Data Science Capstone

Optimizing Coffee Shop Sales through Time-Based Analysis and Predictive Modeling

Samantha Morello, Ruhie Bakshi, Brianna Derra

Table of Contents

Abstract

Over the years, coffee, in its various forms, has become a staple in morning routines worldwide, and the industry continues to experience rapid change and innovation as increasing demand, advancements in technology, and cultural shifts all contribute to its economic growth and development. For producers in countries such as Brazil, Colombia, and Ethiopia, coffee bean exports heavily contribute to their gross domestic product, generating revenue and creating invaluable employment opportunities. To consumers, these changes mean exposure to more options within the market, and the shaping of their social interactions and lifestyle choices. As

for entrepreneurs, such an uptick in demand for coffee shop chains has diversified product options and created space for several companies, both large corporations and small, family-owned establishments, to grow and thrive amidst heavy competition from household names such as Starbucks. However, in order to achieve success in such a high-demand industry, it is essential to understand the behaviors and needs of the consumer in order to streamline several facets of operational logistics, maximize revenue, and predict market trends.

This analysis, “Optimizing Coffee Shop Sales through Time-Based Analysis and Predictive Modeling” aims to implement data science techniques in order to provide useful insights for a small chain of coffee shops in New York City. The analysis will utilize time-series analysis and multiple predictive models, as well as historical data, to explore the key aspects of customer behavior, product profitability, and sales forecasting, allowing for proactive inventory management, strategic promotional planning, and staffing adjustments. Ultimately, these data-driven insights empower the coffee shops to make informed decisions, improve operational efficiency, and enhance customer satisfaction.

Project Plan

66% of Americans say that they include coffee in their morning ritual every day (French, 2024). Taking into account consumers’ busy lifestyles and varying tastes, it’s no wonder that the coffee industry, which includes national and global brands such as Starbucks, Dunkin’ Donuts, and Dutch Bros., as well as local cafes and make-at-home brands such as Keurig, Folger’s, and Maxwell House, have become so successful. In fact, a report published by analysts with the S&P Global market projects that the coffee industry will soon exceed $200 billion in annual revenue (S&P Global).

Any successful business, regardless of industry, must have a way to not only analyze and visualize their data, but also a way to forecast trends and changes. This is a challenge that is solved with the use of data science tools.

For the purpose of this research project, we will focus on transactional data from a small chain of coffee shops based in New York City. This dataset includes information gathered from three distinct locations. It contains roughly 150,000 observations and offers a wide variety of variables including specific product details, transaction dates, unit prices and much more. We plan to thoroughly explore this data set with the goal of uncovering different patterns in profitability, predict and optimize future sales, and visualize customer purchases. Our goal with this project is to deliver data driven recommendations in order to enhance overall profitability and operational efficiency.

Research Questions

The importance of understanding customer behavior, profitability, and sales forecasting is crucial to driving business success. Therefore, the following questions will be researched in this project in order to optimize performance across the company:

Research Question 1: How does gross revenue vary over time and locations?

Analysis of profitability over time and at certain locations can contribute to many aspects of the company's success. Identifying key traits that affect profitability will in turn maximize revenue. Additionally, this helps the business to make informed decisions about inventory, pricing, and marketing strategies.

Research Question 2: How can predictive models be used to forecast future sales and schedule promotions effectively?

Generating accurate sales forecasts can allow the business to anticipate future needs more effectively. With this it will be possible to manage inventory more efficiently, avoid stock issues, and maximize sales. Having the ability to forecast future sales and schedule promotions effectively will help the business immensely.

Research Question 3: How can we visualize purchase patterns and identify trends in customer behavior across different times and locations?

The creation of an interactive calendar will aid users, such as managers and owners, to view important information with the click of a button. Clear and organized data is crucial to business success and performance. This tool helps the business easily identify sales trends over a large period of time which supports and improves decision making and planning.

Data Documentation/Highlights

Observations: 149,116

Variables:

- Transaction ID
- Transaction Date
- Transaction Time
- Store ID
- Store Location
- Product ID
- Transaction Quantity
- Product Category
- Product Type
- Product Detail
- Size
- Month Name
- Day Name
- Hour

- Unit Price
- Total Bill
- Month
- Day of the Week

Product Types:

- Brewed herbal tea
- Brewed Black tea
- Brewed Green tea
- Brewed Chai tea
- Drip coffee
- Organic brewed coffee
- Barista Espresso
- Gourmet brewed coffee
- Scone
- Pastry
- Premium brewed coffee
- Hot chocolate
- Biscotti
- Sugar free syrup
- Regular syrup
- Herbal tea
- Black tea
- Chai tea
- Green tea
- Drinking Chocolate
- Organic Chocolate
- Clothing
- Housewares
- Espresso Beans
- Green beans
- Organic Beans
- Premium Beans
- Gourmet Beans
- House blend Beans

Locations:

- Astoria
- Lower Manhattan
- Hell's Kitchen

Research Question Summaries

For Research Question 1, we are trying to determine how profitability varies across time and locations. For this we will take a look at different ways to use time and locations as predictors of sales. We will analyze the total sales, dates, and location data. Specifically we will focus on the following variables: Transaction Date, Day Name, Month, Total Bill and Store Location. We will use machine learning tools such as Random Forest to be able to make predictions based on the data. Additionally we will use matplotlib to create clear visualizations of the feature importance and accuracy. With this we aim to identify which features are most useful in helping predict profitability for the business.

For Research Question 2, we are attempting to determine which predictive models can effectively be used to forecast future sales and if they have the capacity to optimize scheduling promotions. We will utilize several variables to extract historical sales data and time-related purchases in order to accurately train predictive models. The variables we will focus on in the question are: Unit Price, Total Bill, Transaction Quantity, Product Category, and Product Type. To do this, we will choose the best predictive model or a combination of models including Random Forests, Decision Trees, Naive Bayes Classifier, etc. Our predictive models will help display sales trends which will ultimately allow us to predict future sales and identify optimal times for running promotions based on past customer behavior and demand patterns.

For Research Question 3, we are interested in developing an interactive calendar to visualize purchase patterns and identify trends in customer behavior across different times and locations. Variables that are important to this question are: Unit Price, Total Bill, Transaction Quantity, Transaction Date, Transaction Time, Month, Day of the Week and Store Location. We will utilize several Python libraries, as well as JavaScript, HTML, and CSS, to create an interactive calendar. We will first sort transaction data by date, time, and store location using the variables mentioned in order to create and display the interactive tool, then identify the top performing item category for each day. This tool will allow users to explore purchasing trends with ease and enable us to clearly visualize trends in customer behavior and day to day profits.

Output Summaries

For research question 1, the analysis will be able to identify different patterns across multiple time formats and the three locations. Multiple graphs and other visualization tools will highlight differences between these categories and the three locations.

For research question 2, predictive models will forecast future sales and provide insights for optimizing promotion scheduling. The results of our predictive models will anticipate and display clear sales trends that will allow us to identify the features having the most significant impact on profitability.

For research question 3, an interactive calendar will display purchasing and customer behavior trends across different times and three locations. The interactive calendar will clearly reveal significant patterns within the data which will allow users to easily identify peak purchasing trends and day to day sales.

Literature Review

Machine learning techniques are frequently utilized to streamline business logistics, visually organize data, and predict future market behavior with sales forecasting. A study of coffee markets in the U.S. was conducted that sought to answer whether regular and differentiated (specialty) coffees exhibited different demand patterns. The study, which collected retail data over the course of 313 weeks, used the Almost Ideal Demand System (AIDS) economic model to estimate parameters, and according to Alamo-Gonzalez (2012), found that the average consumer would pay “33.00% more for differentiated coffees than for the regular coffee, and from 19.00% to 217.00% more for coffees differentiated by country of origin than for the regular coffee.” The ability of any business to predict and respond to demand via sales forecasting is paramount in order to gain a competitive advantage in a dynamic market.

Accurate sales forecasting is essential for businesses to optimize inventory management and maximize profitability. There are a number of strategies and machine learning techniques that can be used to improve accuracy. One paper by Gustriansyah et al. (2022) introduces a novel approach, SalesKBR, for sales forecasting, integrating multiple machine learning techniques. Specifically, it incorporates k-Means clustering, Recency-Frequency-Monetary (RFM) model, and the Best-Worst Method.

Clustering is a common data mining technique that is used to group data points with similar characteristics. In the SalesKBR model, k-Means clustering is used to segment products based on frequency, quantity, and monetary value. Grouping products that have similar sales patterns helps in the prediction of future sales.

The RFM model is a framework in marketing which evaluates the purchasing behavior of customers to give them a value. Recency measures the time since the product was last sold. Frequency measures how often a product is sold over a given period. Monetary measures the total sales generated by a product. Using these three criteria, a score is given which can then be used for future sales predictions.

The Best-Worst Method (BWM) is a tool that is used to determine the criteria that has the largest impact on an outcome, as well as the criteria that has the smallest impact on an outcome. For example, criteria like the frequency of purchases, quantity sold, or monetary value may have a critical role in predicting future sales accurately. BWM allows for the prioritization of the most important sales predictors.

SalesKBR integrates these three components in order to improve sales forecasting accuracy. In fact, using the Symmetric Mean Absolute Percentage Error (SMAPE), the model achieved an average SMAPE of 27.12%, indicating a low error rate. This model serves as an example of how advanced machine learning techniques can be used to optimize sales forecasting, providing a relevant starting point for further exploration.

In addition to improving sales forecasting, machine learning also plays a critical role in helping businesses create and capture value. To build upon the effective machine learning techniques mentioned in the SalesKBR model, Costa-Climent et al. (2023) highlights how different applications of machine learning can drastically improve small and medium-sized enterprises. The study reveals that several machine learning techniques not only enhance forecasting accuracy but also optimize business operations by increasing efficiency, innovation, and profitability. The analysis explains the link between specific value creation and capture, which are essential for improving operational processes and profitability within successful businesses.

Machine learning is described as a valuable tool that can help businesses to identify trends in value creation and capture in order to predict future patterns and drastically improve overall business performance. Costa-Climent et al. (2023) describes the process of value creation as the integration of machine learning tools that help to improve efficiency with the goal of increased overall business performance. When considering value capture, the focus is to transform improvements into tangible outcomes, such as increased revenue, customer retention, cost savings and overall profitability. With the incorporation of machine learning tools, businesses can continuously refine future predictions, ensuring sustained value creation and capture to maximize long-term profits.

According to Costa-Climent et al. (2023), the integration of machine learning technologies is crucial for sales analysis and optimization of small and medium-sized enterprises. This directly correlates to the needs of a business like a coffee shop that is looking to optimize profitability and efficiency. By integrating the ideas of value creation and capture from Costa-Climent et al. (2023), machine learning techniques can be applied to enhance sales optimization, profitability, and operational processes. More specifically, these tools can be used to better manage inventory and stocking issues as well as enable more targeted promotions. With this, a coffee shop has the potential to drastically enhance customer experience and maximize overall financial performance of the business.

Costa-Climent et al. (2023) findings highlight the transformative role of machine learning in small and medium-sized enterprises. By continuously refining machine learning models, and incorporating real-time data, businesses have the ability to forecast future market changes accurately and ensure sustainable long-term profitability. For a coffee shop, the integration of these technologies provide a clear path to sales optimization, operational efficiency, and increased profitability. As machine learning techniques continue to evolve, their benefits will only improve and further empower businesses to achieve greater success.

Exploratory Data Analysis

For this project, the data collected focuses on transactional data from a small chain of coffee shops based in New York City. The dataset used within this study was retrieved from kaggle.com, a web-based platform for data science and machine learning professionals that allows users to compete, collaborate, and learn. This dataset includes information gathered from three distinct locations, contains roughly 150,000 observations and offers a wide variety of variables. Some of the variables we will focus on include product details, transaction dates, unit prices and much more. Through our exploratory data analysis (EDA), we plan to explore this data set with the goal of uncovering different patterns in profitability, predict and optimize future sales, and visualize customer purchases. Our overall goal with this exploration is to gather new insights that will ultimately allow us to deliver data driven recommendations in order to enhance overall profitability and operational efficiency.

Before we can explore the data, we must first have a clear understanding of the different variables that exist within the dataframe, and the values or objects that they represent. For this project, we will analyze the data using the following attributes:

store_location: Whether the transaction occurred at the Astoria, Hell's Kitchen, or Lower Manhattan store location.

transaction_qty: Number of units purchased in a single transaction.

unit_price: Set price per unit.

total_bill: Overall total of the transaction, excluding tax or gratuity.

product_category: Product categories include: Tea, Coffee, Bakery, Drinking Chocolate, Flavours, Loose Tea, Packaged Chocolate, Branded, and Coffee Beans.

product_type: 29 unique subcategories that group each product by characteristic.

product_detail: Additional details pertaining to each product.

size: Size of the product purchased (if applicable). Sizes are: Small, Medium, and Large.

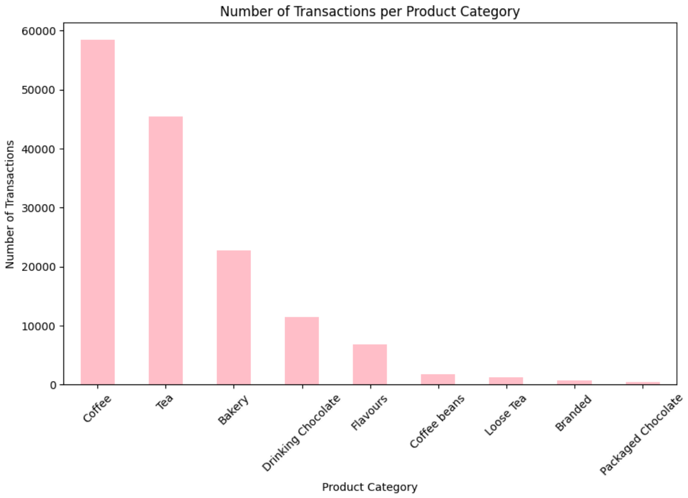
month_name: Month in which the transaction occurred.

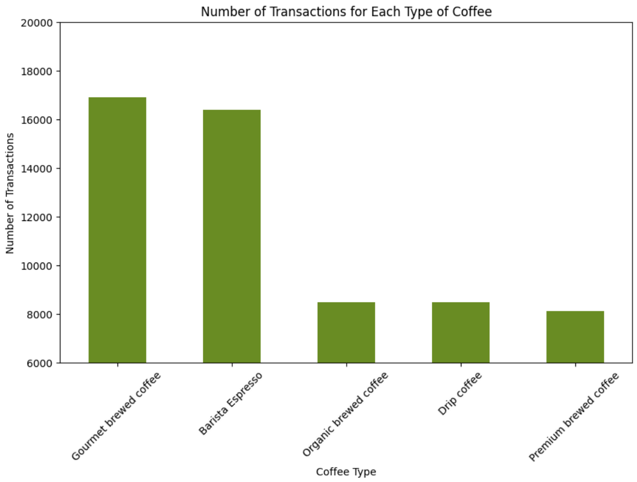
day_name: Day of the week on which the transaction occurred.

hour: Hour in which the transaction occurred.

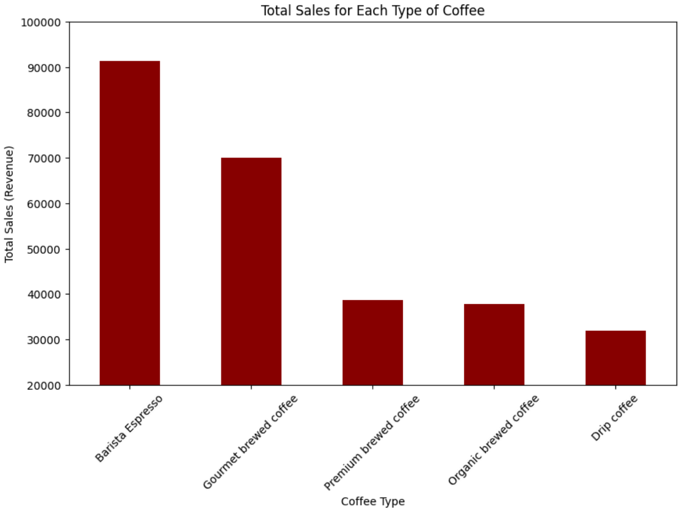
By clearly defining the important variables within our data, we can confidently proceed with our analysis. This ensures that each variable is interpreted accurately, allowing us to uncover meaningful insights and navigate the next steps.

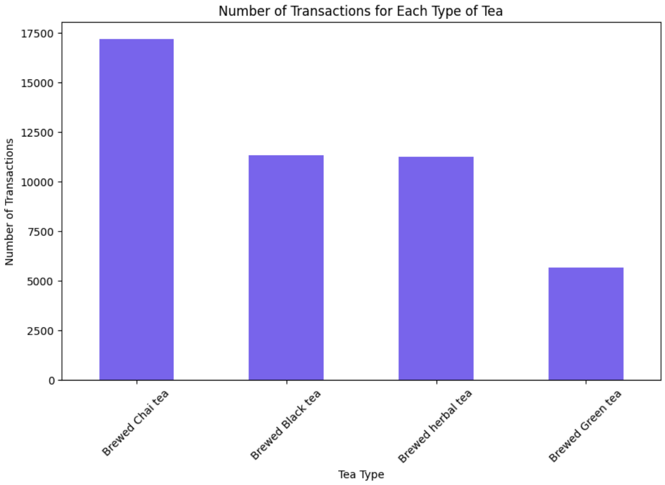
To begin, Exploratory Data Analysis (EDA) is used to analyze and investigate datasets and summarize main characteristics. To better understand the distribution of variables, we will first take a look at the frequency of transactions across each product category, displayed in a bar graph. This visualization is made possible with the use of the Matplotlib library, an extremely useful tool for creating detailed, interactive plots within Python. From the bar graph it was revealed that Coffee has the highest number of transactions, making it the most popular product category by far. While the Packaged Chocolate category has the least number of transactions. Additionally, the graph indicates a strong customer preference for beverages (Coffee and Tea) over other product categories. With that said, marketing efforts could focus on promoting different varieties of coffee or tea to increase average transaction value, given the dominance of these two categories. Since we have determined that coffee and tea are the most popular product categories, let's take a look at the specifics. To do so, we will examine the number of transactions for each type of coffee and the number of transactions for each type of tea, using bar graphs. For types of coffee, we have Drip Coffee, Organic Brewed Coffee, Barista Espresso, Gourmet Brewed Coffee, and Premium Brewed Coffee. The graph for the number of



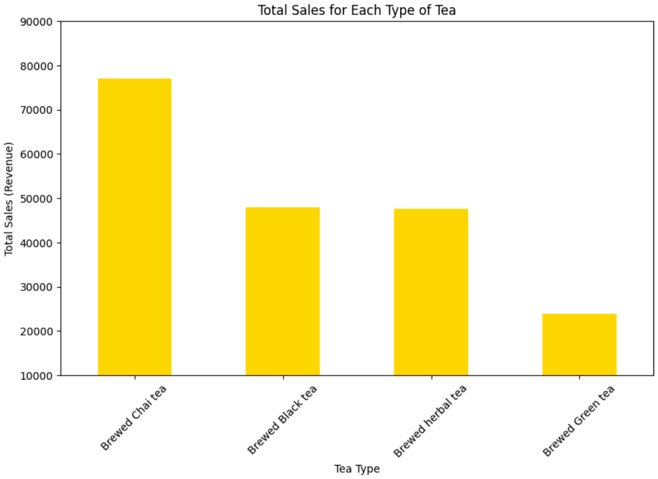


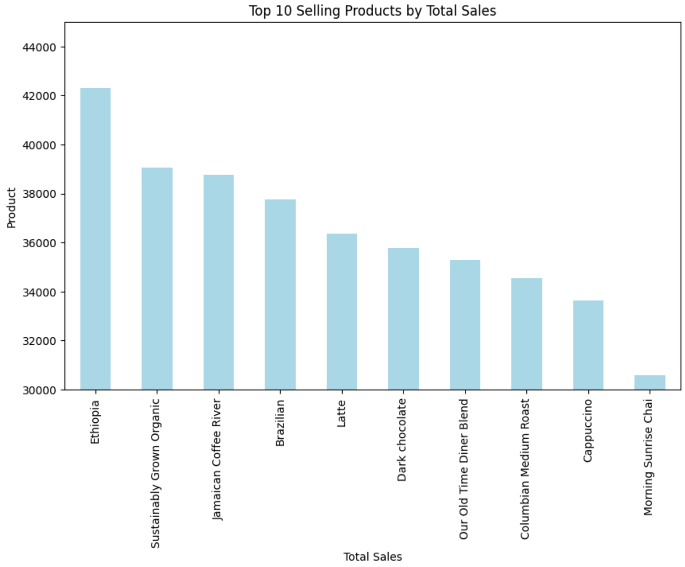
transactions for each type of coffee tells us that Gourmet Brewed Coffee and Barista Espresso have the highest number of transactions, with both exceeding around 16,000 transactions. This tells us that customers favor these two types of coffee the most which suggests that they are the key drivers of overall coffee sales. Additionally we can see that Premium Brewed Coffee has the lowest number of transactions, with around 8,000 transactions. Although it seems to be the least favorite, it is still relatively popular compared to the other beverage categories. Furthermore, if we explore the revenue that is generated through sales of each type of coffee, we can see again that Gourmet Brewed Coffee and Barista Espresso are again the top performers, although this graph highlights Barista Espresso as the more profitable of the two. As the two graphs are still relatively proportional, this is a good indication of a correlation between the number of transactions and the total sales.



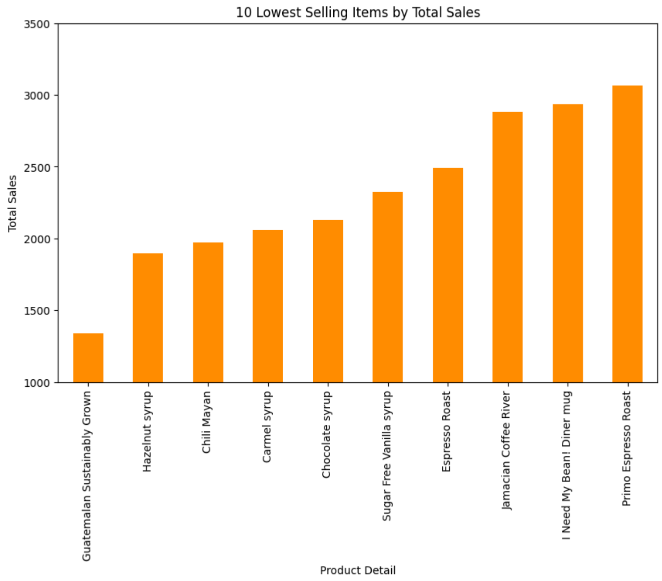


Next let's take a look at a bar graph of the number of transactions for each type of tea. For types of tea, we have Brewed Herbal Tea, Brewed Black Tea, Brewed Green Tea, and Brewed Chai Tea. The graph for the number of transactions for each type of tea tells us that Brewed Chai Tea has the highest number of transactions at nearly 17,500. Brewed Black Tea and Brewed Herbal Tea are tied at second and third most transactions. Brewed Green Tea has the fewest number of transactions, significantly lower than the other other tea types. Again, looking at total sales of each type of tea, we can see the same proportional relationship reflected visually in the graphs, with Brewed Chai Tea. Because our Brewed Black Tea and Brewed Herbal Tea aren't too far behind the top seller, Brewed Chai Tea, promotions could be focused around them in hopes of elevating their sales to the top.



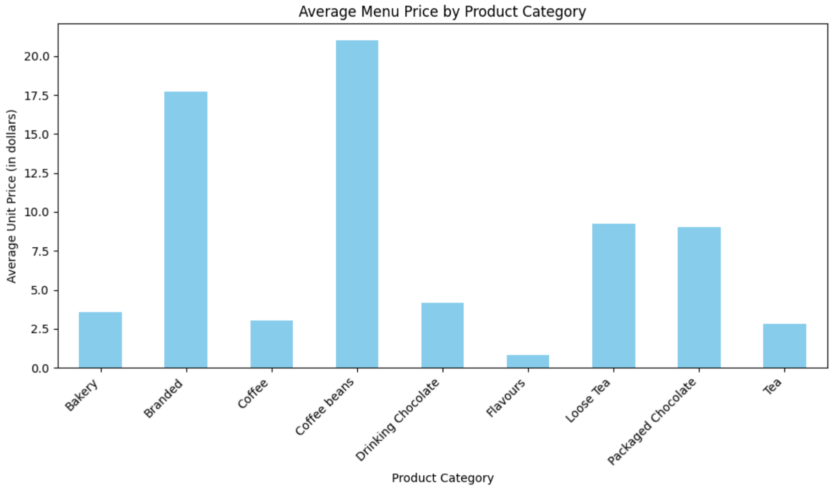


To get even more specific into the product sales, we can look at the best and worst selling products by total sales. The highest-selling product is Ethiopia coffee, generating over $42,000 in sales, followed by Sustainably Grown Organic. Interestingly, the lowest-selling item is Guatemalan Sustainably Grown coffee. This displays the complexity of

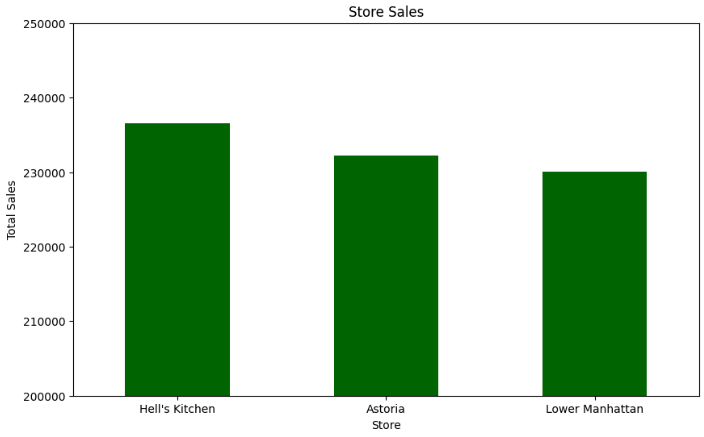


understanding consumer preferences because there is no clear cut answer regarding sustainability preferences. While one sustainable option ranks among the top sales performers, another has the lowest sales.

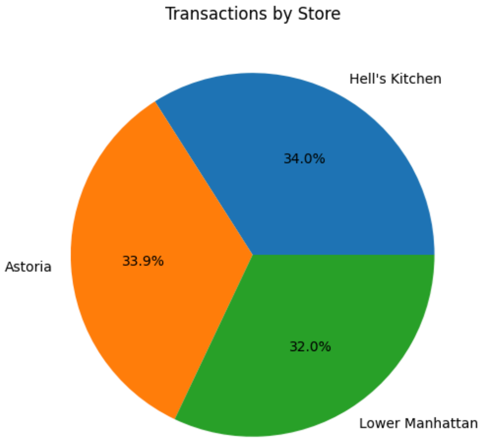
When analyzing revenue data, it is crucial to look not only at the summated revenue data, but also at pricing details for individual categories and items. Doing so allows for a clearer understanding of how each observation within the dataset contributes to the overall revenue generated. This graph paints a picture of the menu prices, demonstrating that sales must be analyzed using multiple approaches, as selling one bag of coffee beans may turn a much larger profit than even nine cups of tea.

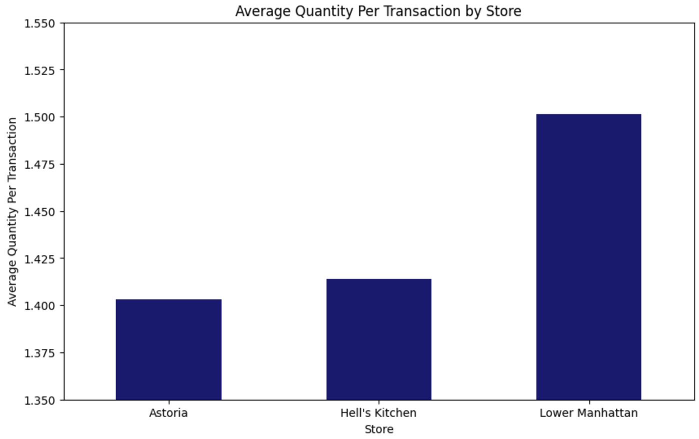


Following this analysis, let's take a look into data across each store location. This will be extremely helpful in determining which stores are performing the best and worst. Additionally, it is important to identify which specific locations are driving the bulk sales, which can help guide inventory decisions and promotional efforts. The three coffee shops based in New York City consist of Astoria, Lower Manhattan, and Hell's Kitchen. From

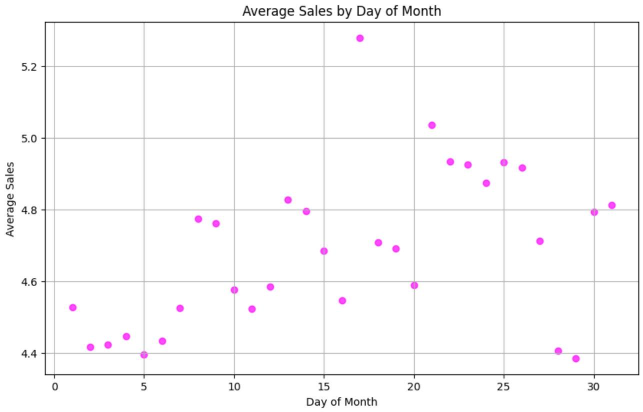


our visualization we can see that Hell's Kitchen has the highest total sales which indicates that this location is the top performer. The Astoria location has the second highest sales, right around 230,000. Finally the Lower Manhattan location appears to have the lowest number of sales, but still contributes a significant amount. In regards to the number of transactions that can be attributed to each store, there seems to be an approximately equal split between the three locations, as shown using a pie chart. From this we can state that no store in specific is underperforming considering that sales figures across all three stores are fairly consistent. Although Hell's Kitchen and Astoria may be higher priorities when it comes to factors like inventory and staffing. On the other hand, because the Lower Manhattan location has the lowest total sales, it could potentially be a great focus for implementing different growth strategies.

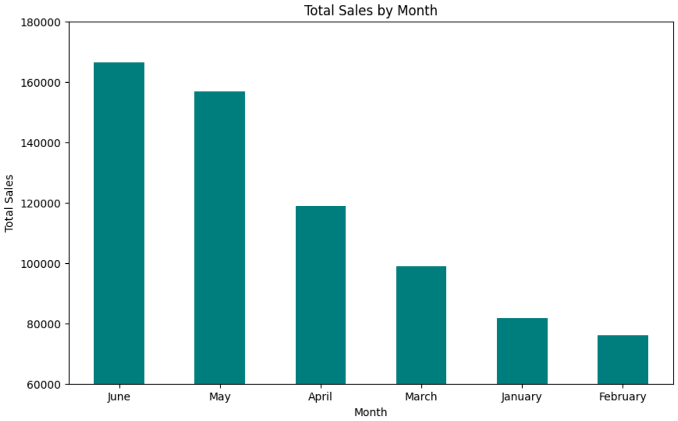


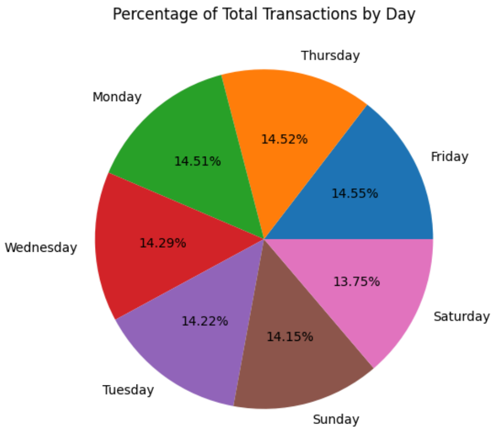


However, if we look at store data on the average quantity purchased per transaction, we can see that Lower Manhattan actually outperformed the other locations, indicating that while the overall number of transactions may be lower, the average patron in Lower Manhattan makes higher quantity purchases than their Hell's Kitchen and Astoria counterparts. This provides valuable insights about how consumer spending patterns differ by store location.

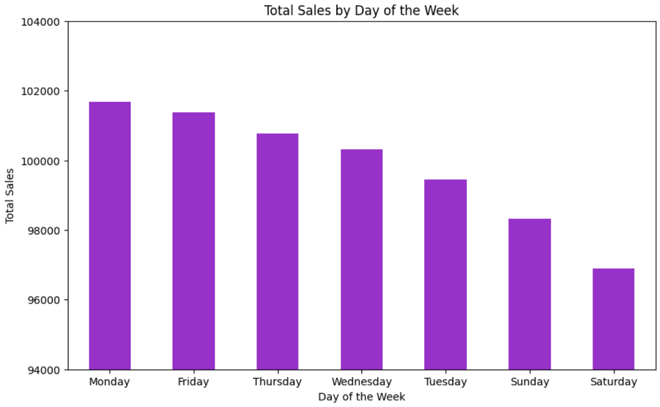


Continuing our sales analysis, let's take a look at the average sales by the day of the month. It is important to understand how sales fluctuate throughout each month in order to uncover patterns and identify potential trends. We will visualize this data through a scatterplot graph, which is also made possible by the Matplotlib library. Some key observations to point out based on this plot is that there is a gradual increase in the average sales within the first half of the month. This reveals that customers seem to spend more during the beginning of the month. We can also see that there is a bit of a spike within the middle of the month, and a drop towards the end of the month. This could be due to several reasons that we will work to uncover throughout our analysis. To understand this plot better, let's focus on patterns over specific periods in hopes of revealing exact sales trends over time.





The dataset used in this project consists of observations recorded during the first two quarters of the 2023 fiscal year, so we will be looking at transactions that occurred between January and June. Across all locations, the busiest time of year seems to be April through June, with June experiencing the highest revenue and number of transactions, followed closely by May. Examining the data on a daily basis, we can deduce that the most foot traffic (number of transactions) occurs on Fridays, followed by Thursdays, and then Mondays, although no one day seems to be underperforming by any substantial margin. Where revenue is concerned, Mondays tend to be slightly more profitable than Fridays or Thursdays, totaling just shy of $102,000 over the course of six months. Looking at the sales data from a temporal perspective provides key details that allow for a more accurate representation of expected future revenue, efficient structuring of daily and seasonal advertisements and promotions, and precise employee scheduling.



This EDA provides plenty of valuable insights which will guide the data analysis for the rest of the project. By understanding the variation in product sales across different categories and locations, we gain information which can aid in maximizing profitability. Analyzing sales data across different time guidelines reveals patterns and gives a strong foundation for developing predictive models for sales forecasting. Additionally, visualizing patterns in customer behavior provides a foundation for creating a tool which culminates a variety of data in a digestible manner, helping to identify trends. Overall, these insights will help guide the next steps in data analysis, as well as impact decision-making for inventory, marketing, and other key areas.

Methodology

In order to successfully achieve the goals defined by our research questions, a wide variety of machine learning techniques were implemented, including Linear Regression, Random Forest, and Decision Tree models. Each model was carefully selected to address the specifics of each task, ensuring accuracy throughout the analysis. The following section provides a comprehensive overview of our methodology, including the specific steps and processes used to address each research question effectively.

Research Question 1: How does gross revenue vary across time and locations?

To investigate how gross revenue varies across time and store locations, we employed machine learning strategies to forecast daily revenue for the coffee shop. The model's purpose was to identify key factors influencing revenue, such as day of week or holiday trends, as well as location.

First, some data cleaning and transformations were performed to prepare the dataset for predictive modeling. Transaction dates were converted to a consistent format and revenue was aggregated from the daily totals. Additional features were created to attempt to capture potential patterns affecting revenue. This included time-based features such as weekdays vs weekends, as well as incorporating holiday features (day of, pre, and post holidays). Additionally, rolling averages over 3, 7, and 14 days were included to capture recent trends, whereas a 30-day lagged revenue feature was used to capture cyclic patterns.

Both decision trees and random forest models were originally developed in order to find the optimal model. They were trained on the daily revenue data using the engineered features. The initial evaluation using Mean Squared Error and R-squared scores provided insight into how well each model fit the data. Ultimately, the random forest model was chosen due to its ability to handle more complexity than the decision tree model. Random Forest models are able to handle non-linear relationships and importance-based feature selection, making it the perfect model for this job. From there, RandomizedSearchCV was used to optimize the hyperparameters such as n_estimators, max_depth, min_samples_split, and max_features. This search improved the model performance as it was able to find a balance between complexity and fit. The model was trained using the best parameters. Additionally, the most influential features were identified and ones that were not influential were removed. This resulted in a Mean Squared Error (MSE) of 34749.48 and R-squared (R2) Score of 0.7978.

The most important features in predicting daily revenue were visualized. From there, partial dependence graphs were made to understand the specific impact of top features on predicted revenue sales. As for the model itself, an Actual vs. Predicted Revenue model was created to visually evaluate the fit. Residual analysis was also performed to indicate any problem areas.

This methodology revealed that recent revenue trends and day of week patterns are strong predictors of daily revenue, whereas location is not. These insights can be used to draw recommendations for optimizing promotional strategies and pricing in order to improve gross revenue.

Research Question 2: How can predictive models be used to forecast future sales and schedule promotions effectively?

In exploring how predictive models can be used to forecast future sales and schedule promotions effectively, we implemented a Decision Tree model. A Decision Tree is a supervised learning algorithm that can be used for both classification and regression tasks. The model works by splitting a given dataset into smaller and smaller subsets, also known as nodes, in order to create a model that predicts the value of a target variable by learning simple decision rules based on the features within a dataset. This model is used primarily to capture interactions between features within the dataset and make accurate predictions on unseen data, which will aid in the ability to predict future sales and schedule promotions effectively. Before implementing the Decision Tree model, a Naive Bayes model was initially experimented with. However, we found that Naive Bayes was not well suited for this specific analysis due to its assumption of feature independence. The assumption of feature independence would limit the model's ability to understand relationships within the data, which led to poor performance in the model. This limitation led us to implement a Decision Tree model, which is better equipped to capture the complex relationships among features when forecasting sales and promotional schedules.

In our implementation, we modeled the Decision Tree to predict sales categories (Low Sales, Medium Sales, and High Sales) based on the key influencing features within our dataset. Because our dataset only offers a Total_Bill variable, some data preparation was needed in order to analyze sales at the daily level. We first began by converting the transaction_date column from a string format into a datetime format. This allowed us to calculate key metrics for each day, including the total sales. In order to gather our total daily sales, we aggregate our Total_Bill variable (Overall total of the transaction, excluding tax or gratuity). Additionally, we aggregated our key features: total quantity sold (sum of all items sold,), average unit price (mean of item prices), as well as the most frequent store location and product category for each day.

We then classified our daily total sales into three categories, Low Sales, Mid Sales, and High Sales, which is defined as "sales category", our target variable. The creation of these categories, or bins, first involved the calculation of the 33rd percentile and 66th percentile of our

Total_Sales, this is what divides our data evenly. To do this we implemented a function called classify_sales() which defined:

- High Sales as days with total sales above the 66th percentile,
- Mid Sales as days with total sales between the 33rd and 66th percentile,
- Low Sales as days with total sales below the 33rd percentile.

After defining our target variable (y), we shifted our focus to feature selection. Once again, the key features selected for this model include total quantity sold, average unit price, store location and product category (X). Because machine learning cannot directly work with categorical data, our store location and product category columns had to be encoded into numerical values with the use of LabelEncoder. Label Encoding is a very simple and effective way to convert categorical variables into numerical variables for use in machine learning algorithms. It is also important to mention that following the aggregation of our feature variables and encoding, our new variable names are Total_Quantity, Avg_Unit_Price, Store_Location_Encoded, and Product_Category_Encoded.

With our target and feature variables prepared we began our implementation of the Decision Tree model. The data was split into training and test sets, with 70% for training data and 30% for testing. This means that the model is to be trained on 70% of our data and to be evaluated on 30% of unseen data. Now we can begin building and training the Decision Tree model to predict our sales categories. Before training the model, it was initialized with a maximum depth of 5 to prevent overfitting. Following this, the model was now ready to be trained on the training data using the .fit() method. This ensures that the model learns the relationships between the features and the sales categories. Finally we can predict! Our trained model was used to predict the sales categories on the test set. We were then able to find that our model achieved an accuracy score of 87.3%, a strong test accuracy. This means the model correctly predicted the sales category (Low Sales, Mid Sales, and High Sales) for 87% of the cases in the test set. We conducted further evaluations and found high precision, recall, and F1-scores for each sales category, which further the effectiveness of the model. Overall, the Decision Tree model appeared to perform very well and effectively predicts daily sales categories which provide valuable insights for forecasting future sales and scheduling promotions.

Research Question 3: What purchase patterns exist, and how do we identify trends in customer behavior across different times and locations?

One of the most quintessential steps when performing any sort of data analysis, but most specifically in the realm of business analytics and sales forecasting, is the ability to clearly and effectively communicate pertinent discoveries in a format that can assist with decision-making. We have done this by utilizing Python's Calendar, Display, and Widgets libraries to display historical day-by-day sales for the first half of 2023 beginning on January 1st (fiscal quarters Q1 and Q2) in a virtual GUI embedded within Google Colab.

To create this helpful visualization, we first installed and imported the necessary packages: 'calendar', 'ipywidgets', and 'iPython.display'. The calendar module offers tools that generate a basic text calendar, while the display and widgets modules offer additional interactive features, such as dropdown lists, clickable dates, and color coding. We will use all of these features to identify potentially high, medium, and low-performing days at a glance, and to view details about each store's performance.

After importing the modules that create the basic framework for our calendar's interface, the next step is to define what we consider to be high, medium, or low performance in our sales predictions. For the purposes of this project we are considering “high performance” to be above the 66th quantile, “medium performance” to be between 33% and 66%, and “low performance” to be below the 33rd quantile.

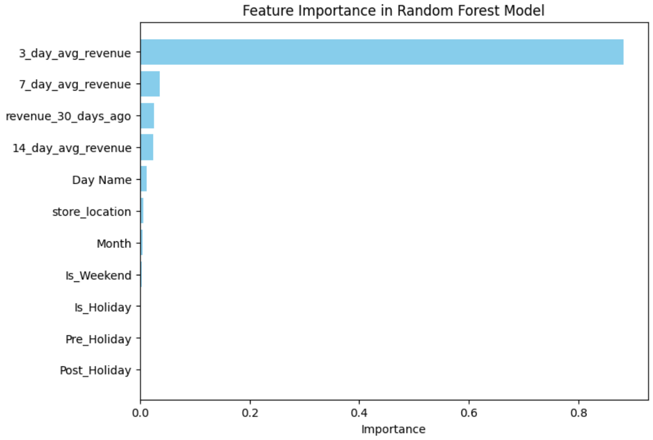
The calendar draws inferences from the Decision Tree model to automatically categorize daily sales totals for each day into high, medium, or low performance. To display this, the code iterates through each individual date in the calendar, relying on the logic from the Decision Tree model to bin the total sales figure of each day compared to the overall daily performance of Q1 and Q2. From there, we define the stylistic HTML and CSS attributes according to the conditional formatting for each day on the calendar: light red for low performance, light yellow for medium performance, and light green for high performance. Further customization with JavaScript allows us to view specific details about each day, including top product category and individual store performance, in a clickable format. We can also click several days simultaneously to compare multiple variables at once.

Of course, as our goal is to show analytics over a large time period, we need to be able to view more than just one single month without the interface becoming overwhelming or hard to navigate. For this reason, we have integrated the dropdown widget functionality that Python offers in order to streamline the look and user-friendliness of our GUI. As our data is focused on transactions for the first half of the year, the displayed options for the month dropdown range in value from 1 (January) to 6 (June), and the year dropdown only shows an option to display 2023. In a real-world scenario, this interactive calendar could easily be updated to display a much larger date range and include more detailed information about store locations, product types, and transaction quantities, yet still be intuitive enough for someone without a background in data science to obtain invaluable insights and identify room for improvement according to a number of different factors, then consequently tailor factors such as promotions, employee scheduling, and advertisements to their business's specific needs.

Data Visualizations

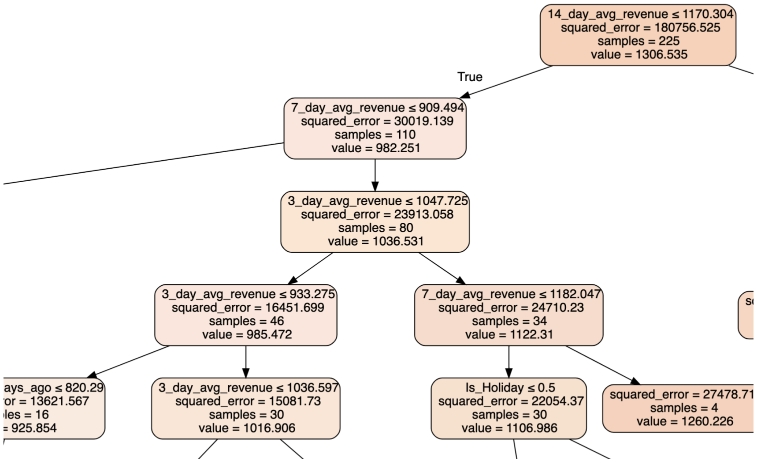
To gain a deeper understanding of our findings, we utilized several data visualization techniques specifically tailored to each model implemented. These visualizations provide clear insights and help to communicate our findings effectively. Additionally, the following data visualizations allow us to uncover key patterns and trends within our data. The following section presents the visualizations generated, along with the steps taken to implement such visualizations.

Research Question 1: How does gross revenue vary across time and locations?

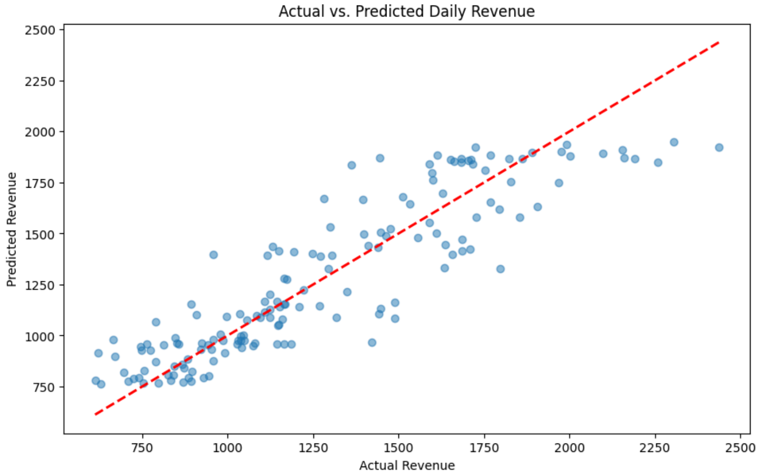


The first key visualization used in evaluating the random forest model was a feature importance plot. This plot was instrumental in understanding which features the model deemed most and least influential in predicting the daily revenue. This visual includes several time-based and trend variables that we thought would play a role in predicting sales. The holiday indicators were included to assess the effect of holidays and the surrounding days on revenue. The weekend feature indicated where a day was a weekend to account for differences in weekday and weekend revenue. These were found to be less important features and were later removed in order to simplify the model. On the other hand, the short-term rolling day averages like the 3 and 7 day revenue, which used the previous day's sales as predictors, seemed to play a significant role in the predictions. This plot not only enhanced our understanding of the model but also clarified which features truly drive revenue predictions.

This understanding of feature importance was further clarified by visualizing a decision tree structure. Although random forests use many trees, looking at one provides some insight into how the predictions are made. This photo is just a small portion of one of the decision trees that is too large to show here. However, it is still clear to see that the tree splits are primarily based on the 3, 7, and 14 day rolling revenue averages, indicating that they are still the driving features at determining the daily revenue.



In order to determine the accuracy of the model, the actual daily revenue values were compared to the predicted revenue values. If the predictions were perfect, all points would be on the red line. This scatter plot shows a strong correlation between the actual and predicted values since the points are generally clustered around the line. This indicates that the model performs well in most cases. However, there are some deviations, with the model having a particularly hard time predicting the higher revenue days.



This shows an area where more refinement could be done. Overall though, this plot confirms that the model makes reasonable predictions.

Research Question 2: How can predictive models be used to forecast future sales and schedule promotions effectively?

Creating a visualization of our Decision Tree model will allow us to better understand how the model makes decisions, and help us to identify which features are most important in predicting the sales categories we have created. To do so we will use Matplotlib, a Python plotting library. The Matplotlib library is commonly used to create a wide variety of static, interactive, and animated visualizations. It is useful for the visualization of our Decision Tree due to its ability to generate a clear, detailed, and static visual representation of the tree structure. Through the use of this library, we will be able to clearly visualize the model's complex decision making process, which will make it easier to interpret and understand. Additionally, the image provided through Matplotlib will help to provide insights into how features interact and influence outcomes within our data.

Before jumping into our visualization, it is important to understand how to read and interpret a Decision Tree. A Decision Tree has boxes, known as nodes, which represent decision points based on features. Within these nodes there are details including what feature is being used for splitting as well as the condition for the split. The details following this include:

Gini: A measure of how mixed the classes are at that node, or the node purity

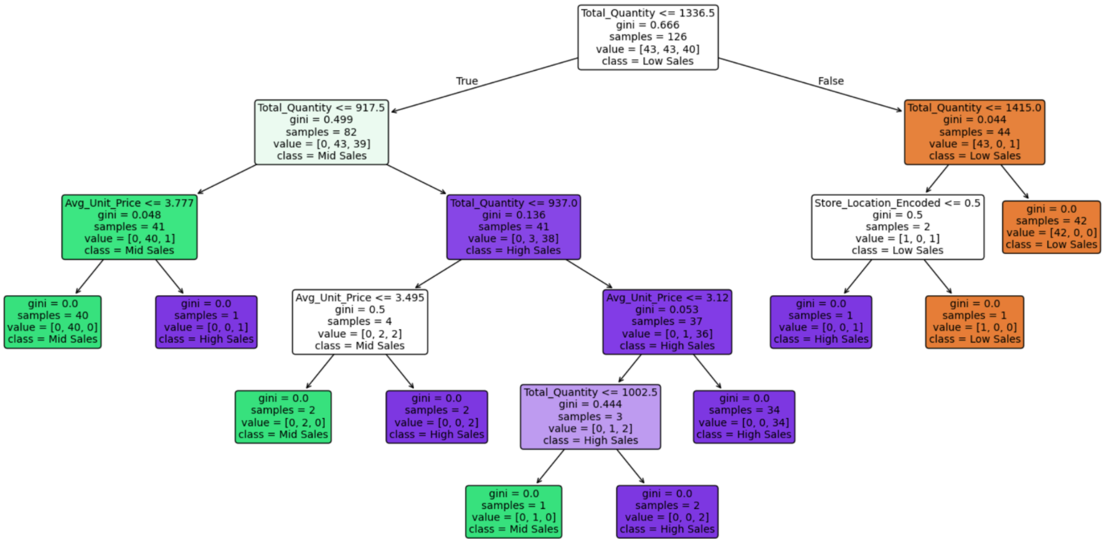
Sample: The number of data points that reach that specific node

Value: The distribution of data points across classes

Class: The predicted class of the node (Low Sales, Mid Sales, High Sales)

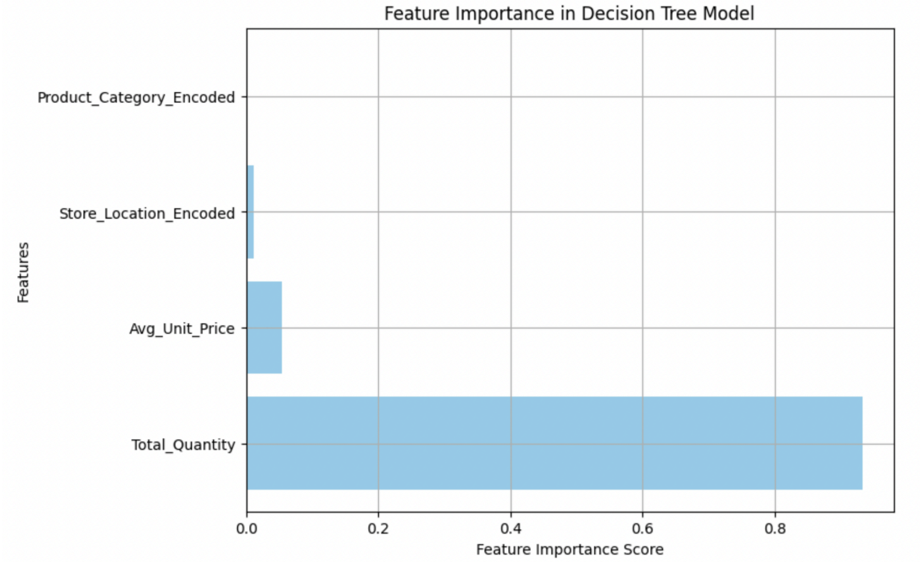
The very first node, at the very top, is known as the root node and represents the first decision rule the model applies to the data. Following this root node, are the branches of the tree which can be thought of as our decision paths. To correctly read a Decision Tree, first begin at the root node and process downward, following these branches. As you move down the tree, you follow yes/no decisions or True/False at each node. These decisions are used to split the data based on the most important features. Continue this process until you reach a leaf node, or the end of the decision tree path. The leaf node is where the model makes a final prediction. By following this process, you can see exactly how the Decision Tree Model reaches its decisions and which features are most influential.

To include some background on the process of building our Decision Tree model, we first began by importing Matplotlib as 'plt'. We include the names of the features used to train the model, 'Total_Quantity', 'Avg_Unit_Price', 'Store_Location_Encoded', 'Product_Category_Encoded'. These will be shown on the tree to indicate which feature splits are made. We must also include the sales categories, 'Low Sales', 'Mid Sales', and 'High Sales' to label the output classes. Additionally, we made sure to color the nodes of the tree based on the majority class, which makes it easier to see which class each node represents. Now let's take a look at our Decision Tree Model.



Analyzing the Decision Tree we can clearly see our root node, branches and leaf nodes. The root node splits the data based on the ‘Total_Quantity’ feature by determining if the total quantity sold is less than or equal to 1336.5. From this we can see that the model proceeds to the left branch (True) or the right branch (False). By following the branches all the way down we are able to gather several useful insights. It is clear that the most influential feature at the root is Total_Quantity. This means that the quantity of items sold per day is the strongest indicator for determining sales categories. We can see that ‘Avg_Unit_Price’ is also a significant factor and that ‘Store_Location_Encoded’ is much less significant in the model. Additionally, we can see that 'Product_Category_Encoded' is not a significant predictor within the model.

Furthermore, to ensure that we fully understand which predictive features are most important within the Decision Tree model, we generated the following bar graph.



Again we can see that the strongest predictive feature is ‘Total_Quantity’. With ‘Average_Unit_Price’ coming second and the 'Store_Location_Encoded', and the 'Product_Category_Encoded' features having little to no impact on the model.

With the visualization of our Decision Tree we can see that our model clearly captures the interactions between features, which lead us to strong predictive accuracy. The depth and splits reflect how different factors like quantity, price, and store location contribute to determining whether a day will result in low, mid, or high sales.

Research Question 3: What purchase patterns exist, and how do we identify trends in customer behavior across different times and locations?

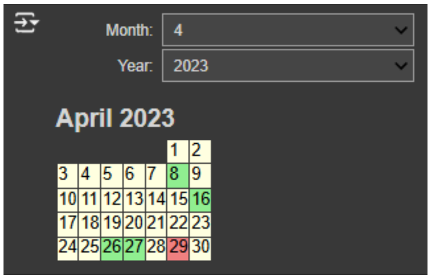
One of the most important goals of data visualization is to optimize comprehensibility, regardless of one's knowledge or background in data science. For this project, we've dynamically assembled the data into a streamlined interactive calendar display that shows both month-at-a-glance and daily sales information. With this visualization, we are able to easily determine the overall performance of each store as well as the business as a whole, and deduce patterns in consumer behavior that might be otherwise missed by only looking at the raw data.

In order to build this interactive calendar, we will need to import the ipywidgets, iPython display, calendar, and datetime modules in order to access the necessary tools to build our framework. Each of these components plays an important role in the finished product: ipywidgets gives us widget tools for customizability (in this case, it powers the month and year dropdown menus), the calendar module generates simple text-based calendars, datetime allows us to specify and manipulate values as specific date/time formats, and display allows us to view the calendar nested into the project code. Next, we create the 'display_calendar()' function, passing the following as arguments: 'year', 'month', and 'df'. This ensures that all necessary data frames and parameters can be accessed within the function and that we can reuse them to appropriately extract and compile sales information.

When generating discrete data for each day in our calendar application, we need to iterate over each date listed in the data frame, and generate aggregate gross revenue figures for each store location across all dates. An overall gross revenue total is then calculated by summing each store's daily sales figure. We use this total figure and the logic from the Decision Tree model to categorize that day's sales performance into one of three bins: low, mid, or high sales based on the full range of overall gross revenue values. The 'sales_category' variable corresponds to each day's color on the calendar, with light coral indicating low sales, light yellow indicating mid-level sales, and light green indicating high sales days. This conditional formatting is set up with a combination of HTML, CSS, and JavaScript, as well as the existing 'sales_category' variables that were previously defined. Additionally, we apply 'value_counts()' and 'idxmax()' to the 'product_category' and 'product_type' column from the 'df' data frame to calculate the most popular product category and product type for each day.

Javascript and HTML code snippets also power the calendar's ability to expand and collapse when each day is clicked, toggling a daily report of the revenue, sales category, and top product category and type values that we calculated previously. Allowing the user to show or hide daily reports when needed creates a more visually appealing application and provides a way to compare values for multiple specific days.

To set up the dropdown menu so that it displays the calendar that corresponds to the selected month, we create the 'year_dropdown' and 'month_dropdown' widgets, and define the function 'update_calendar()'. This function's purpose is twofold: to clear the calendar currently being shown, and to update with a different calendar that displays a different month or year according to the selected dropdown widgets. Finally, the line 'display(HTML(html_calendar))' prompts Python to display the fully customized interactive calendar as pictured in the example below.

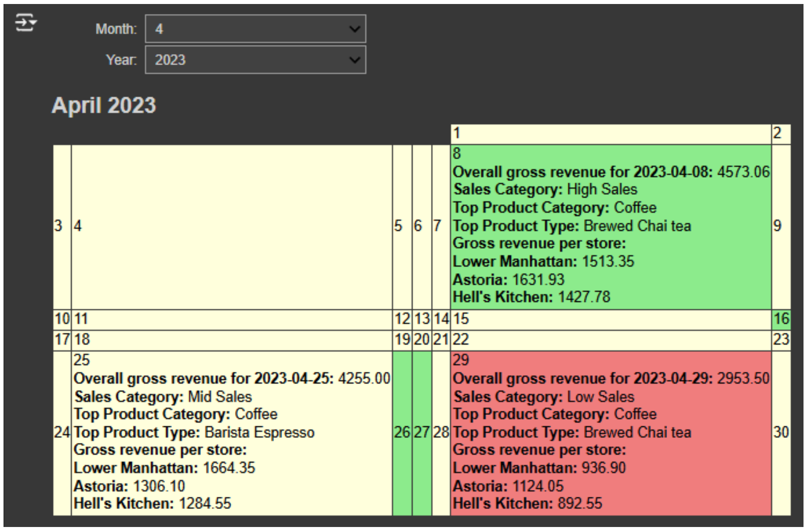


Month: 4

Year: 2023

April 2023

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30



Month: 4

Year: 2023

April 2023

					1	2
3	4	5	6	7	8 **Overall gross revenue for 2023-04-08:** 4573.06 **Sales Category:** High Sales **Top Product Category:** Coffee **Top Product Type:** Brewed Chai tea **Gross revenue per store:** **Lower Manhattan:** 1513.35 **Astoria:** 1631.93 **Hell's Kitchen:** 1427.78	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25 **Overall gross revenue for 2023-04-25:** 4255.00 **Sales Category:** Mid Sales **Top Product Category:** Coffee **Top Product Type:** Barista Espresso **Gross revenue per store:** **Lower Manhattan:** 1664.35 **Astoria:** 1306.10 **Hell's Kitchen:** 1284.55	26	27	28	29 **Overall gross revenue for 2023-04-29:** 2953.50 **Sales Category:** Low Sales **Top Product Category:** Coffee **Top Product Type:** Brewed Chai tea **Gross revenue per store:** **Lower Manhattan:** 936.90 **Astoria:** 1124.05 **Hell's Kitchen:** 892.55	30

For the entirety of January, February, and March, all sales categories are listed as either 'Low Sales' or 'Mid Sales', indicating that these are not particularly busy months for the coffee shops. However, business begins to see a slight uptick in revenue during the month of April, then even more so in May and June, suggesting that people spend more time and money in the store in the spring and summer, and that it may be wise to increase scheduled employees during those time periods to accommodate an increase in traffic. While there does not seem to be a clearly top performing store, Hell's Kitchen almost always seems to be less busy than its counterparts, though the difference is marginal and no store is severely underperforming. As far as the products that contribute the most to each day's revenue, coffees of various types are consistently popular, but the 'Brewed Chai tea' cannot be discounted, as it is a product type that is frequently purchased. These analytics are all invaluable segments of information when considering the logistics of scheduling, product orders, and marketing, and together they can help the business reach its full potential of success.

Ethical Recommendations

In this analysis of “Optimizing Coffee Shop Sales through Time-Based Analysis and Predictive Modeling," we explore how data-driven insights affect and improve coffee shop gross revenue across product categories and locations, improve sales forecasting and promotion scheduling, and reveal trends in customer purchase patterns through the use of predictive modeling. While predictive modeling offers opportunities for business optimization, it is important to address the potential ethical concerns in order to ensure responsible and fair use of the models.

With our implementation of predictive models for forecasting future sales and scheduling promotions effectively, we must consider potential ethical concerns. While predictive modeling offers clear benefits, there is one ethical consideration that stands out when considering that the model is used responsibly. The ethical issue that arises when using sales forecasting to schedule promotions is the impact on fair pricing. If this model is misused, it could lead to manipulative pricing strategies. For example, if the model predicts high demand on a specific day, or on the contrary, a notably low demand day, the coffee shop owner may be tempted to increase prices to maximize profit and offset costs. These price adjustments could negatively impact and exploit customers. This is an example of price gouging, unfairness, and transparency, which is major when considering ethics of machine learning models. As noted in a recent paper discussing AI ethics, “Promoting fairness and nondiscrimination is an essential theme within the context of AI… AI-enabled systems for customer prioritization require companies to protect consumers’ privacy and autonomy and promote justice and nondiscrimination” (Spanish Journal of Marketing, ESIC). To ensure fairness and transparency in sales and pricing, it is important to maintain transparent pricing strategies, limit price increases during peak times, and ensure that the scheduling of promotions are balanced across high and low demand periods. These ethical recommendations will allow the business to take advantage of the models insights while promoting fair and responsible pricing.

The consequences of price gouging are numerous. While an increase in price might yield better revenue on a short-term basis, in the long run it is an incredibly harmful practice that creates issues for both the consumer and the business. For example, if the price of a barista espresso, a consistently popular item, were raised drastically, it might price out certain economic groups of consumers who are unable to consistently afford to overpay for their favorite latte, and thus be forced to go out of their way to buy a coffee elsewhere. This would eventually lead to a damaged reputation for the coffee shops and a decrease in overall business due to the ethical implications of their business practices, even if the loss in sales is offset by an increase in individual product pricing.

In addition to the potential harm of misusing the data insights in a way that negatively impacts sales, customers, and suppliers, deciding not to use the research insights also has its own set of ethical considerations. In fact, the coffee shop would miss opportunities to optimize revenue, improve customer satisfaction, and make informed resource allocation decisions. For example, not using the insights could lead to wrong supply decisions. If there's a surplus, there could be waste involved, decreasing revenue. If there's a shortage, there may be a failure to meet the customer demands, decreasing customer satisfaction and limiting the revenue potential.

In general, ignoring research data could lead to lost opportunities for creating a more efficient and well-loved company. The business will be less capable of understanding and addressing customer preferences and adapting to demand, ultimately harming the company's reputation. It is important to maintain a balanced approach, where data is used responsibly to optimize revenue without sacrificing ethical business practices. With this in mind, the coffee shop company will be able to earn the trust and loyalty of its customers while still improving profits.

Works Cited

Alamo-Gonzalez, C. (2012, December 1). *Implications of product differentiation in food demand: The case of coffee in the United States*. DSpace Repository, Texas Tech University Libraries. https://ttu-ir.tdl.org/items/ed0c3fe7-c712-4788-82b7-1a8337d22d84

Artificial intelligence and predictive marketing: An ethical framework from managers' perspective. (n.d.). *Spanish Journal of Marketing - ESIC*. https://www.emerald.com/insight/content/doi/10.1108/SJME-06-2023-0154/full/html

Costa-Climent, R. (2023, February 15). *Using machine learning to create and capture value in the business models of small and medium-sized enterprises*. Science Direct. https://www.sciencedirect.com/science/article/pii/S026840122300018X

French, R. (2024, May 14). Americans drinking record amount of coffee. *Supply Side Food & Beverage Journal*. https://www.foodbeverageinsider.com/beverage-development/coffee-consumption-hits-record-high-in-us

Gustriansyah, R., et al. (2022). *An approach for sales forecasting*. *Expert Systems with Applications, 207*, 118043. https://doi.org/10.1016/j.eswa.2022.118043

Light_Shot. (2024, March 30). *Coffee shop sales analysis*. Kaggle. https://www.kaggle.com/datasets/divu2001/coffee-shop-sales-analysis?resource=download

S&P Global. (n.d.). Sharing insights elevates their impact. https://www.spglobal.com/commodityinsights/en/ci/products/food-commodities-food-manufactu ring-softs-coffee.html#:~:text=Coffee%20producing%20countries%20export%20the,estimated%20to%20exceed%20%24200%20billion

Appendix

See ‘Capstone Project.ipynb’ and ‘Capstone_Project.html’ files for source code.