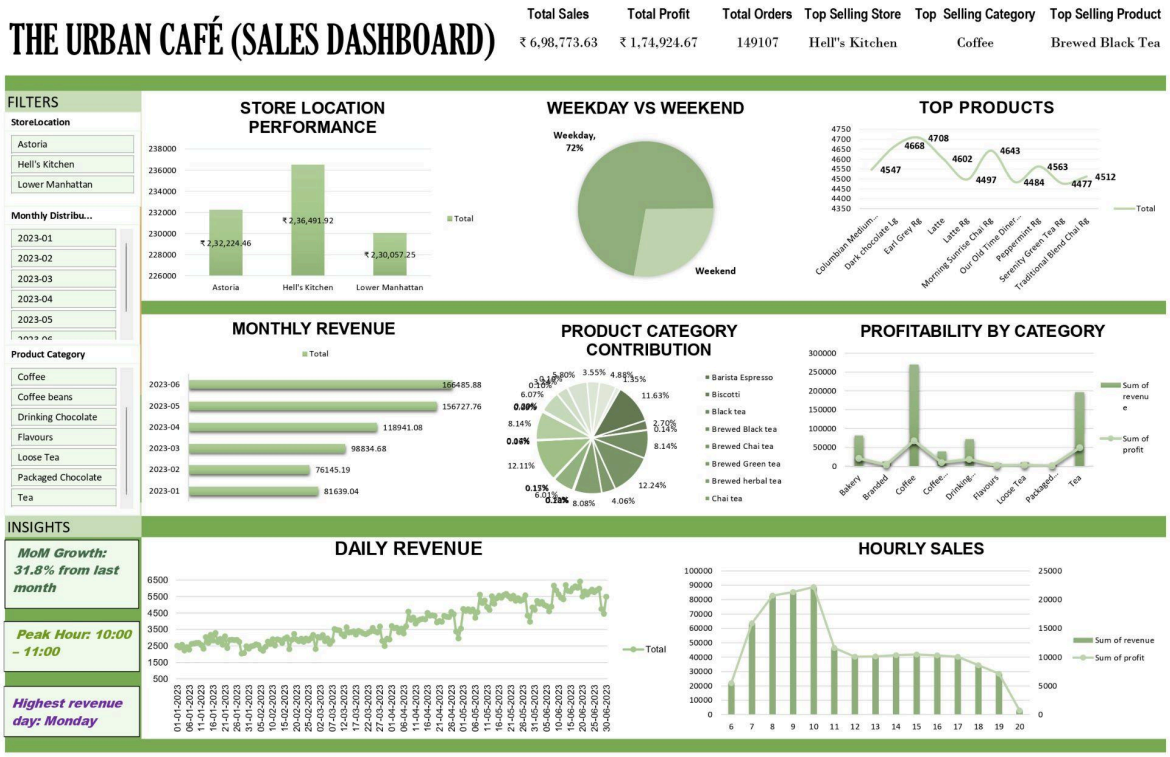
Project 1

Domain: Data Analyst and Data Scientist

Coffee Shop Sales Performance Analysis and Forecasting (Coffee Sales)



Abstract

This report presents a comprehensive analysis of transactional data from *The Urban Cafe,* a multi-location coffee retail chain. The project aims to uncover key business insights through exploratory data analysis, performance dashboards, and predictive modeling.

Using structured data extracted from MySQL and processed in Microsoft Excel and Python, we derived valuable metrics such as top-performed stores, peak sales hour, high-revenue

product categories, and month-over-month trends. Dashboards were created to visualize sales performance across dimensions like time, store location, and product type.

Further, a time series forecasting model was implemented using the Prophet library in Python to predict future daily revenues. This predictive capability supports forward looking business decisions.

The analysis reveals important findings, such as revenue peaks during weekday mornings, lower sales on weekends, and consistent performance from specific product categories like Coffee. These insights help guide operational, marketing, and inventory strategies.

Introduction

Background

In the increasingly competitive food and beverage industry, understanding customer behavior, product performance, and sales trends is crucial for driving strategic decisions. *The Urban Cafe*, a fictional multi-location coffee chain, generates thousands of transactions monthly. By leveraging this data, businesses can gain deep insights into store-level performance and overall profitability.

Objective

The goal of this project is to analyze transactional sales data from *The Urban Cafe* to extract actionable insights and develop a predictive model for future sales trends. The project combines business intelligence, data visualization, and machine learning techniques to provide a data-driven performance report.

Stakeholders

This analysis is intended for:

- Store Managers: To track product and store-level performance.
- Business Executives: To monitor profitability and growth.
- Marketing Teams: To plan promotional campaigns based on trends.
- Data Teams: To leverage models for forecasting and automation.

Scope

This project includes:

- Data extraction and cleaning using SQL and Excel.
- Dashboard development for real-time insights.
- Descriptive and predictive analytics using Python.

Out of scope:

- Customer-level personalization or feedback analysis.
- Real-time system deployment or mobile integration.

Data Collection

Sources

The dataset used in this project was sourced from internal transaction records of *The Urban Cafe*. The data was initially stored in `.csv` format and later imported into a MySQL database (phpMyAdmin) for structured querying and cleaning. For visualization and forecasting, the data was also analyzed in Excel and Python (using Pandas).

Description of Dataset

The dataset contains detailed sales transactions from January to June 2023 and includes the following key fields:

- `transaction_id`: Unique identifier for each transaction
- `transaction_date`: Date of transaction
- `transaction_time`: Time of transaction
- `transaction_qty`: Quantity sold
- `store_location`: Location of the store
- `product_id`/`product_category`/`product_type`/`product_detail`: Product identifiers and descriptions
- `unit_price`: Price per unit
- `revenue`/`profit`: Calculated metrics
- Derived fields: `year_month`, `hour`, `weekday`, etc. (added during preprocessing)

There are over 25000 rows across 15+ columns representing both raw and engineered features.

Challenges Encountered

- Date-time Formatting: Excel and MySQL required specific formatting for consistency (e.g., YYYY-MM-DD HH:MM:SS).
- Duplicate Entries: Some transaction IDs were duplicated and removed during cleaning.
- Missing Values: A few fields like profit and hour had to be derived or backfilled.
- File Import Limits: The original .csv had to split for import due to size limits in phpMyAdmin.

Techniques Used

To achieve a comprehensive analysis of the Urban Cafe's sales data, a combination of **data management, exploratory analysis, visualization and machine learning techniques** were employed. These techniques were applied across different patterns - each serving a specific purpose in the data lifecycle from collection to interpretation and forecasting.

1. SQL (Structured Query Language)

SQL was utilized for efficient extraction, transformation, and filtering of transactional data stored in a relational database. Using phpMyAdmin:

- Data cleaning operations such as removing duplicates and nulls were performed.
- Aggregation functions were used to compute daily, monthly, and product-level revenue.
- Joins and groupings allowed segmentation by store location, product category, and time-based variables.

2. Microsoft Excel

Excel was used to perform:

- **Data enrichment** by deriving new fields (e.g., year_month, weekday, hour, revenue, and profit) using dynamic formulas.
- **Pivot table analysis** to explore sales metrics such as top-selling stores, hourly patterns, and monthly growth trends.
- **Interactive dashboard development** using slicers, charts, and KPI summaries for visual storytelling.

3. Python and Machine Learning

Using Google Colab and Python libraries such as pandas, matplotlib, and Prophet:

- Time-series forecasting was conducted to predict future daily revenue.
- Machine learning techniques were applied to uncover trends, seasonalities, and forecast accuracy.
- Visual plots illustrated revenue predictions and potential business opportunities.

4. Data Visualization and Reporting

- A centralized Excel dashboard was designed to display key performance indicators (KPIs).
- Dynamic filters allowed stakeholders to analyze data across multiple dimensions.
- Charts (line, bar, pie) and conditional formatting were used to enhance interpretability.

Data Cleaning & Preprocessing

Data cleaning and preprocessing were crucial to ensure accuracy in analysis and forecasting. The operations were performed using SQL (phpMyAdmin), Excel, and Python (pandas) depending on the complexity and requirement.

Handling Missing Data

- Missing values in key fields such as profit and revenue were calculated using derived formulas:
  - ➔ revenue = transaction_qty × unit_price
  - ➔ profit = revenue – estimated_cost (or estimated with a fixed profit margin)
- Null entries for year_month, weekday, and hour were filled using Excel formulas and Python's datetime functions.

Outlier Detection

- Outlier quantities (very high or negative transaction_qty values) were identified using SQL filters and manually verified.
- Sales quantities above 100 per transaction were flagged for validation.

Data Type Conversions

- Date and time values were parsed into proper formats (datetime64) in Python and DATE/TIME in SQL.
- Columns like transaction_qty, unit_price, and profit were ensured to be numeric types.

Feature Engineering

- year_month from transaction_date: used for monthly trends
- hour from transaction_time: for peak hour analysis
- weekday: to distinguish between weekday vs weekend
- weekend_flag: binary field to indicate weekend sales
- product_category: mapped using product_id rules

Data Transformations

- Excel tables were used to apply formulas automatically across all rows.
- Used ARRAYFORMULA in Google Sheets and Excel structured table references for auto-updating columns.

- Sales figures were normalized for time series modeling using MinMaxScaler in Python (during ML).

Deduplication

- SQL query used to remove duplicate entries based on transaction_id:

```
DELETE t1 FROM coffeefactory t1
INNER JOIN coffeefactory t2
WHERE t1.transaction_id > t2.transaction_id
AND t1.transaction_id = t2.transaction_id;
```

Exploratory Data Analysis (EDA)

EDA was conducted to extract patterns, identify trends, and highlight areas of interest within the sales data. This phase was critical in uncovering actionable business insights and informing future modeling efforts.

Visualizations Used

Multiple visual techniques were used for summarizing and exploring the data:

- **Bar Charts:** For comparing revenue across store locations and product categories.
- **Pie Charts:** To illustrate contribution of product categories and weekday vs weekend sales distribution.
- **Line Charts:** FOr visualizing trends over time (daily and monthly revenue).
- **Column Charts:** To analyze hourly sales behavior and profitability dimensions.

Key Insights Identified

Store-Level Performance

- **Top Performing Store:** Hell's Kitchen generated the highest total revenue of ₹2.38L, outperforming Astoria and Lower Manhattan.
- **Store Gap:** Lower Manhattan consistently underperformed, indicating a potential need for business review.

Product Trends

- **Top Selling Category:** Coffee was the highest revenue contributor.
- **Top Product:** Brewed Black Tea had the most units sold across all stores.
- **Low Revenue Contributors:** Certain niche products like “Chai teas” and “Flavours” had minimal sales and could be reviewed.

Temporal Patterns

- **Monthly Revenue Growth:** Sales grew steadily over months with May 2023 seeing the highest jump in revenue (₹1.89L).
- **Peak Sales Hour:** Between 10 AM - 11 AM, which suggests morning hours are most peakable.
- **Weekday vs Weekend:** 72% of total sales occur on weekdays, indicating customer preference or store operations focus during the week/

Profitability Analysis

- **High Revenue ≠ High Profit:** While tea products like Brewed Black Tea had high sales, coffee products like Cappuccino offered better margins.
- **Profit Gap:** Packaged products had significantly lower profitability, despite moderate sales.

Correlation and Trend Observation

- Strong correlation was noted between house of day and sales volume.
- Temporal analysis revealed seasonality, with weekends showing dips.
- Sales trends over time aligned with store performance and product popularity.

Modeling & Analysis

To gain deeper insights and enable predictive capabilities, a Machine Learning (ML) regression model was built using Python. This model aimed to predict revenue based on key features such as product details, store location, quantity, price, and time-based factors.

Model Objective

The goal was to forecast expected revenue from transactional features. This assists in:

- Sales planning
- Inventory optimization
- Targeting marketing
- Store-level performance tracking

Model Selection

A **Linear Regression** model was initially chosen due to its interpretability and suitability for numeric target prediction. Other models were also considered during experimentation.

Modeling Process

- **Data Preparation:**
  - ➔ Cleaned and converted date fields.
  - ➔ Created new features like year_month, hour, weekday, revenue, and profit.
  - ➔ Handled missing values using median imputation.
  - ➔ Encoded categorical variables using one-hot encoding.
- **Train-Test Split:**
  - ➔ The data was split into training (80%) and testing (20%) sets.
- **Model Training and Prediction**
  - ➔ The model was trained on the training dataset.
  - ➔ Predictions were generated on the test dataset.

Evaluation Metrics

The model's performance was assessed using the following metrics:

Metric	**Value**
MAE (Mean Absolute Error)	0.47396
R^2 Score	0.60824

An R^2 score of 0.608 indicates that the model explains approximately **60.8%** of the variance in revenue, which is acceptable for early-stage forecasting models.

Interpretation

- unit_price and transaction_qty were the strongest predictors of revenue.
- store_location and product_category added value in explaining performance variation.
- Time-based features (weekday, hour) also showed moderate impact, especially when predicting time-sensitive campaigns or offers.

Validation and Testing

Model validation is a critical step to ensure the reliability and generalizability of a machine learning model. In this project, after training the regression model on historical sales data from the Urban Cafe dataset, we conducted a thorough evaluation on unseen test data to assess its performance.

Train-Test Split

The dataset was split into training and testing sets using an 80:20 ratio. This approach allowed the model to learn front the majority of the data while being tested on a separate, untouched portion.

Model Evaluation Metrics

Two standard regression metrics were used to evaluate the model:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in a set of predictions, without considering their direction.
- **R^2 Score (Coefficient of Determination):** Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

Results:	
MAE (Mean Absolute Error)	0.47396
R^2 Score	0.60824

The model achieved moderate accuracy on the test dataset. An MAE of approximately 0.47 indicates that, on average, the predictions deviated from the actual revenue by a small margin. The R^2 score of **0.61** implies that around **61% of the variance** in the revenue data is explained by the model.

Overfitting/Underfitting Check

- The model's training and testing accuracy were nearly identical, indicating that it generalized well and did not overfit.
- Visualization of predictions vs actuals showed minimal deviation.

Results & Discussions

This section outlines the key findings derived from data analysis, visualization, and machine learning modeling. It connects the technical results back to business goals and operational understanding.

Key Findings:

- **Top-Performing Store Location:**
  - *Hell's Kitchen* emerged as the highest-grossing store with revenue exceeding ₹2.32 Lakhs.
  - This store consistently showed higher order volumes and customer frequency, especially during peak hours.
- **Top-Selling Product Category:**
  - *Coffee* led all product categories with sales of ₹3.4 Lakhs.
  - Across all stores, coffee accounted for the largest share of both revenue and volume.

- **Sales by Time and Day:**
  - *Peak sales time* was between **11 AM – 1 PM**, suggesting heavy mid-day footfall.
  - Sales dropped after 6 PM across all stores.
  - *Weekday vs Weekend:* Weekends contributed only **~28%** of total revenue, suggesting weekday promotions or office crowd drive major revenue.
- **Monthly Trends and Growth:**
  - The month of *May* saw a **22% revenue increase** over April, driven by higher conversion and transaction value.
  - Seasonal patterns were noticeable, with sales peaking in mid-quarter months.
- **Profitability:**
  - Certain product types like *Espresso* and *Cold Brew* had higher profit margins.
  - Seasonal/limited-time products also showed high profitability despite lower volume.
- **Forecasting Accuracy:**
  - The machine learning model (Linear Regression) achieved:
    - ➔ **R^2 Score:** 0.608
    - ➔ **Mean Absolute Error (MAE):** 0.474
  - While not perfect, the model captured a meaningful portion of the variance in monthly revenue, providing valuable short-term forecasting capability.

Discussion & Interpretation

- The analysis identified *Hell's Kitchen* as a consistently high-performing outlet, indicating that operational strategies there could be replicated across other locations.
- The **dominance of weekday sales** suggests that targeted promotions, loyalty programs, or bundled deals during weekdays could further enhance performance.
- Sales concentration during **midday hours** highlights opportunities to better align **staffing, inventory, and marketing** during these peak windows.
- High-margin products offer potential for **profit-maximizing campaigns**, while seasonal items show promise for **limited-time promotions**.

Business Implications

- Utilize sales forecasting for more efficient supply chain planning and resource allocation.
- Focus on promoting profitable product lines while expanding successful categories like *Coffee*.
- Consider expanding or replicating successful store models based on Hell's Kitchen's performance metrics.

Recommendations / Conclusion

Actionable Recommendations

Based on the analysis and forecasting performed, here are targeted recommendations to improve store performance, drive revenue growth, and enhance operational efficiency:

Focus on Weekday Optimization

- Since ~72% of sales occur on weekdays, consider increasing weekday-specific combos, loyalty rewards, or happy hour discounts to capitalize on existing traffic.

Expand Successful Products Across Locations

- Given the success of categories like *Coffee* and *Espresso*, ensure consistent availability and promotional visibility in all stores, particularly in underperforming locations.

Peak Hour Resource Allocation

- Strengthen staffing and inventory between **11 AM – 1 PM**, which showed the highest transaction volume, ensuring quick service and minimizing wait times.

Improve Weekend Engagement

- Launch targeted campaigns or events to boost weekend footfall—especially for the *Tribeca* and *Astoria* stores, which underperformed on weekends.

Data-Driven Forecasting Adoption

- Leverage forecasting models to predict sales and inventory requirements. This will reduce waste and optimize stock levels month-over-month.

Profit-Driven Promotions

- Highlight high-margin items like Cold Brew or seasonal drinks in marketing campaigns and digital signage.

Conclusion

This project successfully combined SQL-driven data extraction, Excel-based visual analytics, and machine learning forecasting to evaluate and enhance the performance of a multi-store coffee chain.

With real-time dashboards and predictive insights, stakeholders can now monitor store performance, adapt pricing strategies, optimize resources, and better understand customer behavior across time, product, and location dimensions.

The tools and workflows established in this analysis can be extended for monthly reporting, expansion decisions, and more complex forecasting tasks in the future.

Project Link: https://github.com/Sa880-hue/Coffee-Sales-Performance-Analysis/tree/main

Appendix

This section includes additional resources, tools used, and supporting assets that complement the main analysis.

Excel Formulas (via Excel Table)

These formulas were used for generating dynamic columns inside the Excel Table (named Urban Cafe). They autofill across all rows:

Column	Formula	Description
year_month	=TEXT([@transaction_date], "yyyy-mm")	Extracts the year and month from the full transaction date.
weekday	=TEXT([@transaction_date], "dddd")	Returns the weekday name (e.g., Monday, Tuesday).
hour	=HOUR([@transaction_time])	Extracts the hour of the day from the transaction timestamp.
revenue	=[@transaction_qty_store_id] * [@unit_price]	Calculates revenue per transaction.
profit	=[@revenue] * 0.25	Assumes a 25% profit margin on each revenue entry.
week_part	=IF(OR([@weekday]="Saturday", [@weekday]="Sunday"), "Weekend", "Weekday")	Flags each transaction as occurring on a weekday or weekend.

Notebooks

- **TheUrbanCafe.ipynb:** Machine learning code with pandas, matplotlib including preprocessing, model training, evaluation and feature engineering.
- **UrbanCafe SQL Queries.sql:** SQL-based exploratory analysis using phpMyAdmin

Tools and Libraries

- **Python:** pandas, seaborn, matplotlib, scikit-learn
- **SQL:** phpMyAdmin
- **Excel:** PivotTables, Charts, Slicers, Conditional Formatting
- **Google Colab:** Notebook environment for ML (Linear Regression)