How Do Consumer Spending Patterns Differ Across New York City Coffee Shops? A Statistical Evaluation of Transaction Patterns, Pricing, and Product variation.

Adriana Cole, Mike Maeda, Mariia (Mariam) Swedan, Johnston Brandon, Duru Katranci

AU College of Business, Alfred University

1 Saxon Drive, Alfred, NY 14802

Course Number: BUSI113-01

Sengupta Ayush

December 11, 2025

Abstract

In this study, we empirically examine consumer spending behavior across three New York City coffee shop locations: Astoria, Hell's Kitchen, and Lower Manhattan. Using a transactional dataset from January 2023, we analyze whether transaction value, pricing, time of day, and product category vary across locations. The analysis is conducted using one-way and two-way analyses of variance, linear regression, and chi-square tests of independence. Results show that average transaction values do not differ significantly across locations or time periods, while product category preferences vary by store location. These findings suggest that although overall spending behavior is stable, location plays an important role in shaping consumption preferences.

1. **Introduction**

Coffee is one of the most widely consumed beverages in the world, (Organization, 2023) influencing daily routines, retail markets, and global trade flows. Whether through small community kiosks or massive business-to-business supply lines, coffee is woven into the financial and cultural fibers of societies around the world. The coffee sector has been characterized in recent years by steep growth in consumer demand and substantial disruption, prompted by shifts in consumer preferences, technological innovation in the retail sector and growing interest among consumers for gourmet products. At the retail level, coffee shops collect vast quantities of transactional data, capturing information about purchasing patterns such as product choice, time of purchase, pricing strategies, and store-specific characteristics. (Research, 2021) These data are an invaluable source of understanding of the coffee market, at both an operational and consumer-behavior level.

Our dataset, "Copy of Coffee Shop Sales.xlsx," (Kaggle, 2023) provides detailed product-level sales records from multiple store locations. It includes variables such as transaction date and time, quantity purchased, store location, unit price, product category (such as coffee, tea, or drinking chocolate), product type, and specific product details. These variables allow us to examine micro-behavioral patterns: for example, identifying which beverages are most popular at certain times of day, how pricing interacts with product mix to influence sales, or how store location shapes consumer preferences (National Coffee Data Trends 2023 Report., 2023). Such transactional data helps reveal operational challenges that cafés face in managing inventory, setting prices, tailoring product offerings, and predicting customer flow across different regions.

This dataset (Kaggle, 2023) also offers the opportunity to explore broader consumer behavior trends that have been highlighted in previous research, such as the rising demand for specialty and gourmet coffee, the persistence of regional variation in preferences for tea or chocolate beverages, and the importance of product differentiation in retail performance. While existing literature has examined consumer behavior and pricing strategies in coffee shops, many studies rely on limited or aggregated data. The detailed granularity of this data set provides a unique opportunity to conduct more nuanced empirical analysis of purchasing decisions and store-level performance.

Given the diversity of findings in the existing literature on consumer behavior and retail operations, this research is important because it provides a focused empirical investigation using comprehensive transactional data. By analyzing this dataset, we can identify factors that shape individual purchasing decisions, understand variation in sales performance across store locations, and evaluate how product mix and pricing strategies influence revenue outcomes. Earlier studies have shown that consumer preferences depend on temporal patterns such as morning vs. afternoon demand, while other research has emphasized the role of product variety and store characteristics in shaping customer behavior. (Wedel, 2016), However, many of these insights remain limited due to a lack of detailed point-of-sale data, underscoring the need for more granular analysis.

This project aims to address that gap through an empirical study of daily retail-level consumption patterns using the "Coffee Shop Sales" dataset. (Kaggle, 2023) The results of this analysis will provide valuable viewpoints into how cafés might optimize operational decision-making, how retail managers may adapt product offerings according to customers, and how price strategies can be aligned with customer behavior. Above all, this research emphasizes the importance of micro-level transactional data in understanding the complexities of consumer behavior in the modern coffee retail environment.

In Section 2, we present the dataset discussing its structure, sources and key variables. Section 3 describes the statistical and analytical models that were used for studying dependencies in the data. Section 4 discusses empirical results, model validation, and related interpretations. Ultimately, in Section 5 we sum up our findings and discuss implications for industry stakeholders and policymakers, as well as future research.

2. **Data**

The dataset being used in this project contains data found from Kaggle on daily coffee transactions at three different locations in New York City. (Kaggle, 2023) The website also has information on a wide variety of topics with their datasets on coffee sales varying. This dataset measures details on transaction information, the type of product, and the unit price of the product. We collected data from January 1st, 2023, to January 31st, 2023, (Kaggle, 2023) and split the information up based on the location of Hell's Kitchen, Astoria, and Lower Manhattan. The location is our unit of analysis, and our dependent factors are the unit price, product type, and transactions. Prior to testing our research, we conducted simple statistical tests to measure things such as averages of product categories, unit price, and transaction quantity. We then created boxplots to see if there were any outliers in the data that could potentially skew our data, we did this for the unit price based on location:

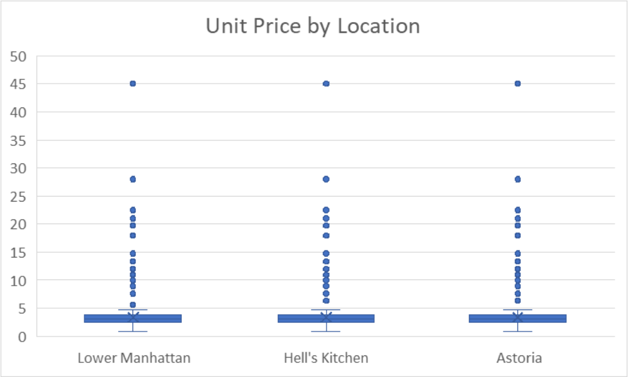

Figure 1: Distribution of Unit Prices by Store Location

We also created a boxplot for our data on transaction quantity to see if there were any outliers that could potentially skew this data, as well as visually showing our data distribution:

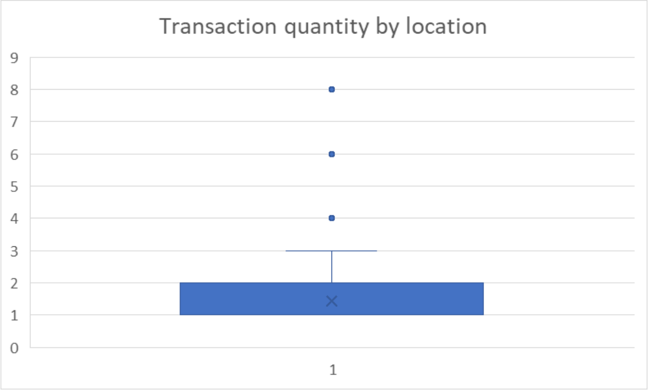

Figure 2: Distribution of Transaction Quantity by Store Location

As we can see in both boxplots, there are outliers which could be due to things such as bulk or specialty orders.

3. **Models**

The models used in this project are designed to analyze how transaction values and purchasing patterns vary across store locations, time-of-day periods, and product categories. Because the dataset contains detailed point-of-sale records from three New York City café locations, we define the key variables before presenting the formal models. Let Yi represent the transaction value for transaction i. Each transaction is linked to a categorical location variable with three levels: Astoria, Hell’s Kitchen, and Lower Manhattan. Time of day is recorded into three blocks, Morning, Afternoon, and Evening, based on the recorded transaction timestamp. Additional variables include the product category purchased, the numerical calendar day of the month, and aggregated daily values such as total revenue and number of transactions per day for each store. These variables allow us to examine both individual level purchasing behavior and broader daily performance trends.

The first model investigates whether the average transaction value differs across the three store locations. To test this, we apply a one-way analysis of variance (ANOVA) (Anderson, 2020) using location as the

factor. The model takes the form

$$Yij = \mu + \tau j + \varepsilon ij,$$

Where μ is the overall mean transaction value, τj is the effect of location j, and εij is the error term. This model assesses whether the three neighborhoods serve customers with meaningfully different spending patterns. We assume the usual ANOVA conditions used in class, including independent transaction observations and reasonably similar variation across groups, which is standard for this type of retail data.

The second model extends the analysis by jointly examining the influence of location and time of day on transaction value. Here we use a two-way ANOVA with interaction to evaluate whether purchasing behavior depends not only on where a store is located but also on when customers visit. The model is written as

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk,}$$

where α_j represents the effect of location, β_k represents the effect of time block, and $(\alpha\beta)_{jk}$ captures whether the pattern of morning, afternoon, and evening spending differs across neighborhoods. This model mirrors techniques used in class for examining two categorical factors simultaneously. We treat the usual class assumptions as reasonable for this context, including the idea that transactions within a location block are comparable.

To explore differences in product mix across stores, we also use a chi-square test of independence. We construct a contingency table of product categories by store location and test whether the distribution of categories is independent of the neighborhood. This model helps determine whether certain stores sell disproportionately more tea, chocolate beverages, or specialty coffees compared to others. The interpretation follows the same reasoning as the chi-square examples used in class, relying on counts of observed versus expected frequencies.

To connect individual transactions to store-level performance, we aggregate the data and model the relationship between daily revenue and daily transaction count using simple linear regression. For each store s and day $d,$ let $R_{d,s}$ denote the total daily revenue and $T_{d,s}$ denote the number of transactions. The regression model

$$R_{d,s} = \beta_0 + \beta_1 T_{d,s} + \varepsilon_{d,s}$$

allows us to estimate the average revenue generated per additional transaction. This approach follows the linear regression methods taught in class and provides insight into whether increases in foot traffic produce proportional increases in revenue. Finally, to examine whether sales changed over the course of January, we introduce a simple trend model relating daily revenue to the day of the month.

Let $DayIndex_d$ represent the numerical day from 1 to 31. The model below captures whether revenue tended to rise or fall over time.

$$R_{d,s} = \gamma 0 + \gamma 1 DayIndex_d + U_{d,s}$$

A positive $\gamma 1$ indicates an upward trend, while a negative value reflects declining sales. This time-trend model parallels the simple regression frameworks used earlier but applies them to sequential daily data. Across all models, categorical variables are incorporated using indicator variables when needed. For example, if Astoria serves as the baseline location, then Hell's Kitchen and Lower Manhattan are represented using dummy variables in the regression equations.

4. Analysis

Using one-way Anova:

In this analysis, a one-way ANOVA (Anderson, 2020)was conducted to examine whether average transaction value differs significantly across the three NYC Starbucks locations included in the dataset: Astoria, Hell's Kitchen, and Lower Manhattan. (Kaggle, 2023) Here, the independent variable is Store Location with three levels (the three stores), and the dependent variable is Transaction Value, the amount spent on each individual customer transaction. ANOVA is appropriate because it compares mean differences across more than two groups while accounting for within-group variability. The goal is to determine whether any observed differences in average spending between the stores are statistically meaningful or simply due to random variation.

Hypotheses:

Null Hypothesis (H_0):

There is no significant difference in mean transaction value across the three store locations.

$$H_0 = \mu_{Astoria} = \mu_{Hells\ Kitchen} = \mu_{Lower\ Manhattan}$$

Alternative Hypothesis (H₁):

At least one store location has a different mean transaction value.

$$H_a: At\ least\ one\ \mu\ differes$$

Dataset Used: For each store location, all transaction values in the dataset were extracted:

- Astoria → 5913 observations
- Hell's Kitchen → 5868 observations
- Lower Manhattan → 5533 observations

This produced three independent samples representing customer spending behavior.

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Astoria	5913	27313.66	4.619256	9.420251
Hell's Kitchen	5868	27820.65	4.741079	52.71639
Lower Manhattan	5533	26543.43	4.797294	10.7707

Table 1:Consumer spending behavior

The average transaction values across stores appear close (≈ 4.6–4.8), but Hell's Kitchen shows much larger variance, meaning purchase amounts fluctuate more at that location.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	95.56345332	2	47.78173	1.948237	0.142556	2.996251
Within Groups	424563.1024	17311	24.52563			
Total	424658.6658	17313				

Table 2:One-way ANOVA comparing mean transaction value across store locations.

Now, since p = 0.1426 > 0.05, We **fail to reject the null hypothesis**.

There is no statistically significant difference in the mean transaction value between Astoria, Hell's Kitchen, and Lower Manhattan. Although the stores have slightly different average spending values, these differences are not large enough to conclude that location impacts on how much customers spend per transaction. In real terms, Customer spending behavior is consistent across all three New York City Starbucks locations in the dataset (Kaggle, 2023)

Using Two Way Anova

A two-way ANOVA was used in this analysis to investigate whether the location of the coffee shop: Astoria, Hell’s Kitchen, and Lower Manhattan, and time of day: Morning: 6am-11:59am, Afternoon: 12pm-5:59pm, Evening: 1pm-8:59pm, influenced the unit price of a product, along with consumer spending. This model was also used to analyze if transaction quantity or the number of products consumers bought were affected by these factors. A two-way ANOVA is appropriate for this analysis as it is typically used when there are two independent variables which are the location and time, and one dependent variable, with unit price and transaction quantity being the two separate variables. This analysis is attempting to prove if there is a correlation between independent and dependent factors. We are also using two-way ANOVA without replication as there is only one value per cell.

Hypothesis:

Null Hypothesis (H_0):

There is no difference in mean transaction quantity across the three store locations

H0A : μAstoria =μHell’s Kitchen =μLower Manhattan

Alternative Hypothesis (H_1):

There is a difference in mean transaction quantity in at least one store location

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Astoria	3	4.262434	1.420811	0.000401
Hell's Kitchen	3	4.328112	1.442704	0.000572
Lower Manhattan	3	4.394055	1.464685	0.000601
Morning	3	4.330817	1.443606	0.000878

Afternoon	3	4.35582	1.45194	0.001651
Evening	3	4.297964	1.432655	0.000208

Table 1: Transaction quantity by location and time of day

The mean transaction quantity across the three store locations does not equal each other, and the p-value is equivalent to 0.22. This means that $p < \alpha$, or $0.22 < 0.05$ meaning we reject the null hypothesis and there is no statistically significant difference in means. Looking at this data, we can interpret that on average, there are more transactions in the afternoon. Knowing that the afternoon is technically the largest time frame in this analysis, this comes to no surprise.

Null Hypothesis (H_0):

There is no difference in mean unit prices across the three store locations

$$\boldsymbol{H_0}\textbf{A: }\boldsymbol{\mu}\textbf{Astoria =}\boldsymbol{\mu}\textbf{HK =}\boldsymbol{\mu}\textbf{LM}$$

Alternative Hypothesis (H_1):

There is a difference in unit price in at least one store location

$$\boldsymbol{H_1}\textbf{A: At least one }\boldsymbol{\mu}\textbf{ differs}$$

Null Hypothesis (H_0):

There is no difference in mean unit prices throughout the day

$$\textbf{H0B: }\boldsymbol{\mu}\textbf{Morning =}\boldsymbol{\mu}\textbf{Afternoon =}\boldsymbol{\mu}\textbf{Evening}$$

Alternative Hypothesis (H_1):

There is a difference in unit price in at least one time of the day

H1B: At least one μ differs

Summary	Count	Sum	Average	Variance
Astoria	**3**	10.24182	3.413939	0.001143
Hell's Kitchen	**3**	10.26791	3.422637	0.000285
Lower Manhattan	**3**	9.989992	3.329997	0.011914
Morning	**3**	10.25621	3.418737	0.001371
Afternoon	**3**	10.18098	3.393661	0.003202

Evening	**3**	10.06253	3.354175	0.013443

Table 2: Unit price by location and time of day

The mean unit price across the three-store locations does not equal each other. The p-value is equivalent to 0.32, which means p< α meaning we reject the null hypothesis. This also means that there is no statistically significant difference between the unit prices. As for the time of day, the means still do not equal each other meaning we reject the null hypothesis again. Looking at this data, it appears that although unit price does not vary much across location or time of day, the amount people spend on items tends to peak in the afternoon and looking at Table 1, we can see that this is also when the transaction quantity also peaks showing a correlation between the two. We can also make the connection that since the average transaction quantity is highest in Hell's Kitchen, it is reasonable to believe that this is connected to the fact that the unit price is higher in this location as well.

Chi – Square Test Analysis.

The aim of this analysis is to determine whether product category preferences differ by store location in a retail data set. The dataset contains transaction counts across three stores, Astoria, Hell's Kitchen, and Lower Manhattan, and ten product categories (Bakery, Branded, Coffee, Coffee Beans, Drinking Chocolate, Flavors, Loose Tea, Packaged Chocolate, and Tea).

To test whether product preference is independent of store location, we perform a Chi-Square Test of Independence.

Why?

- The data counts/frequencies, not averages.
- Both variable store location and product category are categorical.
- We want to know if there is a relationship (dependence), not prediction.
- It is the correct test for observed vs. expected frequencies on a contingency table.

Hypotheses.

Null Hypothesis (H₀):

There is no association between store location and product category.

Product sales distribution is the same across all stores.

- Alternative Hypothesis (H₁):

There is an association between store location and product category.

Product sales distribution differs between stores.

Location/product	**Bakery**	**Branded**	**Coffee**	**Coffee beans**	**Drinking Chocolate**	Flavors	**Loose Tea**	**Packaged Chocolate**	**Tea**	**Grand Total**
Astoria	850	42	2321	59	509	180	43	15	1894	5913
Hell's Kitchen	868	11	2339	91	436	248	57	24	1794	5868
Lower Manhattan	925	43	2131	69	388	348	41	18	1570	5533
Grand Total	**2643**	**96**	**6791**	**219**	**1333**	**776**	**141**	**57**	**5258**	**17314**

Data Preparation: A pivot table was created. This produced the Observed Frequency Table (O), the actual counts of products sold in each store-category combination.

Using the expected frequency formula;

$$E_{ij} = \frac{(Row\ Total_i)(Column\ Total_j)}{Grand\ Total}$$

Table 3;Observed Frequencies for Product Category by Store Location

Meaning: Multiply the total sales in a store by the total sales in a category then Divide by total transactions across all stores.

These expected values represent what the counts would look like if the null hypothesis were true

Loc/Pr	**Bakery**	**Branded**	**Coffee**	**Coffee beans**	**Drinking Chocolate**	**Flavours**	**Packaged Chocolate**	**Tea**
Astoria	902.6255631	32.78549	2319.232	74.79190251	455.2402102	265.0161	19.46638558	1795.689

Hell's Kitchen	902.6255631	32.78549	2319.232	74.79190251	455.2402102	265.0161	19.46638558	1795.689
Lower Manhattan	902.6255631	32.78549	2319.232	74.79190251	455.2402102	265.0161	19.46638558	1795.689

Table 4;Expected Frequencies Computed Using the Independence Model

For EACH cell, the contribution to chi-square was made using the equation:

$$X_{ij}^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Then, for all cells, summed all the contributions.

$$X^2 = \Sigma \frac{(O-E)^2}{E}$$

Chi-Square Contributions	**Bakery**	**Branded**	**Coffee**	**Coffee beans**	**Drinking Chocolate**	**Flavours**	**Packaged Chocolate**	**Tea**
Astoria	3.068215667	2.589779	0.001348	3.334374129	6.34854947	27.2728	1.024771656	5.382399
Hell's Kitchen	1.328269075	14.47615	0.168493	3.512444737	0.813165625	1.092561	1.055853928	0.001588
Lower Manhattan	0.554621369	3.182389	15.27716	0.448526291	9.931560901	25.9846	0.110461527	28.36538

The total chi-square value is (X^2) is 158.5649

Having the degrees of freedom.

$$df = (r-1)(c-1) = (3-1)(10-1) = 18$$

Applying CHISQ.DIST. RT (158.5649,18) to find the P- value,

Table 5;Chi-Square Cell Contributions for Store Location × Product Category

the p – value obtained is $\mathbf{1.\ 58292} \times \mathbf{10^{-25}}$,which is a very small value

Interpretation:

Because the P – Value < 0.05, **We reject the null hypothesis** because:

- There is a statistically significant relationship.
- Product category sales depend on the store location.
- Preferences are not evenly distributed across stores, and each store has a unique purchasing pattern.

Finally, the chi-square test of independence was conducted to examine whether product category distribution differed across three store locations. Using observed transaction counts and calculating expected frequencies under the assumption of independence, a chi-square statistic of 158.56 was obtained with 18 degrees of freedom, yielding a p-value of 1.58×10^{-25}.

Because the p-value is far below the significance threshold of 0.05, we reject the null hypothesis and conclude that store location and product category are not independent. This means customers in different store locations exhibit significantly different purchasing preferences across product categories. The retail strategy should therefore consider location-specific product demand patterns.

Using Regression

In this analysis, we are trying to determine if transaction value and unit price are related by plotting the points on a scatter plot and finding a regression line using the regression equation:

$$\boldsymbol{y = b1_x + b_o}$$

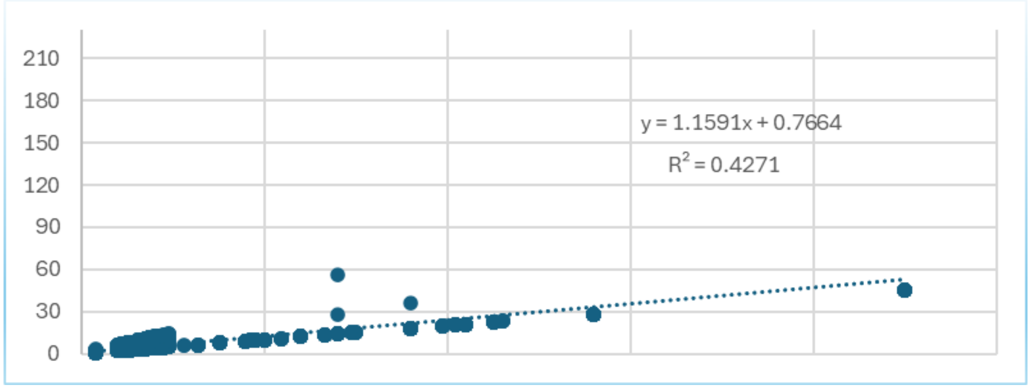

Table 1: Scatterplot for transaction value(y) and unit price(x)

The estimated regression equation that we found was y=1.1591x + 0.7664, and the two variables show a linear relationship. Based on this graph, for every $1 that the unit price increases, the transaction value increases by nearly $1.16. Since the relationship is positive and linear, this means there is a direct relationship between the unit price and transaction value, meaning people spend more when the unit price is higher. The 0.7664 or b_0 shows that if the unit price was $0, the transaction value would be $0.77, which is unrealistic since a unit price would not be $0. The R^2 represents that the relationship between these two variables is moderately strong as it sits at roughly 0.43 meaning 43% of the transaction value is determines by the unit price. This leaves the other 56% of transaction values influenced by other factors.

5. **Concluding Remarks**

The analysis of the "Coffee Shop Sales" dataset (Kaggle, 2023) provided insights into consumer behavior and retail performance across three New York City cafe locations: Astoria, Hell's Kitchen, and Lower Manhattan. Using detailed transactional data-including transaction value, unit price, product category, time of day, and location-several statistical techniques were applied to understand spending patterns. One way ANOVA results indicated that average transaction values were similar across all three stores, with no statistically significant differences, suggesting that overall customer spending behavior is consistent across locations despite minor variations. Two-way Anova further revealed that either store location nor time of day significantly affected unit prices or transaction quantity, although both transaction quantity and spending peaked in the afternoon, reflecting a temporal pattern in customer purchasing.

A chi-square test of independence highlighted that product category preferences varied significantly by store location. With a very small p-value, the test confirmed that customer preferences were not distributed across the three cafes, emphasizing that each store exhibited unique purchasing patterns. This finding underscores the importance of tailoring inventory and product offerings to the specific preferences pf a stores customer base rather than applying a uniform strategy across all locations. Regression analysis additionally showed a positive linear relationship between unit price and transaction value, indicating that higher price items tend to generate larger transaction amounts, although 57% of transaction value variability remained influenced by other factors.

Overall, this analysis demonstrates that while spending per transaction is fairly constant across locations, store specific preferences and peak purchasing times should guide operational and marketing decisions. Retail managers can use these insights to optimize product mix, pricing strategies, and staffing schedules according to the temporal and location specific trends identified. Furthermore, the study highlights the value of granular transactional data in understanding consumer behavior, enabling cafes to make data-driven decisions that improve revenue performance and align offerings with customer demand.

References

(n.d.).

Anderson, D. R. (2020). *Statistics for business & economics.*

Association, N. C. (2023). *National Coffee Data Trends 2023 Report.* Retrieved from https://www.ncausa.org/

Kaggle. (2023). *Coffe shop sales dataset.* New York. Retrieved from https://www.kaggle.com/datasets/keremkarayaz/coffee-shop-sales

Organization, I. C. (2023). Retrieved from World coffee consumption trends.: https://www.ico.org/

Research, N. E. (2021). Retail data analytics in the coffee industry. *Journal of Consumer Market Studies*, 33-49.

Wedel, M. &. (2016). Marketing analytics for data-rich environments. *Journal of Marketing,*. Retrieved from https://doi.org/10.1509/jm.15.0413