**Coffee Shop Chain Sales Data**

Business Implications

**Class:** IST 652

**Instructor:** Professor D. Landowski

**Student Group:**

Adit Shah

Keeley Ables

**Table of Contents:**

Introduction .......................................................................................................... 3

a. Business Questions …............................................................................. 3

Data Source and Description …............................................................................. 3

Pre-Processing …................................................................................................... 6

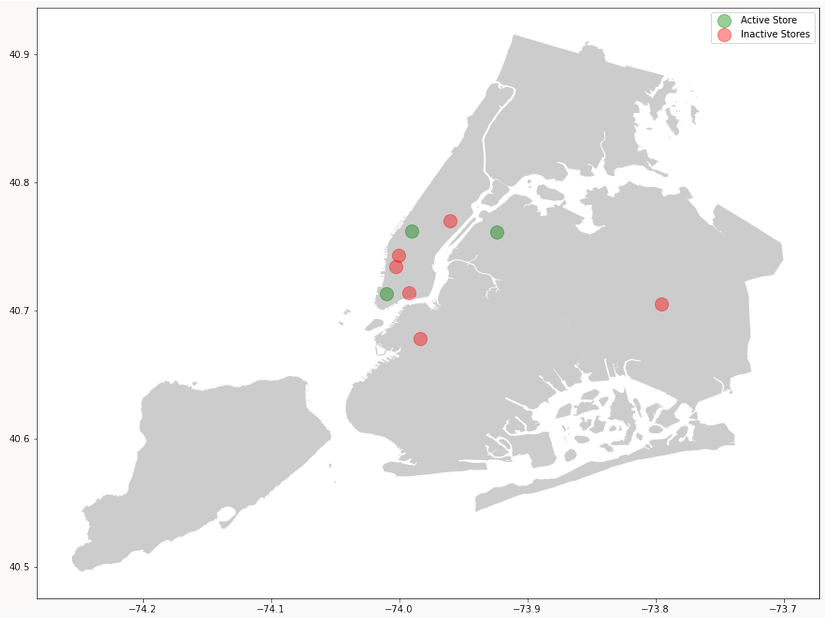
Analyses and Interpretations ................................................................................ 8

Conclusion …......................................................................................................... 16

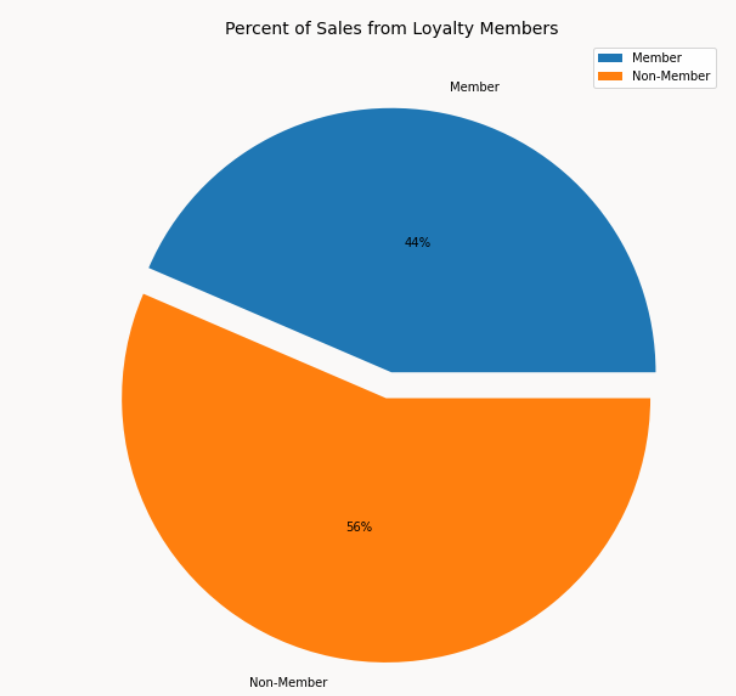
Limitations of our Analyses …............................................................................... 17

Appendix …........................................................................................................... 18

Introduction

The focus of this initiative was a small, struggling coffee shop chain in New York. It was this team’s objective to analyze sales and potential avenues for the allotment of advertising resources. The rationale for our objectives is displayed in the bubble map below. As seen, this coffee shop chain’s sales are such that it was necessary to close five out of eight of their branch locations. As indicated by the legend on the map, the green bubbles represent the active locations, while the red bubbles represent stores that have been closed due to inadequate sales.

To further emphasize the need for support, we have included the below pie chart displaying the percent of sales attributed to loyalty program members, versus non-members. This led us to question why the majority of our customers were non-members.



To support this business, and drive sales, we would need to concentrate on aspects such as advertising, price point, and customer loyalty. The answers to the following questions address these focus areas.

**Business Questions**

1. How should the business appropriate its advertising resources?

* On which products?
* Which products enjoy the highest sales? Is there an evident reason for their popularity?
* What is the typical product price point, and is there evidence of price elasticity?
* On which locations?
* Which location has the highest concentration of sales?
* On which customer demographics?
* Who are our target customers in the loyalty program?
* Can we predict membership?

Data Source and Description

The team analyzed two months of data describing various facets of the business. The links below lead the reader to download eleven data sets.

<https://www.kaggle.com/ylchang/coffee-shop-sample-data-1113>

<https://community.ibm.com/accelerators/catalog/content/Small-coffee-chain-sales>

These data sets include sales receipt data, loyalty member customer data, sales targets, transaction dates, weather patterns, inventory and spoilage data, store outlet data, and staff data.

The following are brief descriptions of each dataset. Please refer to the appendix section of this report to see tables displaying specific fields, descriptions, and examples from each data set.

**Sales Receipt Data**

Sales Receipt Data houses the metadata concerning each sale. Information concerning each product sold, when and where it was sold, and who sold it can be found in this dataset. It is the primary dataset from which other data concerning the business branches.

**Customer Loyalty Membership Data**

This data set consists of demographic and contact information for each loyalty program member. It names each member, records their loyalty card number, gives each an identification number, as well as their birth day/year.

**Daily Sales Targets**

Metadata concerning daily sales goals (per product type) by store location can be found in the Sales Targets data set.

**Dates**

This data is composed of the day, week, month, quarter, and year in which sales occurred.

**Generations**

The generations data set classifies each birth year by generation. For example – A birth year of 1989 would mean people born in that year are “Older Millennials”

**Weather**

The Weather data set reports the ranges of temperature, amount of precipitation, and wind speed by date

**Pastry Inventory and Spoilage**

This data set tracks the amount of pastry inventory. It tracks inventory at the start of the day, inventory sold, and the amount wasted (due to non-sales), by date and product.

**Product**

The product data set places each product into a product type, group, and category. It gives the retail and wholesale price, as well as the unit measurement it is stored in. This data also displays whether or not it is tax exempt, whether it has associated promotions, and whether it is new or not.

**Sales Outlet Data**

Sales Outlet Data displays the location, type, contact information, and active status of each store, by store number.

**Staff Data**

Finally, this data set identifies each staff member, and discloses each employee’s position, location of work, and start date. Positions range from CEO to Barista

Pre-Processing

**Packages used:**

***Analysis***

* Numpy
* Pandas
* Scikit-learn

***EDA & Pre-processing***

* Datetime
* OS

***Visualization***

* Matplotlib
* Seaborn
* Geopandas
* Geoplot
* Shapely

***Data Explorations & Pre-processing:***

Steps taken for each RAW file

Step 1: Verified the data import occurred correctly via “.head()” function

Step 2: Checked data types of each column via “.info()” function

Step 3: Created a function (“EDA\_func”) to perform basic EDA steps and create a dataframe summarizing the results. Resultant columns include:

* Column name
* Data type
* Total rows
* Duplicate values
* Unique values
* Mean
* Min
* Max

The code that contains this function was designed to accept a directory from the user. The code would then loop through all csv files in the directory, perform the above steps, and append to a master EDA file (Refer to EDA\_Code.ipynb)

Step 4: Results for most files were “as expected”. Most of the files had a primary key and had almost no missing values. Some files (like “staff.csv”) had columns that contained values of multiple datatypes, for example – strings and numbers in the same column. The impact of such issues was relatively limited, as those files/columns were not used in the analysis. Also, approximately half of the values in the customer\_id column was 0’s or blank. We interpreted this to mean that those customers were not part of the loyalty program. If you look at the “customer.csv” file, you will notice that each customer in that file has a loyalty\_card\_number. This, in turn, implies that this coffee shop chain only assigns a customer id for people who are part of the loyalty program.

· Step 5 (pre-processing):

o During the analyses, date and time fields had to be reformatted from strings to dates and times respectively.

o Blank values were turned to 0 in instances where customer ID was needed for the analysis.

o Columns with “Y”/”N” were turned into 1/0

o For the “Precipitation” column in the Weather data frame, values had to be shifted 1 cell “up”. The original column contains “yesterday’s” precipitation amount. To see if weather had any impact on that day’s sales, we had to shift the values 1 day.

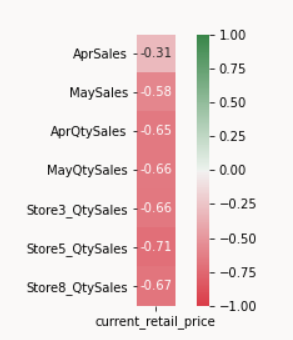
**Analyses and Conclusions**

* *Which products enjoy the highest sales? Is there an evident reason for their popularity?*
* *What is the typical product price point, and is there evidence of price elasticity?*

Notice that we addressed the above questions simultaneously, as there is somewhat of an overlap in the scope of each question. As a result, visualization support for both analyses could easily be shown in the same graph.

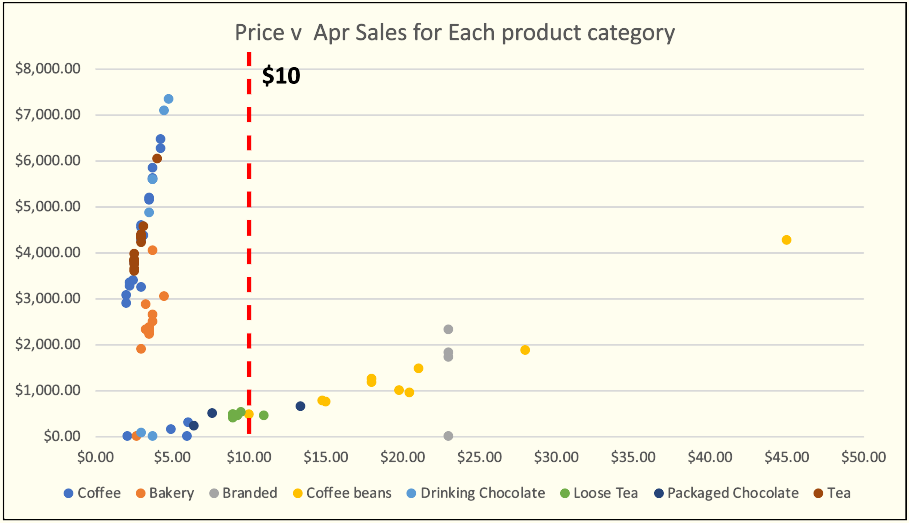
To answer these questions, we performed a correlation analysis between sales and price point. Through this visualization, we would be able to see best-selling products, as well as price elasticity. We justified that seeing this correlation with price may give us insight as to the reason behind a product category’s high sales. The line of thinking would be perhaps the product sells well because of its low price point.

We began our process by, of course, reading the sales receipt data into python. We chose to keep only the columns helpful to the analysis. New columns were created for April and May sales, quantity of sales for both months, and quantity of sales by store location. These columns were created for future aggregation. Next, unnecessary columns were again filtered out and sales were aggregated by product ID. The Product table was then merged with the newly modified Sales Receipt table on the product ID field. From this merged table, the correlation heat map below was created.



While this heat map does serve to answer the second portion of our question, it was now apparent that there is a correlation between sales amount and price point, further investigation was needed to fully answer both inquiries. As a result, we created a scatter plot displaying the effect of price on sales.

The scatterplot would require a table with product ID, product category, retail\_price, and sales. Therefore, product\_id, current\_retail\_price, and April sales were the columns imported. In addition, product\_id and product\_category were also imported then joined with first three imported columns. An empty plot was created and each product category, with its data was assigned to the plot. From there, the scatterplot was formatted to include a legend with the different product categories. The y axis label would become ‘Sales’, and the x axis label would become ‘Retail Price.’ The title assigned would be ‘Sales v Retail Price per Product Category.’ Please find the resulting scatterplot below.



This scatterplot visualizes the overlap between our initial two questions which are the objective of this correlation analysis. The visualization of sales by individual product was confusing, as there are eighty-eight of these products. However, the above scatterplot does a more effective job communicating the same message. From this graph, we were still able to see what kinds of products were selling. Moreover, the scatterplot suggested the elasticity of each product category. This comparison between price point and sales could provide insight as to why some of these products were selling.

*Interpretation*

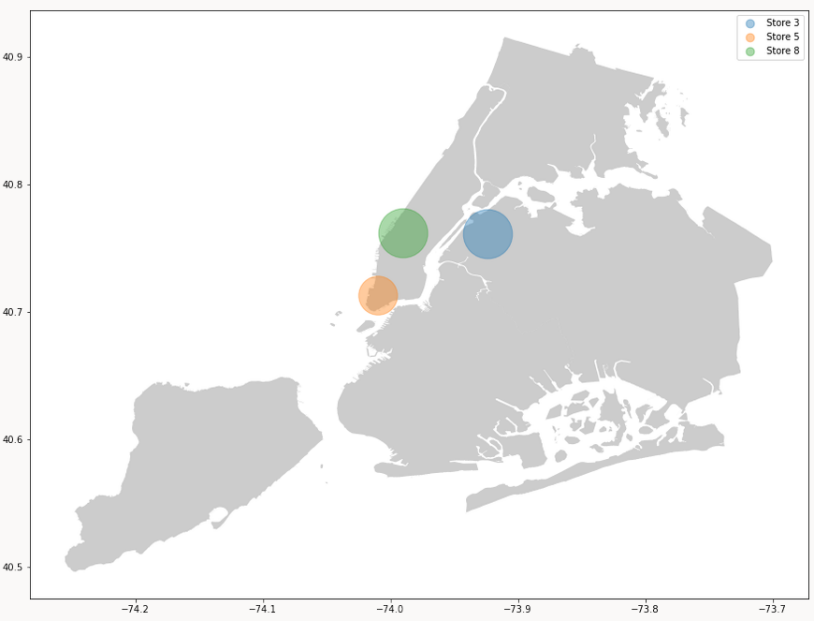
From the correlation analysis, the team learned that the three most popular products are found within the drinking chocolate product category, the coffee product category, and the tea category. To clarify, when sales are aggregated by product category, the coffee category does have the highest sales. However, when considering sales at the product level, the best-selling product is from the drinking chocolate category.

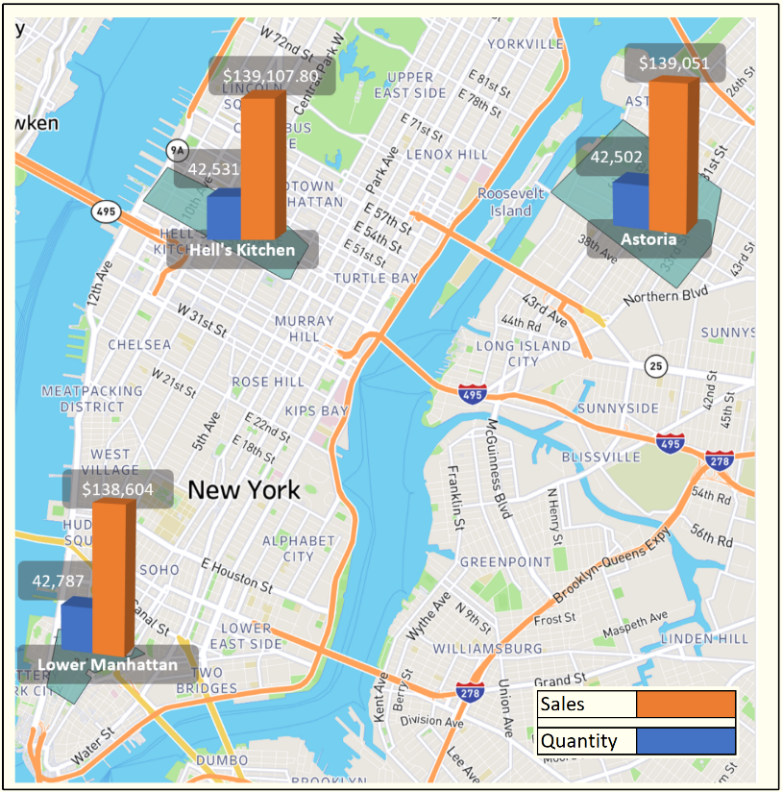
Being a coffee shop business, it may be intuitive that beverages would be the top-selling products in the chain. However, considering the business specialized in coffee products, it may be unexpected to see that a product from the drinking chocolate category would return the highest sales. A limited justification could be the inexpensive price point of the top-selling product, as evidenced by the Sales v Retail Price per Product Category scatterplot. However, this reasoning is contradicted by the second highest selling product from the coffee category. This product is slightly more affordable, yet those sales (on an individual product level) are notably less. From our available data, it is difficult to interpret the cause of this. Further investigation would be helpful to find the reasoning behind this observation. Although not as high as the top-seller, second place is held by a product from the coffee category. Again, the high sales could be considered instinctive, as the business specializes in coffee. Intuition continues to suggest justification for popularity when considering third highest sales are returned by a product from the tea category. When seeking caffeinated beverages, tea is one popular option which many people drink. This is the most obvious answer for the popularity of the tea product at our coffee shop chain. Once again, the scatterplot referenced above also displays the low price point of this tea product found in third place. It is likely that that the price point is a contributing factor to the sales of this product. Beyond the provided justifications for the chain’s top sellers, nothing more specific or conclusive was found in the available data.

To answer our second question, the typical product price point is approximately ten dollars. The scatterplot confirms the answer to the second question. Yes, these products are sensitive to price point. As evidenced by the graph, products with a lower price point then ten dollars typically boast higher sales. Adversely, products costing more than the ten-dollar benchmark tend to have lower sales. The notable exception to this rule being coffee beans, showing a fair amount of resilience to price.

* *Which location has the highest concentration of sales?*

To resolve this question, it was necessary to use geolocation data for the active coffee shop branches. We then visualized the data using a bubble map. Our process for this analysis involved using Geopandas, Geoplot, Pandas, Shapely, Matplotlib, and OS. In this process the sales receipt and sales outlet data were read in. From these tables, transaction, sales outlet, and coordinate columns were kept, creating the visualization. A map of the New York City boroughs was created using geoplot. Next, sales, quantity sold, and number of orders were aggregated by store location. A data frame with store geolocation data was created and the bubble map was plotted and formatted. Below is the resulting bubble map which displays the sales by active location.



In addition to the bubble map created in Python, we supplement with another map visualization to support a clear, comprehensible view of sales by location, as well as a more specific, close-up view of active store location.

Notice that the more specific map quantifies not only sales, but quantity sold by location.

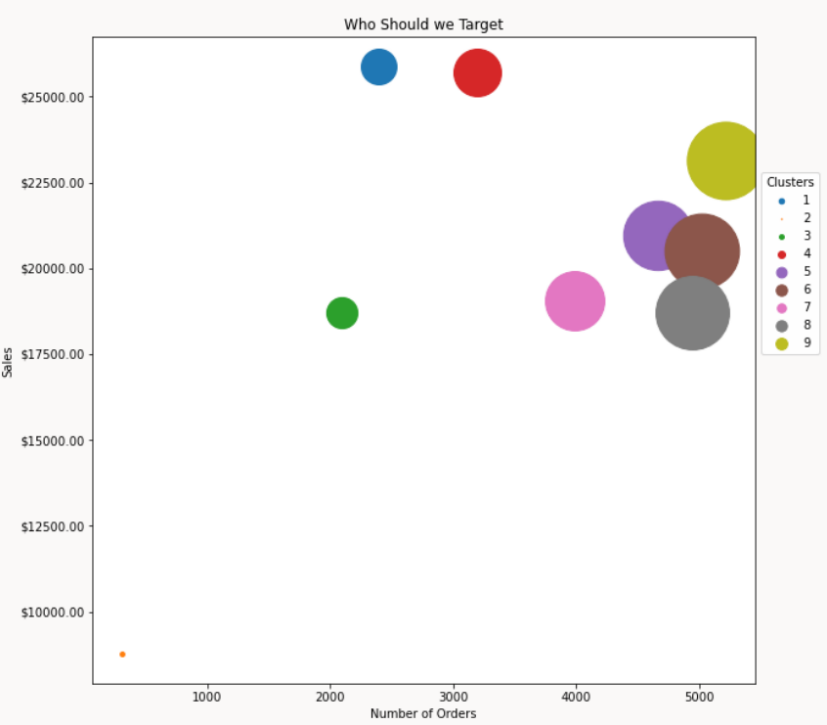
*Interpretation*

Referring to our visualizations, one could see that although the stores exhibit a similar sales performance, the Hell’s Kitchen branch has the highest sales. This is interesting because according to the magnified map, the Manhattan location sold the highest quantity of products. This means that the Hell’s Kitchen location is slightly more successful at selling higher priced items. Also notable about this observation is Manhattan’s high-end reputation. One may predict that the Manhattan location would sell more expensive products. This observation implies that possibly the Hell’s Kitchen branch is putting more of a concerted effort into upselling customers than other stores. Additionally, another inference that could be gleaned from this data is, given the similar sales and quantity metrics of each active location, it is possible that the same performance mistakes that are hindering the coffee shop chain’s success, are happening at all three locations.

* *Who are our target customers in the loyalty program?*
* *Can we predict membership?*

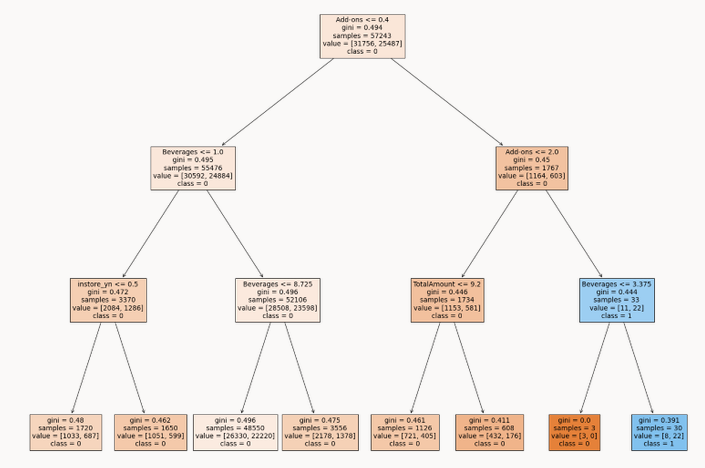
In effort to acquire a wider, more loyal customer base, we focused on determining who our customers were. Furthermore, who should we be targeting in customer acquisition? In answer to this question, we ran a KMeans cluster analysis on our Loyalty members.

The process of this analysis began by reading in the sales receipt csv file into Python. A conditional statement was used to filter out customers who were not members. A weather data set was read in as well so that we could later determine what kind of weather our customers favored when out visiting our shops. In addition, we read in product data to see what our customers were buying. From these datasets, we kept columns concerning transaction data, date, temperature and other weather columns, as well as product columns. Next, the weather data was formatted, as the precipitation column needed to be shifted. This is because the precipitation column quantified the inches of precipitation from the previous day. Dummy columns were then created for sales\_outlet\_id column and the column showing if the customer bought instore or not. The hour value from the time stamp was extracted, as it was the only necessary value, and the others would hinder the KMeans analysis. Once this processing was completed, the data frames were merged, and a pivot table was created to show sales by transaction and product group. An aggregated table was created from the remaining columns not included in the pivot table. They were grouped by transaction ID. Subsequently, the aggregated table was merged with the pivot table and the KMeans algorithm was implemented using Euclidean distance. From the analysis, nine clusters emerged, as seen in the following visualization.



Observe the largest cluster was cluster nine, and the smallest was cluster two. Furthermore, the cluster who spends the least was also cluster two, while the cluster with the highest sales was cluster four.

From the pie chart visualization displaying the low ratio of Loyalty member customers to total customers presented earlier, had questioned if we could predict which customers would choose to become members. To answer this question, we performed a Decision Tree analysis.

To perform this analysis, sales receipt data was read into Python. From there, zeros were inserted into the customer ID column. This is because only Loyalty members had ID numbers. Therefore, if a customer was not a loyalty member, there was no value was assigned to that customer in the customer ID column. So, we filled these empty values with zeros. Next, we discretized this column, converting to a binary variable. If there was a customer ID assigned to the customer in that row, the value was filled with a one. If the there was a zero left from the previous step, it remained. Next, weather and product tables were read into Python so that some of those fields could be used as variables for our analysis. Only necessary columns were kept. These columns included transaction data, sales outlet data, date, precipitation, and wind data, as well as product data. These columns became the variables for our Decision Tree analysis. The precipitation column in the weather data frame was, once again, shifted, and dummy variables were created for the store number column and instore\_yn column. Furthermore, the timestamp column was converted to a datetime data type and only the hour value was kept, as not to hinder the Decision Tree analysis performance. The modified tables were then merged, and a pivot table was created which showed the sales by transaction and product group. Subsequently, earlier merged table was aggregated and merged with the pivot table. Subsequently, the data was split into training and testing data at an 80/20 split. The max\_depth was limited to three to limit the model nodes. This is because without this limit, the model was quite complex. Output of the Decision Tree analysis is seen below.

When considering this model, it is important to note that the data used would have had to contain data describing members, as well as non-member customers. Knowing this, we realized that there was no generation or birthdate data collected on these non-member customers. As a result, we could not use these features as predictors of membership registration. The attributes appropriate for use in this analysis were limited to the transaction and weather data described in the above process, for these customers. We continued with the analysis, as we predicted that we may see a relationship between customers with higher sales and Loyalty members.

*Interpretations*

**KMeans**

The KMeans cluster analysis led us to choose three target clusters of loyalty customers.

\*Note: *Cluster numbers may differ from the presentation, as numbering changes with each run of the K-Means algorithm*

**Cluster 1 (The equivalent of cluster 8 in the presentation):**

Our reasoning behind the focus on cluster one because their top sales are representative of the top sales across the nine clusters of customers. Additionally, this cluster was composed primarily of Millenials and some Baby Boomers. They preferred plain beverages and bakery goods. This cluster was more likely to buy at the Astoria location in the late afternoon between 2:00pm and 5:00pm.

**Cluster 4 (The equivalent of cluster 9 in the presentation):**

Our justification for this choice was the quantity of sales. This cluster had the highest total sales. Most people in this cluster were born after 1989. These younger customers composed the bulk of this cluster. These customers preferred specialty drinks and bakery items. This cluster primarily bought in the late morning between 8:00 and 10:00am. This cluster tended to respond to promotions and sales.

**Cluster 2 (The equivalent of cluster 4 in the presentation):**

When referring to the bubble map above, cluster 2 may seem a counter-intuitive choice to target for advertising. It is the smallest cluster with the least sales. It may be tempting to target a cluster that has already shown some response to the product. However, we observed that this cluster is the one that is most likely to buy branded goods. This group is largely composed of younger people born after 1989. This cluster prefers to shop at our commercial locations in the late morning between 8:00am and 10:00am. These locations may be in their workplace vicinity, and it is possible that these morning hours are the start of their workday.

**Decision Tree Interpretation**

The extensive depth of the Decision Tree, combined with the low accuracy on the testing set, suggested that the model was overfitted to the training data. In addition, there was no statistically significant relationship between the buying patterns of the customers and membership registration. From this analysis, we concluded that more data needs to be collected on non-members. This information should include generation, birthdate, and emails/contact information. In the available data, it is important to mention that not even contact information for members had been collected. Furthermore, the fact that magnitude of customer sales is not a significant predictor of membership registration suggested that customers who were spending the most at our stores were unaware of customer loyalty program. This also suggests that sales staff should be suggesting membership with every sale.

**Conclusion**

To address which products the business should target advertising resource appropriation, we performed the correlation analysis between sales and price point. We found that our product sales were indeed influenced by price point, the outlier being coffee beans. As this product sold at a high price point relative to the rest of the products, it showed resiliency to price point. As a result, we concluded that coffee beans should be a target for advertising. In addition, we observed that although the ten-dollar benchmark that was the average product price point, sales were quite low at this price. In fact, as evidenced by the Price v April Sales scatterplot presented earlier, one can see that beyond the five-dollar mark, sales drop dramatically. Therefore, we suggest targeting some of the higher-priced items within the highest performing product categories for advertising in hopes of closing the wide gap between average price point of sales and average product price point. Our final suggestion addressing the focus on products for advertising include investing in professional development for staff, training them to upsell these targeted products in the store.

When considering which locations to target advertising, we used geolocation data from our active store locations to create a bubble map displaying sales by location. We found from our analysis that store eight boasted the highest sales and the second highest sales amount per item sold. We found it interesting that despite Manhattan’s reputation for being a posh location, not only did this location experience the lowest sales performance, but also the lowest sales amount per items sold. This was our justification for choosing store five (Manhattan) as the primary focus of advertising resources. It should be noted, however, that the sales and quantity data was similar among all the active locations. This suggests that all three stores require additional advertising support.

Finally, the question of who the focus of advertising should be was answered using the KMeans clustering algorithm, as well the Decision Tree model. From the KMeans analysis we learned that our customers fit into nine clusters. We concluded that clusters one, two, and four should receive primary focus of advertising resources. Because our customer clusters have different attributes, we suggest that advertising be differentiated. Avenues for differentiated advertising could be emails, texts, or the investment in an app for the coffee shop chain. Any of these channels for advertising would target the interests and preferences of each cluster. Additionally, contact information needs to be collected on all customers for this reason. To encourage customers to give their contact information, we suggest possibly collecting it in the form of a short survey in exchange for a free incentive, such as a free coffee or pastry item or branded item, etc. The coffee shop chain may find investing in a new POS (point of sale) system helpful for housing the data.

Our Decision Tree analysis implied that if we had more data on all customers, perhaps a more efficient prediction model would be possible. More data would mean more variables for prediction and a better chance that a significant relationship would be found between these new variables and Loyalty membership. In addition, we found a bit counter-intuitive that magnitude of customer spending was not a predictor of membership. We interpreted this to mean that customers may be unaware of the Loyalty membership program. We therefore suggest that advertising be directed at customer loyalty as well. Sales staff should be suggesting membership with every sale, and exclusive promotions and sales should be directed at Loyalty members to incentivize registration.

**Limitations of our Analyses**

Our coffee shop chain sales data was a collection of datasets describing a fictitious business. This fact may have made some of our analyses more difficult. For example, the sales data per store location was close to an even distribution among the three stores. Another example of this limitation was seen while conducting the Decision Tree analysis. The lack of data for the “non-member customers,” made this analysis difficult. Despite the limitations, we made an effort to continue with each analysis as though we had actual data. We then interpreted the results as though it were real-world data and assigned the same business conclusions to the results as we would if we saw this with actual data in the workplace.

**Team Member Roles**

Adit Shah and I are the team members who worked on this project. We both proposed options for data sets to work with, then eventually settled on the data sets describing the coffee shop chain.

**Adit Shah’s Tasks**

Adit ran the KMeans analysis in Python

The correlation between price and sales in Python

Decision Tree analysis in Python.

Wrote the Analysis section of the Project Proposal

Helped make modifications to the Final Project Presentation and Project Report

**Keeley Ables’ Tasks**

Wrote the rest of the Project Proposal

Created new data frame in Python that modified fields from several imported tables, merged them, and aggregated them by transaction ID, then summarized the new data frame.

Worked on Decision Tree Analysis

Created Presentation

Wrote the Final Project Report

**Joint Tasks**

Collaborated through email and Teams on nearly a daily basis.

Made decisions about which analyses to use

**Appendix**

**Sales Receipt Data**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| transaction\_id | Id number for each sale | 12 |
| transaction\_date | Date of sale | 4/27/2019 |
| transaction\_time | Time of the sale | 9:53:55 |
| sales\_outlet\_id | Id number of the coffee chain location | 8 |
| staff\_id | Staff Id of person that made the sale | 42 |
| customer\_id | Loyalty member number of purchaser, if any | 8232 |
| Instore\_yn | Instore purchase? Y/N | N |
| order | Order number | 18 |
| product\_id | Id of product sold | 38 |
| quantity | Number of each product sold | 2 |
| line\_item | Amount charged by transaction and product id | 7.5 |
| Unit\_price | Amount charged for one unit of product | 3.75 |
| Promo\_item\_yn | Is the item a promotional item? Y/N | Y |

**Customer Loyalty Membership Data**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| customer\_id | Loyalty member number of customer | 3001 |
| home\_store | Location where the customer registered for a loyalty membership | 3 |
| customer\_name | First/Last name of loyalty member | Kelly Key |
| customer\_email | Email of loyalty member | Ina@non.gov |
| Customer\_since | Date customer registered to be a loyalty member | 1/4/2017 |
| loyalty\_card\_number | Id on customer loyalty membership | 908-424-2890 |
| birthdate | Customer loyalty member’s date of birth | 5/29/1950 |
| gender | Gender of customer loyalty member | M/F |
| birth\_year | Birth year only for possible analysis purposes | 1950 |

**Daily Sales Targets**

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| sales\_outlet\_id | Id number of the coffee chain location | 3 |
| transaction\_date | Date of sales | 4/1/2019 |
| beverage\_target | Goal for beverage sales by date | 1800 |
| food\_target | Goal for food sales by date | 456 |
| merchandise | Goal for merchandise sales by date | 48 |
| beans\_target | Goal for coffee bean sales by date | 96 |
| daily\_target | Goal for daily total sales | 2400 |

****

**Dates**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| transaction\_date | Date of sales | 4/28/2019 |
| Date\_ID | ID number to distinguish unique dates | 20190401 |
| Week\_ID | ID number unique for each week | 14 |
| Week\_Desc | Week number | Week 14 |
| Month\_ID | ID number unique for each month | 4 |
| Month\_Name | Name of the particular month | April |
| Quarter\_ID | ID number for the quarter | 3 |
| Quarter\_Name | Gives the quarter and year | 2Q19 |
| Year\_ID | Year only for possible analysis purposes | 2019 |
| Day\_of\_Month | Day of month only for possible analysis purposes | 1 |
| Day\_of\_Week | Day of week only for possible analysis purposes | Monday |

**Generations**



|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| birth\_year | Birth year of at least one customer loyalty member | 1946 |
| generation | Generation description of at least one customer loyalty member | Baby Boomers |

**Weather**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| date | Month/day/year of weather recording | 4/1/2019 |
| Max\_temp | Temperature high by day in degrees Fahrenheit | 8 |
| Min\_temp | Temperature low by day in degrees Fahrenheit | 1 |
| Avg\_temp | Average temperature for that day in degrees Fahrenheit | 4.5 |
| Precip\_inches\_yesterday | Amount of precipitation in inches from the previous day | 0 |
| wind\_mph\_highest | Highest wind speed of the day in miles per hour | 23 |
| wind\_mph\_avg | Average wind speed of the day in miles per hour | 4.8 |

**Pastry Inventory and Spoilage**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| sales\_outlet\_id | Store number | 3 |
| transaction\_date | Date of sales | 4/26/2019 |
| product\_id | ID specific to each product | 70 |
| start\_of\_day | Amount of inventory kept in the store at the start of the day, by product | 18 |
| quantity\_sold | Amount sold by day and product | 10 |
| waste | Spoilage-Amount of inventory not sold and thrown out by day and product | 8 |
| % waste | Percentage of inventory wasted | 44% |

**Product**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| product\_id | ID specific to each product | 26 |
| product\_group | General type of product | Whole Bean/Teas |
| product\_category | Type of product | Coffee beans |
| product\_type | More specific type | Organic beans |
| product | Name of product | Our Old Time Diner Blend |
| product\_description | Product slogan | It's like Carnival in a cup. Clean and smooth. |
| unit\_of\_measure | Unit of measure a product is stocked by | 12 oz |
| current\_wholesale\_price | Amount stores would pay to buy products for their shelves | 14.4 |
| current\_retail\_price | Amount coffee chain sells product to customers | 18 |
| tax\_exempt\_yn | Is the product taxed? Y/N | N |
| promo\_yn | Is the product a promotional product? Y/N | Y |
| new\_product\_yn | Is the product new? Y/N | N |

**Sales Outlet Data**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| sales\_outlet\_id | Store number | 2 |
| sales\_outlet\_type | Store type | retail |
| store\_square\_feet | Square footage of the store location | 3400 |
| store\_address | Building number and street of store location | 164-14 Jamaica Ave |
| store\_city | City of store location | Brooklyn |
| store\_city\_province | State where the store is located | NY |
| store\_telephone | Phone number for the store location | 972-871-0402 |
| store\_postal\_code | Zipcode by store location | 11106 |
| store\_longitude | Coordinates by store location | -73.795168 |
| store\_latitude | Coordinates by store location | 40.761196 |
| Neighorhood | Neighborhood by store location | Jamaica |
| active\_yn | Is this an active location? Y/N | Y |

**Staff Data**

****

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Example** |
| staff\_id | ID number unique to each staff member | 10 |
| first\_name | First name of staff member | Sue |
| last\_name | Last name of staff member | Tindale |
| postion | Job position of staff member | CFO |
| start\_date | Date of the staff member’s start of employment | 8/3/2001 |
| location | Store number or building of the location | HQ (Headquarters)  WH (Warehouse)  3 (Store number) |