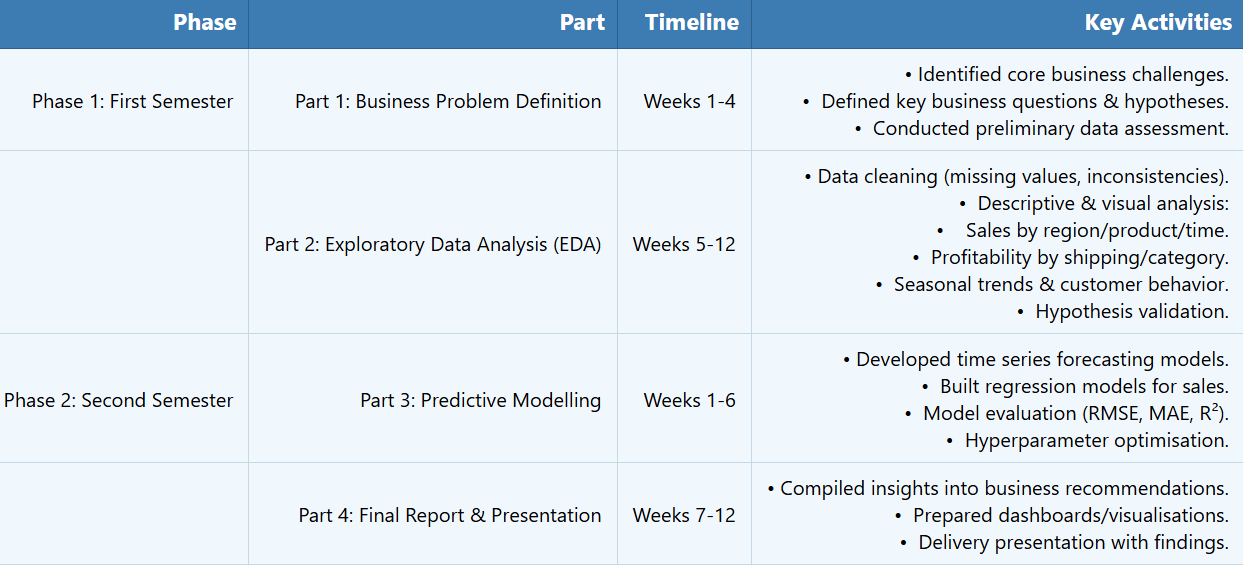


Figure - Project Scoop and Management.

# Data Sources

The main dataset used in this project is the Candy Sales Dataset, which comes from the Maven Analytics Data Playground. (Analytics, 2024)

This dataset includes real-world sales data from a candy distributor based in the United States. It provides a range of useful information, such as:

* Locations of both customers and manufacturing factories
* Details of sales orders, including the types of products sold and how much profit each one made
* Information about shipping methods and how long deliveries took

This dataset was chosen because it includes all the essential details needed to explore trends, understand customer behaviour and build predictions. Its variety of data points makes it especially useful for analysing how different factors, like location, product type and shipping, affect sales performance.

# Ethical Considerations

This project followed clear ethical guidelines to use data in a fair and responsible way.

**Data Privacy and Legal Compliance**

It used a public dataset that did not include personal information such as names, addresses, or payment details. If any sensitive data had been present, it would have been removed or made anonymous to follow GDPR and other privacy rules. This helped the project stay within legal and ethical standards. (Irwin, 2024)

**Fair and Clear Analysis**

It made sure not to favour any region, product, or group unfairly. Known factors that could affect the results, like seasonal trends or local differences, were taken into account. Every stage, data cleaning, analysis, and modelling, was carefully recorded so that others could check or repeat the work. Any assumptions were explained clearly to avoid confusion.

**Responsible Use of Results**

Machine learning models were used only to understand sales trends, not to make important decisions without human input. The project explained where the models might fall short, so people would not rely too heavily on them. All sources were properly referenced, and the results were shared honestly to support good decision-making.

# Data Understanding

1. **Dataset Overview**

This dataset contains daily sales records from a US-based company that sells sweets and confectionery. It covers a four-year period, from 2020 to 2024. The information has been collected to help us better understand how the business operates and how its products perform over time.

The dataset includes several important types of information:

* Sales Transactions: This includes details such as the date each order was made, how many items were sold, how much revenue was earned, the cost of the products, and the profit made from each sale.
* Product Information: Each product is listed with its category and product line (such as chocolate bars, jelly sweets, etc.), along with how much each item costs the company to produce.
* Customer and Factory Locations: The data tells us where customers and factories are based, including the country, state, and city. This can help us spot regional trends or differences in sales.
* Shipping Details: This part of the data includes the method used to deliver the products, how long deliveries take, and which factory sent the goods.
* Altogether, this dataset gives a detailed view of how products are made, sold, and delivered – which is essential for finding ways to improve sales, reduce costs, and better meet customer needs.

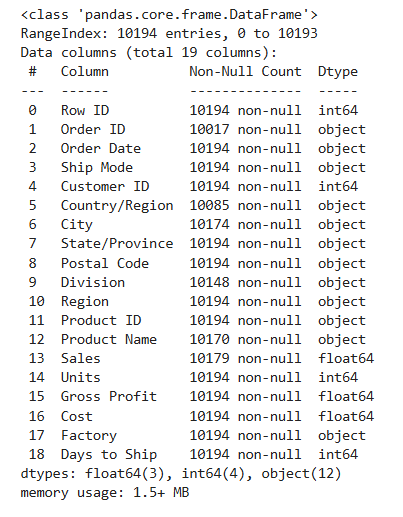


Figure - Candy Dataframe .info

**2. Data Structure**

The main dataset, called Candy\_Sales.csv, is made up of several key parts that help us look at the data from different angles. This structure allows us to explore:

* Time-based analysis: It looks at how sales change over time by checking monthly and yearly trends. This helps spot busy periods and quiet seasons.
* Geographical analysis: It examines sales in different regions and the locations of factories to show which areas perform well and where improvements are needed.
* Product analysis: It identifies which products sell the most, which sell the least, and how much profit each one makes.
* Operational analysis: It uses the data to check how efficiently goods are delivered and the costs of logistics, helping to find savings or problems.

**3. Initial Data Quality Assessment**

Before preprocessing, the following checks were performed:

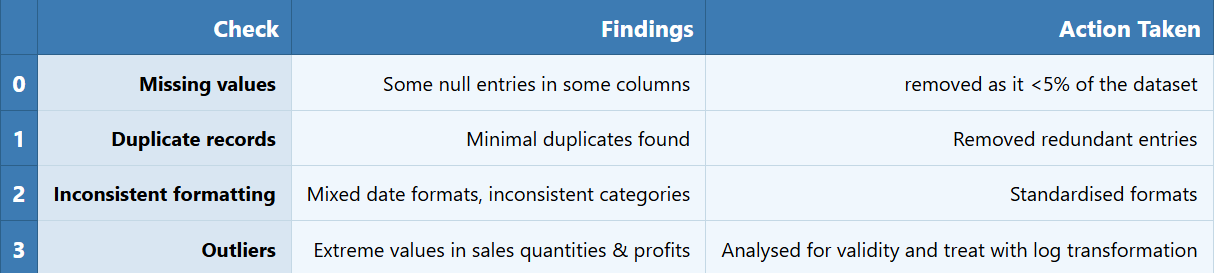
****

Figure - Dataset Assessment before preprocessing.

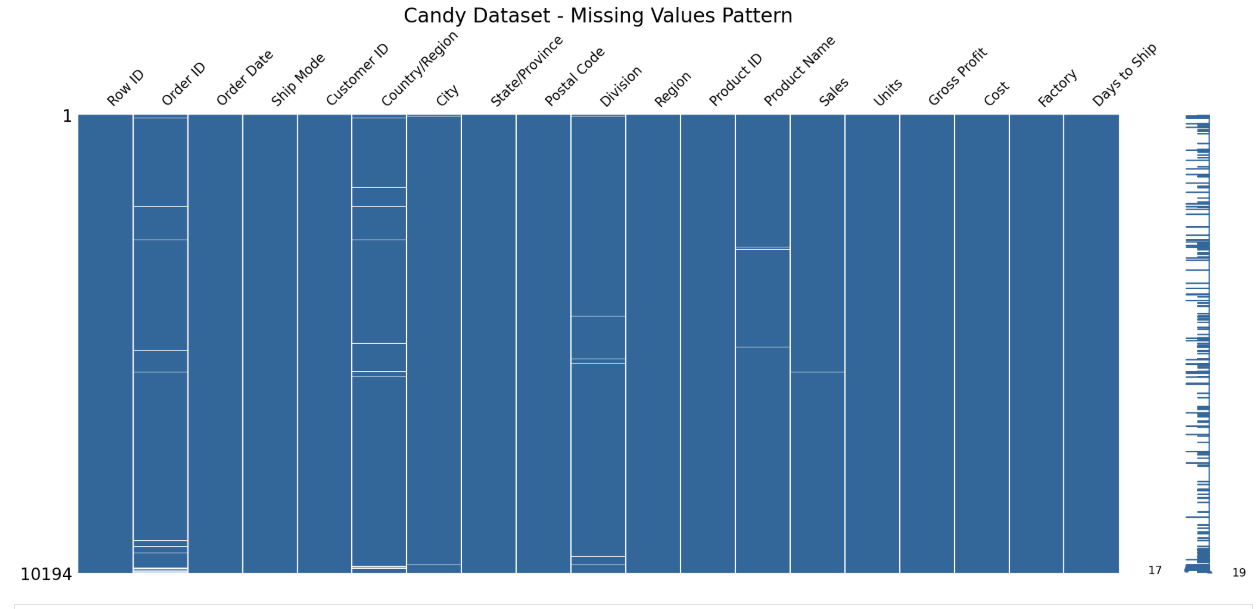
****

Figure - Missing values Pattern.

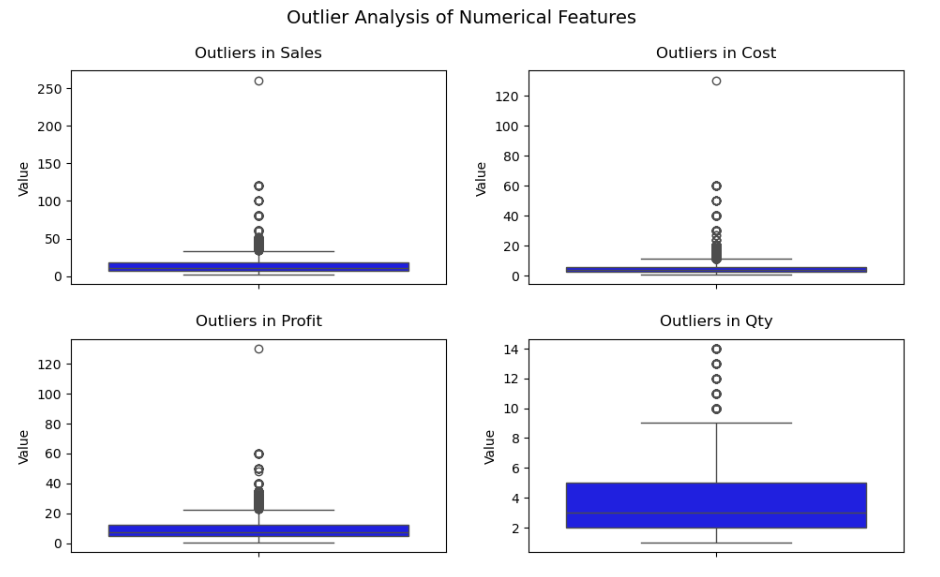
****

Figure – Numerical features outliers, before applying Log transformation

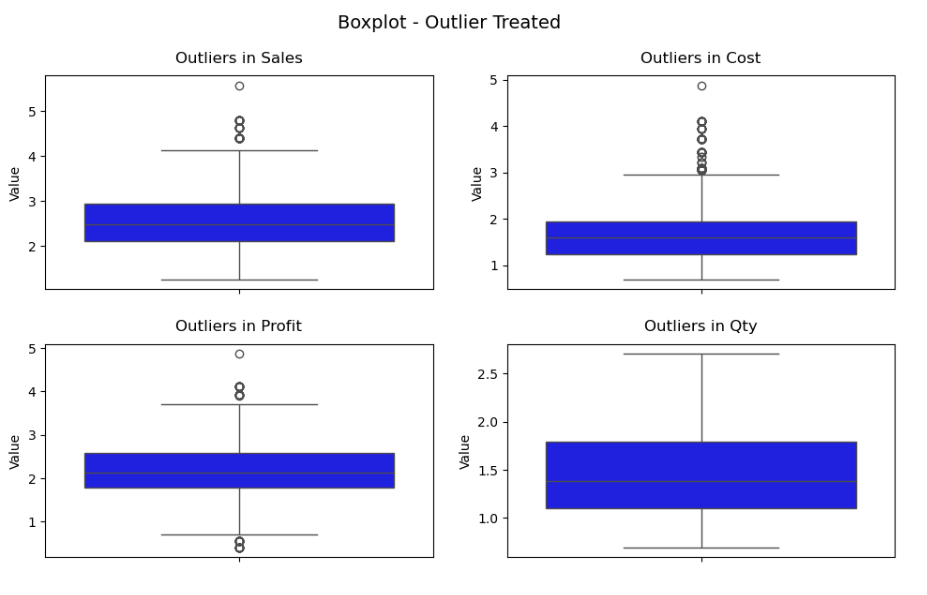
****

Figure - Outliers treated after applying log transformation

**4. Key Variables & Relationships**

**Target Variables:**

* Sales Amount: Used as the primary target for machine learning-based sales prediction.
* Quantity (Qty): Used for time-series demand forecasting models.

**Predictor Variables:**

* Product category
* Geographical location (e.g., region, city)
* Time-based features (month, day of week, seasonality indicators)
* Shipping method (e.g., express, standard)
* Factory (production source)

**5. Initial Insights from EDA**

**Initial Sales Insights**

The average sale is $14.44, but values vary widely ($2.50 to $260), with most transactions around $10.80. Similarly, costs average $4.95 (ranging from $1 to $130), while profits average $9.46 (from $0.50 to $130). The higher averages compared to medians suggest a few large transactions skew the data.

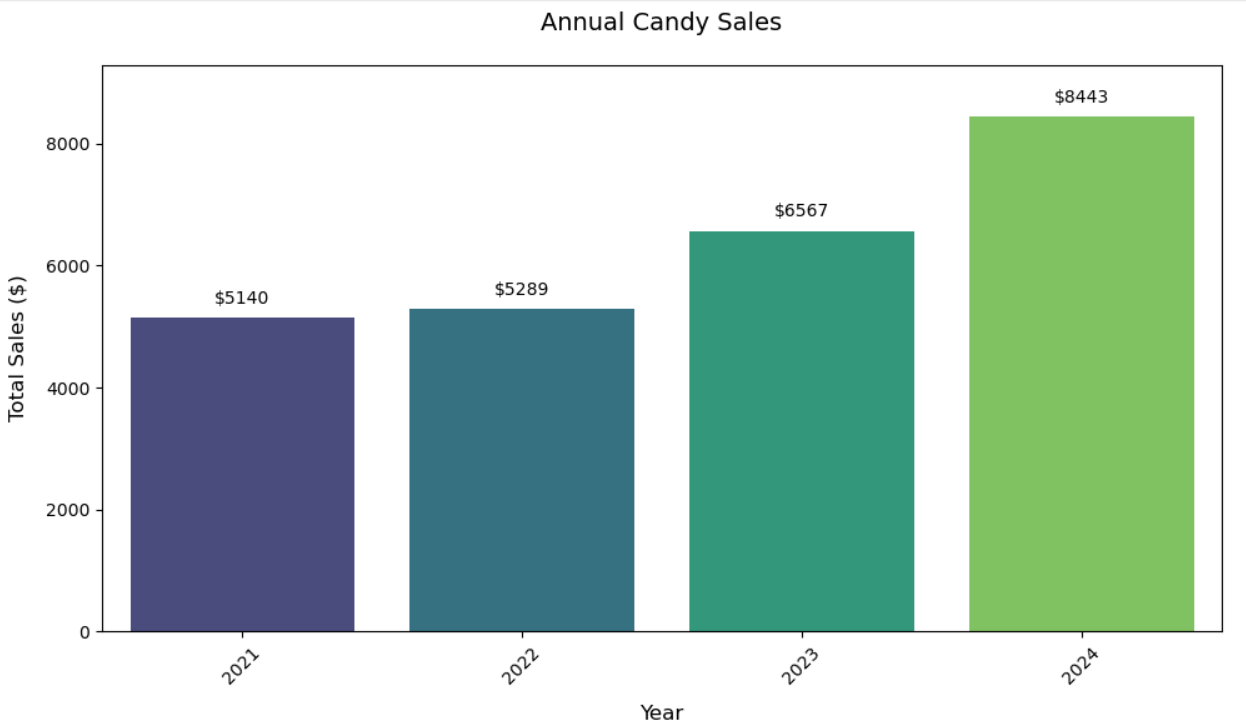


Figure - Annual Candy Sales 2021-2024

**Quantity and Trends**

Most orders are small, averaging 3.87 units per sale (median: 3), with bulk purchases up to 14 units. The wide ranges in sales, costs, and profits indicate a mix of everyday and high-value orders, likely from different customer types or promotions.

**Customer and Order Patterns**

The data includes 8,318 orders, mostly from the US (9,695 orders), particularly California. "Standard Class" shipping is the most popular (5,943 orders), and chocolate products dominate (9,614 orders), especially the "Wonka Bar Scrumdiddlyumptious" (2,049 sales). The top customer placed 19 orders.

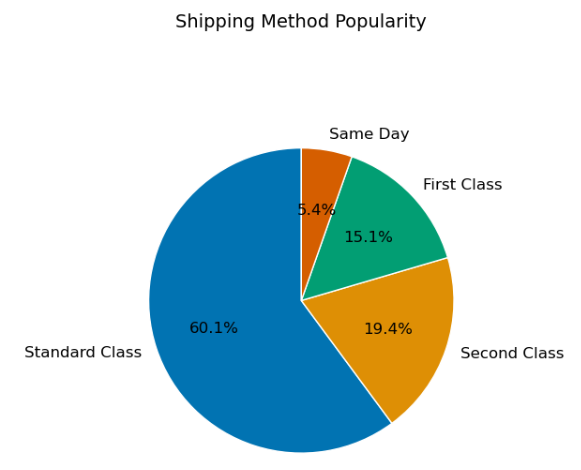
****

Figure - Pie chart - Shipping Method Popularity

Standard Class is by far the most popular option, accounting for the majority of orders (5,943), followed by Second Class (1,921) and First Class (1,491), while Same Day delivery is the least utilised (531 orders). Customers prioritise cost-effectiveness over speed, with expedited shipping methods (First Class and Same Day) representing less than 25% of total orders**.**

**Key Takeaways**

The business relies heavily on US sales, with the Pacific region and California as top markets. Chocolate products, mainly made by "Lots ONuts," drive most revenue. The skewed profit data suggests occasional large orders significantly impact averages.

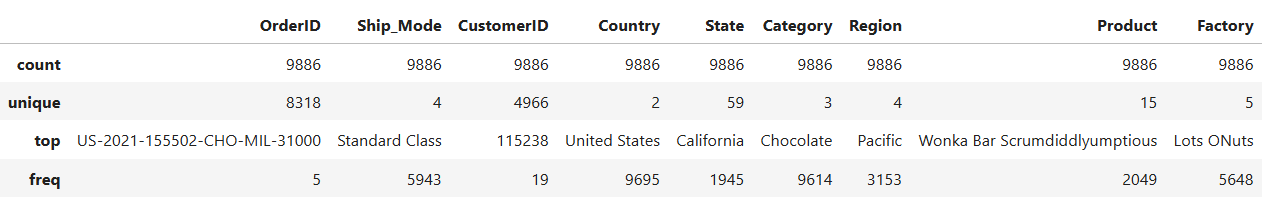


Figure - Statistical Analysis Of Categorical Features

# Technology/Methodology

## Libraries

The project used Python libraries for data processing, visualisation, and machine learning. Core libraries included NumPy and Pandas for data manipulation, Matplotlib and Seaborn for visualisations, and Scikit-learn for machine learning. For time series forecasting, Statsmodels was used to implement ARIMA and SARIMA models. Additional tools like Missingno helped identify missing data patterns, while SciPy supported statistical hypothesis testing.

## Machine Learning

Five regression models were tested to predict sales and profitability. Linear and Ridge Regression served as baseline models, while Decision Trees, Random Forests and Gradient Boosting (GBR) captured non-linear relationships. Tree-based models outperformed linear models due to their ability to handle complex interactions between features, such as seasonal trends and bulk orders. GBR showed strong generalisation, making it a key candidate for optimisation.

## Time Series

Time series forecasting was used to predict candy demand for the next 12 months to improve stock planning and reduce waste. Unlike simple methods that treat data points separately, this approach looks at how demand changes over time, spotting important patterns and trends.

The SARIMA model was chosen because it handles both long-term trends and seasonal changes well. Candy sales often rise during holidays like Christmas or Easter, and SARIMA can detect these regular patterns to make better predictions.

SARIMA breaks demand into three parts:

* Trend: The general increase or decrease over time.
* Seasonality: Regular ups and downs at certain times of the year.
* Random noise: Unpredictable changes without a clear pattern.

By separating these parts, the model creates clearer, more accurate forecasts, revealing patterns in the sales data that simpler methods might miss

## Hypothesis Test

To answer key business questions, two statistical tests were carried out to find useful patterns in the data. This helped the company make smarter decisions by showing if certain factors truly affect profits, rather than results happening by chance.

The first test used ANOVA to see if different shipping methods change profit margins. If differences were clear, the company might need to rethink its delivery options. For example, faster shipping may cost more but could increase profits by keeping customers happy or speeding up sales.

The second test used a t-test to compare profits from cheaper and more expensive products. This showed whether the pricing strategy was working well. If costly items earned less profit, prices might need adjusting to boost overall returns.

## Hyperparameter Tuning and Cross Validation

To improve model performance, it uses GridSearchCV to find the best settings, called hyperparameters. This means testing different options, such as tree depth, learning rate, and number of trees. By trying these combinations, it aims to give more accurate and reliable results. However, some models only improve slightly, which shows that improving the input data—known as feature engineering—is often more important than just changing settings.

For the SARIMA model, designed for time-series data like candy demand, it also uses grid search to find the best forecasting settings. SARIMA works well because it considers trends, seasonal patterns, and past values. Testing many settings helps it choose the most accurate model.

In summary, adjusting settings helps, but it is not enough by itself. Better results come from combining tuning with good data preparation and choosing the right features. This shows that careful data work is key to strong predictions.

# Results and Discussion

## Business Questions

The analysis centred on the main concerns and questions raised by the candy distributor. The goal was to identify key findings that would help tackle their biggest challenges and offer valuable insights.

**How do sales and profits compare between different countries, and what patterns can be seemed in their relationship?**

The results show that the United States performs significantly better than other countries. It achieves both higher sales and higher profits. This indicates that customers in the US tend to buy candy more regularly, providing a steady flow of income. In contrast, customers in Canada purchase less often but tend to buy larger amounts each time they do. This difference in buying habits is important for understanding how to approach each market effectively.

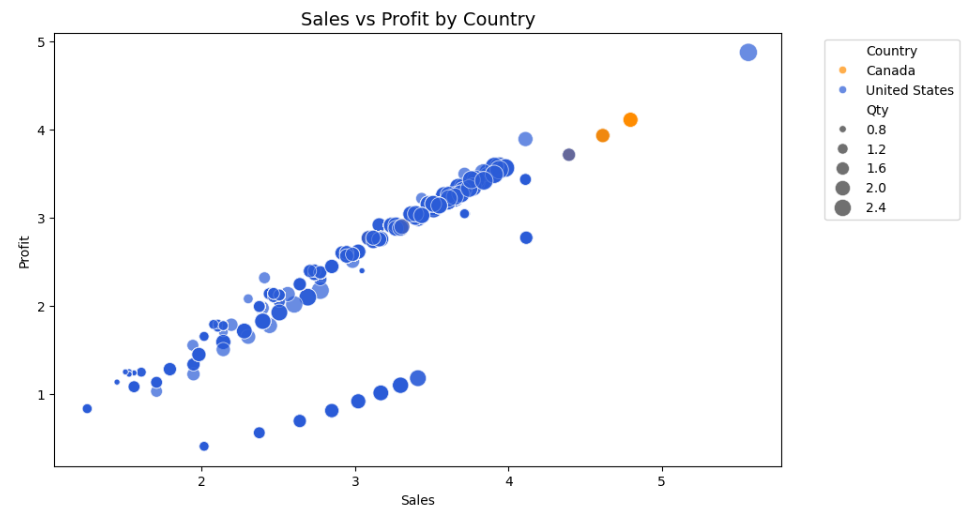
****

Figure - Comparing Sales and Profits Between Different Countries

*The North American candy market demonstrates robust growth driven by strong consumer preferences for confectionery products across the United States, Canada, and Mexico. The region benefits from well-established distribution networks, innovative product launches, and a strong presence of major manufacturers.   
(Intelligence, 2025)*

**How do sales volume and profitability margins vary across states, and are there significant regional differences?**

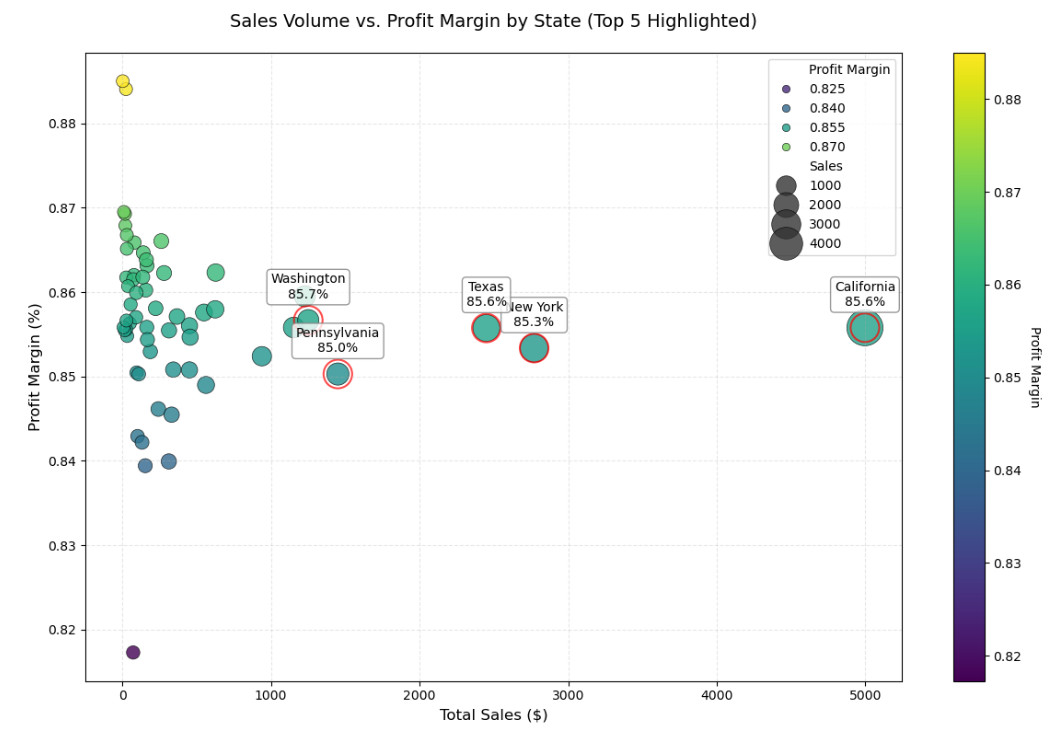
****

Figure - Sales volume and profitability margins variation across states

The chart compares sales and profits in several US states, showing some clear trends. California has the highest sales, nearly 5,000, with profits over 4,000. But when it comes to profit from each sale, Washington leads, keeping 85.7% of its sales as profit, the highest margin of all.

New York and Texas have profit margins very close to California’s, about 85.5%, showing these big states manage their costs and pricing well. Pennsylvania’s margin is slightly lower at 85%, suggesting it could improve cost control.

All five states keep more than 85% of their sales as profit, which means they are good at turning sales into profit, despite differences in total sales. Washington stands out for having the highest profit margin with smaller sales, hinting at effective cost management that others might learn from.

In short, while higher sales increase total profit, how well a state manages costs is just as important for making money.

**Which are the 5 lowest-selling products?**

Fun Dip and Nerds are selling much worse compared to the other products. For instance, Laffy Taffy sells for $14, which is three and a half times higher than Fun Dip and Nerds. This large gap in sales shows that the company should consider taking action to boost these products. Possible steps could include running special promotions or discounts, improving the quality or packaging of the products, or lowering their prices to attract more customers. Addressing these issues could help increase their popularity and sales.

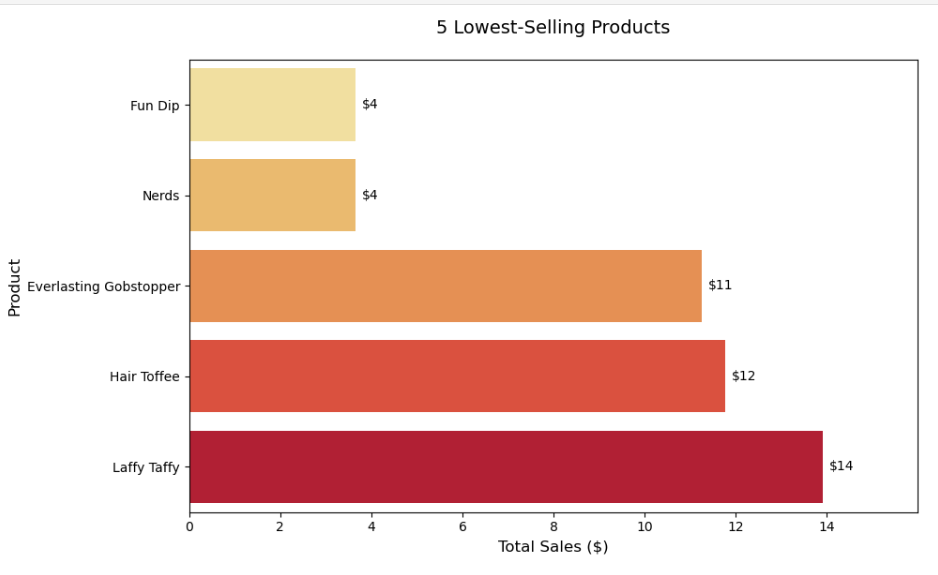
****

Figure - Lowest Selling Products

**Which factory generates the highest total sales, and are there statistically significant differences in performance across factories?**

The Wicked-Chacys factory clearly leads in sales, earning between $120,000 and $140,000 each year for the last four years. This is much higher than the Secret Factory and The Other Factory. The big difference suggests Wicked-Chacys is doing something right, perhaps making more products, offering better quality, or being in a busier location. The other factories could learn from Wicked-Chacys’ methods to improve their own sales. The consistent gap shows Wicked-Chacys’ success is real, not just luck.

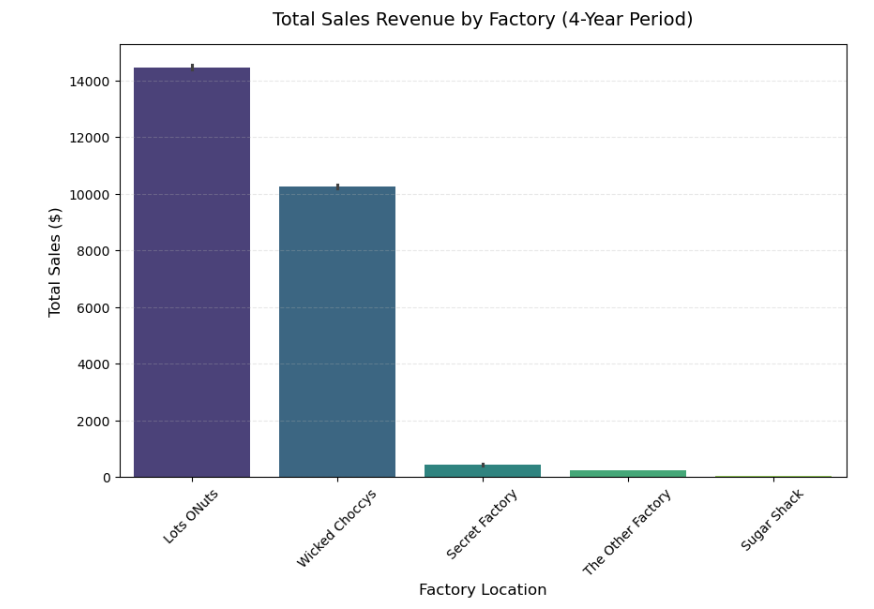


Figure - Revenue by Factories in the last four years

**How do factory performance metrics vary across different Factories and Countries?**

The chart shows clear sales differences between factories and countries. In the US, Wicked Choccys and Lots ONuts sell much more than others. In Canada, sales are low everywhere, even for the top seller. Also, Sugar Shack and The Other Factory do not sell in Canada at all. This suggests the US market is bigger and more active, or these factories may need to boost their marketing and sales in Canada.

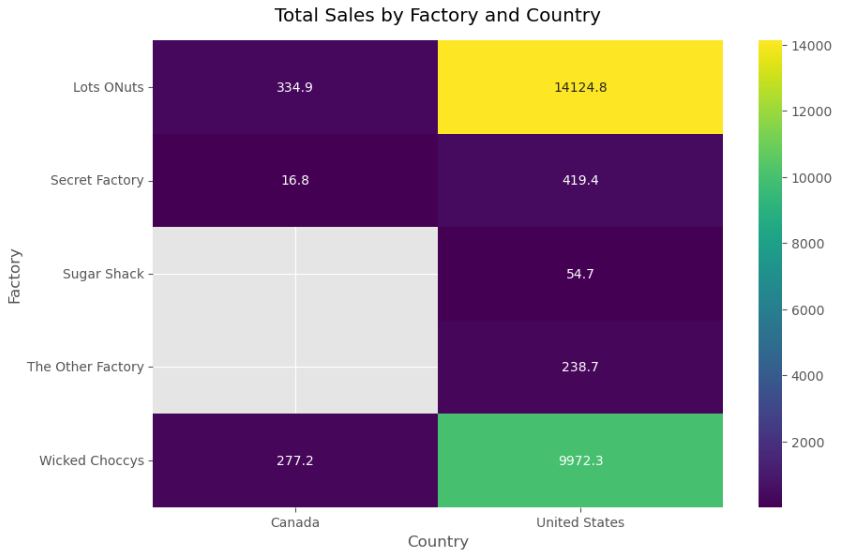


Figure - Factory Performance Across Countries and Factories

## Hypothesis Testes

**Hypothesis 1: Does the Choice of Shipping Method Affect Profitability?**

The main aim here was to find out if using faster delivery options helps the company earn more money. Understanding this would guide decisions on whether it is worth investing in quicker shipping or adjusting prices depending on delivery speed. The findings could reveal the best way to manage deliveries and set prices that benefit the business.

**What Was Tested**:

* The shipping method has no effect on profit (this is called the null hypothesis).
* Faster shipping leads to higher profits (this is the alternative hypothesis).

**What Was Found:**

The analysis showed no strong evidence to reject the first idea (the null hypothesis). In other words, there was no clear difference in profit margins between faster and slower shipping methods. This suggests that the current pricing likely already covers the delivery costs well. Even though faster shipping costs more to provide, customers appear willing to pay extra for it, so profits remain similar regardless of the shipping speed. On the other hand, slower shipping does not seem to reduce profits as much as might have been expected.

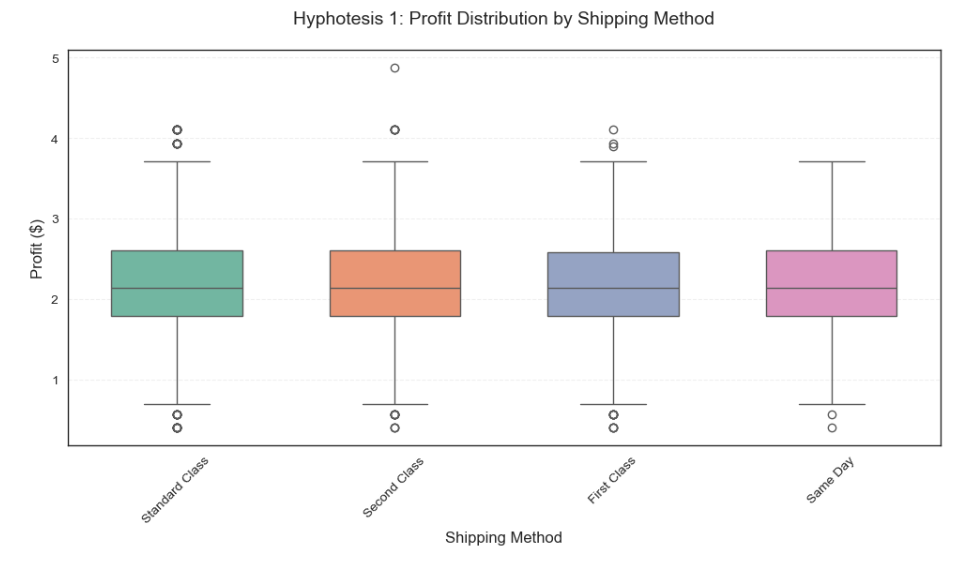


Figure - Hypothesis Test 1. Profit Distribution by Shipping Method

**Hypothesis 2: Does the Candy Distributor Need to Change Prices or Operations Based on Product Costs?**

The company wanted to see if products that cost more to make or buy have different profit margins compared to cheaper products. This would help decide if they need to change prices or how they manage different products.

**What Was Tested:**

* The assumption that there is no difference in profit margins between low-cost and high-cost products (null hypothesis).
* The possibility that profit margins differ between these groups (alternative hypothesis).

**What Was Found:**

The data showed a clear difference in profit margins between cheaper and more expensive products, so the idea that they were the same is rejected. Specifically, cheaper products earned slightly higher profits, around 86%, compared to 84% for the expensive ones.

Although this 2% difference might seem small, it could represent a significant amount of money when considering all sales combined. The lower profit margin on expensive products is likely because they cost more to make or purchase. On the other hand, cheaper products may sell in larger volumes or have more flexible pricing, which helps increase their overall profitability.

**Conclusion:**

These findings suggest that cheaper products currently perform better in terms of profit percentage. This insight could be important for Candy Distributor when deciding whether to adjust pricing or operations depending on the cost tier of their products.

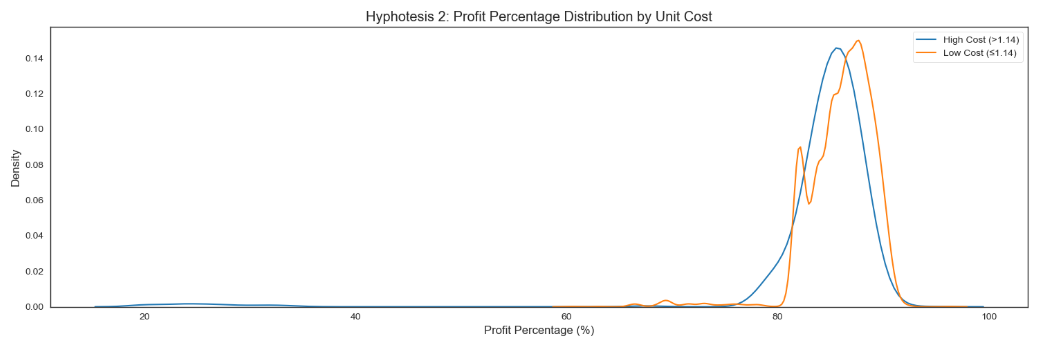


Figure - Hypothesis Test 2. Profit by Unit Cost

## ML Regressor Models

To identify the most suitable model for the task, five different machine learning models were tested. Each model has its own strengths and weaknesses, making them better suited to different types of data and problems.

**Linear Regression**

This is a simple and straightforward model that assumes a direct, straight-line relationship between input and output. It serves as a basic starting point. Since our data was non-linear, this model was expected to perform less well. No adjustments or fine-tuning of settings (called hyperparameters) were needed for this model.

**Ridge Regression**

This model is an improved version of linear regression. It adds a penalty to the size of the model’s coefficients to prevent it from fitting the training data too closely (a problem known as overfitting). However, it still relies on the assumption that the relationship between variables is linear, which can limit its effectiveness with more complex data.

**Decision Tree Regressor**

This model works by splitting the data into smaller groups based on certain rules, allowing it to capture more complex, non-linear relationships. However, without careful limits, it can become too closely tailored to the training data, making it less reliable on new data.

**Random Forest Regressor**

This approach combines many decision trees to create a more stable and accurate prediction. By averaging the results of multiple trees, it reduces the chance of mistakes caused by noise in the data. As a result, it usually performs better than a single decision tree.

**Gradient Boosting Regressor**

This model builds trees one after another, with each new tree trying to fix the errors made by the previous ones. This step-by-step correction process helps it achieve very strong performance, especially when the data contains important interactions between features.

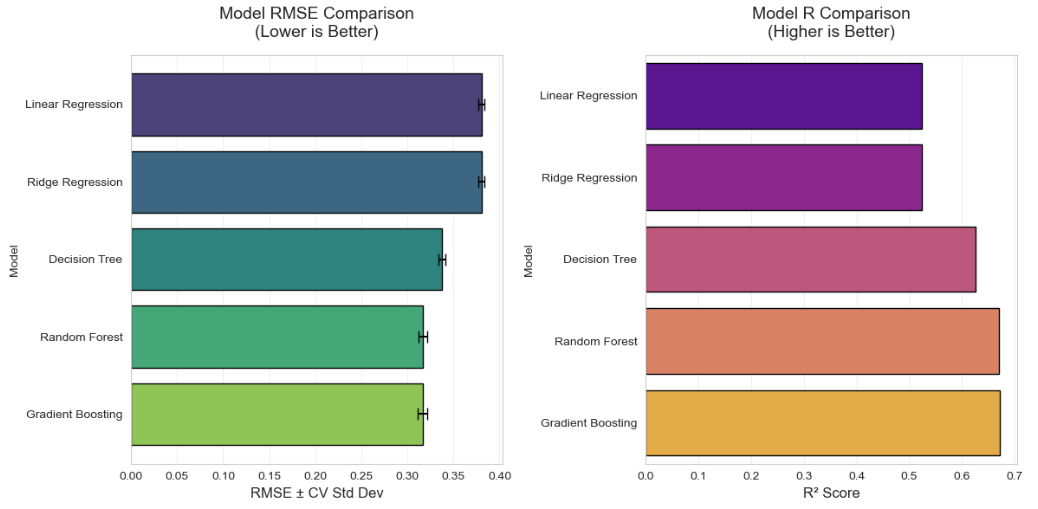


Figure - ML Models Performance

**Model Performance Analysis**

This section looks at how well different models explain and predict the data, using simple measures:

* R (shows how much of the data variation the model explains)
* RMSE (shows how far off predictions are, on average).

**Linear Regression**

Explains about 44% of the data’s variation and scores 46.5% on new data, meaning it generalises fairly well. The error (RMSE 0.38) is low and consistent, but the model’s simple straight-line approach may miss complex relationships in the data.

**Ridge Regression**

An improved version of linear regression that adds a penalty to avoid overfitting when predictors are linked. It explains about 52% of the variation on both training and test data, showing better generalisation. The error (RMSE 0.38) is stable, thanks to handling related variables better.

**Decision Tree**

Fits training data well (69% explained) with low error (RMSE 0.30), but performance drops on new data (63% explained, RMSE 0.34). This suggests overfitting, the tree learns training details too closely and struggles to predict new data accurately. Some tuning could reduce this issue.

**Random Forest**

Combines many decision trees to reduce overfitting. It performs well on both training (70% explained) and test data, with stable error (RMSE 0.32). This shows good reliability and generalisation. Further tuning could improve it more.

**Gradient Boosting**

The best performer, with strong, consistent results on training and test data (around 67% explained, RMSE 0.32). It balances prediction accuracy without overfitting, making it the most reliable model here.

Summary

Linear and ridge regression are simple and work reasonably but have limits. Decision trees can overfit. Random forests and gradient boosting perform best, with gradient boosting slightly better for practical use due to its strong and stable predictions.

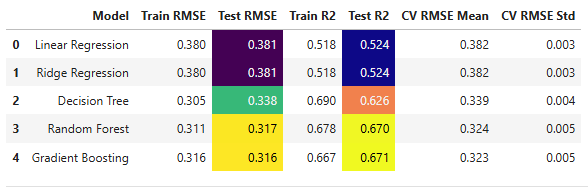


Figure - ML Models Performance dataframe

## Hyperparameter Tuning and Cross Validation for tree-based models

The tree-based models – Decision Tree, Random Forest, and Gradient Boosting, show better results than the linear models. Because of this, it focuses on these tree-based models and uses Hyperparameter Tuning to improve them further. Hyperparameter Tuning is a process where it adjusts key settings in the models to make them more accurate and perform better.

Below are the results after this optimisation process:

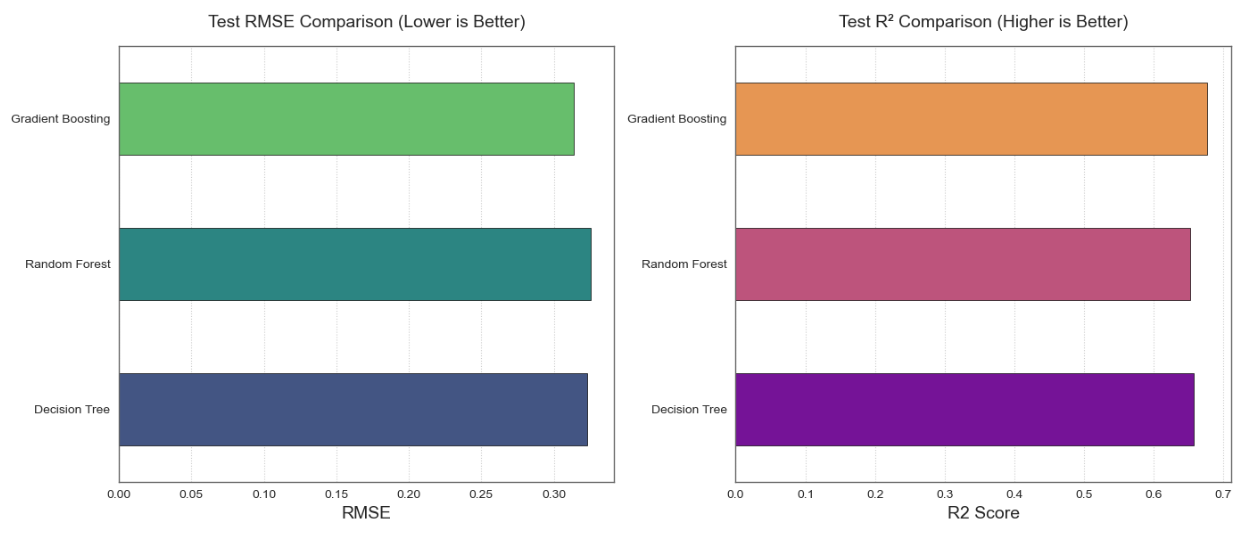


Figure \_ Models Performance After Applying Hyperparameter Tunning

**Model Performance Comparison After Applying Hyperparameter Tuning and Cross Validation**

The Gradient Boosting model performs well on both training and test data, showing a good balance. Its training and test errors (RMSE 0.30 and 0.31) are very close, meaning it learns the data without overfitting. It explains about 69% of the variation in the data (R² scores), which is strong. Cross-validation confirms its reliability with a consistent error (RMSE around 0.32). Overall, it is accurate and generalises well to new data.

The Random Forest model fits the training data very closely (low RMSE 0.26), but its test error is higher (RMSE 0.33), suggesting some overfitting. The drop in R² from 0.77 (training) to 0.65 (test) supports this. Cross-validation also shows less consistency than Gradient Boosting.

The Decision Tree model fits the training data less well (higher RMSE 0.32) but performs similarly on test data, showing it avoids overfitting. However, it explains less of the data variation (R² about 0.65) and is less accurate overall. Cross-validation shows more variation, meaning less stable performance.

In summary, Gradient Boosting is the most reliable and accurate model. Random Forest overfits slightly, while Decision Tree is simpler but less powerful. For best results, Gradient Boosting is the top choice.

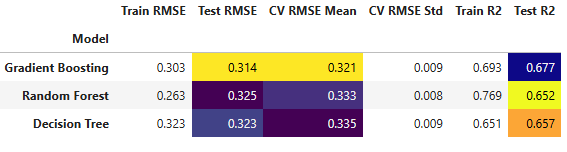


Figure - Models performance- Hyperparameter Tunning Dataframe

## Feature Importance

Gradient Boosting Regressor shows that the most important factor for predicting sales is how much each customer spends, called Sales\_Per\_Customer. This means that the model mainly looks at individual customer spending to make its forecasts. The second key factor is whether an order is a bulk purchase or not, which also affects the predictions but less than customer spending. Other features, like price per unit, have only a small impact. Overall, the model focuses mostly on customer buying habits, especially higher spending and bulk orders, to predict sales.

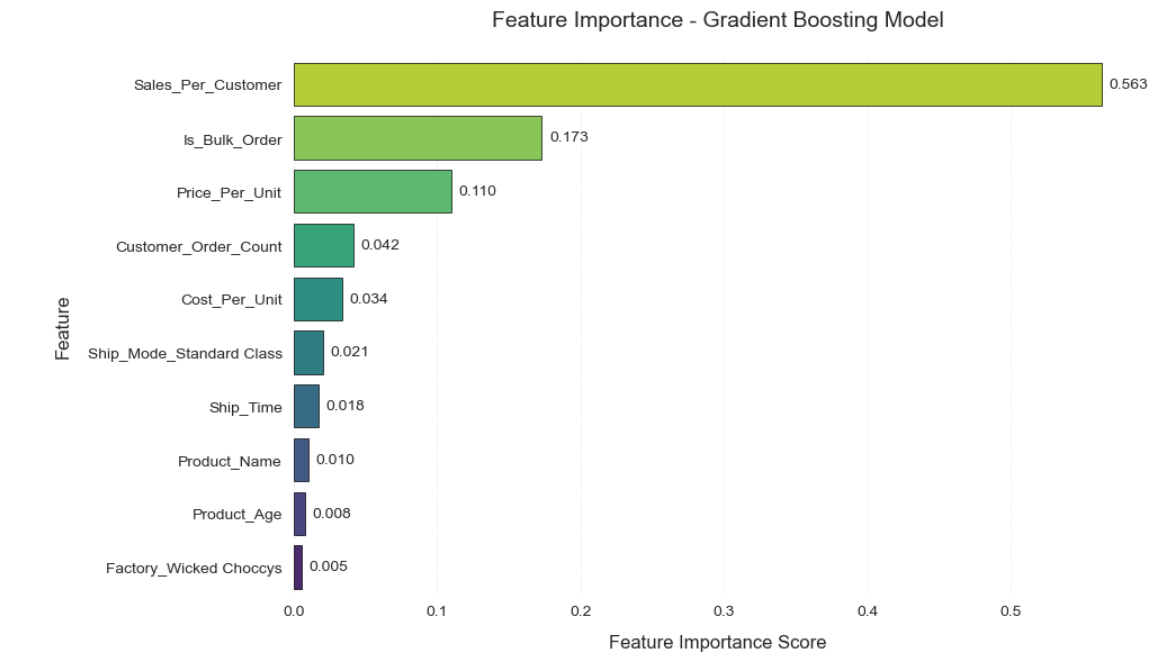


Figure - Feature Importance Using Gradient Boosting

# Time Series Analysis

The aim was to predict future order quantities using the ARIMA model. The analysis showed clear seasonal trends: demand was highest in September, November, and December, meaning the end of the year is much busier. Conversely, January and February saw the lowest demand, indicating quieter months at the start of the year. This yearly pattern, with orders rising towards the end of the year and falling in early winter, helps businesses plan better for changes in demand throughout the year.

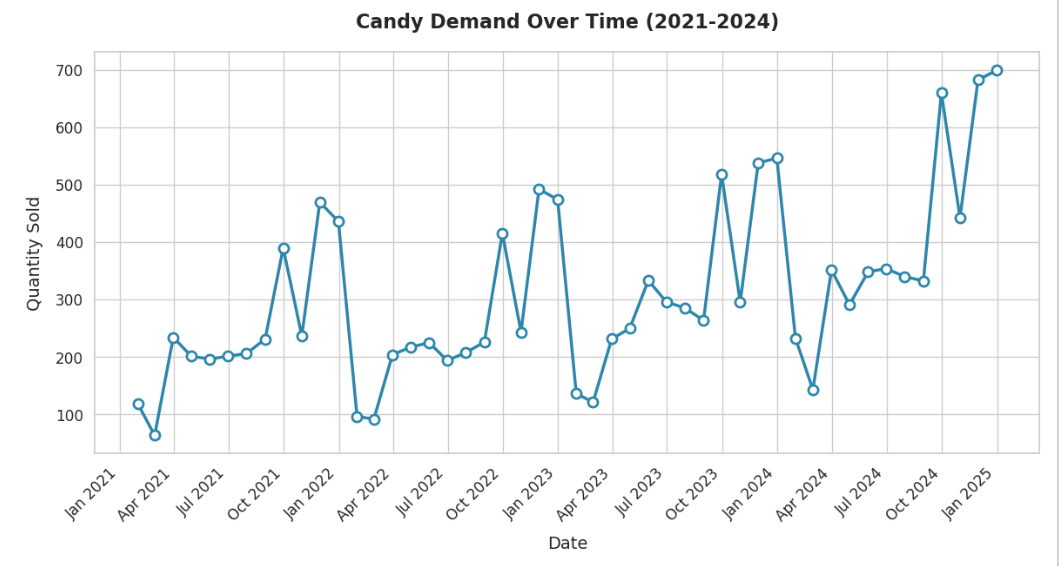


Figure - Candy Demand trends over time

The data reveals a clear and noticeable rise in candy sales over the last three years. In 2022, sales increased by a steady 3.42%, showing gradual growth from the year before. However, in 2023, sales jumped sharply by 23.58%, indicating a much faster increase in demand. This sudden growth could be due to changing consumer tastes, effective advertising, economic recovery, or people spending more after the pandemic. Understanding these changes helps businesses plan their production, manage stock, and create better promotions. Paying attention to this trend is important for companies to stay competitive in the candy market.

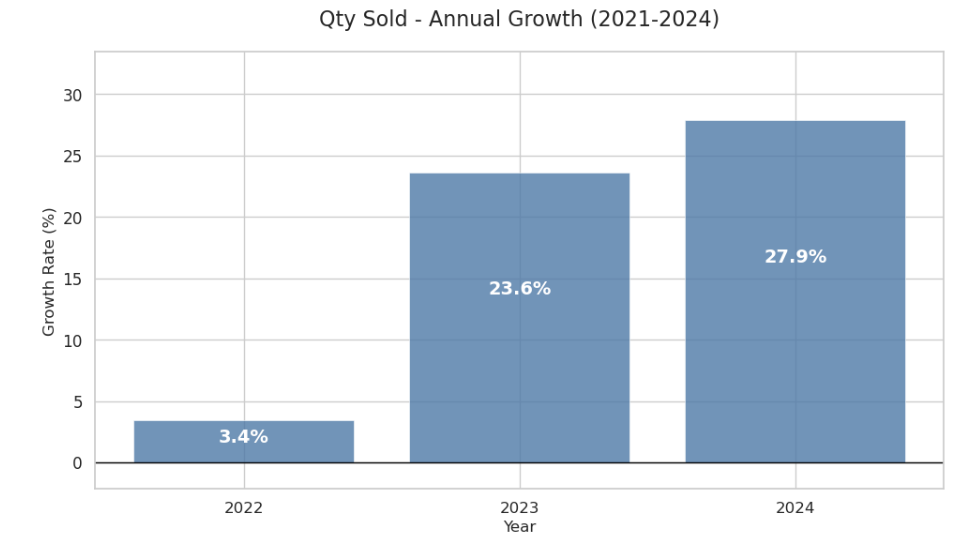


Figure - Candy Annual Growth 2021-2024

**Fitting seasonal SARIMA Model Using First Parameters Found:**

After breaking down the time series, it wastested if the data was stationary, meaning it did notshow trends or seasonal patterns, using the Augmented Dickey-Fuller (ADF) test. If the test’s p-value is below 0.05, it shows the data is stationary, the idea that it changes over time can be rejected. Stationarity is important because forecasting models like ARIMA and SARIMA work best when the data is stable. Since the data had clear seasonal patterns, it wasused seasonal differencing (subtracting values from 12 months earlier) to remove these repeating effects. This step helps make the data steady and ready for accurate forecasting.

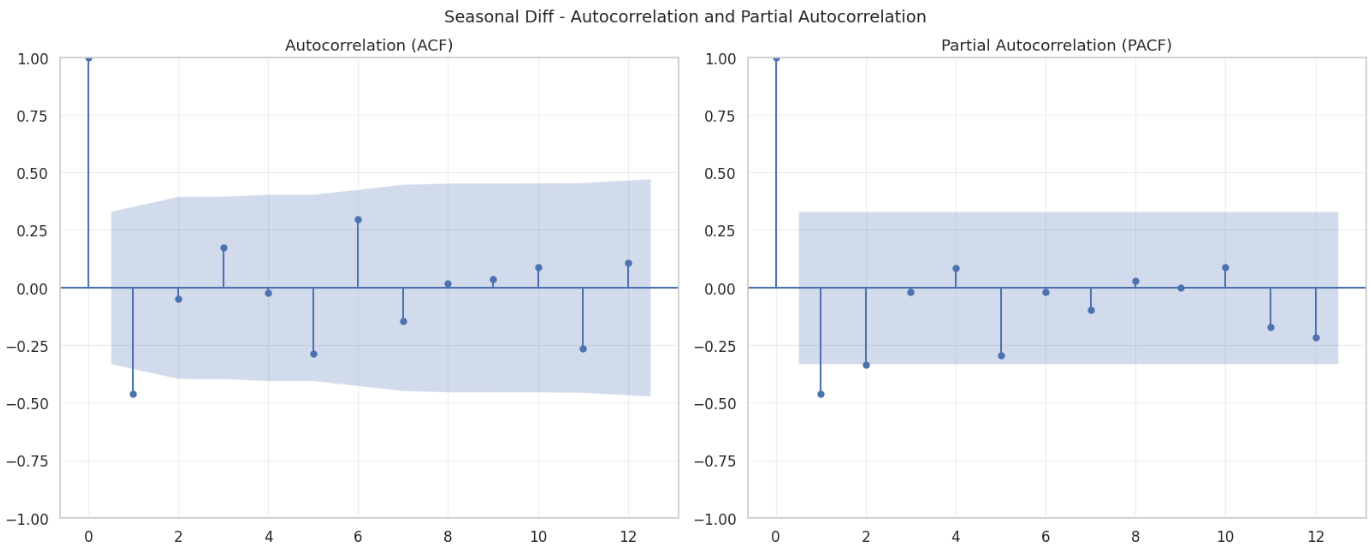


Figure - Seasonal Differentiation - ACF and PACF

**Interpreting SARIMA Parameters:**

The initial SARIMA parameters were selected based on patterns observed in the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots. The parameters indicate:

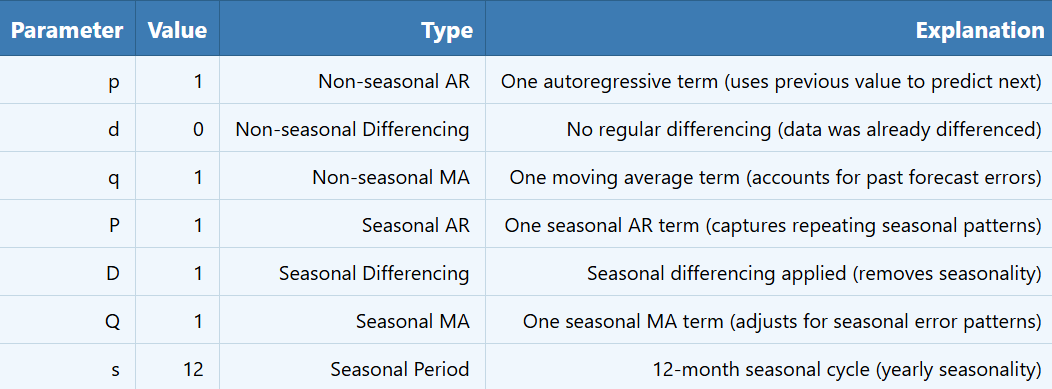


Figure - Interpreting SARIMA Parameters

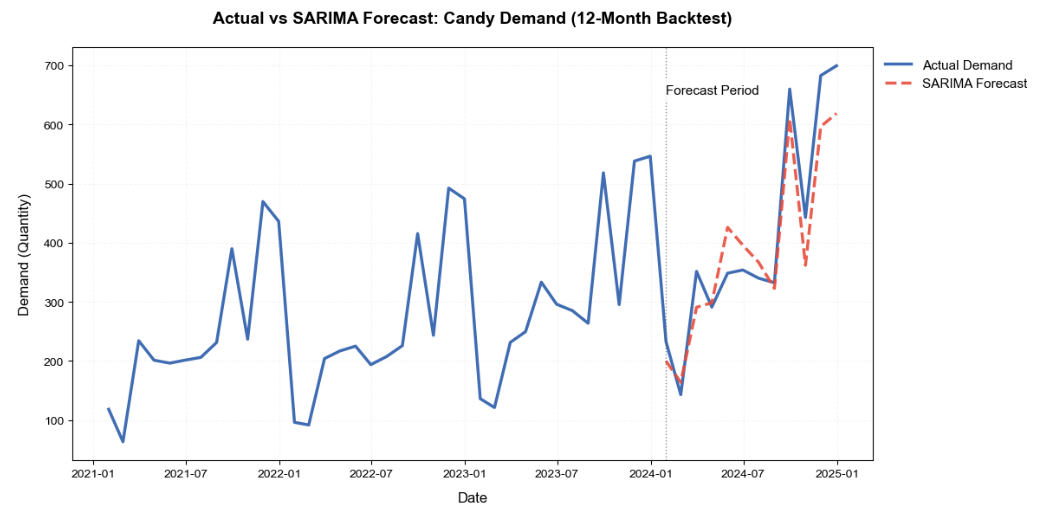
****

Figure - Demand Forecasting - SARIMA Model

The chart above compares the real demand for candy with predictions made by a SARIMA model over a 12-month back testing period, from January 2021 to January 2025. The demand is shown in quantities, recorded every six months (in January and July). This comparison helps us see how accurately the model can predict actual demand.

The red line shows the forecast, while the blue line shows the real demand. The two lines generally follow a similar pattern, which means the model does a good job overall. However, there are some differences where the model either overestimates or underestimates demand. The forecast sometimes misses sudden increases or drops, showing that the model struggles to predict unexpected changes or outside influences.

**Fitting the best parameters using Grid Search:**

To make the SARIMA model work better, it uses a method called grid search to fine-tune its settings. This means it tests many different combinations of values to find the ones that give the most accurate forecasts. By checking each option step by step, it finds the best settings to understand data patterns more clearly and improve its predictions.

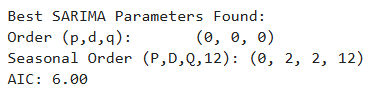


Figure - The best parameters found using Grid Search

**Forecasting using the best parameters found:**

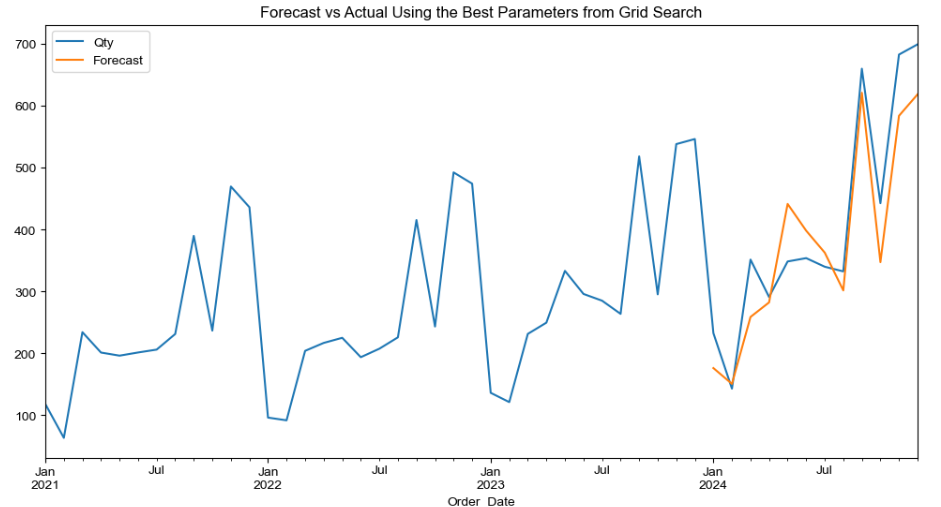
****

Figure - Demand forecasting after applying Grid Search

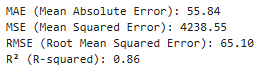
****

Figure - SARIMA Performance

The model shows a strong R value of 86%, meaning it explains 86% of the changes in demand based on past data. This tells us the model fits the historical patterns well and is useful for making forecasts. It captures most of the important trends with only small errors. However, the model may find it hard to predict sudden changes or unexpected events because it smooths out sharp ups and downs. So, while it is good at showing overall trends, some variations remain unexplained, which could affect accuracy in unpredictable situations. Additionally, trying to improve the model with a grid search did not lead to better results.

**Candy Demand Future Forecasting:**

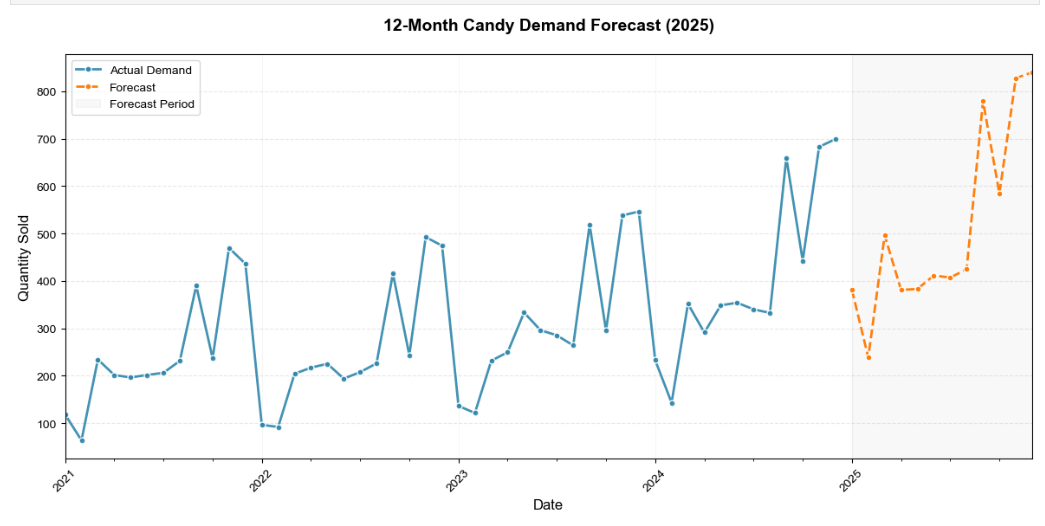
****

Figure - Final Future Demand Forecast

The SARIMA model predicts monthly candy demand for 2025, showing key changes throughout the year. These forecasts help the Candy Distributor plan stock and cash flow more effectively. Demand peaks in September, November, and December due to events like Halloween and Christmas, allowing early stock build-up and promotions. February sees a drop in demand after the holidays, preventing excess unsold stock. Summer months have steady demand, supporting smooth supply. With this insight, the distributor can reduce waste, improve delivery schedules, and boost profits in busy months while lowering costs in quieter periods.

*According to the National Confectioners Association (NCA) report, in 2023, 60% of total candy sales were made during four major seasons, including Valentine’s Day, Winter holidays, Easter, and Halloween* (Candy Market, 2025)*.*

# Conclusion

This Capstone project set out to support a US-based candy distributor in improving their sales performance, managing stock more effectively, and increasing overall profitability by using data analysis and machine learning techniques. Through careful examination of sales patterns, testing important business ideas, and predicting future demand, the project produced useful insights that can help guide better decision-making.

The analysis uncovered clear seasonal trends, with sales reaching their highest points in September, November, and December, coinciding with popular holidays such as Halloween and Christmas. Among the machine learning methods tested, the Gradient Boosting model showed the best results in predicting sales figures. For forecasting future demand, the SARIMA (Seasonal ARIMA) model was particularly successful, as it could accurately capture both general trends and seasonal changes over time.

The key findings also revealed that lower-priced products tend to generate slightly higher profit margins compared to more expensive ones. Interestingly, the method of shipping did not have a significant effect on profitability. Furthermore, looking at sales by region, states like California and Washington stood out as top performers, suggesting valuable opportunities for focused marketing and smarter stock management in these areas.

These insights give the distributor the ability to better predict customer demand, adjust inventory levels accordingly, and fine-tune promotional activities. This helps to reduce waste, avoid overstocking, and ultimately increase profits. By adopting a data-driven approach, the business is better equipped to compete in a changing market while improving operational efficiency and customer satisfaction.

Looking ahead, future work could include developing real-time demand tracking systems and exploring more detailed customer segmentation. These steps would help the business make even more precise predictions and tailor strategies to meet customer needs more effectively.

# References

Irwin, L. (2022). *The GDPR: What exactly is personal data? - IT Governance Blog En*. [online] IT Governance Blog En. Available at: https://www.itgovernance.eu/blog/en/the-gdpr-what-exactly-is-personal-data.‌

Statista. (n.d.). *Confectionery - United States | Statista Market Forecast*. [online] Available at: https://www.statista.com/outlook/cmo/food/confectionery-snacks/confectionery/united-states.

‌Statista. (n.d.). *Confectionery - United States | Statista Market Forecast*. [online] Available at: https://www.statista.com/outlook/cmo/food/confectionery-snacks/confectionery/united-states.

‌Gourmet Pro (2025). *A Deep Dive into the Confectionery Market in the US [2025]*. [online] Gourmetpro.co. Available at: https://www.gourmetpro.co/blog/confectionery-market-in-the-us.

‌Pvt, I. (2025). *Global Confectionery Market to Hit Valuation of US$ 453.64 Billion By 2033 | Astute Analytica*. [online] GlobeNewswire News Room. Available at: https://www.globenewswire.com/news-release/2025/01/15/3010291/0/en/Global-Confectionery-Market-to-Hit-Valuation-of-US-453-64-Billion-By-2033-Astute-Analytica.html.

‌www.futuremarketinsights.com. (n.d.). *Sugar-Free Sweets Market*. [online] Available at: https://www.futuremarketinsights.com/reports/sugarfree-sweets-market.

‌Schwarz, L. (2024). *Inventory Control: Best Practices and Everything You Need | NetSuite*. [online] NetSuite.com. Available at: https://www.netsuite.com/portal/resource/articles/inventory-management/what-are-inventory-management-controls.shtml.

‌30dayscoding.com. (2024). *30 Days Coding*. [online] Available at: https://30dayscoding.com/blog/distributing-candy-a-comprehensive-guide [Accessed 22 May 2025].

‌Fortunebusinessinsights.com. (2024). *Candy Market Size, Industry Share, Forecast, 2032*. [online] Available at: https://www.fortunebusinessinsights.com/candy-market-110265.

‌Haveignition.com. (2024). *Go-to-Market Strategy for candy*. [online] Available at: https://www.haveignition.com/industry-guides/go-to-market-strategy-for-candy.

‌ Mordorintelligence.com. (2019). *Mordor intelligence*. [online] Available at: https://www.mordorintelligence.com/industry-reports/candy-market.

‌Frank, A. and Scott (2015). *Capstone Title Marketing Research and Analytics Database*. [online] Available at: https://coeng.uaa.alaska.edu/innovation/wp-content/uploads/2020/10/2015-Marketing-Research-and-Analytics-Database-Final\_Report\_Complete\_V2.pdf [Accessed 22 May 2025].

‌