Business Analytics and Machine Learning Prediction of a Coffee Vending Machine Sales Dataset

JBrien Cezar Barcoma
(with assistance from Claude 3.5 Sonnet LLM and ChatGPT 4o LLM)
Author

Draft Ver.4 / 29 July 2024
Version / Date

Table of Contents

1. Executive Summary

This analysis examines sales data from a coffee vending machine in Ukraine from March to June 2024, seeking to identify opportunities to minimize expenses and maximize profit. Key findings and recommendations include:

1. **Financial Performance:**
    - Initial investment: ₴236,000 UAH (Approx. $5,760 USD)
    - Monthly recurring costs: ₴11,000 UAH (Approx. $269 USD, excl. Cost of ingredients)
    - Revenue trend: Overall positive, with fluctuations
        - March: ₴7,100 UAH
        - April: ₴6,700 UAH – Decreased 5%
        - May: ₴9,100 UAH – Increased 26% following price cut
        - June: ₴7,800 UAH – Decreased 16%
2. **Pricing Strategy:**
    - A price cut implemented in May led to a substantial revenue increase
    - Recommendation: Consider strategic, periodic discounts to boost sales
3. **Product Mix:**
    - Most profitable drink is Latte, followed by Capuccino and Americano with Milk.
    - Recommendation: Focus marketing efforts on drinks with milk as an ingredient.
4. **Customer Behavior:**
    - Peak sales times: 10am and 7pm on Weekdays (Start of work and over-time), 12noon on Saturdays.
    - Payment preferences: 90% of transactions made with cards, 10% with cash.
    - Customer: 14% of all transactions are done by 2 customers.
    - Recommendation: Optimize the vending machine's tsock and maintenance schedule to align with peak usage times, and to implement a digital loyalty program to reward frequent users and to decrease churn.
5. **Future Outlook:**
    - Financial Trajectory: The business is currently operating at a loss.
    - Price cut improved revenue, leading to a narrowing of losses over the months.
    - Current numer of months is insufficient to reliably predict break-even point.
    - May require significant work in terms of formulating pricing strategies, marketing and adverts to increase sales and make businss profitable.

2. Introduction

This article presents a comprehensive analysis of sales data from a coffee vending machine in Ukraine, covering the period from March to June 2024. By leveraging business analytics and machine learning techniques, we aim to uncover valuable insights that can drive operational improvements and increase profitability.

2.1 Motivation

This article seeks to have a deep understanding of customer preferences, pricing strategies, and operational efficiencies to minimize expenses and maximize profit. With the use of advanced data analytics, we ultimately aim to transforming a simple vending machine into a highly optimized and profitable business venture.

2.2 Problem Statement

Having data on the transactions made with the vending machine gives the business owner the following opportunities to improve the business:

1. Optimizing pricing strategies to balance profitability with customer attraction and retention.
2. Understanding and catering to customer preferences in terms of product mix and timing.
3. Identifying patterns in sales data to inform inventory management and maintenance schedules.
4. Assessing the impact of external factors such as seasonality or local events on sales performance.
5. Determining the most effective payment methods to offer customers.

This analysis aims to turn the above opportunities into profit by answering key questions such as:

- What are the most popular coffee types and at what times?
- How do pricing changes affect overall revenue and customer behavior?
- Are there identifiable patterns in daily or hourly sales that can inform operational decisions?
- What is the current profitability of the operation, and how can it be improved?
- Can we predict future sales trends to inform business planning?

By answering these questions we can obtain actionable insights that can drive measurable improvements in the coffee vending machine's performance and profitability.

3. Dataset Overview

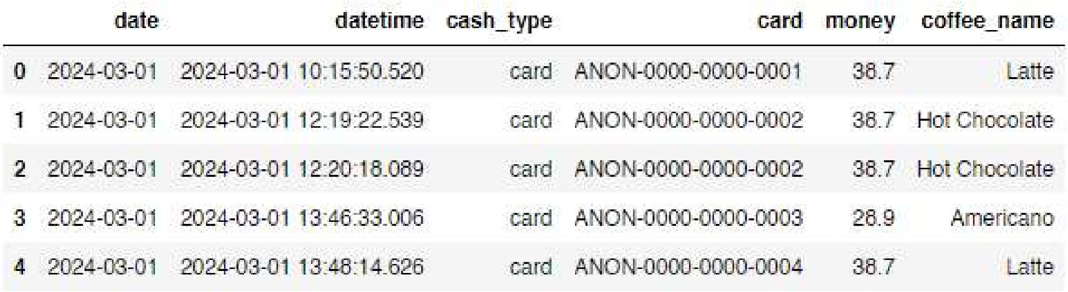
3.1 Data Sources and Description

The primary dataset for this analysis is obtained from Kaggle and contains transactional data from a coffee vending machine located in Ukraine. The data spans four months from March 2024 to June 2024, with monthly updates. The dataset comprises 896 rows and 6 columns, each representing a distinct transaction.

The dataset includes the following attributes:

#	Column	Definition	Data Type
1	date	Date of transaction.	Categorical
2	datetime	Date with time-stamp of transaction.	Categorical
3	cash_type	Whether payment was made in cash or card.	Categorical
4	card	Anonymized card number.	Categorical
5	money	Money spent in Ukrainian Hryvnia (UAH ₴).	Numeric
6	coffee_name	Type of coffee drink bought.	Categorical

Fig 1: Table of Dataset Column Attributes



```
         date                 datetime cash_type                 card money  coffee_name
0  2024-03-01  2024-03-01 10:15:50.520      card  ANON-0000-0000-0001  38.7        Latte
1  2024-03-01  2024-03-01 12:19:22.539      card  ANON-0000-0000-0002  38.7  Hot Chocolate
2  2024-03-01  2024-03-01 12:20:18.089      card  ANON-0000-0000-0002  38.7  Hot Chocolate
3  2024-03-01  2024-03-01 13:46:33.006      card  ANON-0000-0000-0003  28.9    Americano
4  2024-03-01  2024-03-01 13:48:14.626      card  ANON-0000-0000-0004  38.7        Latte
```

Fig 2: Python df.head(5) output of the top 5 rows of the dataset

3.2 Data Quality Assessment

To ensure the reliability of our analysis, a thorough data quality assessment was conducted. The following aspects were evaluated: (link references to visualizations in appendox showing the below was done.)

1. **Completeness**: We verified that there were no missing values in any of the columns, save for when the card number is null for when the transaction utilized cash.[1]
2. **Consistency**: Date and datetime fields were cross-checked to ensure consistency. No discrepancies were found between these two fields.[2]
3. **Accuracy**: The money field was examined for any anomalous values. All amounts were found to be within a reasonable range for coffee purchases.[3]
4. **Uniqueness**: While individual transactions were not assigned unique identifiers, the combination of datetime and card number (for card transactions) provided a reliable way to distinguish between transactions.
5. **Timeliness**: The dataset is updated monthly, ensuring that it remains current for ongoing analysis.
6. **Validity**: All categorical variables (cash_type, card, coffee_name) contained valid entries corresponding to expected categories.[5]

The dataset was found to be of high quality, requiring minimal cleaning or preprocessing.

The main limitation identified was the relatively short time span of four months, which may limit the ability to detect long-term trends or seasonality effects. The lack of geographical or demographic information about customers restricts certain types of analyses.

Despite these limitations, the dataset provides enough information for deriving meaningful insights into the operation of the coffee vending machine business.

4. Methodology

This section outlines the technical approach used to process, analyze, and derive insights from the coffee vending machine dataset. We detail the data pipeline, tools employed, and the analytical methods applied to extract meaningful business intelligence.

4.1 Data Pipeline and Tools

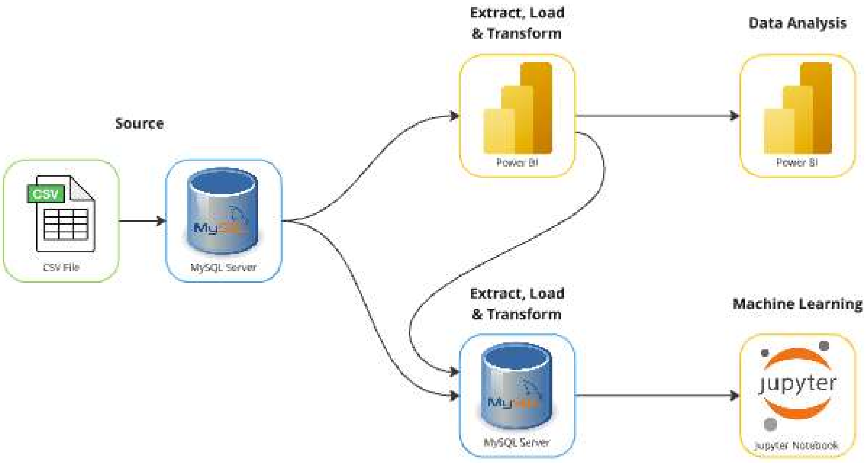
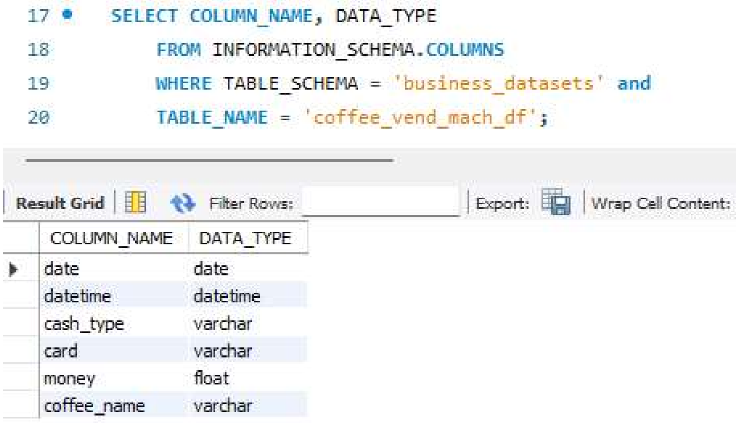


Fig 3: Data Pipeline Flowchart made in Miro

The pipeline consists of the following stages:

1. **Data Ingestion**: The raw dataset was imported into a MySQL server relational database management system (RDBMS). This step leverages on the common availability of SQL as a skill in the data anlytics industry. [6]
2. **Extract, Transform, Load (ETL)**: The ETL process was divided between PowerBI and MySQL server. MySQL was used for initial data structuring and basic transformations, while PowerBI handled more complex transformations and feature engineering. [7]
3. **Data Analysis and Visualization (Part 1, this document)**: PowerBI served as the primary tool for data analytics and visualization. Its user-friendly interface and powerful analytical capabilities made it ideal for exploring the dataset and creating insightful visualizations.
4. **Machine Learning Predictions (Part 2)**: For more advanced predictive analytics, we utilized Python in a Jupyter notebook environment. This allowed us to leverage powerful machine learning libraries such as scikit-learn and pandas for sophisticated data modeling.

This approach focusing on no-code workflows for the data anlytics allows for quick processing of data. Uploading all data into the MySQL server ensures that data is all located centrally, as opposed to loading the files directly onto PowerBI where the source files may be scattered throughout the drive. A limitation of this process is the lack of automation in data ingestion, where python coding and tools such as Apache Airflow may be used to automatically download the updated files with the next months transaction from kaggle.



```
SELECT COLUMN_NAME, DATA_TYPE
    FROM INFORMATION_SCHEMA.COLUMNS
    WHERE TABLE_SCHEMA = 'business_datasets' and
    TABLE_NAME = 'coffee_vend_mach_df';
```

COLUMN_NAME	DATA_TYPE
date	date
datetime	datetime
cash_type	varchar
card	varchar
money	float
coffee_name	varchar

Fig 4: Verifying column names and data types in SQL.

4.2 Analytics Approach

Our analytical approach was structured to address key business questions and extract actionable insights from the data. The analysis was conducted in several stages:

1. **Exploratory Data Analysis (EDA)**: To examine distributions, identify basic patterns, and calculate summary statistics which will inform subsequent more advanced anlysis.
2. **Time Series Analysis**: To identify trends, seasonality, and potential cyclical patterns in sales.
3. **Customer Segmentation**: We analyze customer behavior to segment customers based on factors such as purchase frequency, preferred coffee types, and payment methods.
4. **Pricing Analysis**: To examine the impact of pricing changes on sales volume and revenue, with particular attention to the price cut implemented in May 2024.
5. **Product Mix Analysis**: To evaluate the popularity and profitability of different coffee types to optimize the product offering.
6. **Operational Efficiency Analysis**: To examining temporal patterns in sales, aiming to identify opportunities for optimizing restocking schedules and maintenance timing.
7. **Predictive Modeling (Part 2)**: To use machine learning techniques in developing models to forecast future sales trends and predict customer behavior.

This approach identifies patterns in the data providing a comprehensive understanding of the coffee vending machine's performance as well as translates these findings into practical, actionable recommendations, identifying potential areas for improvement to improve the coffee vending machine business.

5. Business Performance Analysis

5.1. Initial Investment and Recurring Costs

We estimate the initial capital investment for the coffee vending machine business to be $5,760, which includes the cost of the machine, installation, and initial stock. Our analysis also identifies recurring monthly costs of approximately $269, encompassing rent, maintenance, utilities, and other operational expenses.

#	Item	Cost	Type	Frequency	Recurring Monthly Cost
1	Vending Machine Cost	$5760	One-time		
2	Space Monthly Rent	$200	Recurring	Monthly	$200
3	Cost of Ingredients	$300	Recurring	Variable	Variable
4	Vending Machine Maintenance	$120	Recurring	Yearly	$10
5	Utilities	$50	Recurring	Monthly	$50
6	Insurance	$36	Recurring	Yearly	$3
7	Licensing and Permits	$72	Recurring	Yearly	$6
	Sum	$5760			$269

Fig 5: Table of Initial Capital Investment and Recurring Costs

5.2 Coffee Sale Price

Estimates were made on the cost price of each drink in the menu. Subtracting the cost price from the sale price gives us the profit margin set for each drink.

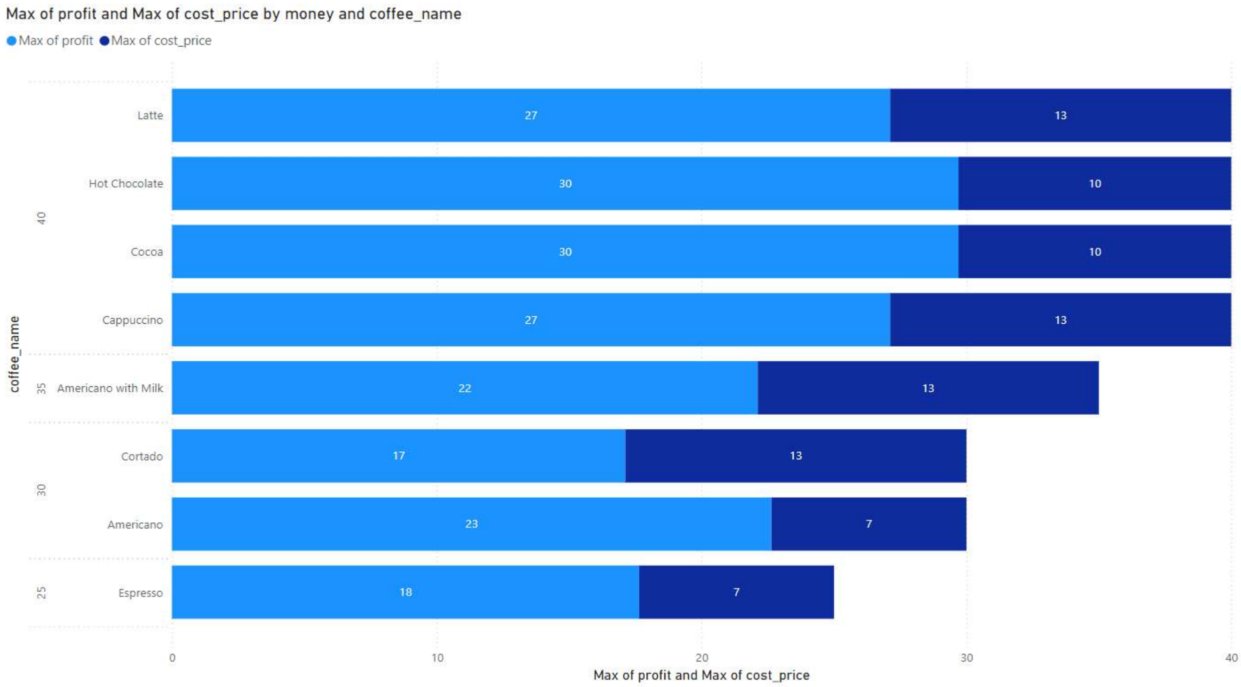


Fig 6: Stacked Bar Chart of Drink Proft Margin and Cost Price

Our analysis includes a price matrix that provides a quick overview of selling prices across different months. This visualization reveals a significant price reduction implemented between April and May 2024. Of particular interest is the impact of this price cut on sales volume.

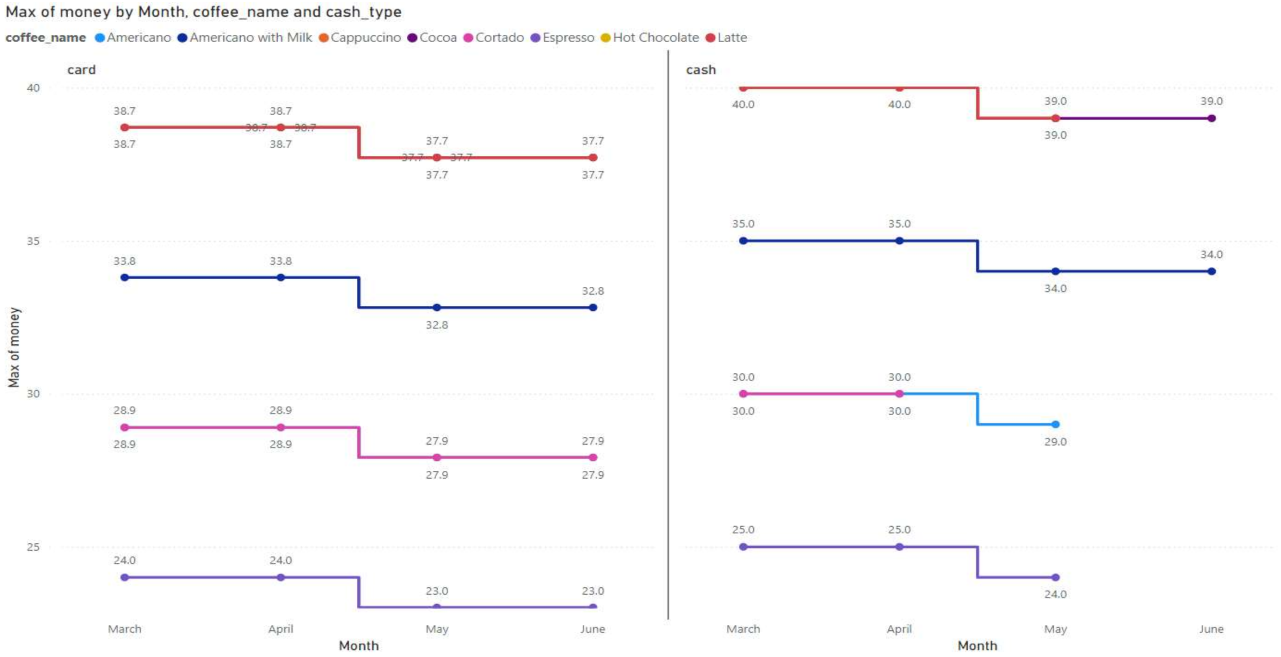


Fig 7: Line Chart of Coffee Selling Prices in Card and Cash over Time

5.3. Revenue Trends

Our examination of revenue trends reveals a generally positive trajectory over the four-month period.

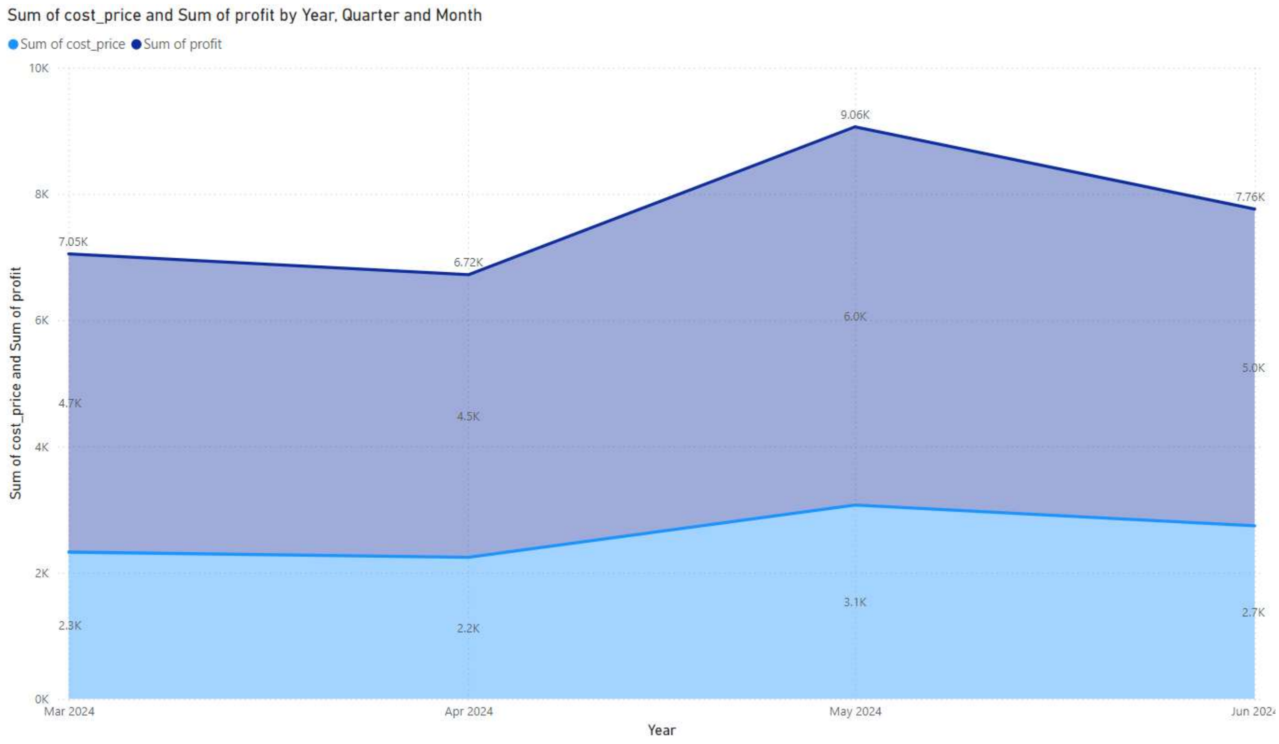


Fig 8: Stacked Area Chart of Revenue (Ingredient Cost + Profit)

The business started with a revenue of UAH 7,100 in March 2024. While there is a slight decrease in April, we observe a significant jump in May, followed by a minor decline in June. This pattern suggests a growing customer base, but also indicates potential external factors influencing sales, which we further investigate in our seasonality analysis.

Key Insights/Next Steps:
The price cut coincides with the observed spike in revenue. Our analysis suggests a strong correlation between the price reduction and increased sales volume, indicating high price elasticity of demand for the vending machine's products. However, we also note that this strategy's long-term sustainability needs further evaluation.

5.4. Product Mix Analysis

We conduct a detailed analysis of the product mix to identify the most popular and profitable coffee types. Our findings show that Latte and other drinks with milk added are consistently the best-selling items and are also the most profiable. This analysis provides valuable insights for inventory management and potential menu optimization.

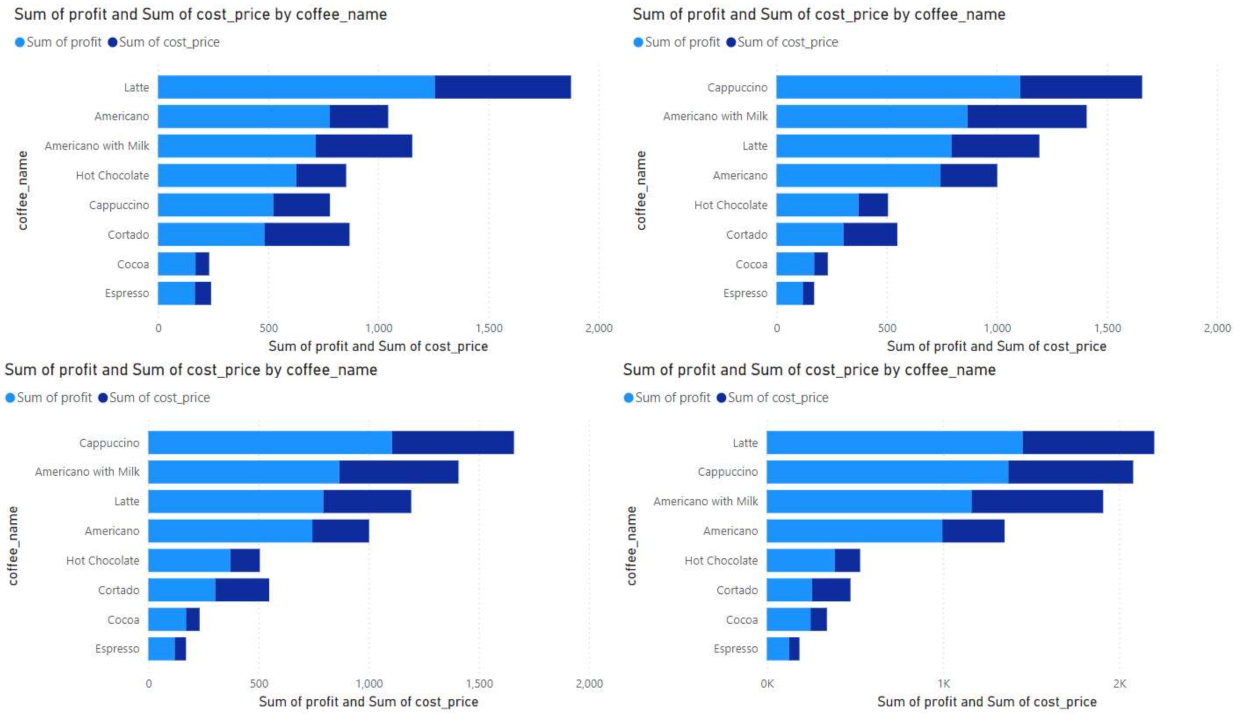


Fig 9: Breakdown of Revenue per Month by Coffee Type

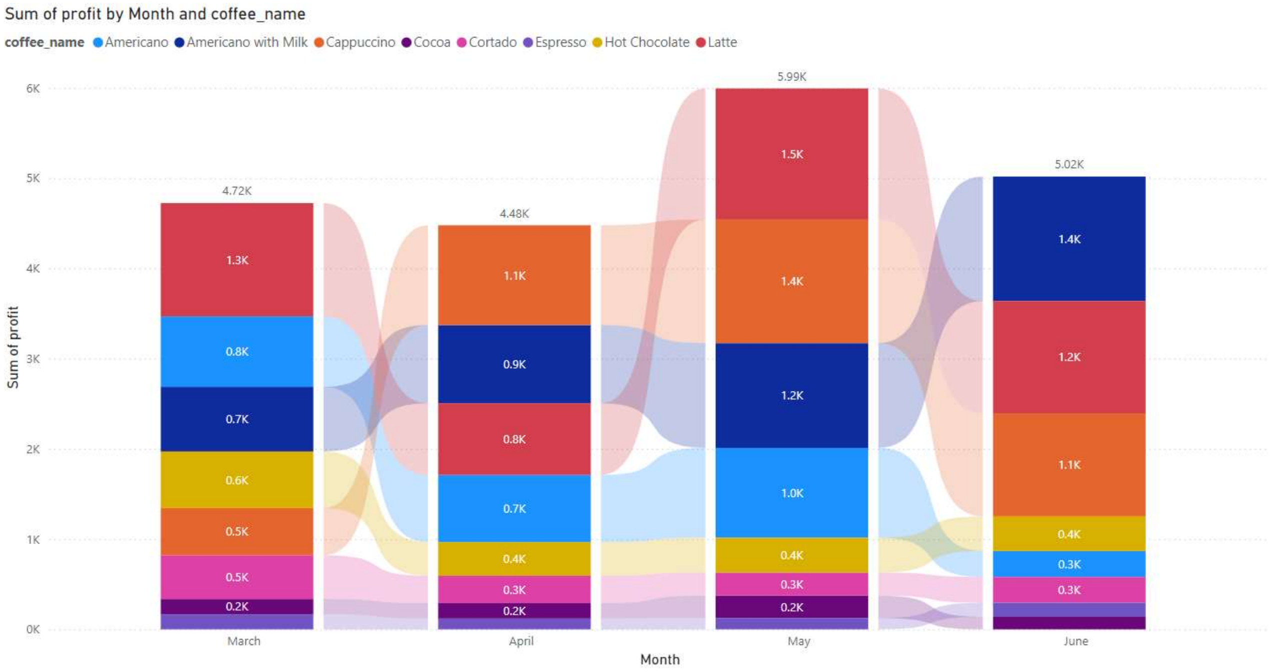


Fig 10: Ribbon Chart of Drink Ranking by Revenue per Month

5.5. Profitability Assessment

Our profitability assessment takes into account the initial investment, recurring costs, revenue trends, and product-specific profit margins. We calculate that the business is operating at a loss since its inception. With only four months of data, we find that there is insufficient information to determine the break-even point. Furthermore, there is significant variation in the revenue for each month due to the changes in pricing strategy.

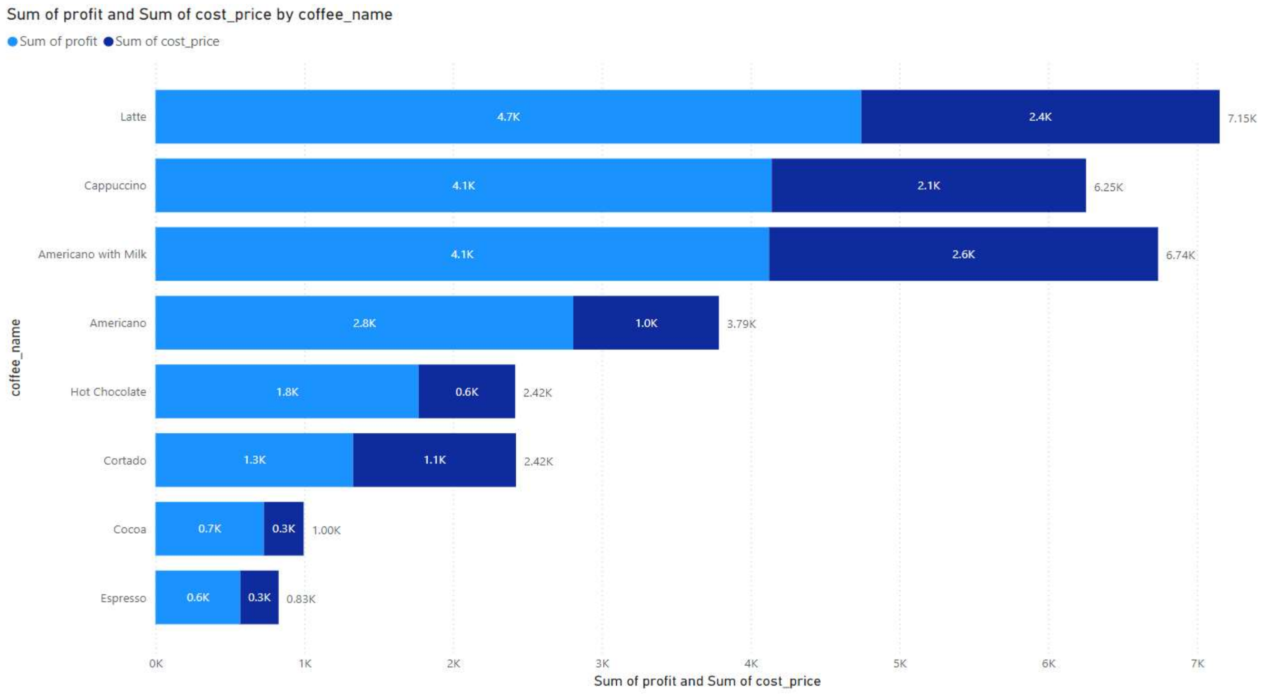


Fig 11: Total Revenue (Ingredient Cost + Profit) by Coffee Drink

This profitability analysis provides a picture of the business's financial health based on the four months of data provided. These insights form the basis for our recommendations on pricing strategies, product focus, and operational efficiencies.

6. Customer Behaviour Insights

6.1. Card vs Cash Payment

Our data indicates a clear preference of paying via cards among customers. 90% of transactions are completed by card, while the remaining 10% use cash.

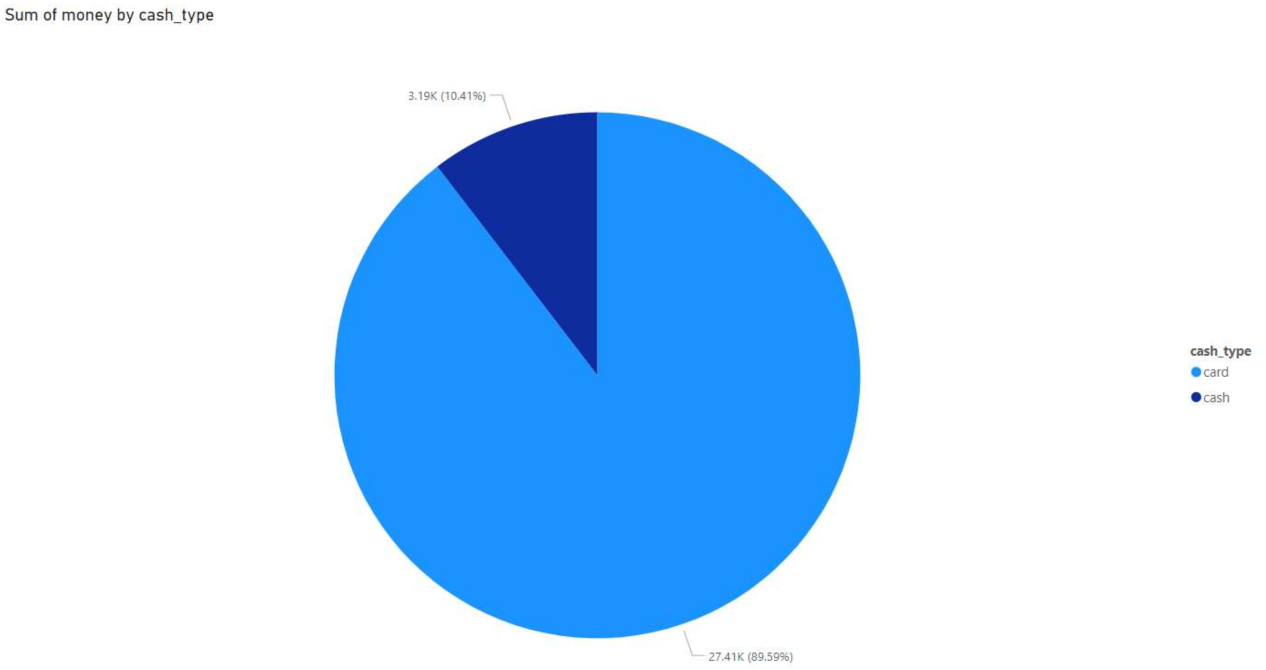


Fig 12: Pie Chart of Card vs Cash Transactions

We note a gradual decrease in cash payments over he four-month period, which may indicate customers taking advantage of the slight discount on prices with card payments.

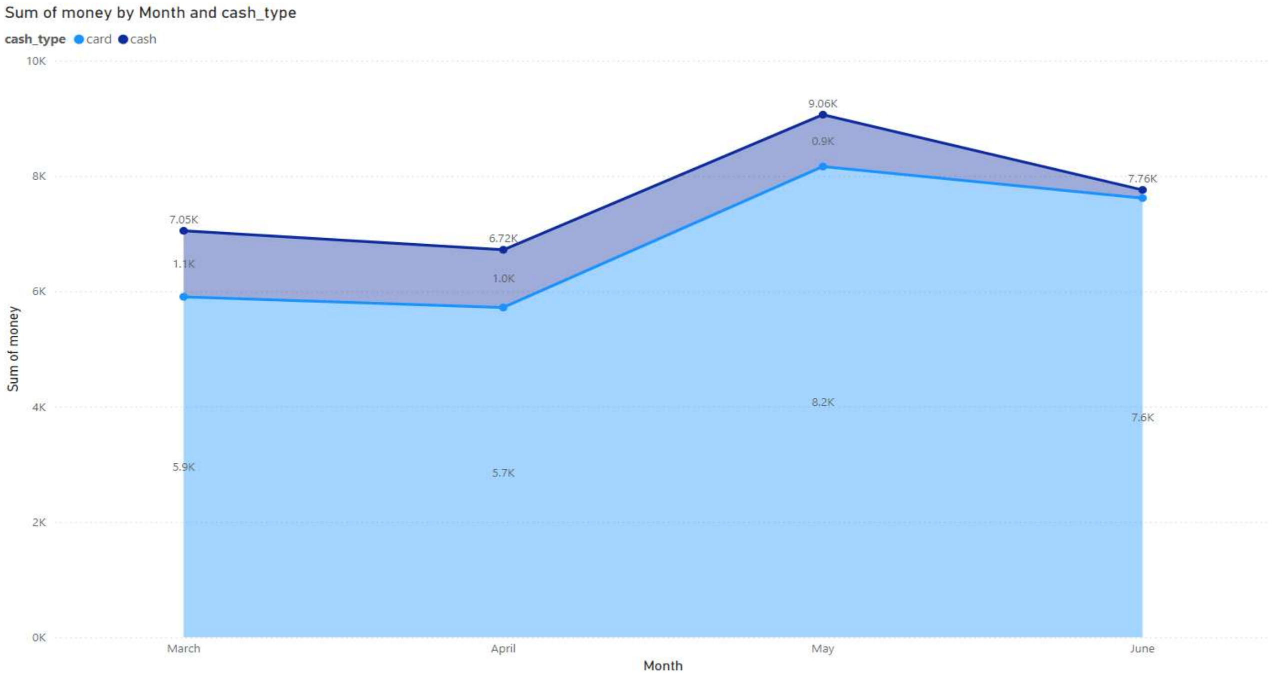


Fig 13: Stacked Area Chart of Card and Cash transaction over four months.

6.2.1. New vs Returning Customers

We observe that a large majority of card customers are returning customers.

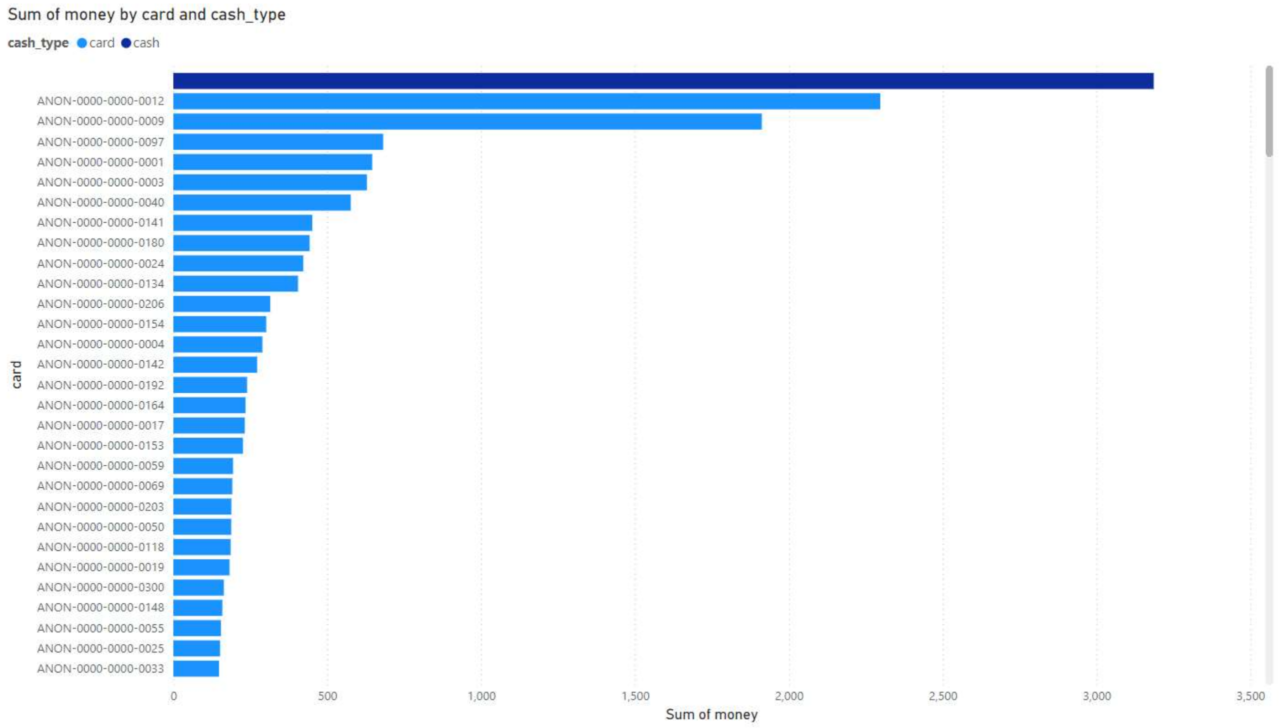


Fig 14: Bar Chart of aggregated revenue of cash payments and individual cards.

6.2.1. Top 10 Card Customers

A look into the drink preferences of the the top 10 spenders show that most purchase a variety of drinks, with the exception of the 4th highest spender, only ordering Latte.

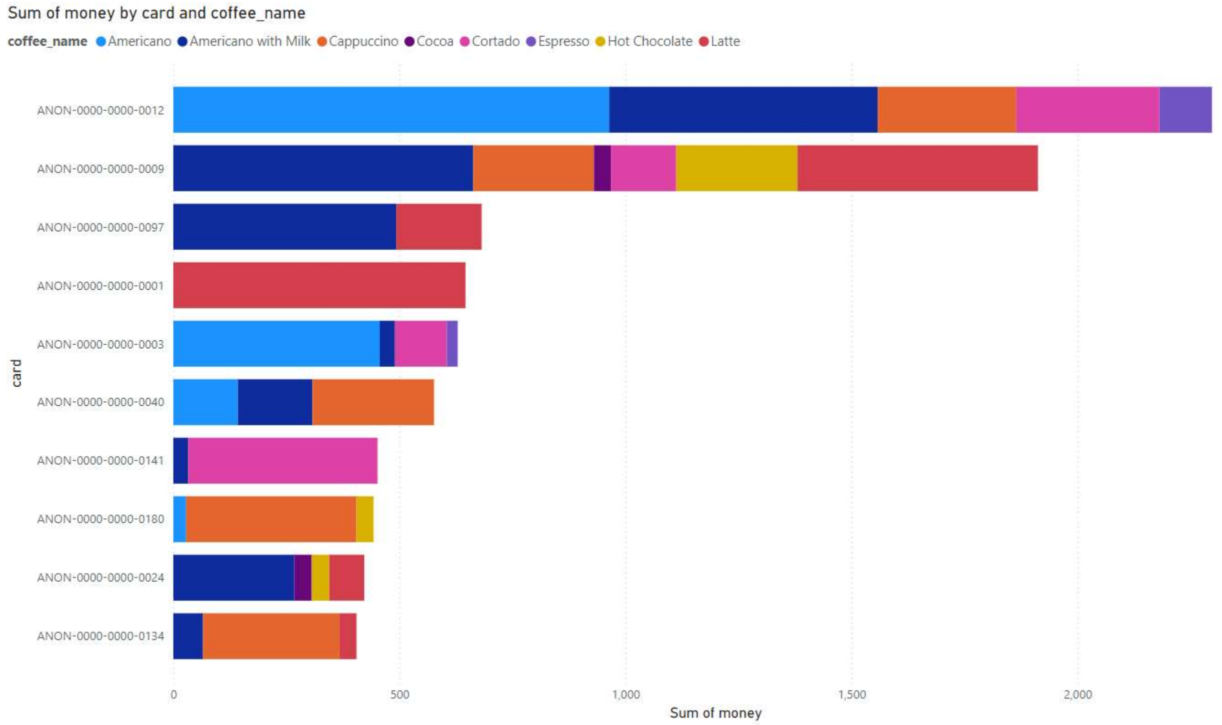


Fig 15: Stacked Bar Chart of Drink Preferences of top 10 card customers.

6.2.1. Top Card Customer

A look into the drink preference of the top spender shows that they purchased an Americano drink the most, followed by Americano with Milk.

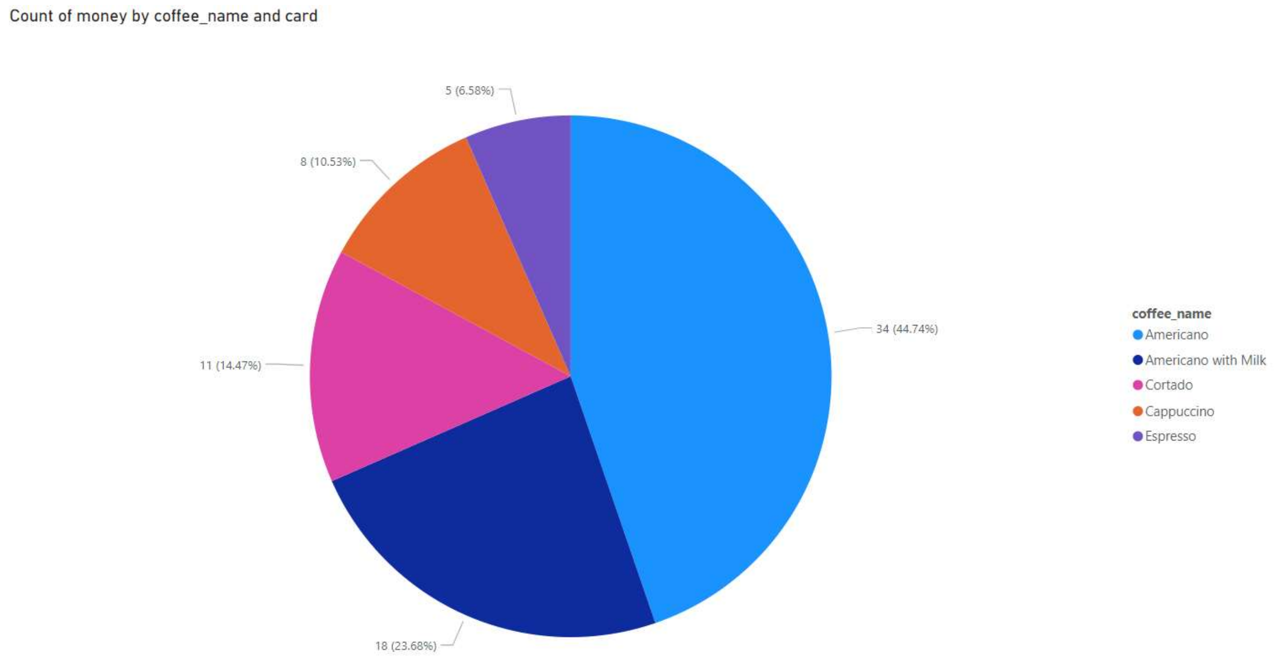


Fig 16: Pie Chart of Coffee Preference by the Top Card Customer

6.2. Temporal Patterns (Weekly, Daily and Hourly)

Examination of transaction timestamps reveals distinct patterns in customer behavior.

6.2.1. Sales by Day of Month

We charted sales by day of the month to see whether there are days when sales would spike, paydays for example. We see a spike in sales around the 20th of each month, which may correspond to the first of the bi-weekly payroll cycle (between 15th and 20th of each month).

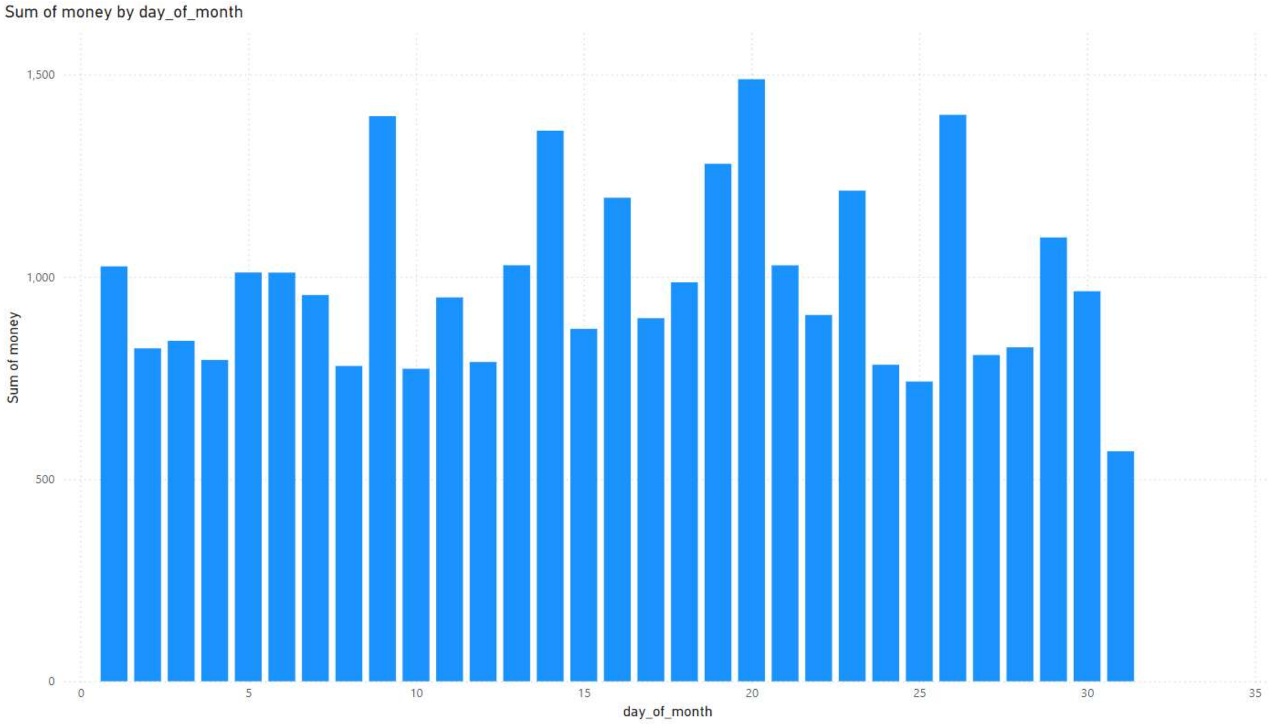


Fig 17: Total Revenue per Day of the Week

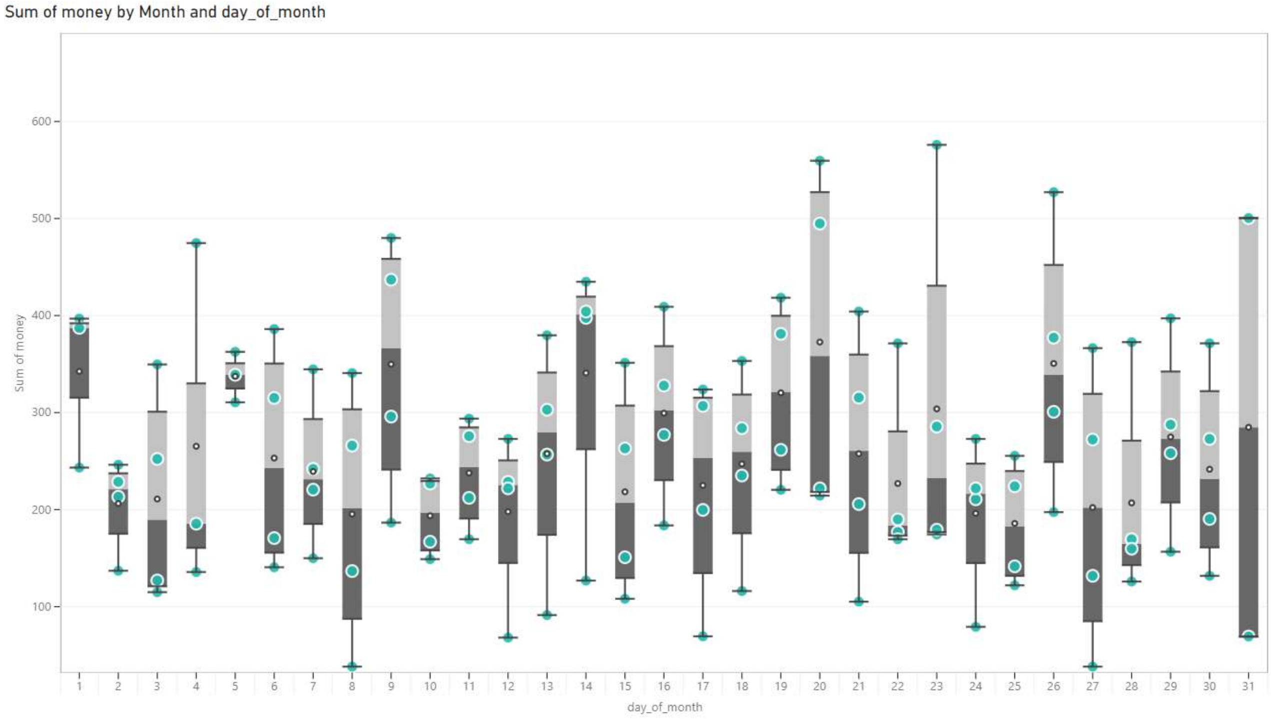


Fig 18: Boxplot of Revenue per Day of the Week

6.2.1. Sales by Day of Week

Aggregating revenue by day of the week shows a v-shaped graph, with wednesdays seeing the lowest sales. Tuesdays and Thursdays are the top revenue days of the week.

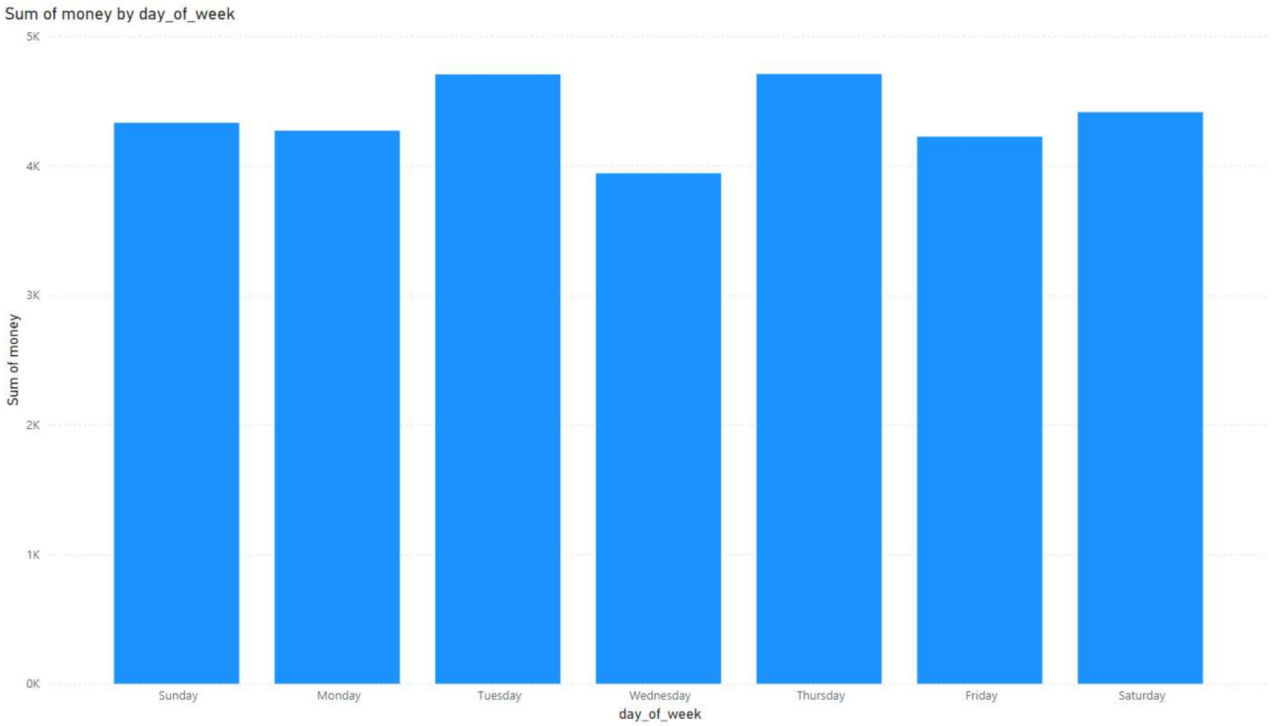


Fig 19: Total Revenue per day of the Week

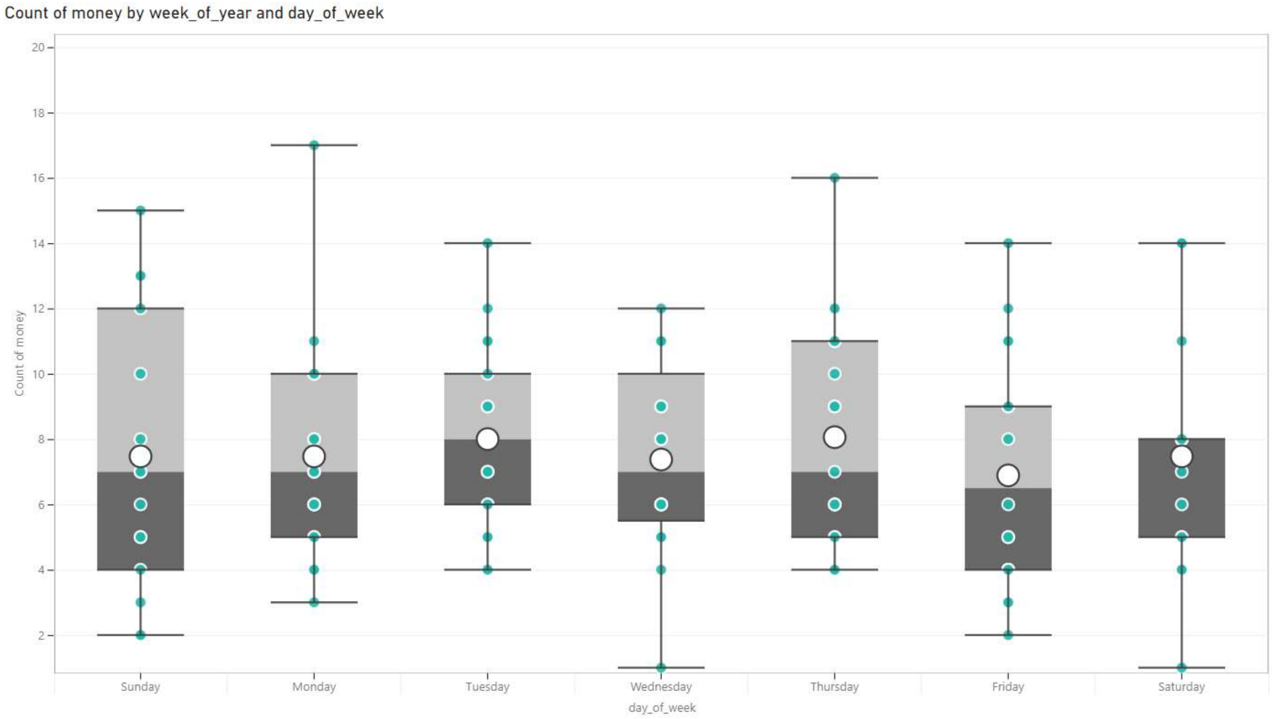


Fig 20: Boxplot of Total Revenue by day of the Week

6.2.2. Sales by Hour of the Day

Hourly analysis shows peak periods at 10am and 7pm, suggesting strong correlation with typical work schedules. There are also no transactions between 11 PM and 7 AM, suggesting the vending machine is in an area that is closed to foot traffic during these timings, a shopping mall near an office complex for example.

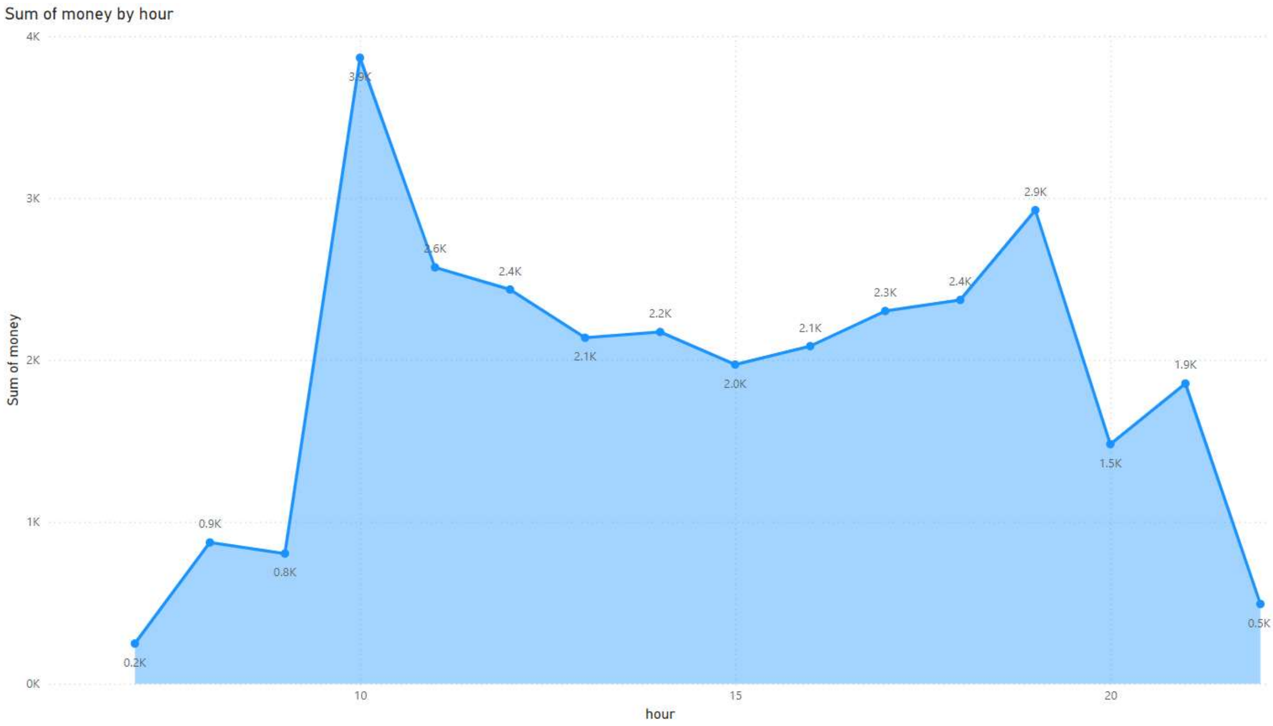


Fig 21: Total Revenue by the hour

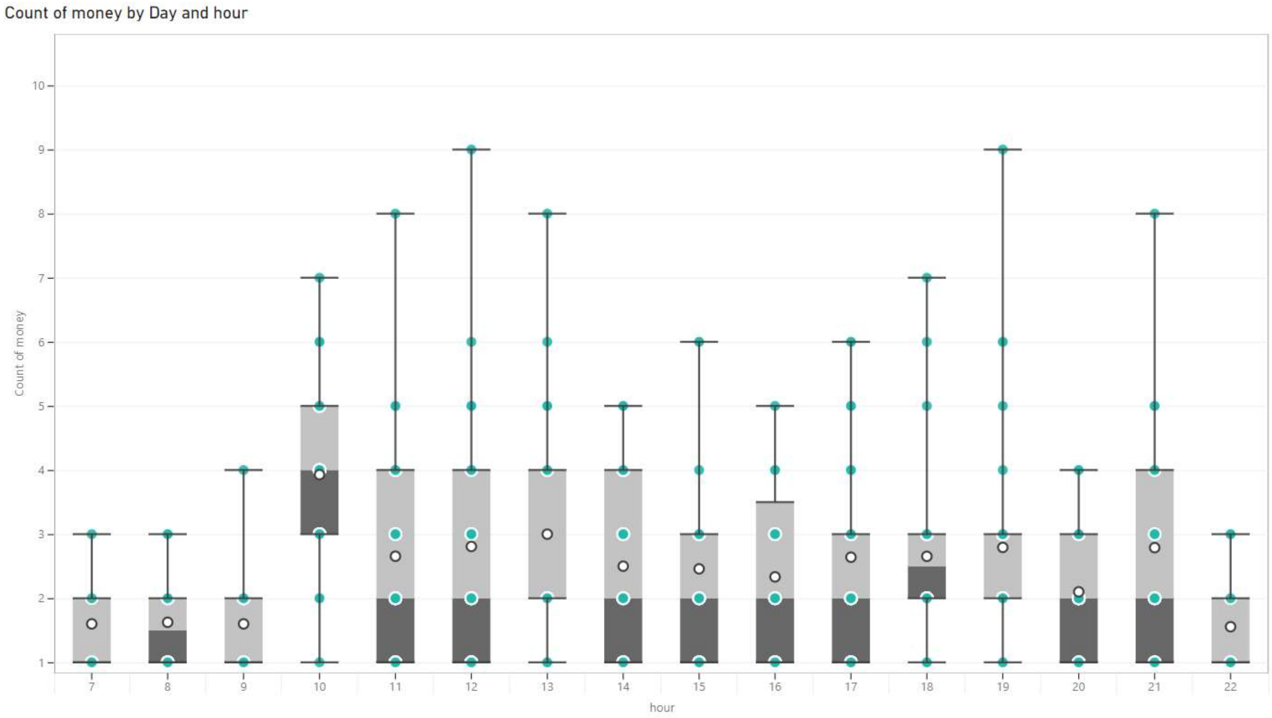


Fig 22: Boxplot of Number of Transaction by the hour

6.2.2. Sales by Hour of the Day by Day of the Week

A breakdown of the hourly sales by the day of the week show slight differences between each day of the week. Weekdays consistently show peaks at 10am and 7pm, with the exception of wednesdays which show subdued revenues. On Weekends, peak sales start later at 10am on Saturdays while Sundays show increase in sales at mid-morning and mid-afternoon.

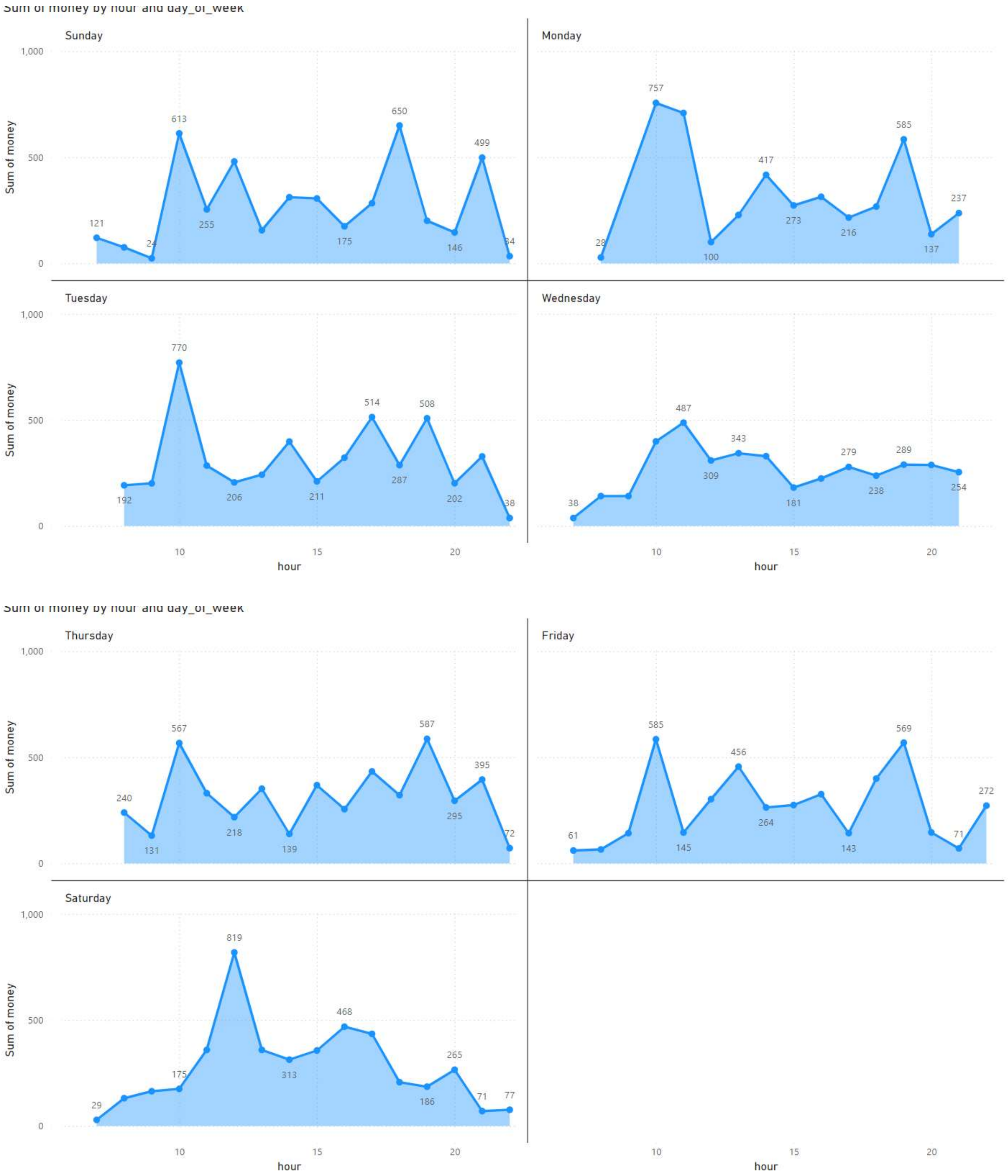


Fig 23 & 24: Area Chart of Total Revenue by Hour by day of the Week

7. Key Findings and Recommendations

7.1. Summary of Insights

1. **Business Performance Analysis**:
    1. Drinks are divided into four price groups, with card payments given a slight discount over cash payments.
    2. A pricecut in early May increased sales by 26% in that month.
    3. Drinks with milk saw the highest sales and are the most profitable.
    4. Revenue is on an overall uptrend due to the price cut, but breaking it down to the start of each new pricing period shows a downtrend immediately the month after.
    5. The business is operating at a loss, but consequently due to the price cut, the loss has narrowed over the four months
2. **Customer Behaviour Insights**
    1. 90% are card payments, 10% are cash payments.
    2. 2 card customers together are responsible for 14% of all transactions.
    3. Weekday peak sales are at 10am and 7pm.
    4. Weekend peak sales are at 12noon on Saturdays.
    5. Wednesday experiences the lowest sales.
    6. There are no sales made from 11pm to 6am on all days.

7.3. Actionable Recommendations

1. **Pricing Strategy:** To consider strategic, periodic price cuts to boost sales while maintaining sustainability of such price cuts.
2. **Restocking Schedule**: To ensure machine is fully stocked and cleaned bfore 10am and 7pm weekdays rushes as well as the Saturday noon peak.
3. **Loyalty Program**: To create a digital loyalty program, either via an App or collaboration with the card issuers, to encourage more sales from the repeat customers as well as to induce loyalty on new customers.
4. **Marketing and Promotion:** To study portential of increasing sales during non-peak hours with marketing and other promotions.

7.4. Limitations and Future Work

This study on coffee vending machine customer behavior has several limitations that could be addressed in future research:

1. **Data collection period:** Our analysis is based on a limited timeframe. A longer-term study could reveal seasonal trends or changes in customer behavior over time.
2. **Geographic scope:** This study does not indicate the exact location in Ukraine and also focuses on a single location. Future work could compare data from multiple locations to identify regional differences in customer preferences.
3. **Limited customer demographics:** While we identified two frequent customers, we lack detailed demographic information. Future studies could incorporate customer surveys to gather more comprehensive data on user profiles.
4. **Payment methods:** While we observed a strong preference for card payments, we did not distinguish between different types of cards or mobile payment methods. Future research could provide a more granular analysis of payment preferences.
5. **Environmental factors:** We did not consider external factors such as weather or local events that might influence vending machine usage due to the lack of information on the exact location of the vending machine. Incorporating these variables could provide more context for usage patterns.
6. **Machine learning applications:** In part 2 we will explore the use of machine learning algorithms to predict peak usage times and optimize restocking schedules.
7. **User experience:** A study on customer satisfaction with the vending machine interface and product quality could provide valuable insights for improving service.

Addressing these limitations in future work would provide a more comprehensive understanding of coffee vending machine usage patterns and customer preferences, potentially leading to improved service and increased profitability.

8. PowerBI Dashboard

To be updated.

9. Conclusion

This study provides valuable insights into customer behavior patterns associated with coffee vending machine usage. Our analysis revealed distinct peak usage times, clear payment preferences, and the presence of highly frequent customers. These findings have several important implications for vending machine operators and businesses.

The identified peak times of 10am and 7pm on weekdays, and 12noon on Saturdays, highlight critical periods when machine maintenance, restocking, and potential staff support should be prioritized. This information can be used to optimize operational efficiency and ensure customer satisfaction during high-demand periods.

The strong preference for card payments (90% of transactions) underscores the importance of maintaining reliable electronic payment systems. Vending machine operators should ensure their machines are equipped with up-to-date, user-friendly card payment options to cater to this customer preference.

The discovery that 14% of all transactions are made by just two customers emphasizes the significance of repeat business. This finding suggests that implementing a loyalty program could be highly beneficial, potentially increasing customer retention and encouraging more frequent purchases from other customers.

Overall, this study demonstrates the value of data-driven decision making in the vending machine industry. By understanding and responding to customer behavior patterns, operators can enhance service quality, improve customer satisfaction, and potentially increase profitability. While there are limitations to this study, it provides a solid foundation for future research and practical improvements in coffee vending machine operations.

10. Appendices

10.1. Use of Large Language Models

In the preparation of this article, large language models (LLMs), specifically Claude developed by Anthropic, were utilized as an advanced research and writing assistant. The use of LLMs in this context is analogous to how a professor might supervise research assistants in the production of academic articles. This approach warrants transparency and explanation.

Methodology: The author maintained full supervisory control over the content creation process. The LLM was used to:

1. Generate initial drafts of sections based on provided outlines and guidelines.
2. Refine and restructure existing content.
3. Suggest improvements in clarity, coherence, and analytical depth.
4. Assist in maintaining consistency of style and tone throughout the document.

The author's role included:

1. Providing the overall structure, key insights, and data points.
2. Critically evaluating and editing all LLM-generated content.
3. Ensuring factual accuracy and alignment with the coffee vending machine dataset.
4. Making final decisions on content inclusion, phrasing, and analytical conclusions, such as the decision to include this section on the use of large language models.

Limitations and Considerations: While LLMs offer significant advantages in terms of speed and linguistic versatility, they also present potential limitations:

1. LLMs may inadvertently introduce subtle inaccuracies, biases or even contradictions.
2. They lack real-world knowledge beyond their training data.
3. They cannot independently verify factual claims or data accuracy.

To mitigate these limitations, all LLM-generated content underwent multiple rounds of review, quality and fact-checking by the author against the original dataset and analysis results.

Ethical Considerations: The use of LLMs in academic and professional writing raises important ethical questions about authorship and intellectual contribution. In this case, the LLM is viewed as a sophisticated tool rather than a co-author. All substantive intellectual contributions, analytical insights, and conclusions stem from the human author's expertise and judgment.

By including this section, we aim to maintain transparency about our methodology and contribute to ongoing discussions about the integration of AI technologies in research and writing processes.

10.2. References

10.2.1 Reference Methods

Our research process involved using Large Language Models (LLMs) to generate preliminary information. These initial facts were then rigorously verified through search engines, primarily Google and Bing. To enhance the relevance and depth of our findings, we employed a Virtual Private Network (VPN) to access Ukrainian servers, conducting our online searches from this vantage point. Additionally, we utilized Google Translate to convert our search queries from English to Ukrainian, further improving the pertinence of our search results.

10.2.2 References

1. https://www.kaggle.com/datasets/ihelon/coffee-sales
2. https://domkofe.com.ua/product-category/kofevarki/12258/
3. https://easyvending.com.ua/ua/i-articles/kak-nachat-biznes-na-kofeynykh-apparatakh/
4. https://rozetka.com.ua/429715844/p429715844/
5. https://www.usemultiplier.com/ukraine/payroll