**Introduction**

The primary objective of this project is to create a set of marketing strategies that the coffee shop can implement to attract potential customers and encourage them to make purchases. The dataset provided contains retail data from a coffee chain with three New York City locations.

**Research Topics**

The research topics I investigated are as follows:

1. forecasting sales
2. predicting the interaction of customers' age and gender on sales
3. measuring the magnitude of change in the quantity due to the change in price
4. predicting the variables that affect customers decision to order online versus in store purchase and
5. identify and select groups of potential customers

**Data Cleaning**

In order to conduct my analysis, I first cleaned the data by dropping null values, replacing blank spaces in variables with hyphens, dropped the unnecessary variables and unnecessary columns etc. I have also standardized some required variables.

**Method**

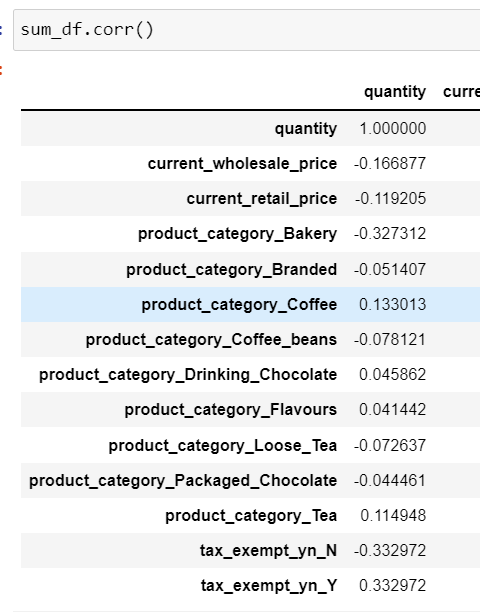
The methods that I have used for the analyses are:

1. to get an initial understanding of the data, I conducted some explanatory analysis mainly evolving around the target variable ‘quantity’ sold in the coffee shop.
2. for forecasting sales, I have used OLS, and KNN.
3. for predicting the interaction effect of gender and age on sales I have used OLS with interaction term
4. for measuring the magnitude of quantity change due to a change in price I have used Price Elasticity method
5. for predicting the variables that affect customers decision to purchase in store I have used Decision Tree, Logistic Regression, SVM, Random Forest and finally
6. I have used Clustering Analysis for customer segmentation.

**Analysis Results and Interpretation**

Explanatory Analysis

* **Positive Correlations**:
  + Coffee: There’s a positive correlation between coffee and quantity sold. Customers seem to purchase more when coffee is involved.
  + Tea: Similarly, tea also shows a positive correlation with quantity sold.
  + Tax Exemption: The presence of tax exemption positively impacts sales.
* **Negative Correlation**:
  + Retail Price: The negative correlation between retail price and quantity sold suggests that higher prices may lead to lower sales.

****

Next, my group by analysis revealed that:

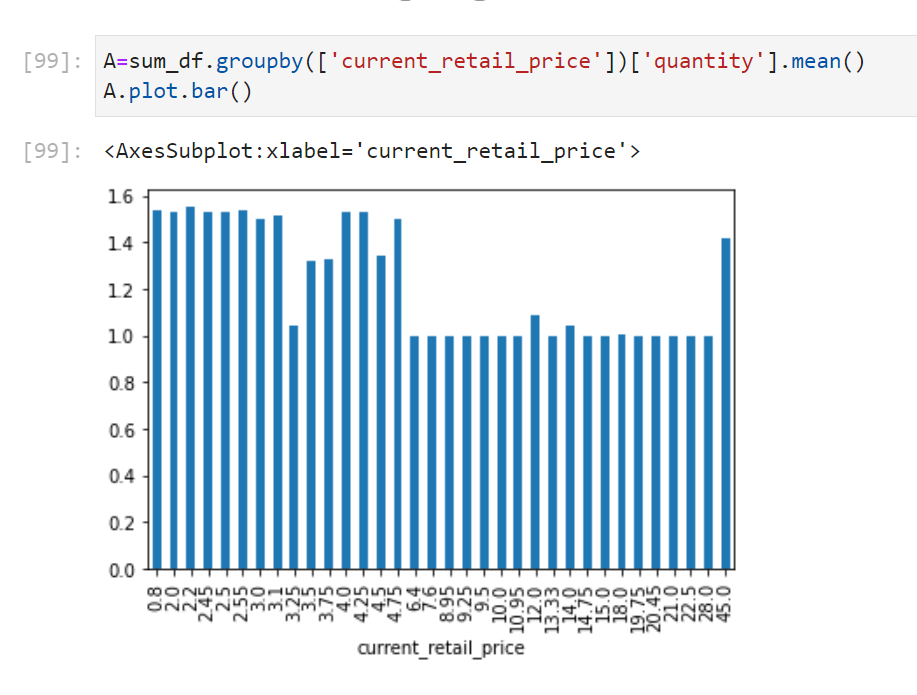
* **Top-Selling Product Categories**:
  + Coffee
  + Drinking chocolate
  + Various flavors
  + Tea
* **Tax Exemption Impact**:
  + When the store receives tax exemption, it appears to positively influence the quantity sold.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

* **Price Range and Quantity Sold**:
  + Between $0.8 and $4.5, the average quantity sold is around 15 units.
  + As the price increases beyond this range, the quantity sold decreases.
* **Exception: Civet Coffee**:
  + At a price of $45, there’s a sudden increase in quantity sold.
  + This unique case is due to Civet Coffee, considered the most expensive coffee globally, extracted from the faces of civet cats. The fermentation process during digestion results in a distinct and rich flavor.



Linear Regression & KNN

OLS Regression Equation:

*Quantity = β0 + β1current\_retail\_price + β2 Bakery + β3 Branded + β4 Coffee\_beans + β5 Drinking\_Chocolate + β6 Flavors + β7 Loose\_Tea + β8 Packaged\_Chocolate + β9 Tea + β10  tax\_exempt\_yn\_Y.*

From the linear regression model, I found:

* Significant variables: All except drinking chocolate and tea.
* R-squared value: Explains 13% of variance in quantity sold.
* Top product categories: Bakery, branded products, and coffee.
* Tax exemption increases sales.
* A screenshot of a computer

  Description automatically generatedPrice change has a negligible effect.

My KNN analysis provided the following insights:

* **Optimal k-values**: Based on your analysis, the optimal values of k were 2 and 6 (as indicated in Appendix G).
* **Test Accuracy**: For both optimal k-values, the test accuracy score was 0.58.
* **Chosen k**: Ultimately, you selected k=2.
* **Overfitting**: The negligible difference between training and test accuracy scores suggests that there are no overfitting issues in the model.

A screenshot of a computer code

Description automatically generated

**OLS with Interaction Terms**

*quantity\_x = β0 + β1gender+ β2 birth\_year + β3 gender\*birth\_year + e*

My analysis suggests that:

* **Statistical Significance**:
  + Gender and the interaction of gender with birth year are both statistically significant (p-value of 0.00).
  + Birth year, although not statistically significant (p-value of 0.678), should still be included due to its impact on the interaction term.
* **Gender Differences**:
  + Male customers purchase around 7 units more than female customers on average.
  + However, as male customers age, their purchasing behavior decreases compared to female customers.
* **Age and Gender Interaction**:
  + Female customers tend to buy more as they age.
  + The positive effect of gender on quantity sold is reduced for older male customers.

A screenshot of a computer

Description automatically generated

**Price Elasticity**

For Price elasticity, I took the log of unit\_price and log of quantity from the file ‘Sales\_Receipt’ and performed price elasticity. The model is:

*lnQuantity\_x = β0 + β1lnunit\_price + e.*

 **R-squared Value**: An R-squared value of 0.023 indicates that only 2% of the variance in the model is explained by the independent variable.

 **Significant Coefficient**:

* The coefficient of ‘lnunit\_price’ is statistically significant (p-value of 0.00).
* With a one-unit increase in the natural logarithm of price, the log of quantity sold decreases by 0.124 units.

 **Elastic Demand**:

* Given that the quantity response is less than half the price increase, you indeed have an elastic demand.
* Elastic demand means that changes in price significantly impact the quantity sold.

A screenshot of a computer

Description automatically generated

**Decision Tree, Logit, SVM, & Random Forest**

The key findings are:

1. **Logistic Regression Model**:

*takeout\_order = β0 + β1quantity + β2* unit\_price*+ β3*birth\_year*+ β4*home\_store\_dum\_5 *+ β5*home\_store\_dum\_8 *+ β6*gender\_M *+ β7*generation\_Gen\_X *+ β8*generation\_Gen\_Z*+ β9generation\_Gen\_Z + β10*generation\_Older\_Millennials *+ β11generation\_Younger\_Millennials + β12product\_category\_Bakery + β13product\_category\_Branded+β14product\_category\_Coffee\_beans+β15product\_category\_Drinking\_Chocolate+β16product\_category\_Flavmys+β17product\_category\_Loose\_Tea+β18product\_category\_Packaged\_Chocolate + β19product\_category\_Tea + β20* tax\_exempt\_yn\_Y *+ e*

* + The model is statistically significant (LLR p-value < 1).
  + However, none of the coefficients have statistically significant p-values.
  + Despite this, you’ve included ‘birth\_year’ to maintain the meaning of the interaction.

A screenshot of a computer

Description automatically generated

1. **Decision Tree Model:**
   * The most important features affecting customers’ in-store purchase decisions are ‘birth\_year’ and ‘product\_category\_Coffee’.
   * These features play a significant role in predicting customer behavior.

A diagram of a network

Description automatically generated

1. **Model Evaluation**:
   * Negligible differences exist between training and test accuracy scores for all four models.
   * Intra-model overfitting is not a concern.
   * Based on test accuracy scores:
     + Decision Tree: 0.509
     + Logit: 0.513
     + SVM: 0.503
     + Random Forest: 0.509
2. **Inter-Model Evaluation**:
   * Precision, recall, and specificity scores vary across models:
     + Decision Tree: Better at predicting positive cases.
     + Logit: Equally good for both positive and negative cases.
     + Random Forest: Better at predicting negative cases.
     + SVM: Good at predicting positive cases but not negative cases.
   * Considering overall performance, Random Forest stands out:
     + Highest precision score (0.511)
     + Highest recall score (0.603)
     + Moderate specificity (0.3022)

**Clustering**

1. **Elbow Method and Clusters**:
   * I used the elbow method to determine the optimal number of clusters.
   * Based on less error, you performed a 5-cluster k-means analysis using random state 10.
2. **Generation and In-Store Purchases**:
   * Cluster 4 had the highest in-store purchases (mean of 0.04).
   * Gen Z customers (birth year mean of 2.65) were prominent in this cluster.
   * Cluster 1 (negative mean for in-store purchases) included Gen X (mean of 1.59).
3. **Top Clusters by Generation**:
   * Top 3 clusters for generation: 0, 1, and 2.
   * Cluster 0 (Gen X) showed a mix of in-store and online purchases, with top products 34, 29, and 36.
4. **Other Clusters**:
   * Cluster 1 (Younger Millennials) leaned more toward online purchases, with top products 74, 36, and 27.
   * Cluster 2 (Baby Boomers) had a mix of online and in-store purchases.
5. **Gen Z Potential Market**:
   * Cluster 4 (Gen Z) stood out as the top in-store customer group.
   * Top products in this cluster: 23, 38, and 57.
   * Both male and female customers were prominent.

A screen shot of a graph

Description automatically generated

***Managerial Implications***

1. **Targeted Promotions**:
   * Focus on bakery products, branded products, and coffee in advertisements.
   * Consider showcasing an aged female customer ordering coffee and cake through the drive-through, alongside a young group of male customers enjoying coffee and branded products in-store.
2. **Store Interiors**:
   * Design the interiors with a 1990s theme, catering to customers born in that era.
   * This aligns with the decision tree results, indicating that these customers prefer in-store purchases.
3. **Gen Z Focus**:
   * Pay attention to Gen Z customers.
   * Consider opening a location near a college campus to attract this generation, known for in-store purchases.
4. **Promotion Strategies**:
   * Employ targeted promotions without significant price reductions.

Conclusion

Thus, to conclude the coffee shop needs to focus more on its young male and female, old female customers and employ more promotion strategies to bring in more customers without needing to reduce the price of its products significantly.