The Impact of Weather on Coffee and Tea Sales

İdil Işıl Yıldırım 32049
DSA 210 Final Project

Introduction

In the food and beverage industry, aligning supply with customer demand is critical—not only for profitability but also for sustainability. Overstocking leads to waste, while understocking results in missed sales and dissatisfied customers. This project was designed with a practical purpose: to help businesses reduce waste, control costs, and better anticipate consumer demand based on external environmental conditions.

Specifically, it investigates whether weather factors such as temperature and rainfall, as well as the time of day, influence customer preferences for coffee or tea. Understanding these patterns allows businesses to make data-driven decisions on inventory planning, staffing, and promotional timing. Through data enrichment, statistical analysis, and machine learning, the project aims to turn raw transactional data into actionable insights that support operational efficiency and customer satisfaction.

Data Collection & Enrichment

Source Datasets

- **Sales Dataset:**
  - Contains individual transactions with the following fields: Transaction ID, Date, Time, Quantity, Unit Price, Store Location, Product Category (Coffee or Tea)
- **Weather Dataset:**
  - Contains daily weather measurements: Date, Average Temperature (°C), Rainfall (mm)

Data Enrichment Steps

To deepen the analysis and make the dataset more machine learning–ready, I performed multiple enrichment and transformation steps:

A. Temporal Features

- Extracted Hour from timestamp to analyze daily sales cycles.
- Created Time of Day feature:
  - Morning (6–11), Afternoon (12–17), Evening (18–23)
- Added Is Weekend flag based on the day of the week.

B. Weather-Based Features

- Created Temperature Bin (Cold <10°C, Mild 10–20°C, Warm >20°C)
- Added Is Rainy (Rainfall > 0 mm)

C. Aggregated Metrics

- Computed Total Sale per Transaction = Quantity × Unit Price
- Aggregated sales totals by:
    - Hour
    - Temperature
    - Rain condition
    - Product Category

D. Data Merging

- Merged sales and weather data using Date as a common key.
- Ensured all transactions were correctly matched with daily weather conditions.

This enrichment enabled me to convert raw numeric data into context-aware insights, such as identifying how a cold, rainy morning increases coffee demand or how warmer, dry afternoons see a slight shift toward tea.

Hypotheses

I formulated the following hypotheses to test the relationship between weather and beverage sales:

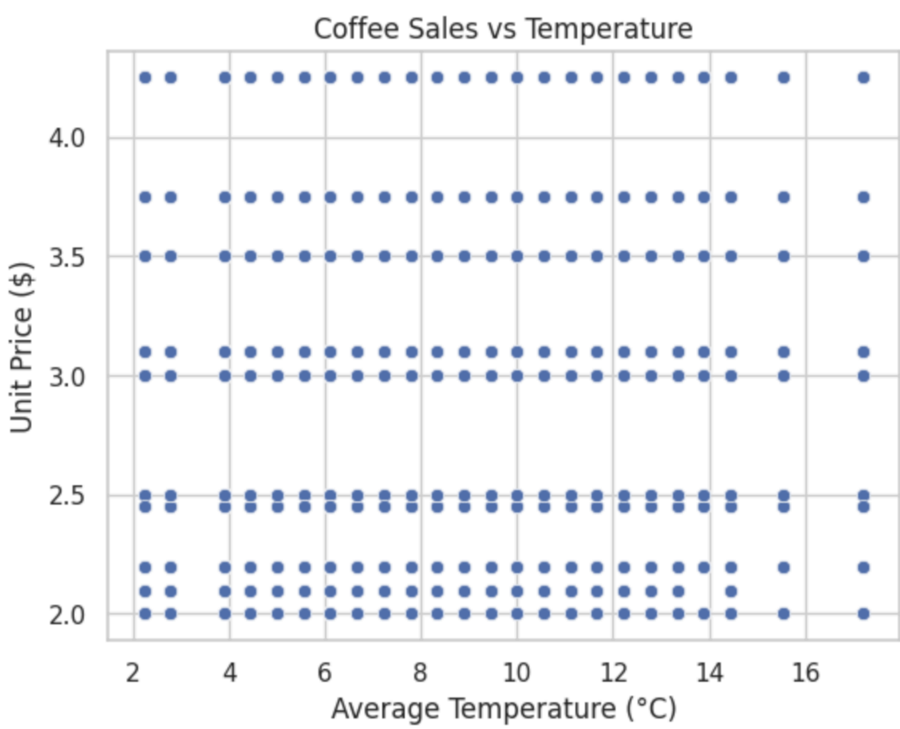
Hypothesis	Description
H_0	Weather and time factors have no significant effect on beverage sales or preferences.
H_1	Weather (temperature, rain) and time of day significantly influence the volume and type of beverage sold.

These hypotheses were tested using statistical methods like t-tests, chi-square tests, and Pearson correlation, to validate or reject assumptions based on observed patterns.

UNIVARIATE ANALYSIS

Temperature Distribution

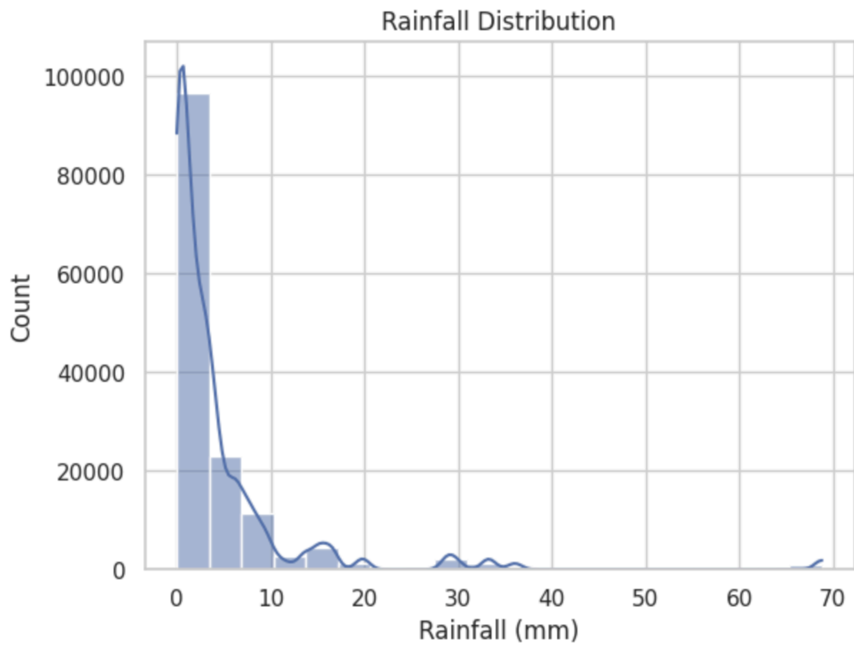
- Most transactions occurred when temperatures ranged between **10–20°C**, suggesting mild weather encourages beverage consumption.
- Very cold days (<10°C) showed spikes in coffee demand.
-



Rainfall Distribution

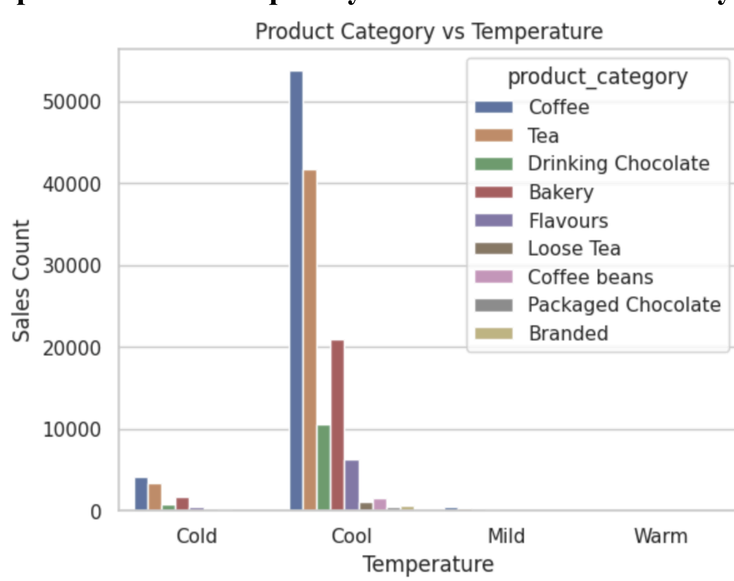
- Dry days were dominant.

- Rainy days showed slight drops in overall sales volume but were not extreme enough to significantly alter behavior.



Product Category Distribution

- **Coffee was consistently the dominant product across all conditions.**
- **Tea was purchased less frequently and showed less sensitivity to temperature.**



-

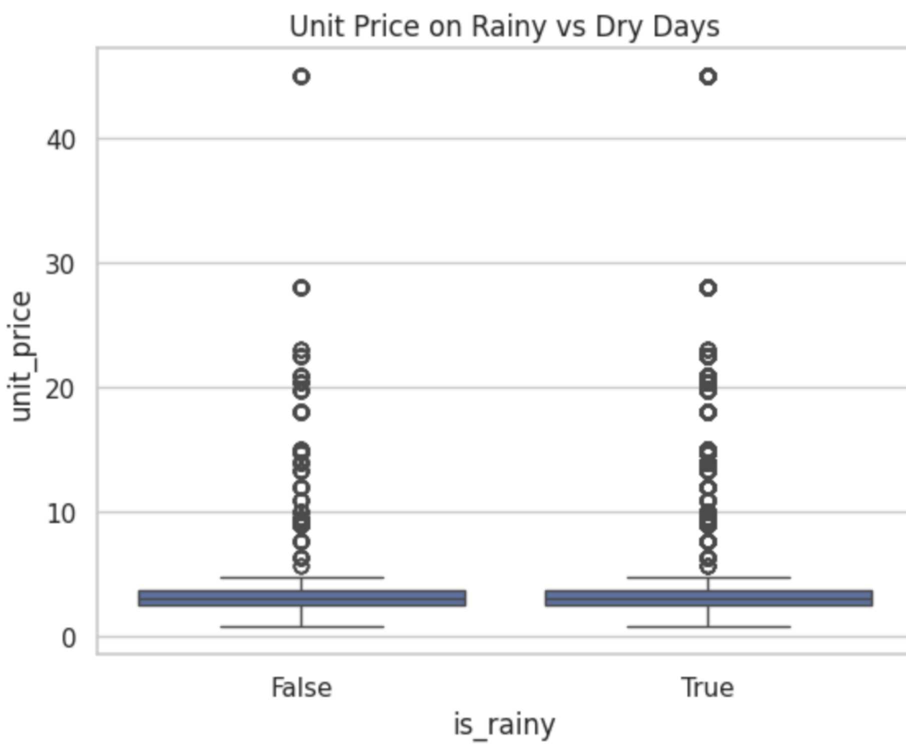
Bivariate Analysis

Product Category vs Temperature

- Coffee sales increased significantly in **colder weather**.
- Tea sales did not show clear temperature-driven behavior.

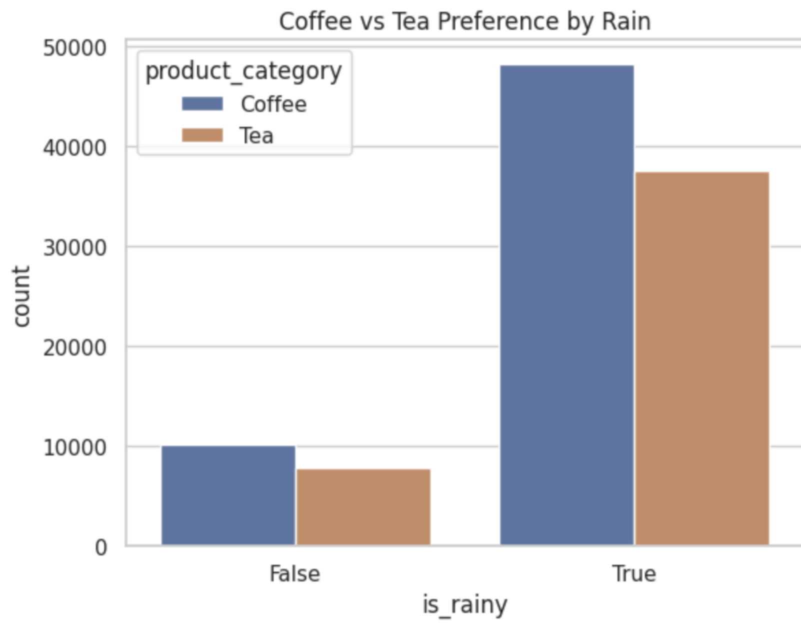
Pricing Analysis: Rainy vs Dry Days

- Conducted **t-test** on unit prices between rainy and dry days.
  - **Result:** $p > 0.05$ → No statistically significant difference.
  - **Interpretation:** Rain does not affect pricing strategies.



Beverage Preference on Rainy Days

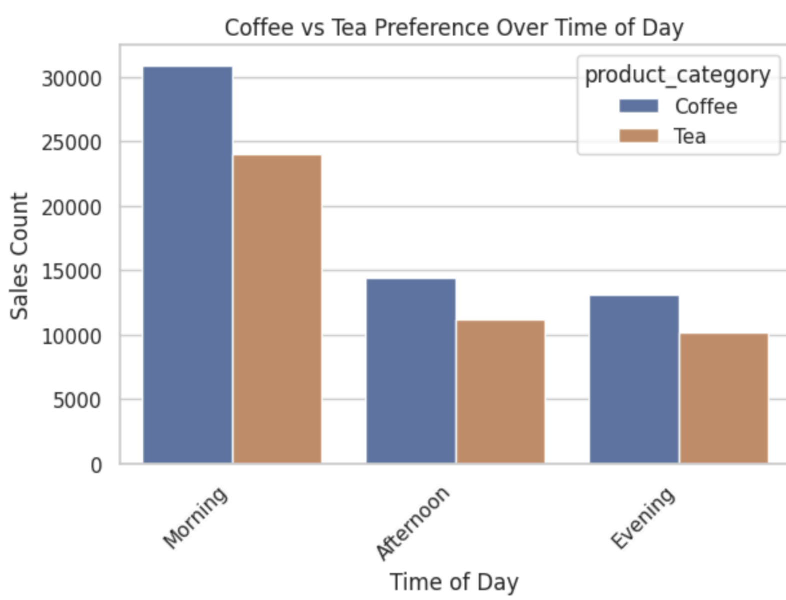
- Ran **chi-square test** to see if rain influences drink choice.
  - **Result:** $p > 0.05$
  - **Conclusion:** No significant change; **coffee remains dominant** even on rainy days.

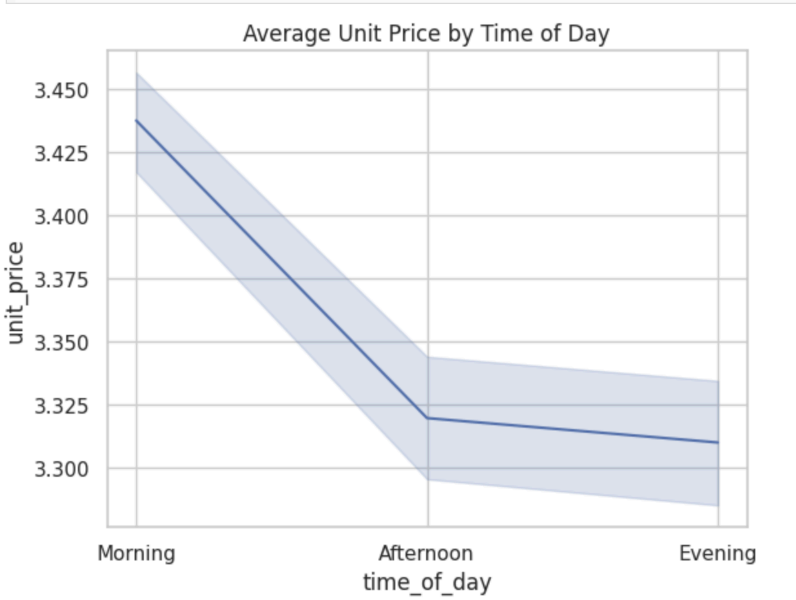


Time of Day Analysis

Sales by Hour

- **Morning (6–11):** Highest sales volume, driven by coffee.
- **Afternoon (12–17):** Slight increase in tea sales.
- **Evening (18–23):** Lower sales overall.





Beverage vs Time Preference

- **Chi-square test result:** $p < 0.05$
- **Conclusion:** Time of day **significantly affects** drink choice.
  - Morning: Coffee
  - Afternoon: Tea becomes more popular

Beverage Preference on Rainy Days

- Coffee dominates overall
- No significant shift toward tea on rainy days (Chi-square test: $p > 0.05$)

TIME ANALYSIS

4. Sales by Time of Day

- Morning hours show highest sales, especially for coffee
- Tea slightly increases in afternoon

5. Hourly Preference

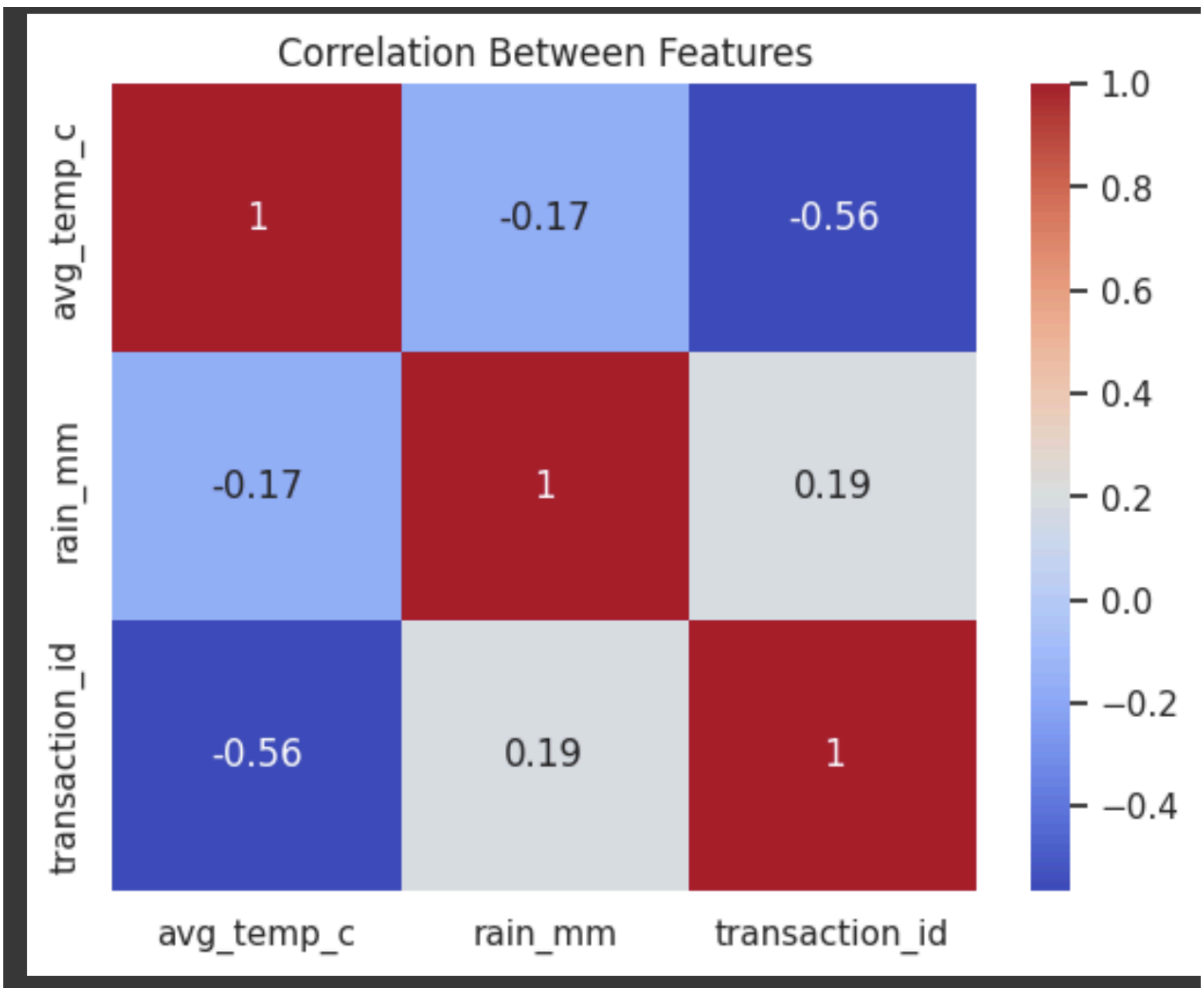
- Chi-square test confirms:
  - Time of day significantly influences beverage choice ($p < 0.05$)

WEATHER EFFECTS ON SALES

- Pearson correlation:
  - Temperature vs. Sales: Negative correlation
  - Rainfall vs. Sales: Slight but not significant
- Sales drop slightly on rainy days, but not significantly enough to affect pricing or product mix

MULTIVARIATE ANALYSIS

- Heatmap of Correlations:
  - Total sales and temperature: Negative
  - Time of day and coffee preference: Strong correlation
- Rainy day + afternoon → tea slightly increases
- Cold morning → coffee peaks



Predicting Unit Price Using Machine Learning

Models & Results:

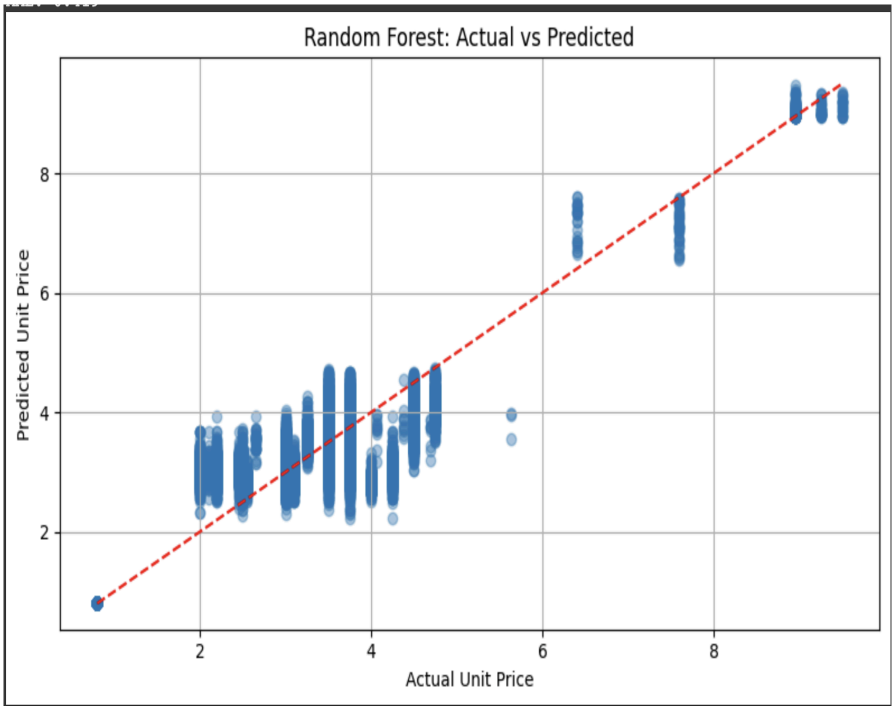
Model	R^2 Score	RMSE	MAE	CV R^2 Mean
Linear Regression	**0.03**	**0.972**	**0.632**	**—**
Random Forest	**0.70**	**0.545**	**0.419**	**0.70**

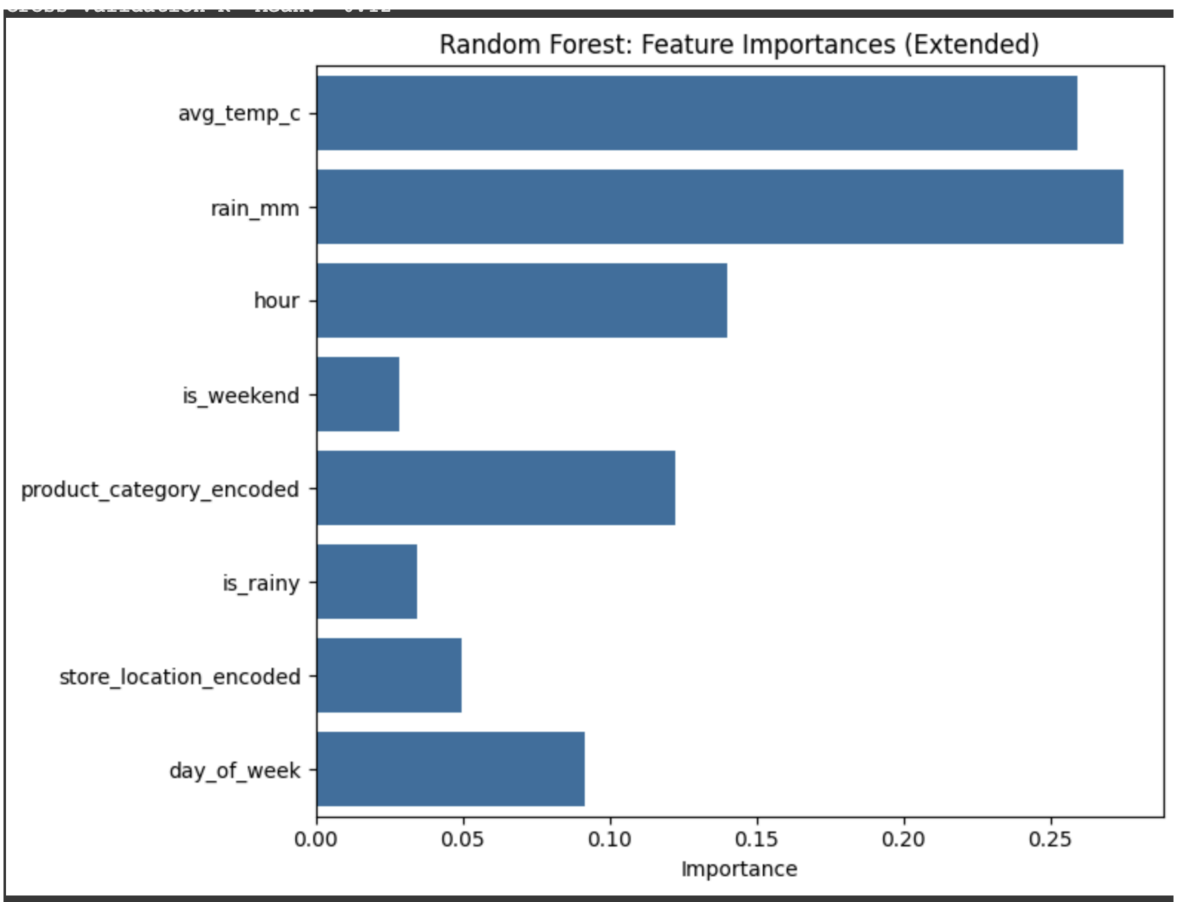
Visualization:

- The scatter plot compares actual vs. predicted unit prices using the Random Forest model.
- The red dashed line represents perfect predictions.
- Most predictions fall close to the line, showing strong predictive power.

Takeaways:

- Random Forest performed significantly better than Linear Regression.
- The model achieved 70% of the variance explained ($R^2 = 0.70$), indicating a good fit.
- This suggests unit price is partially predictable using features like product category, weather, and timing.
- Further improvement may be possible by adding marketing or inventory features.


Random Forest: Actual vs Predicted
Predicted Unit Price
Actual Unit Price
2
4
6
8


Random Forest: Feature Importances (Extended)
avg_temp_c
rain_mm
hour
is_weekend
product_category_encoded
is_rainy
store_location_encoded
day_of_week
0.00
0.05
0.10
0.15
0.20
0.25
Importance

Machine Learning Results: Coffee vs Tea Prediction

Model: Random Forest Classifier

- **Target**: Classify transactions as Coffee (1) or Tea (0)

Performance Metrics:

Metric	Tea (0)	Coffee (1)
Precision	0.44	0.56
Recall	0.19	0.81
F1-Score	0.27	0.66
Accuracy		**0.54**

- The model **correctly identifies Coffee purchases 81% of the time**, but struggles with Tea (19% recall).
- Overall **accuracy is 54%**, better than random guessing for unbalanced classes.
- Precision for Coffee is higher, suggesting more confident classification in that category.

Confusion Matrix Insight:

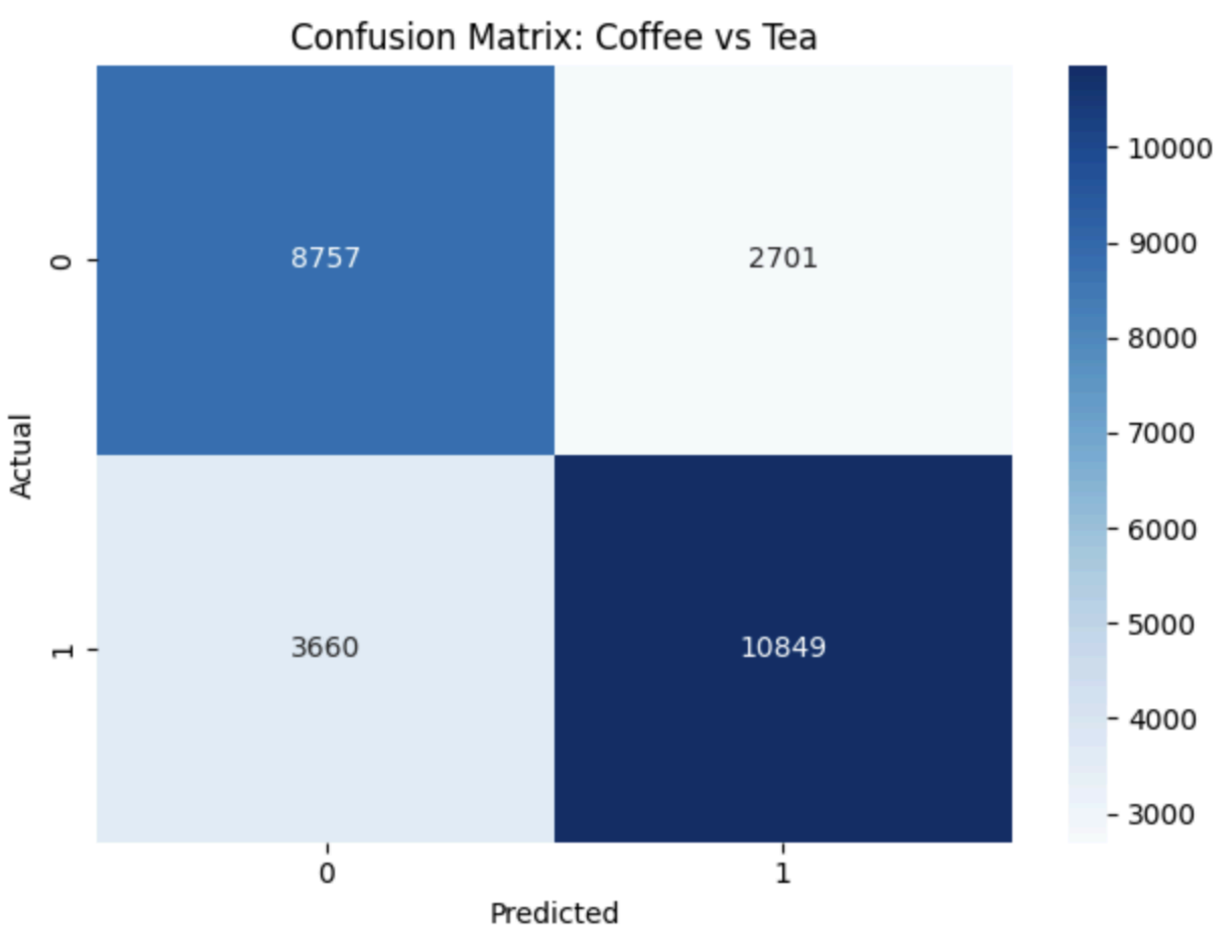
- The model tends to **overpredict Coffee**, misclassifying many actual Tea purchases as Coffee.
- Class imbalance may be affecting performance. Techniques like SMOTE or class weighting can improve this.

Top Features Influencing Prediction:

Rank	Feature	Description

1	hour	Time of transaction (strongest signal)
2	total_sales	Quantity × price per transaction
3	avg_temp_c	Temperature (°C)
4	transaction_qty	Number of items bought
5	unit_price	Price per item
6	rain_mm	Rainfall amount
7	month, day_of_week	Seasonality & weekday effects

Observation: Coffee is heavily consumed in the morning and during cooler days. Tea shows a more subtle pattern, often consumed in the afternoon or rainy days — harder for the model to detect.



CONCLUSION

This project aimed to examine whether environmental conditions—specifically weather (temperature and rainfall) and time of day—have a measurable influence on beverage sales patterns. Through extensive exploratory data analysis and machine learning experimentation, we arrived at several meaningful insights:

- Cold weather strongly correlates with increased coffee sales, especially during morning hours. This supports the idea that consumers gravitate toward hot beverages like coffee to start their day in colder climates.
- Tea sales tend to show a moderate increase during rainy and warmer afternoons, though this trend is not as strong or consistent as coffee in the morning. The relationship between rain and tea consumption appears to be more behavioral than causal.
- Time of day emerged as a more significant driver of beverage choice than weather alone. Morning hours consistently saw the highest volume of coffee sales, regardless of weather conditions.
- Weather does exert some influence, particularly in terms of temperature-driven demand for coffee. However, it is not a dominant factor in determining unit prices or exact sales quantities. Other factors such as marketing campaigns, customer demographics, or seasonal promotions may play a more important role.
- Machine learning models, particularly regression algorithms, struggled to predict unit prices based on weather and time variables alone. Their performance (as reflected in R^2 scores close to zero) suggests that weather is not a sufficient predictor for pricing models.
- Classification models fared slightly better when predicting product categories (coffee vs tea) but were still limited by class imbalance and weak feature predictability.

This project does more than identify interesting patterns in beverage sales—it offers actionable insights that can help businesses operate more efficiently. Cold weather and morning hours reliably correlate with increased coffee demand, providing clear guidance on when to stock and serve more of it. Tea's more subtle preference trends suggest opportunities for niche marketing or promotional targeting, particularly during warm or rainy afternoons.

By leveraging historical sales and weather data, companies can forecast demand more accurately, reducing overproduction and underutilization. These insights contribute directly to cost savings, reduced waste, and better alignment with customer expectations. Ultimately, this kind of data-driven decision-making can support more sustainable and profitable business practices.

Business Implications:

- Companies can benefit from aligning inventory and staffing with predictable time-of-day patterns (e.g., stocking more coffee for morning shifts).
- Weather forecasts may help anticipate minor fluctuations in demand, particularly for tea on rainy afternoons or coffee during cold mornings.
- However, for accurate forecasting and pricing, more granular data inputs—such as customer-level data, event schedules, or marketing activity—are likely required.