Coffee Shop Expansion Analysis

Project Statement

The business aims to expand by opening three coffee shops in India's top three major cities. Since its launch in January 2023, the company has successfully sold its products online and received an overwhelmingly positive response in several cities. As a data analyst, your task is to analyze the sales data and provide insights to recommend the top three cities for this expansion.

Datasets Used

Sales.csv : sale_id, sale_date, product_id, customer_id, total, rating

Customers.csv : customer_id, customer_name, city_id

Products.csv : product_id, product_name, price

City.csv : city_id, city_name, estimated_rent

Tools & Technologies

 Python: Pandas, Seaborn, Matplotlib

SQL: Joins, aggregation

Visuals: Line chart, bar chart, bubble chart, box plot, Histograph

Data Cleaning

```python
import pandas as pd

# 1. Load the raw CSVs
sales_df = pd.read_csv(r'C:\Users\manal\Downloads\sales.csv')
products_df = pd.read_csv(r'C:\Users\manal\Downloads\products.csv')
customers_df = pd.read_csv(r'C:\Users\manal\Downloads\customers.csv')
city_df = pd.read_csv(r'C:\Users\manal\Downloads\city.csv')

# 2. Drop duplicates
sales_df.drop_duplicates(inplace=True)
products_df.drop_duplicates(inplace=True)
customers_df.drop_duplicates(inplace=True)
city_df.drop_duplicates(inplace=True)

# 3. Drop rows with critical nulls
sales_df.dropna(subset=["sale_id", "sale_date", "product_id", "customer_id", "total"], inplace=True)
products_df.dropna(subset=["product_id", "product_name", "price"], inplace=True)
customers_df.dropna(subset=["customer_id", "customer_name", "city_id"], inplace=True)
city_df.dropna(subset=["city_id", "city_name"], inplace=True)

# 4. Format and clean date column in sales
sales_df["sale_date"] = pd.to_date(sales_df["sale_date"], errors="coerce")

# 5. Save cleaned files in different folder
sales_df.to_csv(r'C:\Users\manal\Downloads\Coffee\sales.csv', index=False)
products_df.to_csv(r'C:\Users\manal\Downloads\Coffee\products.csv', index=False)
customers_df.to_csv(r'C:\Users\manal\Downloads\Coffee\customers.csv', index=False)
city_df.to_csv(r'C:\Users\manal\Downloads\Coffee\city.csv', index=False)
```
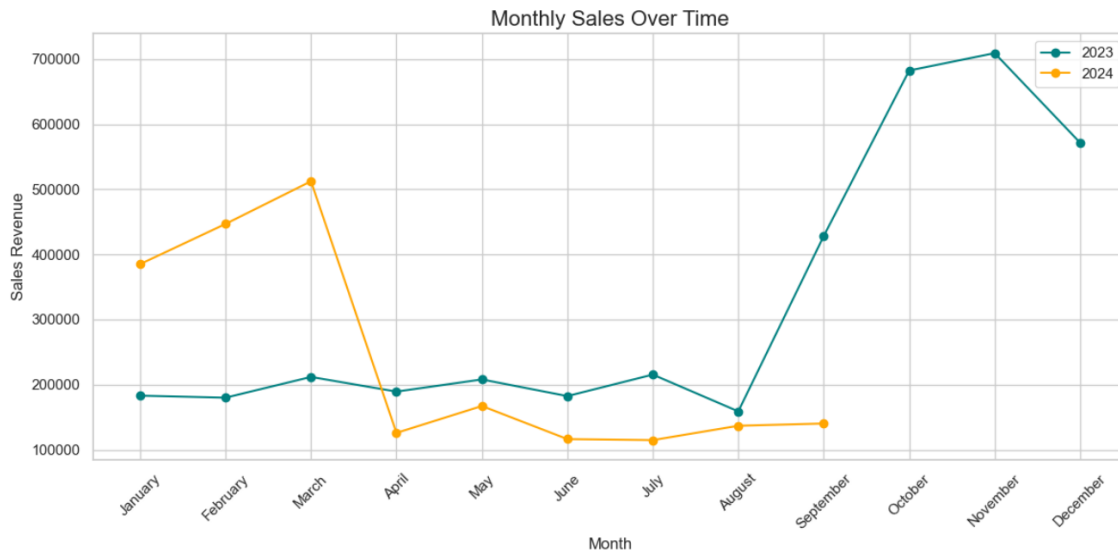
This Python script performs initial data preprocessing using the **pandas** library to prepare raw CSV files for analysis. It involves loading datasets, removing duplicates and nulls, formatting dates, and saving cleaned versions for further use.

EDA

Total sales over time :

The line chart above shows a comparative analysis of **monthly sales revenue** for the years **2023 (first year of operations)** and **2024 (second year)**. This visualization helps evaluate the company's performance over time and identify growth or decline patterns.

In 2023**,** As the company's first year of operations, 2023 demonstrated **consistent and healthy sales growth**, especially in the **final quarter**. Sales remained steady between **₹180K and ₹220K** during the first eight months, increased suddenly and strongly from **September to November**, reaching a peak of **₹710K** in November. This pattern suggests **successful scaling, market acceptance**, and possibly **effective seasonal promotions.**

2024 began with strong sales: **₹385K in January**, rising to **₹515K in March**, indicating **initial growth** and possibly a **larger customer base** carried over from 2023. However, post-March, sales **declined sharply**, falling to nearly **₹130K** in April and staying consistently low throughout the rest of the year. This unexpected drop suggests potential issues such as **marketing slowdown**, **customer churn**, or **operational inefficiencies**.

Revenue by Product :
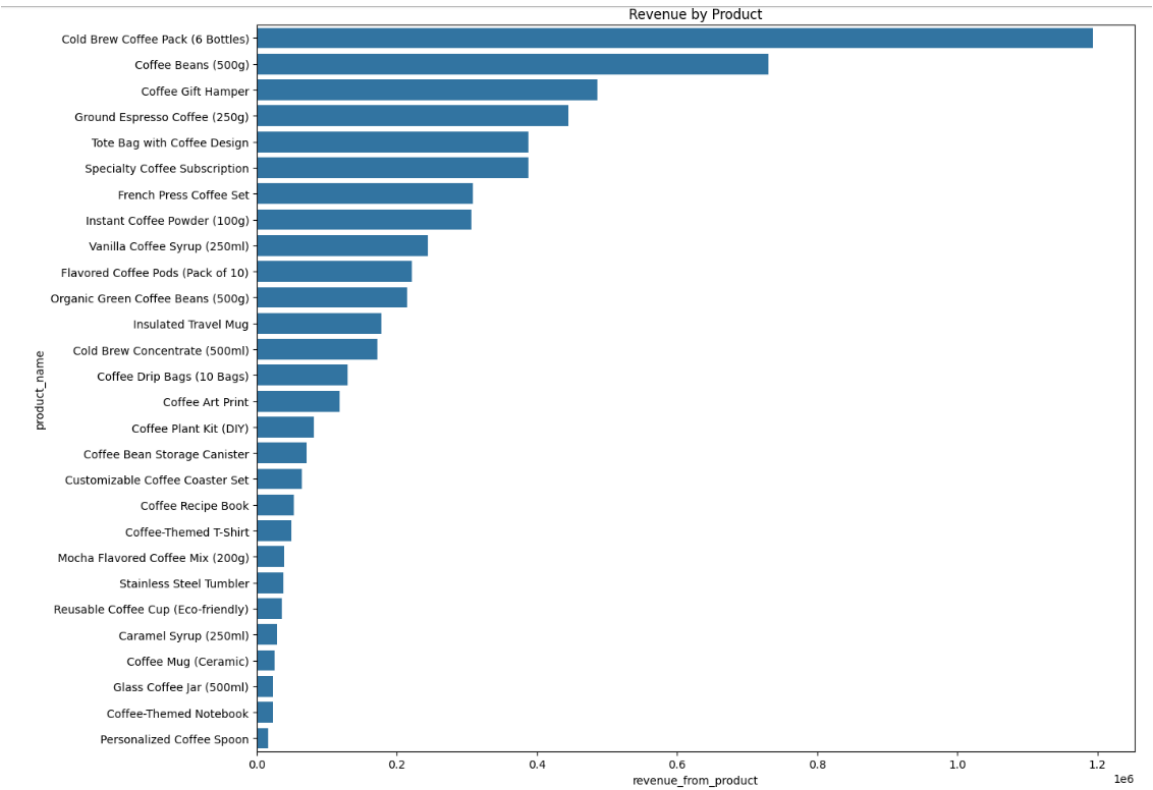
The above bar chart illustrates the **total revenue generated by each product** sold by the company. This visual representation helps identify the **best-performing products** and informs decisions around **inventory planning**, **marketing focus**, and **product bundling strategies**.

The **Cold Brew Coffee Pack (6 Bottles)** is by far the **highest revenue-generating product**, exceeding ₹1.2 million. Its strong performance suggests high customer demand, possibly due to convenience and trendiness.

Other top performers include : Coffee Beans (500g), Coffee Gift Hamper, Ground Espresso Coffee (250g), Tote Bag with Coffee Design

Products like the **Personalized Coffee Spoon**, **Coffee-Themed Notebook**, and **Glass Coffee Jar (500ml)** are among the lowest contributors to total revenue.

Product-wise total unit sold :


Sales_count by Product
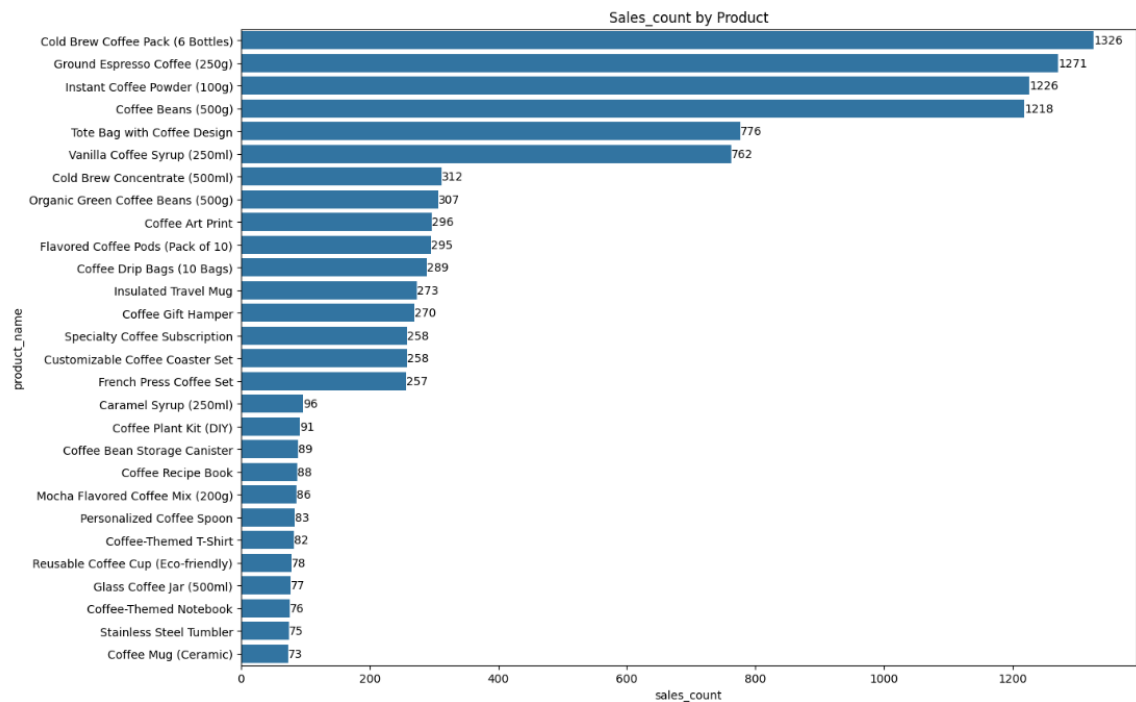
The chart above shows the **sales count** for each product, revealing how frequently individual items were purchased by customers. This analysis helps in identifying the **most popular products**, customer preferences, and potential inventory priorities.

Top-Selling Products : The **Cold Brew Coffee Pack (6 Bottles)** leads in unit sales with **1,326 units sold**, indicating it is both a high-demand and high-conversion product. Following closely are : **Ground Espresso Coffee (250g)** – 1,271 units, **Instant Coffee Powder (100g)** – 1,226 units, **Coffee Beans (500g)** – 1,218 units. These top four items represent the core coffee products and show consistent consumer interest.

Mid-Tier Performers : Items like the **Tote Bag with Coffee Design**, **Vanilla Coffee Syrup**, and **Cold Brew Concentrate** performed moderately well, suggesting they serve as complementary or lifestyle-oriented purchases.

Low Sales Volume Products:Products such as the **Coffee Mug (Ceramic)**, **Stainless Steel Tumbler**, and **Coffee-Themed Notebook** had fewer than 80 units sold each.

**Estimated coffee consumers by city :**
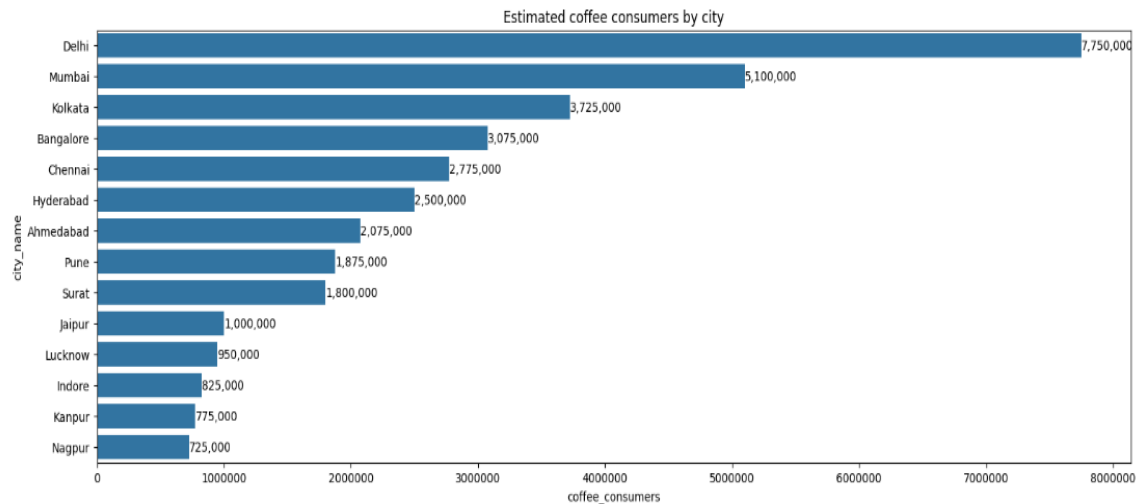
The chart above visualizes the **estimated number of coffee consumers** in major Indian cities. This data-driven estimate is likely based on each city's population and assumed **25 percent** of population are **coffee consumers**. It provides valuable insights into **potential market size** and **target regions** for sales, marketing, and expansion strategies.

Delhi leads with **7.75 million** estimated coffee consumers, making it the largest potential market. **Mumbai** follows with **5.1 million**, while **Kolkata** and **Bangalore** also show strong numbers at **3.72 million** and **3.07 million**, respectively.

Other high-potential cities include **Chennai** (2.77 million), **Hyderabad** (2.5 million), **Ahmedabad**, **Pune**, and **Surat** (1.8M–2.1M range).

Smaller yet notable markets include **Indore**, **Kanpur**, and **Nagpur**, ranging from **725K to 850K** consumers.

Revenue by city :
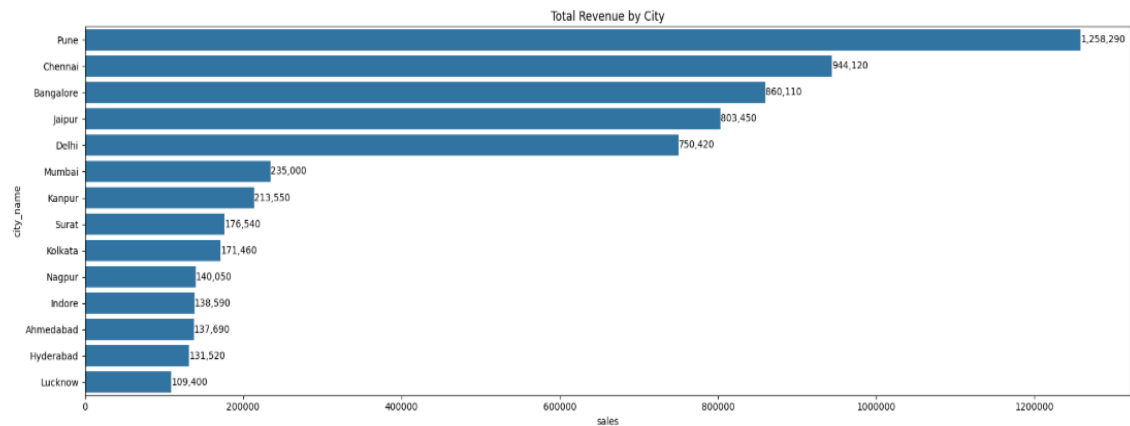
This horizontal bar chart illustrates the **total sales revenue generated from each city**, providing insight into which urban markets are contributing most significantly to the company's earnings.

Pune is the top-performing city with a revenue of **₹1,258,290**, which stands out as the highest among all cities. **Chennai** and **Bangalore** follow with strong figures of **₹944,120** and **₹860,110**, respectively. **Jaipur** (**₹803,450**) and **Delhi** (**₹750,420**) also show high revenue.

Surprisingly, **Mumbai**—India's commercial capital—generated only **₹235,000**, indicating a potential gap between market size and actual revenue capture.Other cities like **Kanpur**, **Surat**, and **Kolkata** fall into a mid-tier revenue range (~₹170K–₹215K).

Hyderabad, **Ahmedabad**, and **Lucknow** are currently under performing in terms of revenue, with figures under **₹140K**.

Average order value by city :
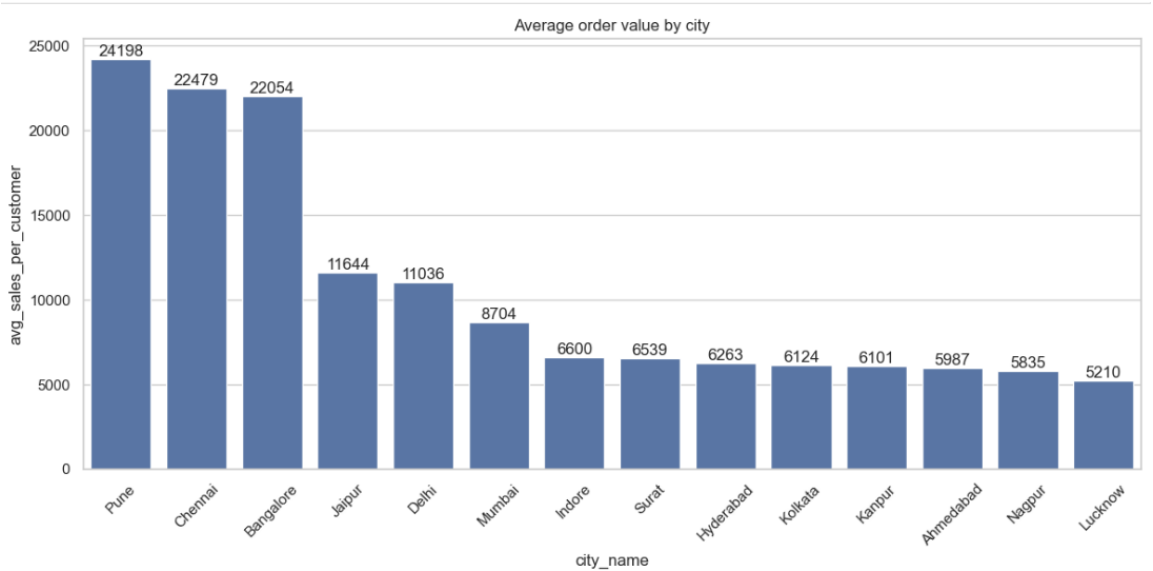
This vertical bar chart shows the **average sales revenue generated per customer** in each city. It provides insight into **how much customers from different regions typically spend**, helping identify high-value markets.

Pune leads with the **highest average order value** at ₹24,198 per customer, indicating a high-spending customer base. **Chennai** (₹22,479) and **Bangalore** (₹22,054) also show strong average spending, positioning them as key markets with premium customer behavior.

Jaipur (₹11,644) and **Delhi** (₹11,036) fall into a mid-tier segment, showing decent average revenue per customer. **Mumbai**—despite being a large metro—has a **lower average value of ₹8,704**, suggesting either smaller basket sizes or less frequent purchases.

Cities like **Lucknow** (₹5,210), **Nagpur**, **Ahmedabad**, and **Kanpur** all report **average values under ₹6,000**, pointing toward price-sensitive customers or limited product adoption.

Customers count by city :
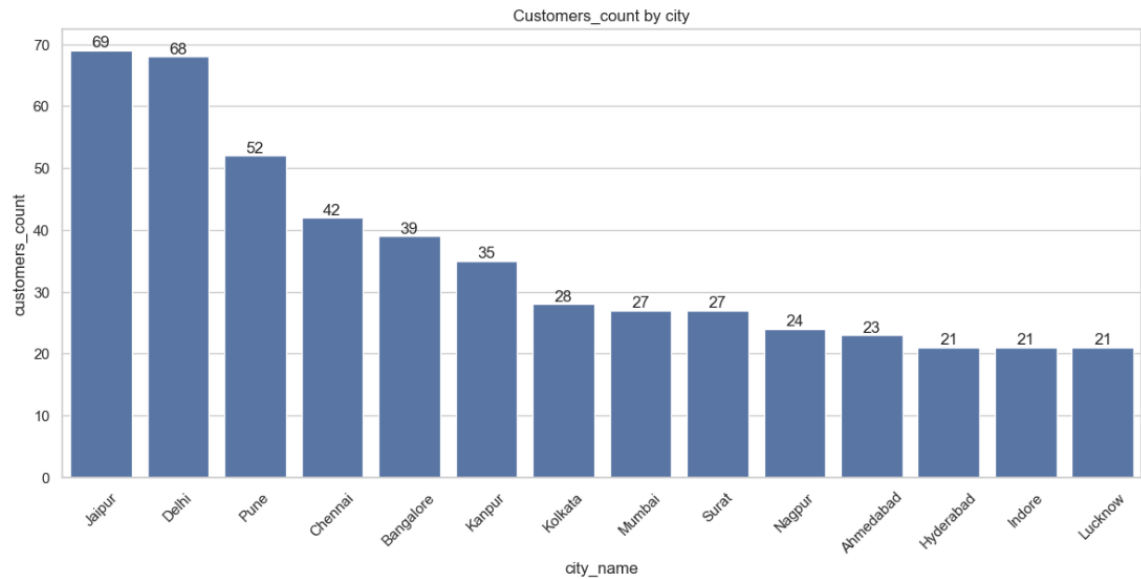
This bar chart displays the **number of unique customers** from each city who made purchases. It highlights which geographic regions have the **largest customer bases** and helps evaluate market penetration and engagement across cities.

Jaipur (69) and **Delhi** (68) have the **highest number of customers**, indicating strong market presence and outreach in northern India. **Pune** (52), **Chennai** (42), and **Bangalore** (39) also show high customer counts, reflecting solid engagement in major metropolitan regions.

Cities like **Kanpur** (35), **Kolkata** (28), **Mumbai** (27), and **Surat** (27) fall in the mid-range.

Nagpur, **Ahmedabad**, **Hyderabad**, **Indore**, and **Lucknow** all have **21–24 customers**, suggesting either under-penetrated markets or lower brand visibility.

Repeat Customer Order Counts (Distribution):

Distribution of Repeat Customer Order Counts
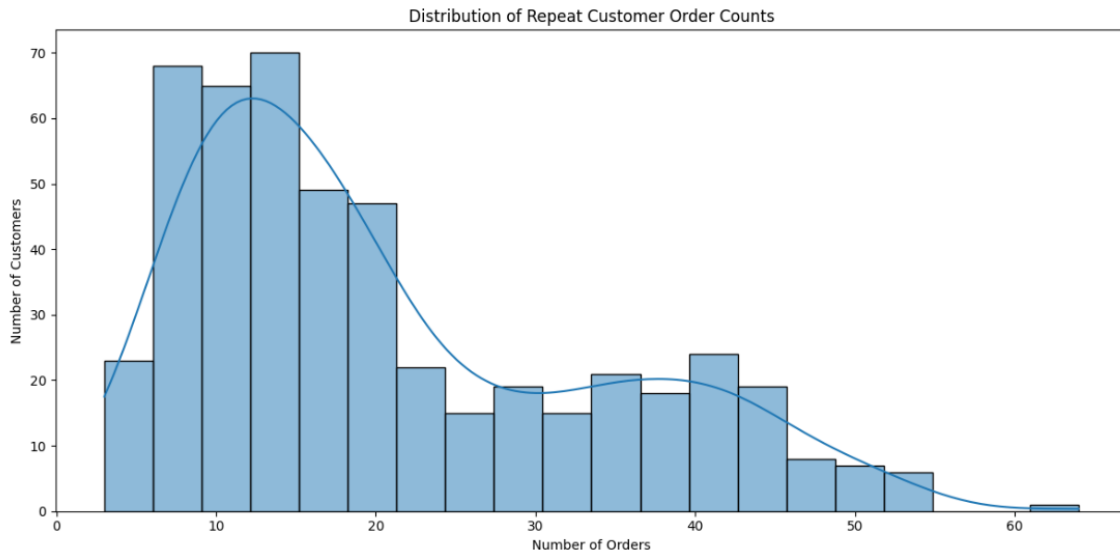
The distribution of **repeat customer orders** is **right-skewed**, indicating that while **most repeat customers** tend to place between **8 and 15 orders**, a smaller group of **highly loyal customers** place **30 or more orders**. This **long-tail pattern** suggests the presence of a **valuable customer segment** with **high purchase frequency**, which can be targeted for **loyalty programs** or **personalized marketing strategies**.

Spend per Customer (Distribution) :


Customer Spending Distribution
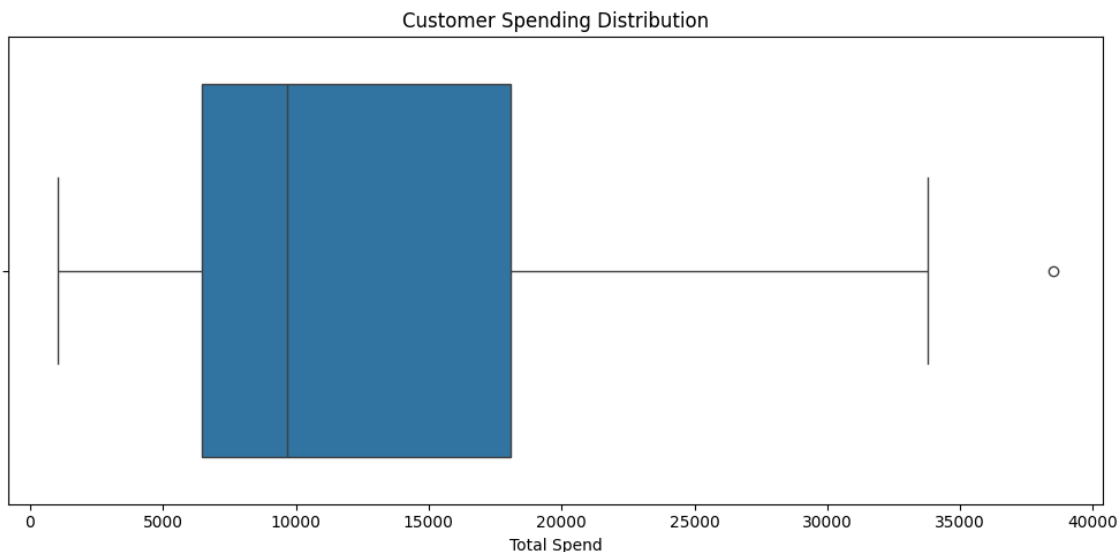
The box plot illustrates that **customer spending** is moderately concentrated between **₹5,000 and ₹20,000**, with a **median spend** around **₹10,000**. A small number of customers spend **considerably more**, with one **outlier** nearing **₹38,000**, indicating the

presence of **high-value customers**. This suggests an opportunity to **segment and target** these customers for **premium offerings** or **exclusive loyalty rewards**.

Products popularity by city :
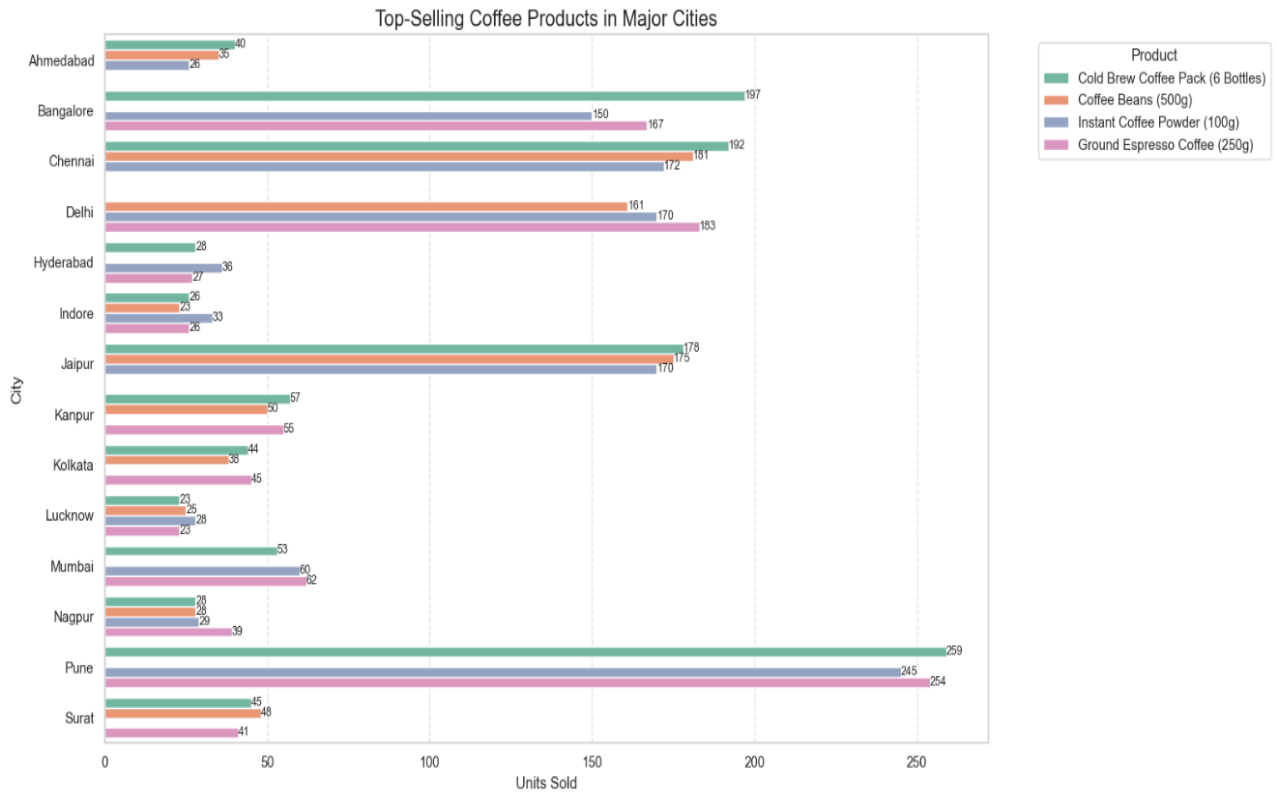
The chart reveals that **Pune** leads in coffee consumption with the **highest sales across all products**, indicating a **diverse and mature coffee market**. **Bangalore, Chennai, Delhi, and Jaipur** also demonstrate **strong demand** for both **premium brews (Cold Brew, Espresso)** and **convenience options (Instant Coffee)**. In contrast, cities like **Nagpur, Indore, Lucknow, and Hyderabad** show **lower but emerging demand**, suggesting **growth potential** through **targeted marketing and product education**. Each city exhibits a **unique blend of preferences**, highlighting the importance of **localized strategies** in product placement and promotion.

City-wise sales vs Estimated rent :
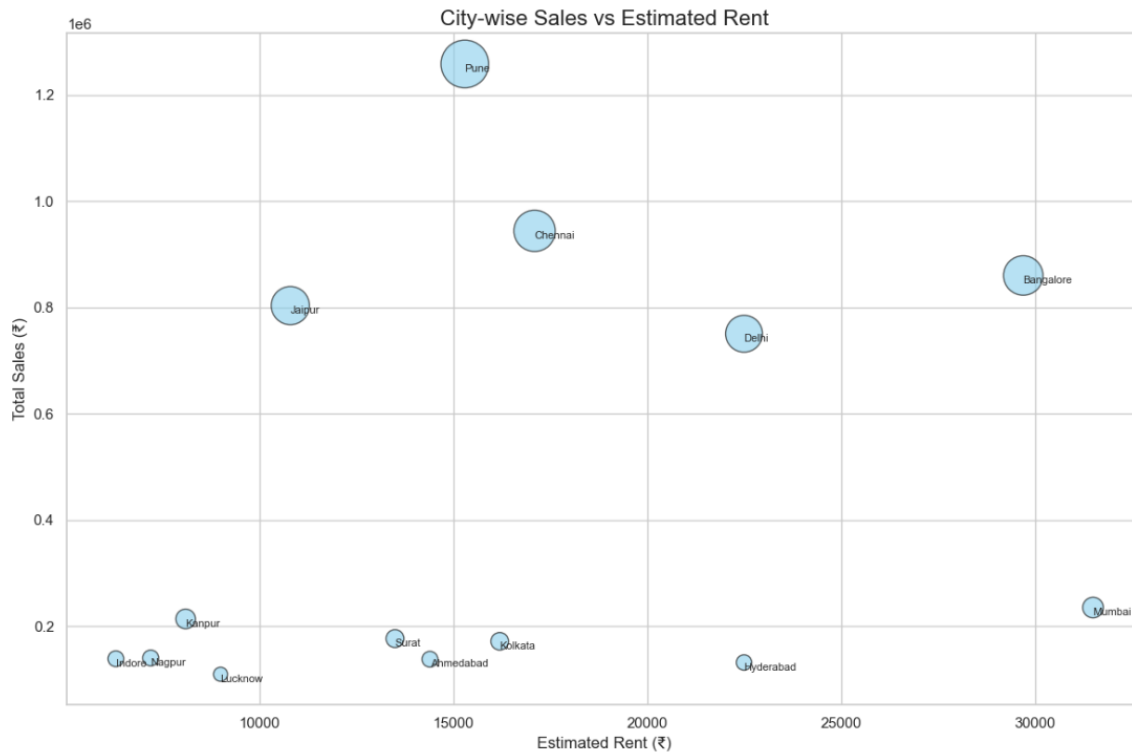
High Revenue & Low Rent Cities (Most Attractive):

Pune stands out as the top performer:

- Highest sales (~₹12.5 lakhs)
- Moderate rent (~₹16,000)
- Large consumer base (large bubble)

Chennai and Jaipur also have:

- High sales
- Moderate rent
- Large bubbles (high potential)

High Rent with High Sales:

Bangalore and Delhi:

- Rent is high (~₹22,000–₹30,000)
- Still manage to generate high sales

High Rent but Low Sales:

Mumbai
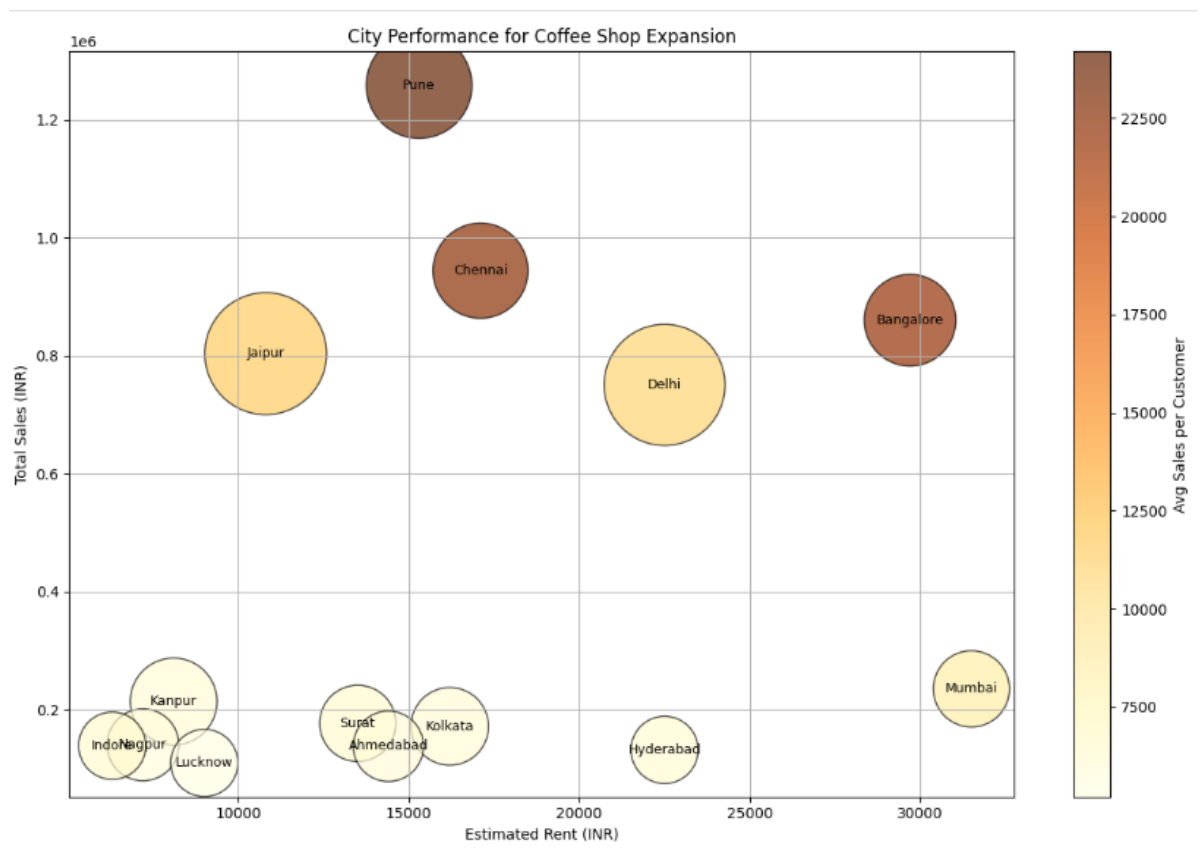
- Highest rent (~₹32,000)
- Low sales (~₹2.3 lakhs)
- Small bubble size (lower consumer conversion)

Low Rent but Low Sales:·

Cities like **Lucknow, Nagpur, Indore, and Ahmedabad**:

- Have low rent (~₹8,000–₹14,000)
- But also low revenue and small consumer base

City Performance for Coffee Shop Expansion :

Top Performing Cities (Expansion Recommended):

Pune

- Highest sales (~ ₹1.25 million)
- Moderate rent (~₹20K)
- Large bubble → Strong customer base

Chennai

- High sales (~ ₹950K)
- Lower rent than Bangalore and Delhi
- Good customer base

Jaipur

- Affordable rent (~₹11K)
- Decent sales (~ ₹850K)
- Big bubble = Good customer base

Cities to Watch (Mixed Potential):

Delhi

- Good sales
- Moderate rent
- Large customer base
- Average spend is not the highest

Bangalore

- Good revenue & spend per customer
- **High rent** may reduce margins

Lower-Priority Cities:

Lucknow, Indore, Nagpur, Kanpur

- Small bubbles (low customer count)
- Low revenue
- Low to average spend

Mumbai

- Highest rent

- Low sales for cost