# COFFEE SALES ANALYSIS REPORT

**1. Introduction:**
The purpose of this project is to analyze coffee sales data using PySpark for distributed data processing and Matplotlib for visualization. The analysis explores transaction volume, coffee type preferences, time-based sales patterns, and payment trends. The project aims to derive actionable insights that can support business strategies, improve revenue, and identify peak sales periods.

**2. Dataset Overview:**
• Entries: 3547 sales transactions
• Columns:
  o hour_of_day – Time of purchase (24-hour format)
  o cash_type – Payment mode (e.g., card)
  o money – Transaction amount (in USD)
  o coffee_name – Type of coffee purchased
  o Time_of_Day – Morning, Afternoon, or Night
  o Weekday – Day of the week
  o Month_name – Month of transaction
  o Date, Time – Timestamp of transaction

The dataset is clean and complete, containing accurate timestamps and categorized transaction records suitable for analytical modeling.

**3. Key Findings:**
a) **Sales and Revenue Overview:**
• Total Revenue: $112,245.58
• Total Transactions: 3,547
• Average Spend per Transaction: $31.65

b) **Top Performing Coffee Types:**
• Latte – $26,875.30
• Americano with Milk – $24,751.12
• Cappuccino – $17,439.14
• Americano – $14,650.26
• Hot Chocolate – $9,933.46
These results show that milk-based beverages dominate the sales distribution.

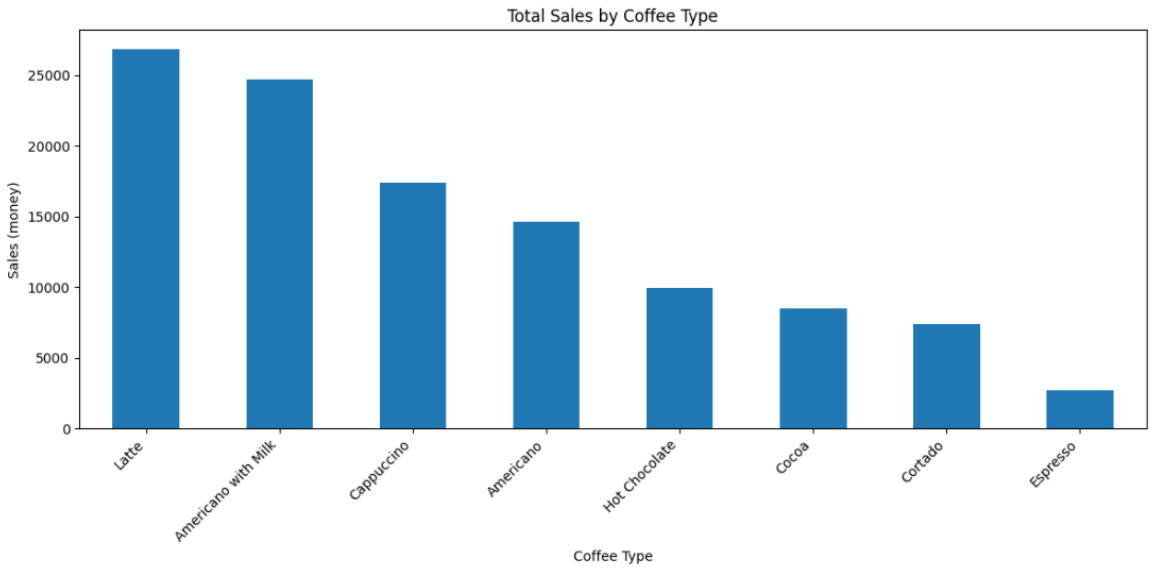
c) **Hourly and Temporal Trends:**
• Highest sales hours: 10 AM, 11 AM, 4 PM, 7 PM, and 5 PM.
• Morning (10–11 AM) and Afternoon (4 PM) are peak revenue periods.
• Weekdays outperform weekends in total revenue, with Tuesday leading ($18,168.38).
• October and March recorded the highest monthly sales volumes.

d) **Payment Method Analysis:**
• All transactions were made using cards, ensuring faster processing and traceability.

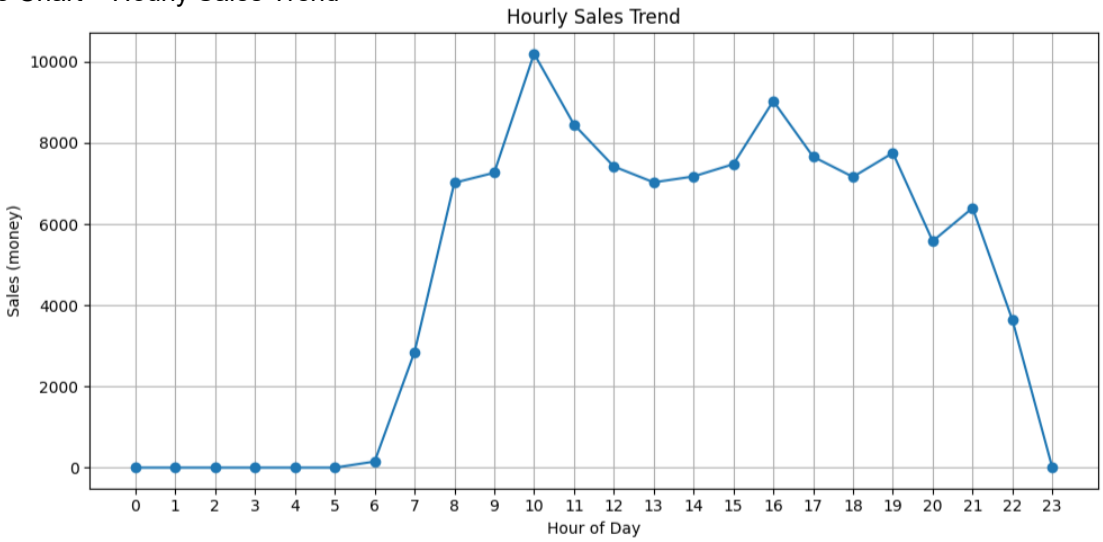
e) **Predictive Modeling:**
• A simple Linear Regression model was built to predict sales amount using *hour_of_day* and *Weekdaysort*.
• Model metrics: $R^2 = 0.0588$ and RMSE = 4.67.
Although weak correlation, the model indicates slight influence of time and weekday on transaction value.

**4. Data Visualization (DV):**
The following visualizations were created using PySpark and Matplotlib:

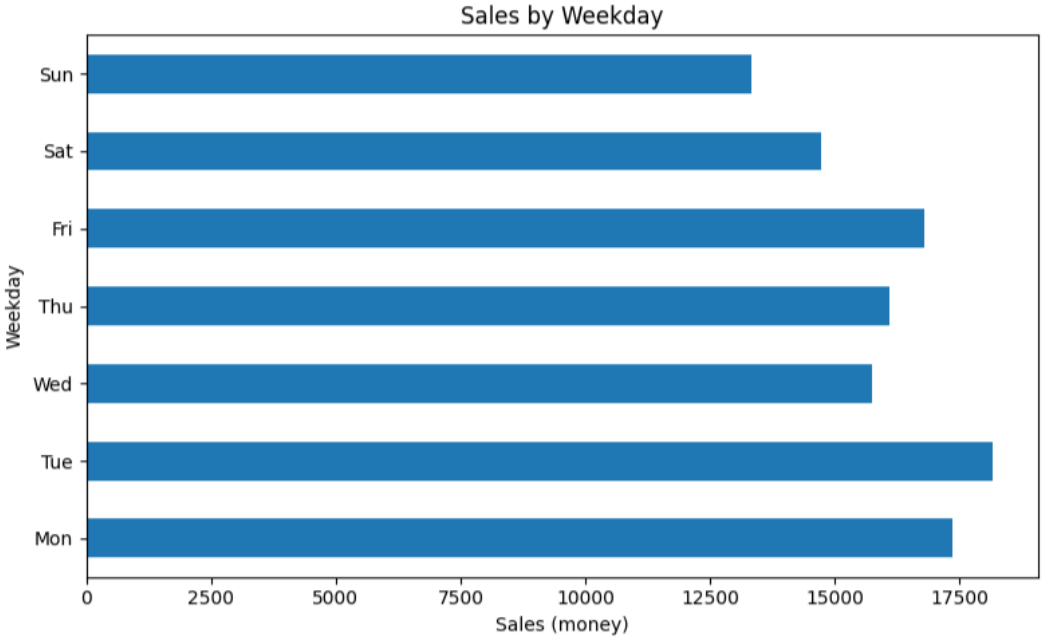
• Bar Chart – Total Sales by Coffee Type



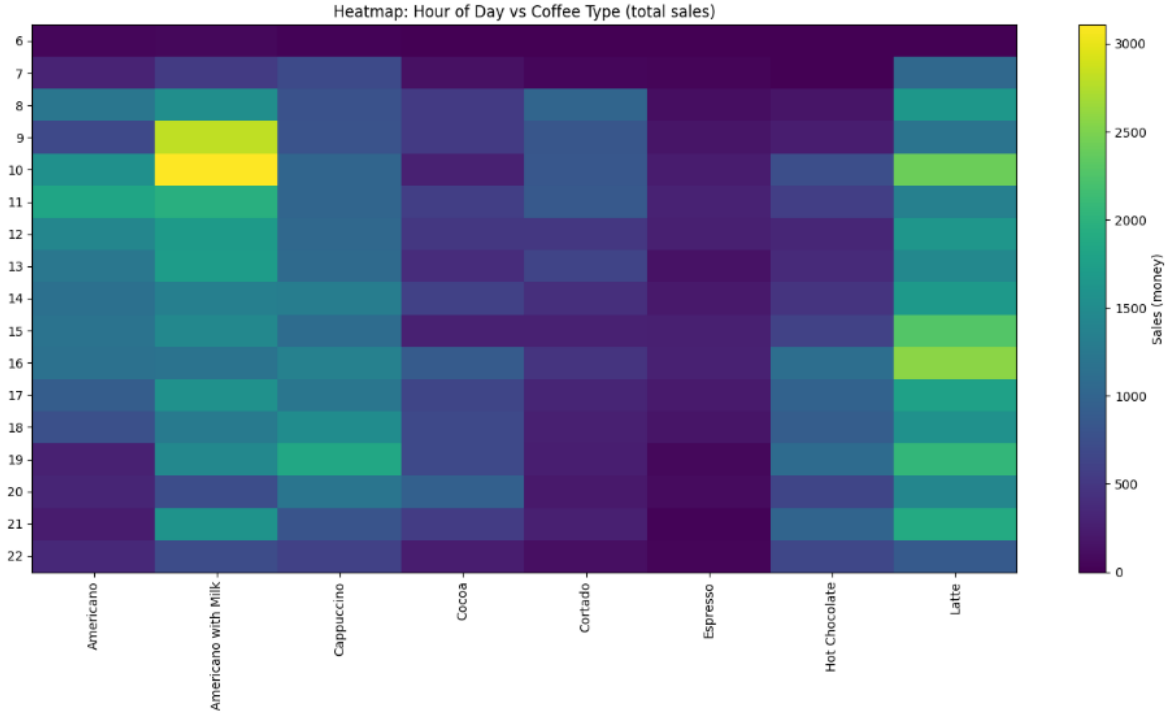

• Line Chart – Hourly Sales Trend




• Horizontal Bar Chart – Sales by Weekday



• Heatmap – Hour of Day vs Coffee Type (Total Sales)




These visuals clearly highlight revenue distribution, busiest time slots, and beverage popularity trends.

**5. Conclusion:**

The analysis confirms that:

• Milk-based beverages (Latte and Americano with Milk) are top revenue drivers.

• Sales peak during mid-mornings and late afternoons, suggesting ideal times for promotions.

• Card transactions dominate all purchases, indicating customer preference for digital payment modes.

• Linear regression showed limited predictability, but revealed mild dependency of revenue on time patterns.

**Recommendations:**

• Offer combo deals and loyalty programs during high-traffic hours.

• Introduce festive discounts in high-performing months like October and March.

• Use time-based sales forecasts to optimize staffing and inventory planning.