

# **Coffee Shop Sales Analysis & Daily Revenue Prediction**

**(Coffee Sales Dataset – 149k Transactions)**

**Student: Abdulwahab Almutlak**

**Program: AI Bootcamp – Week 2 Project**

**Dataset: *Coffee Shop Sales* (Kaggle)**

**Tools: Python, Pandas, Matplotlib, Seaborn, Scikit-learn,  
Streamlit, Plotly**

**Date: November 2025**

# Table of Contents

Title Page ..... 1

Table of Contents ..... 2

Introduction (Dataset & Problem Definition) ..... 3

Data Exploration ..... 4

Data Cleaning ..... 5

Exploratory Data Analysis (EDA) ..... 6

Dashboard Overview (Screenshots) ..... 8

Insights & Conclusion ..... 9

References ..... 11

## Introduction

This project utilizes a real world dataset which includes over 149,000 coffee shop transactions collected at three different corners in the city of New York. The dataset contains product categories, product types, transaction time-stamps, store locations and sales revenue for each sale.

To help coffee shop chains to make better decisions on staffing, product mix, marketing efforts and operational planning, they need to know how daily sales are performing. But, as is, it's usually hard to make sense of transaction data alone.

## Problem Definition

The main goal of this project is to:

Investigate and examine the sales dataset to identify significant trends

Determine the best-selling items, the busiest retailers, and the periods of highest sales

Create an interactive Streamlit dashboard that makes it simple for users to filter, view, and comprehend the data

Create a machine learning model that can forecast daily income using date-related characteristics (day, month, year, weekday, weekend)

This project offers thorough insights that can assist business owners and decision-makers in increasing operational effectiveness and revenue forecasting by fusing exploratory analysis, interactive visualization, and predictive modeling

# Data Exploration

In this section, we performed an initial exploration of the coffee shop dataset to understand its structure and key characteristics. The dataset contains sales records from three different store locations, including information such as product categories, transaction dates, quantities sold.

I examined the number of rows and columns, checked data types, reviewed sample records, and identified basic patterns such as the most common products and the distribution of transactions across stores.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 149116 entries, 0 to 149115
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   transaction_id         149116 non-null  int64
1   transaction_date       149116 non-null  datetime64[ns]
2   transaction_time       149116 non-null  object
3   transaction_qty        149116 non-null  int64
4   store_id              149116 non-null  int64
5   store_location        149116 non-null  object
6   product_id            149116 non-null  int64
7   unit_price            149116 non-null  float64
8   product_category      149116 non-null  object
9   product_type          149116 non-null  object
10  product_detail         149116 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(4), object(5)
memory usage: 12.5+ MB
```

	transaction_id	transaction_date	transaction_qty	store_id	product_id	unit_price
count	149116.000000	149116	149116.000000	149116.000000	149116.000000	149116.000000
mean	74737.371872	2023-04-15 11:50:32.173609984	1.438276	5.342063	47.918607	3.382219
min	1.000000	2023-01-01 00:00:00	1.000000	3.000000	1.000000	0.800000
25%	37335.750000	2023-03-06 00:00:00	1.000000	3.000000	33.000000	2.500000
50%	74727.500000	2023-04-24 00:00:00	1.000000	5.000000	47.000000	3.000000
75%	112094.250000	2023-05-30 00:00:00	2.000000	8.000000	60.000000	3.750000
max	149456.000000	2023-06-30 00:00:00	8.000000	8.000000	87.000000	45.000000
std	43153.600016	NaN	0.542509	2.074241	17.930020	2.658723

# Data Cleaning

In this project, the coffee sales dataset contained several issues that required correction to ensure accuracy and reliability

## 1-Fixing Column Names

Removed extra spaces and replaced spaces with underscores to make column names clean and usable in Python.

## 2-Converting Date & Time Columns

A new column `transaction_datetime` was created by combining:

`transaction_date`

`transaction_time`

This allowed extraction of useful features like hour, day, month, and weekday.

## 3-Creating New Useful Columns

These columns help with analysis and prediction:

hour: hour of the transaction

day\_name: name of the day

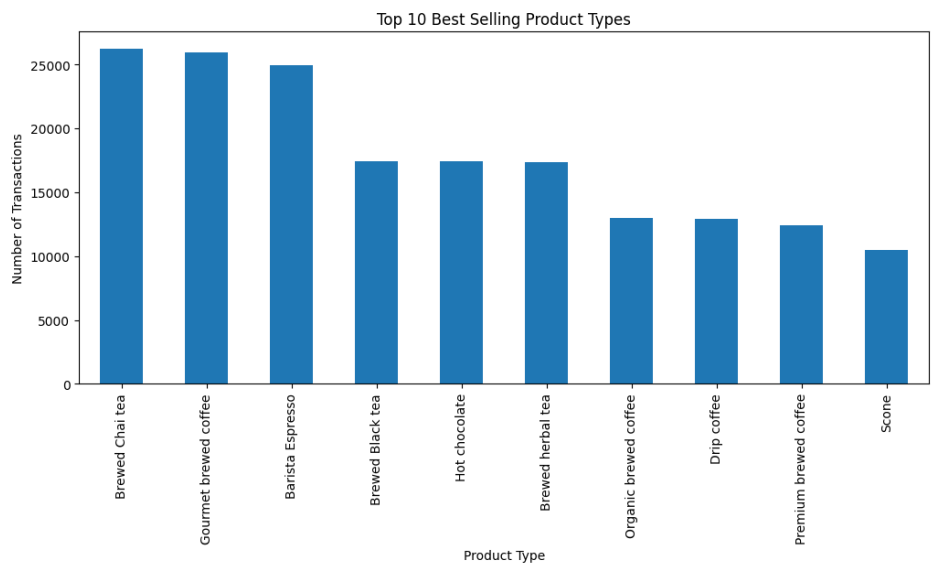
month: month number

day\_of\_week: number representing day of week

is\_weekend: 1 if weekend, 0 otherwise

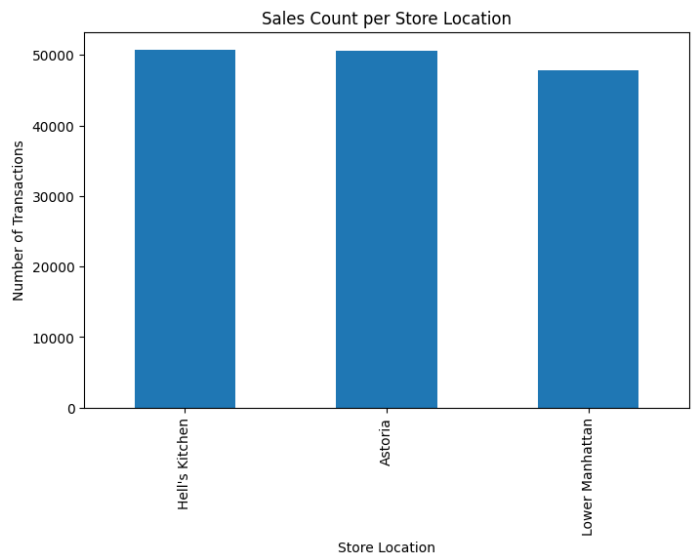
This step is partly cleaning and partly feature engineering, but both contribute to preparing data for modeling.

# Exploratory Data Analysis (EDA)



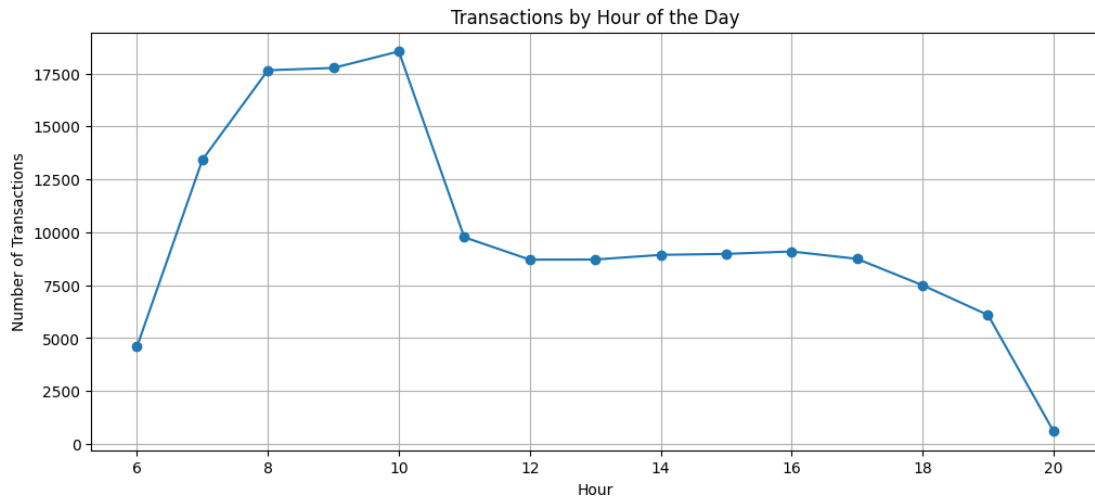
**Figure 1: Top 10 Best-Selling Product Types**

This bar chart shows the top 10 best-selling product types based on the total number of transactions. *Brewed Chai Tea* and *Gourmet Brewed Coffee* are the most popular items, lower-ranking items such as *Scone* and *Premium Brewed Coffee* show significantly fewer transactions



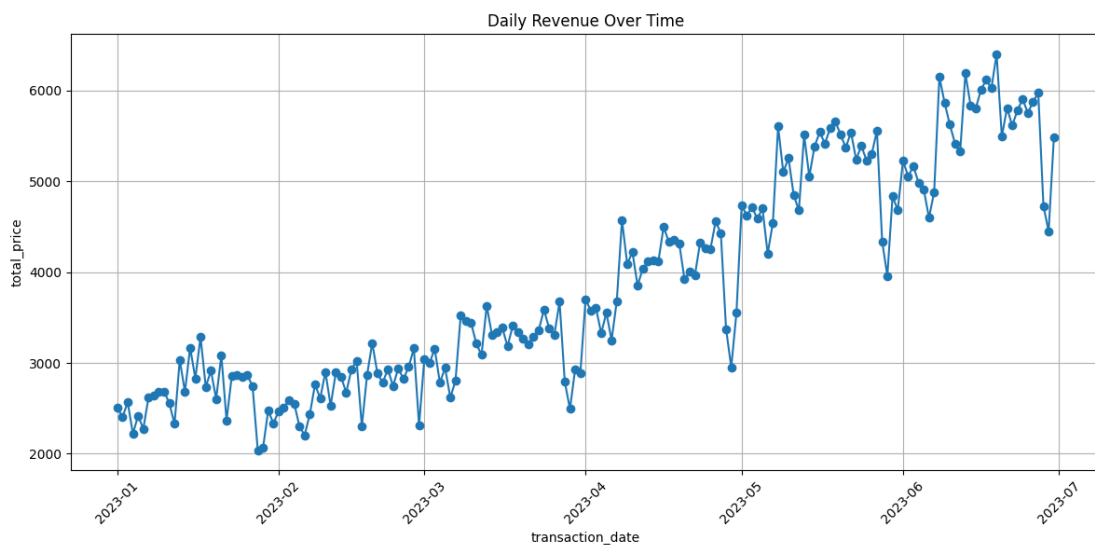
**Figure2: Sales Count per Store Location**

This bar chart compares the number of transactions across the three coffee shop branches: *Hell's Kitchen*, *Astoria*, and *Lower Manhattan*. The results show that Hell's Kitchen and Astoria have nearly identical transaction volumes, while Lower Manhattan records slightly fewer but still competitive sales



**Figure 3: Transactions by Hour of the Day**

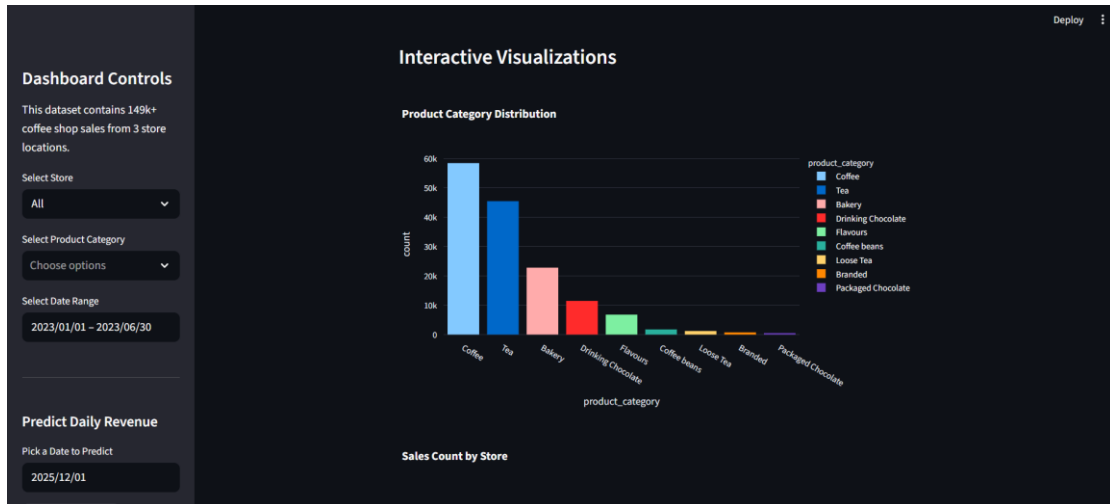
This line chart shows how customer activity changes throughout the day. The number of transactions peaks between 8 AM and 10 AM, which indicates the morning rush when customers typically buy coffee. After 11 AM, transaction volume drops and remains moderate during the afternoon, suggesting lower customer traffic compared to morning hours.



**Figure 4: Daily Revenue Over Time**

This line chart shows the daily revenue generated across all stores between January and June 2023. The trend reveals a gradual increase in revenue over the months, with noticeable fluctuations on certain days. The rise in later months may indicate increased customer demand, seasonal effects, or promotional activities.

## Dashboard overview



This is the main interface of the interactive Streamlit dashboard. On the left sidebar, users can filter the dataset.

The main area displays interactive visualizations such as product category distribution, sales count per store.

These dynamic charts update instantly based on the selected filters, allowing users to explore the dataset efficiently.

The screenshot shows the "Predict Daily Revenue" form. It includes a title "Predict Daily Revenue", a label "Pick a Date to Predict", a date input field containing "2024/12/01", a "Predict Revenue" button, and a green box displaying the "Predicted Revenue: \$5,117.44".

The user selects a date and clicks "Predict Revenue" after which the model (Random Forest Regressor) estimates the expected revenue for that day.

The predicted value is displayed clearly in a green box, making it easy for users to interpret and use the result for decision-making.

## Insights and conclusion

### Key Insights

Based on the exploratory data analysis and dashboard visualizations, several important insights were identified:

Coffee beverages dominate sales, especially *Brewed Chai Tea*, *Gourmet Brewed Coffee*, and *Barista Espresso*, which consistently show the highest transaction counts.

Astoria and Hell's Kitchen stores generate the largest number of transactions, while *Lower Manhattan* shows slightly lower activity.

Peak transaction hours occur between 8 AM and 11 AM, indicating strong morning demand.

Daily revenue shows a clear upward trend over the months, suggesting business growth or seasonal effects.

### What I Learned

Perform full data cleaning and feature engineering on a real dataset.

Conduct EDA using various plotting techniques (bar charts, line plots, time-series analysis).

Build an interactive dashboard using Streamlit.

Train prediction models (Linear Regression + Random Forest) and compare performance.

Export a trained ML model and integrate it inside a live dashboard.

### Dataset Limitations

While the dataset was rich, some limitations exist:

The dataset covered only a specific time period.

No weather and seasonal and external factors were included that might affect sales.

## Overall Conclusion

The business is stable with strong morning demand and clear best-selling categories.

Coffee and Tea should remain the primary focus, while low-selling items could be bundled or promoted to improve performance.

## References

[Coffee Shop Sales](#)

[Streamlit • A faster way to build and share data apps](#)