King Saud University
College of Computer and Information Sciences
Department of Information Technology

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

IT 326 Course Project

Semester-2, 1446H

Analyzing the Key Factors Behind High-Rated Cafés in Riyadh

Prepared by

<Lamya Alnahdi, 443202647>
<Norah Alfaheed, 444200779>
<Leen Alqahtani, 443200591>
<Hissah Alotaibi, 444200349>

Table of Contents

1. Introduction

Cafés have become an integral part of modern social and cultural life, offering much more than just coffee. Their widespread popularity stems from several factors, including ambiance, customer experience, and brand identity. However, despite the overall success of cafés, competition among brands continues to intensify, making it challenging for businesses to stand out.

The key issue lies in understanding the real factors that influence a café's success. While many cafés offer similar products and services, some achieve remarkable popularity, while others struggle to succeed. This study aims to explore the essential elements that attract customers and enhance their overall experience. By identifying these factors, café owners and stakeholders can develop strategies to improve their appeal in an increasingly competitive and dynamic market.

The main research question guiding this study is:
"What are the key factors that make cafés highly rated in Riyadh?"
The study will explore several critical aspects, including customer preferences, service quality, rating, and the number of reviews.

2. Data sources

We utilized a web scraping method to extract data from Google Maps, as BS4 was blocked from accessing its content. To overcome this, we employed the **Instant Data Scraper** extension, manually selecting each café to extract reviews and ratings. The tool, available here, scans webpage elements and exports data into an Excel file, allowing us to bypass restrictions. However, this approach required manual intervention, resulting in challenges such as formatting inconsistencies, the need for data cleaning, and the potential for human error in data entry. The collected data includes reviews and ratings for multiple cafés, with extraction dates ranging from **February 1 to February 4, 2025**, and is available in shared **OneDrive links** for each café:

- **Flame Cafe** (1/2/2025): Link
- **Margin Cafe** (1/2/2025): Link
- **OnOff Cafe** (1/2/2025): Link
- **Ghamra Cafe** (1/2/2025): Link
- **HAI Cafe** (1/2/2025): Link
- **Namq Cafe** (1/2/2025): Link
- **Volume Cafe** (3/2/2025): Link
- **Kiltura Cafe** (3/2/2025): Link
- **Iota Cafe** (3/2/2025): Link
- **Croi Cafe** (3/2/2025): Link
- **Tnfis Cafe** (3/2/2025): Link
- **Folds Cafe** (4/2/2025): Link
- **Groovy Cafe** (4/2/2025): Link
- **Baa Cafe** (4/2/2025): Link
- **Tul Cafe** (4/2/2025): Link
- **Nafees Cafe** (4/2/2025): Link
- **Khussa Cafe** (4/2/2025): Link
- **Mjaz Cafe** (4/2/2025): Link

- **Deers Cafe** (4/2/2025): Link
- **Era Cafe** (4/2/2025): Link
- **M Dee Cafe** (4/2/2025): Link
- **Dopa Cafe** (4/2/2025): Link

To improve our analysis, we later developed a custom scraper script to append **busy hours** data for each café, which will be crucial for the next phases of analysis.

The dataset includes information about coffee shops, their reviews, and operational details. Below is a detailed description of the data and an evaluation of its potential biases:

1. **Description of the Data:**
    a. **Observations:** The dataset contains information about coffee shops. Each record represents a single coffee shop.
    b. **Features and Data Types:**
        - **Coffee:** Names of the coffee shops (Qualitative- Nominal).
        - **Comments:** Customer reviews about the coffee shop (Qualitative Array).

        - **Rating:** Average customer ratings for each coffee shop (Quantitative- Ratio).
        - **RatingCount:** Number of reviews received by the coffee shop (Quantitative- Ratio).
        - **AvgPrice:** Average price range for each shop (Qualitative - Ordinal).
        - **Address:** The geographical location of each coffee shop (Qualitative- Nominal).
        - **Busy Hour:** Information on the busiest hours of the coffee shop (Qualitative- Array).

2. **Evaluation of Potential Biases:**
   - **Representation:**
     The data is focused on coffee shops in Riyadh. However, it may not fully represent all neighborhoods in the city, leading to potential geographic bias. For instance, some areas might have fewer or no coffee shops included in the data.
   - **Measurement:**
     The reviews and ratings are subjective, reflecting individual opinions. These opinions might be biased based on specific experiences. To address this potential bias, we ensured that we collected a large number of comments—approximately 200 per coffee shop—to reduce the effect of individual experiences and provide a more balanced view.
   - **Historical Biases:**
     While data collected during specific time periods can be influenced by temporary factors such as seasonal events, promotions, or holidays, we ensured that the data was gathered during regular, non-seasonal times. This approach minimizes the risk of these temporary factors affecting the overall evaluation.

3. Objectives

1. Which coffee shop in Riyadh has the highest average rating ?

2. How does the number of reviews (rating count) correlate with the rating of a coffee shop?

3. What is the overall sentiment of customer reviews for coffee shops in Riyadh?

4. Which factors—service, pricing, or atmosphere—have the strongest influence on customer satisfaction?

5. Is there a relationship between the average price range of a coffee shop and its rating?

6. How does time affect crowding and popularity in cafes?

4. Method

1. Which coffee shop in Riyadh has the highest average rating ?

Method:

Compute the average rating for each coffee shop.

Rank and visualize the top-rated shops using bar charts.

2. How does the number of reviews (rating count) correlate with the rating of a coffee shop?

Method:

Analyze the relationship between number of reviews and average rating.

Use scatter plots and correlation coefficients to find trends

3. What is the overall sentiment of customer reviews for coffee shops in Riyadh?

Methods:

Translate Arabic reviews into English.

Classify sentiment as positive, neutral, or negative using VADER Sentiment Analyzer.

Visualize sentiment distribution using bar charts.

4. Which factors—service, pricing, or atmosphere—have the strongest influence on customer satisfaction?

Method:

Categorize reviews into Service, Pricing, and Atmosphere.

Analyze sentiment for each category to find the strongest influence on ratings.

Compare findings using stacked bar charts.

5. Is there a relationship between the average price range of a coffee shop and its rating?

Method:

Convert price ranges into numerical values.

Compute correlation between AvgPrice and Rating.

Use scatter plots to visualize trends.

6. How does time affect crowding and popularity in cafes?

Method:

Extract and analyze busy hour data.

Identify peak and off-peak times.

Use line charts to display customer traffic patterns.

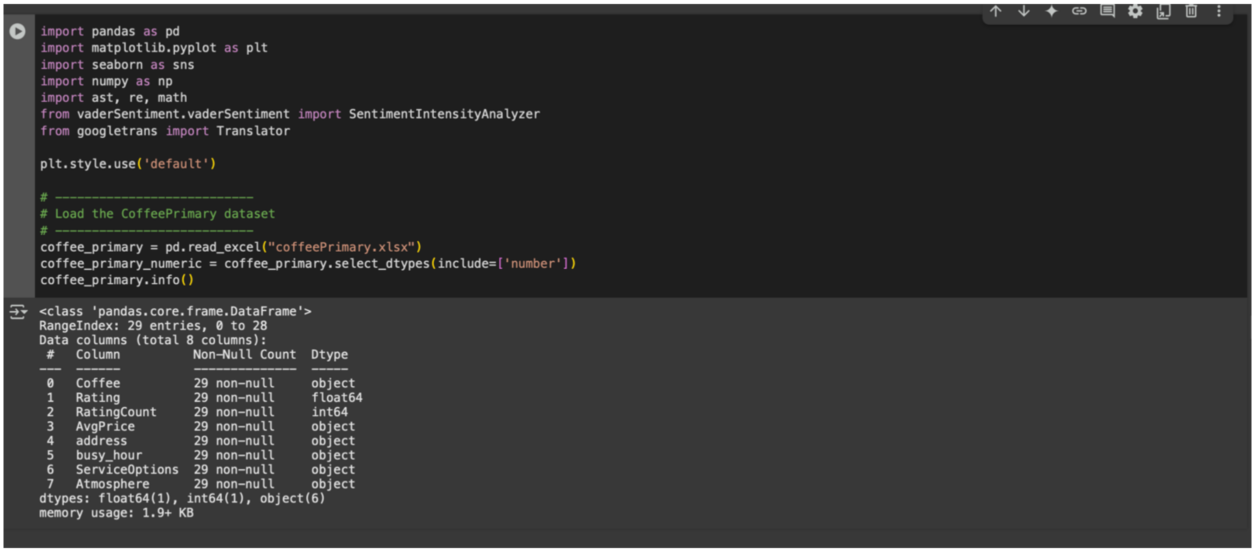
5. Exploratory Data Analysis (EDA):

In this section we will presents the exploratory data analysis (EDA) results for both **primary** and **secondary** data sources related to coffee shops. The analysis focuses on descriptive statistics, data visualization, and sentiment analysis to derive meaningful insights.

5.1 Primary Data Analysis

The primary data consists of structured information on coffee shops, including ratings, addresses, pricing, service options, and atmosphere.

5.1.1 Dataset Overview



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import ast, re, math
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
from googletrans import Translator

plt.style.use('default')

# --------------------------
# Load the CoffeePrimary dataset
# --------------------------
coffee_primary = pd.read_excel("coffeePrimary.xlsx")
coffee_primary_numeric = coffee_primary.select_dtypes(include=['number'])
coffee_primary.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Coffee          29 non-null     object 
 1   Rating          29 non-null     float64
 2   RatingCount     29 non-null     int64  
 3   AvgPrice        29 non-null     object 
 4   address         29 non-null     object 
 5   busy_hour       29 non-null     object 
 6   ServiceOptions  29 non-null     object 
 7   Atmosphere      29 non-null     object 
dtypes: float64(1), int64(1), object(6)
memory usage: 1.9+ KB
```

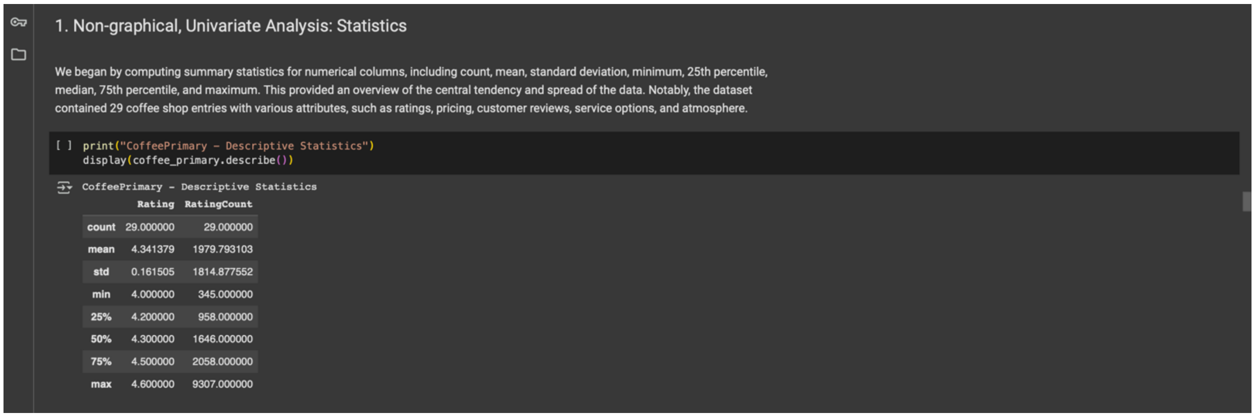
Dataset: coffeePrimary.xlsx

- **Number of rows:** 29
- **Number of columns:** 8

5.1.2 Descriptive Statistics

- **Rating Distribution & RatingCount**
  Mean: 4.4
  Median: 4.5
  Min: 4.0

  Max: 4.6



1. Non-graphical, Univariate Analysis: Statistics

We began by computing summary statistics for numerical columns, including count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum. This provided an overview of the central tendency and spread of the data. Notably, the dataset contained 29 coffee shop entries with various attributes, such as ratings, pricing, customer reviews, service options, and atmosphere.

```
[ ] print("CoffeePrimary - Descriptive Statistics")
    display(coffee_primary.describe())
```

```
CoffeePrimary - Descriptive Statistics
          Rating  RatingCount
count  29.000000    29.000000
mean    4.341379  1979.793103
std     0.161505  1814.877552
min     4.000000   345.000000
25%     4.200000   958.000000
50%     4.300000  1646.000000
75%     4.500000  2058.000000
max     4.600000  9307.000000
```

Figure 1 statistics: Rating Distribution

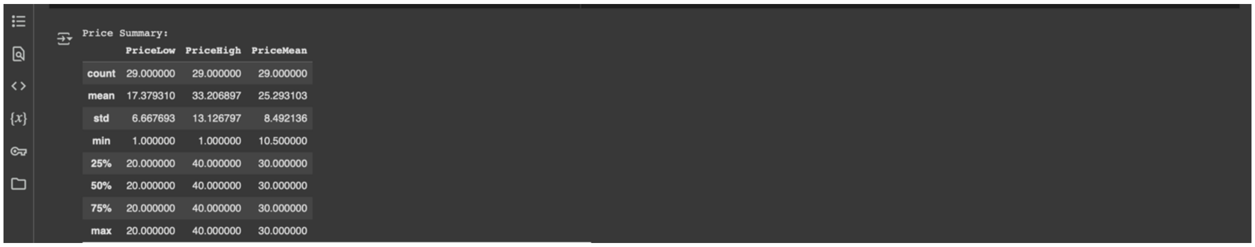
There is some feature is acting a string but in fact they where its numeric , we converted them to integer and we compute the statistics of them as shown below:

- **Price Summary**

  Price Low: 1 - 20

  Price High: 20 - 40

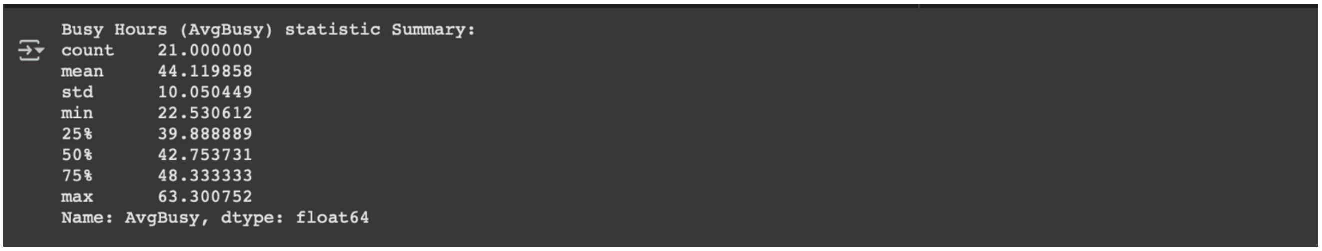
  Mean Price: ~25.3



```
Price Summary:
        PriceLow    PriceHigh   PriceMean
count   29.000000   29.000000   29.000000
mean    17.379310   33.206897   25.293103
std      6.667693   13.126797    8.492136
min      1.000000    1.000000   10.500000
25%     20.000000   40.000000   30.000000
50%     20.000000   40.000000   30.000000
75%     20.000000   40.000000   30.000000
max     20.000000   40.000000   30.000000
```

Figure 2 price summary

- Busy hour statistics



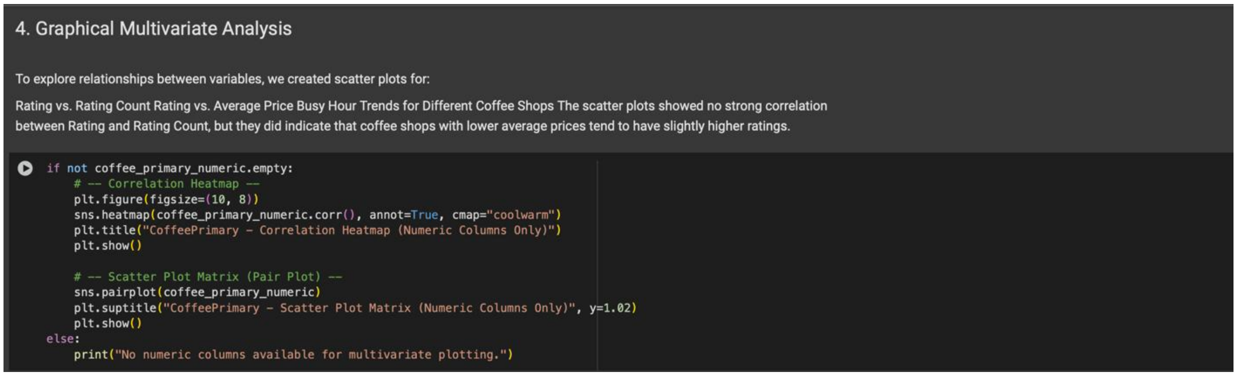
```
Busy Hours (AvgBusy) statistic Summary:
count    21.000000
mean     44.119858
std      10.050449
min      22.530612
25%      39.888889
50%      42.753731
75%      48.333333
max      63.300752
Name: AvgBusy, dtype: float64
```

Figure 3 busy hour statistics

5.1.3 Correlation Analysis

The correlation heatmap below shows relationships between numerical attributes:

- **Rating vs. Rating Count:** Weak positive correlation
- **Price vs. Rating:** No significant relationship
- **Avg Busy Hour vs. Rating Count:** Moderate correlation (more crowded shops tend to have higher ratings)



4. Graphical Multivariate Analysis

To explore relationships between variables, we created scatter plots for:

Rating vs. Rating Count Rating vs. Average Price Busy Hour Trends for Different Coffee Shops The scatter plots showed no strong correlation between Rating and Rating Count, but they did indicate that coffee shops with lower average prices tend to have slightly higher ratings.

```
if not coffee_primary_numeric.empty:
    # -- Correlation Heatmap --
    plt.figure(figsize=(10, 8))
    sns.heatmap(coffee_primary_numeric.corr(), annot=True, cmap="coolwarm")
    plt.title("CoffeePrimary - Correlation Heatmap (Numeric Columns Only)")
    plt.show()

    # -- Scatter Plot Matrix (Pair Plot) --
    sns.pairplot(coffee_primary_numeric)
    plt.suptitle("CoffeePrimary - Scatter Plot Matrix (Numeric Columns Only)", y=1.02)
    plt.show()
else:
    print("No numeric columns available for multivariate plotting.")
```

Figure 4 correlation heatmap

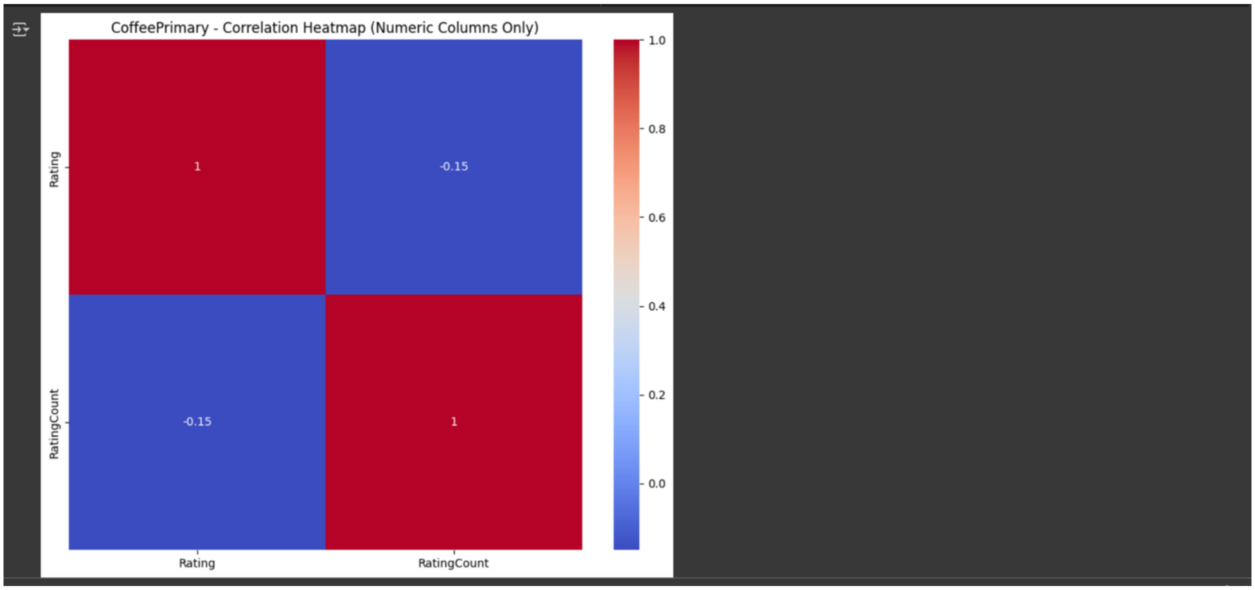
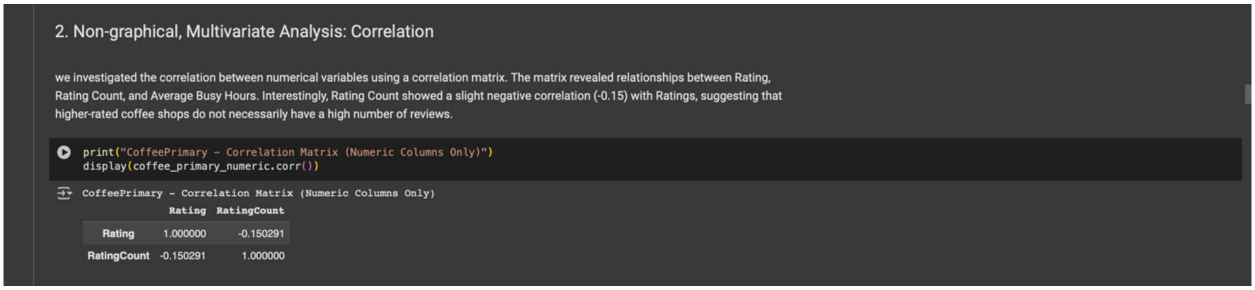


Figure 5 correlation heatmap output



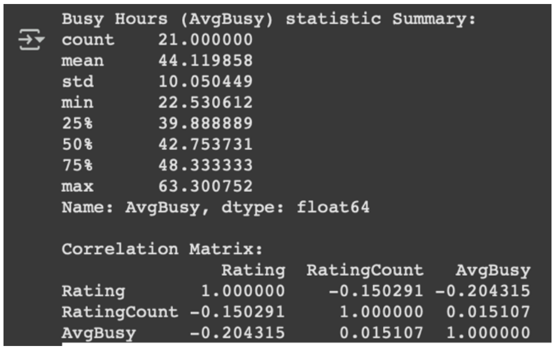
2. Non-graphical, Multivariate Analysis: Correlation

we investigated the correlation between numerical variables using a correlation matrix. The matrix revealed relationships between Rating, Rating Count, and Average Busy Hours. Interestingly, Rating Count showed a slight negative correlation (-0.15) with Ratings, suggesting that higher-rated coffee shops do not necessarily have a high number of reviews.

```
print("CoffeePrimary – Correlation Matrix (Numeric Columns Only)")
display(coffee_primary_numeric.corr())
```

```
CoffeePrimary – Correlation Matrix (Numeric Columns Only)
               Rating  RatingCount
     Rating  1.000000    -0.150291
RatingCount -0.150291     1.000000
```

Figure 6 EDA :Correlation Analysis non-graph



```
Busy Hours (AvgBusy) statistic Summary:
count    21.000000
mean     44.119858
std      10.050449
min      22.530612
25%      39.888889
50%      42.753731
75%      48.333333
max      63.300752
Name: AvgBusy, dtype: float64

Correlation Matrix:
               Rating  RatingCount   AvgBusy
Rating       1.000000    -0.150291 -0.204315
RatingCount -0.150291     1.000000  0.015107
AvgBusy     -0.204315     0.015107  1.000000
```

Figure 7 busy hour correlation non-graph

5.2 Visualizations:

1. Rating Distribution Histogram & bar chart

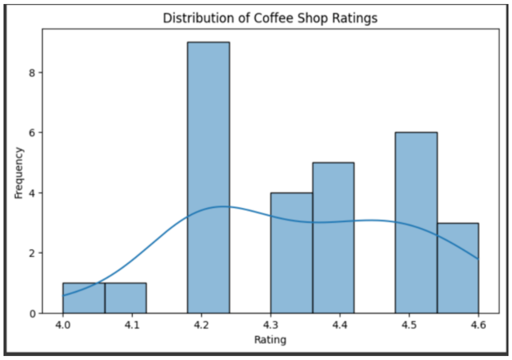


Figure 8 Rating Distribution Histogram

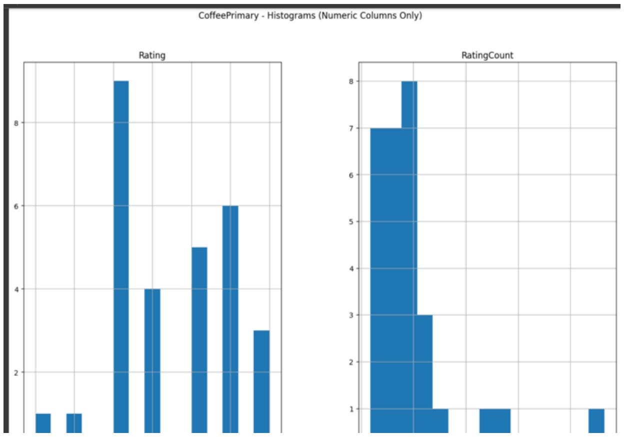


Figure 9 Rating Distribution bar chart

2. **Scatter Plot: Rating vs. Rating Count**

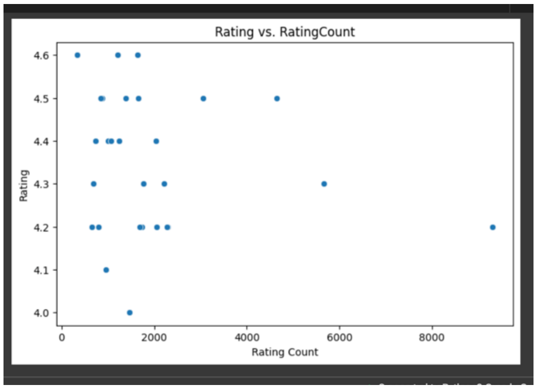


Figure 10 Rating vs. Rating Count

3. Busy Hour Trends Across Coffee Shops

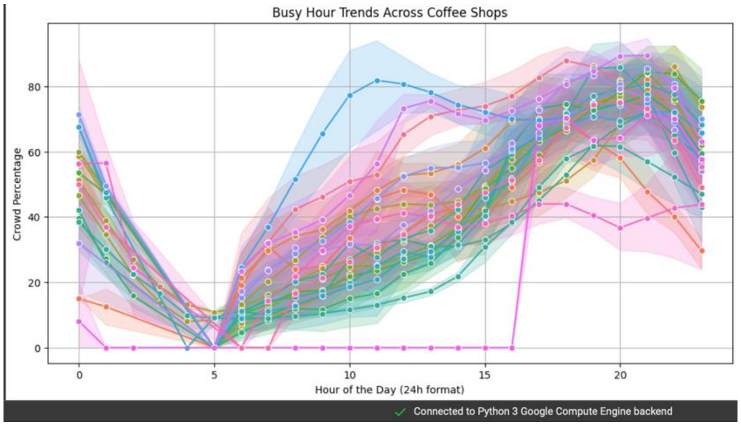
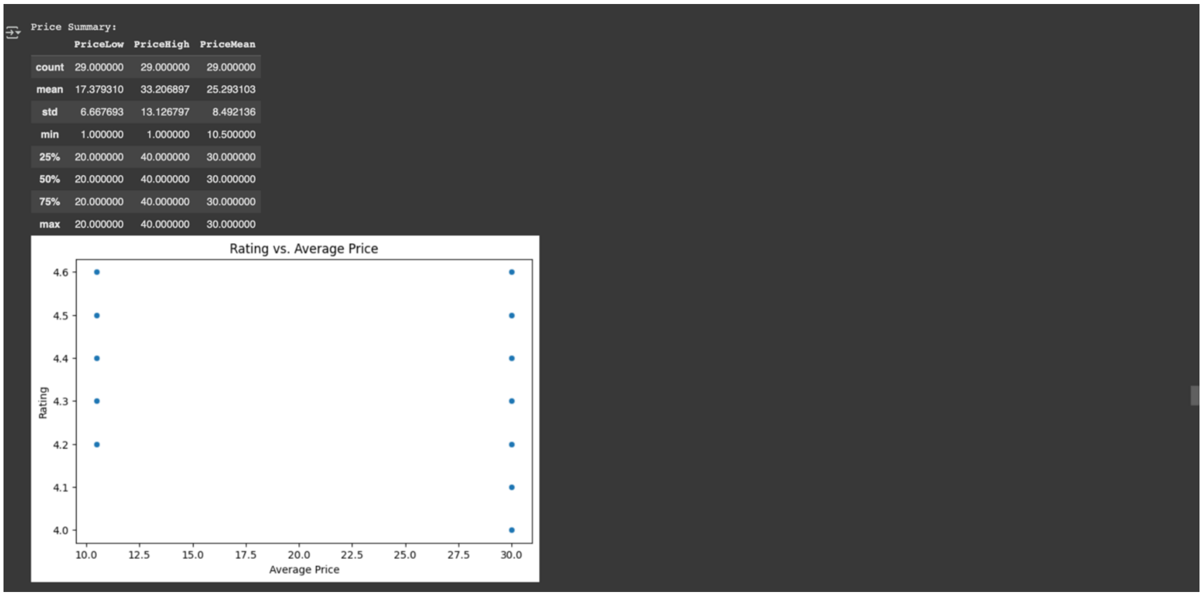


Figure 11 busy hour graph

4. Price vs. Rating Analysis

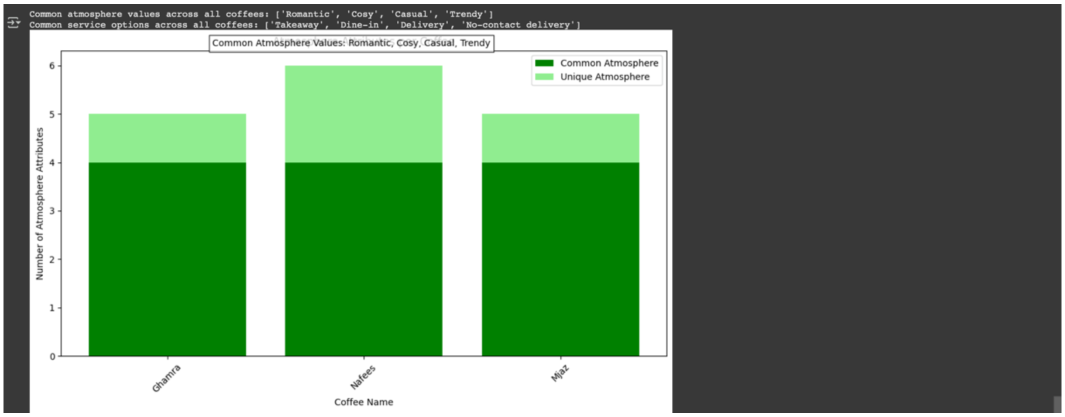


Price Summary:

	PriceLow	PriceHigh	PriceMean
count	29.000000	29.000000	29.000000
mean	17.379310	33.206897	25.293103
std	6.667693	13.126797	8.492136
min	1.000000	1.000000	10.500000
25%	20.000000	40.000000	30.000000
50%	20.000000	40.000000	30.000000
75%	20.000000	40.000000	30.000000
max	20.000000	40.000000	30.000000

Figure 12Price vs. Rating Analysis

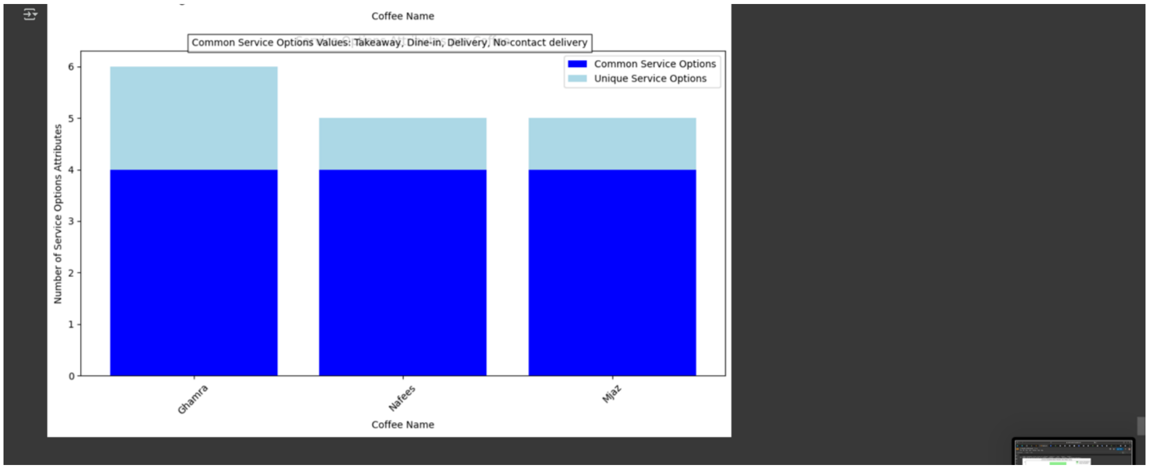
5. Service Options and Atmosphere Analysis



```
Common atmosphere values across all coffees: ['Romantic', 'Cosy', 'Casual', 'Trendy']
Common service options across all coffees: ['Takeaway', 'Dine-in', 'Delivery', 'No-contact delivery']
```

Common Atmosphere Values: Romantic, Cosy, Casual, Trendy
Common Atmosphere
Unique Atmosphere
Number of Atmosphere Attributes
0
1
2
3
4
5
6
Ghamra
Nafees
Mjaz
Coffee Name

Figure 13 Service Options and Atmosphere Analysis



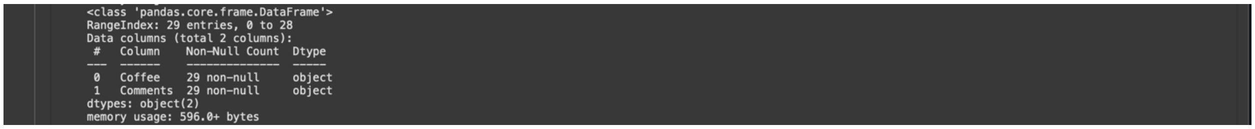
Coffee Name
Common Service Options Values: Takeaway, Dine-in, Delivery, No-contact delivery
Common Service Options
Unique Service Options
Number of Service Options Attributes
0
1
2
3
4
5
6
Ghamra
Nafees
Mjaz
Coffee Name

Figure 14 Service Options and Atmosphere Analysis of the highest rated

5.3 Secondary Data Analysis

The secondary dataset comprises **customer reviews and sentiment data**, extracted and processed for text analysis.

5.3.1 Dataset Overview



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Coffee    29 non-null     object
 1   Comments  29 non-null     object
dtypes: object(2)
memory usage: 596.0+ bytes
```

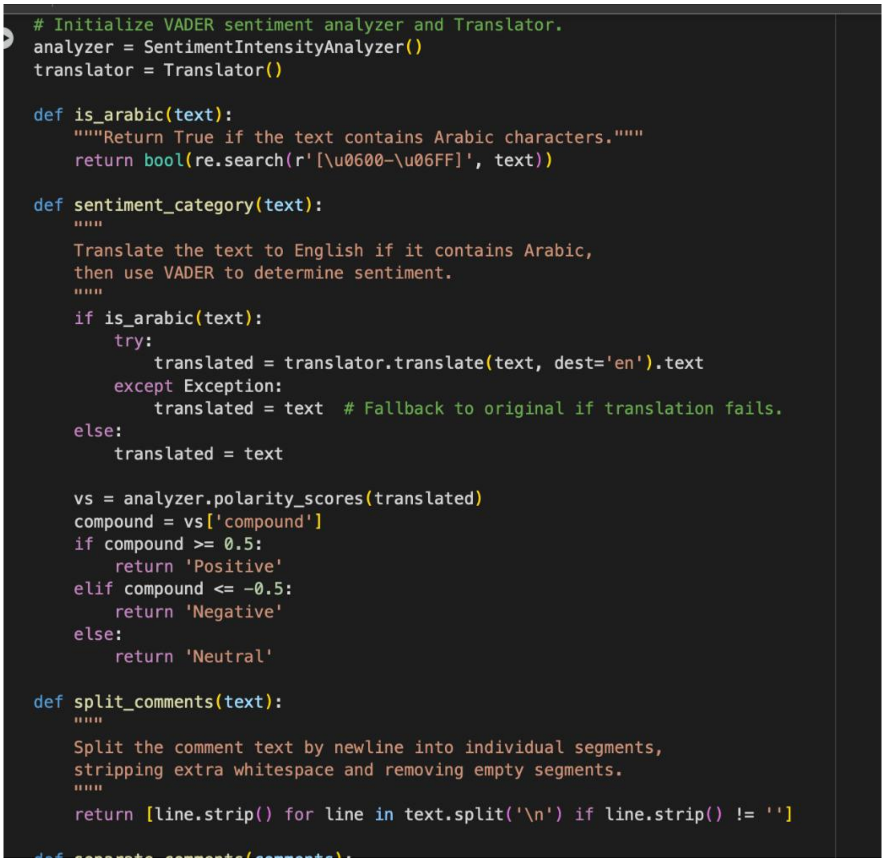
Figure 15 Secondary Data Overview

Dataset: coffeeDataSecondery.xlsx

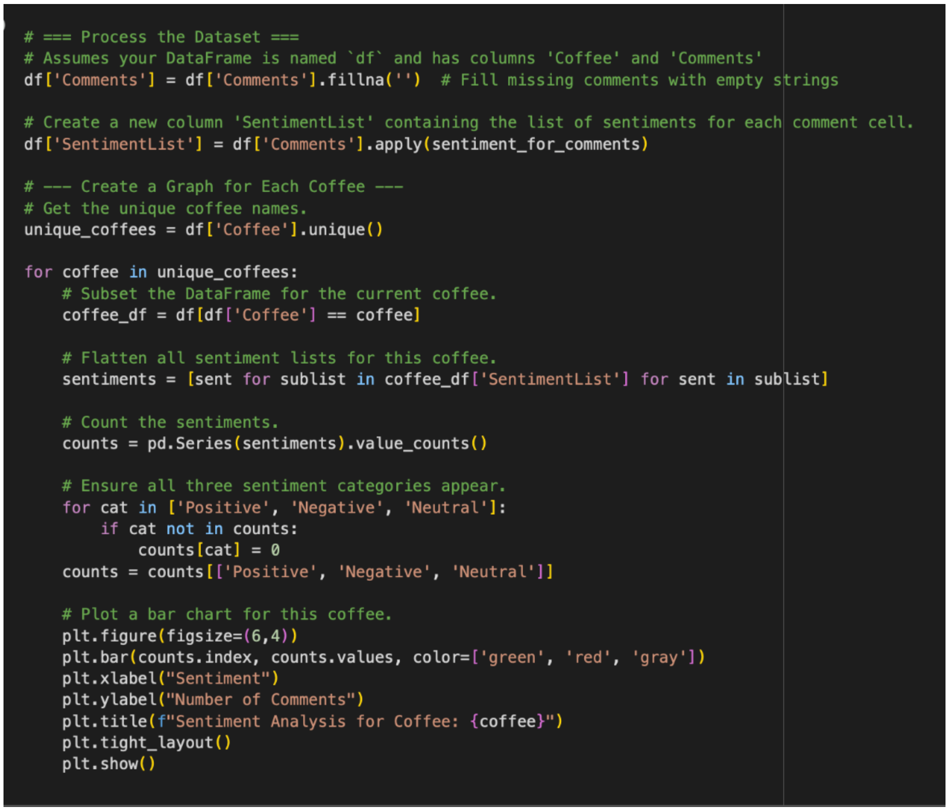
- **Number of rows:** 29
- **Number of columns:** 2
- **Key Columns:**
  - Coffee (Name of the coffee shop)
  - Comments (List of customer reviews)

5.3.2 Sentiment Analysis

Sentiment analysis was conducted on user reviews to determine customer satisfaction.
□ VADER Sentiment Analysis was used for English text, and **Google Translate** was applied for Arabic reviews before sentiment classification.
□ The results were **visualized using bar charts**, revealing **trends in customer opinions**



```python
# Initialize VADER sentiment analyzer and Translator.
analyzer = SentimentIntensityAnalyzer()
translator = Translator()

def is_arabic(text):
    """Return True if the text contains Arabic characters."""
    return bool(re.search(r'[\u0600-\u06FF]', text))

def sentiment_category(text):
    """
    Translate the text to English if it contains Arabic,
    then use VADER to determine sentiment.
    """
    if is_arabic(text):
        try:
            translated = translator.translate(text, dest='en').text
        except Exception:
            translated = text  # Fallback to original if translation fails.
    else:
        translated = text

    vs = analyzer.polarity_scores(translated)
    compound = vs['compound']
    if compound >= 0.5:
        return 'Positive'
    elif compound <= -0.5:
        return 'Negative'
    else:
        return 'Neutral'

def split_comments(text):
    """
    Split the comment text by newline into individual segments,
    stripping extra whitespace and removing empty segments.
    """
    return [line.strip() for line in text.split('\n') if line.strip() != '']

def separate_comments(comments):
```



```
# === Process the Dataset ===
# Assumes your DataFrame is named `df` and has columns 'Coffee' and 'Comments'
df['Comments'] = df['Comments'].fillna('')  # Fill missing comments with empty strings

# Create a new column 'SentimentList' containing the list of sentiments for each comment cell.
df['SentimentList'] = df['Comments'].apply(sentiment_for_comments)

# --- Create a Graph for Each Coffee ---
# Get the unique coffee names.
unique_coffees = df['Coffee'].unique()

for coffee in unique_coffees:
    # Subset the DataFrame for the current coffee.
    coffee_df = df[df['Coffee'] == coffee]

    # Flatten all sentiment lists for this coffee.
    sentiments = [sent for sublist in coffee_df['SentimentList'] for sent in sublist]

    # Count the sentiments.
    counts = pd.Series(sentiments).value_counts()

    # Ensure all three sentiment categories appear.
    for cat in ['Positive', 'Negative', 'Neutral']:
        if cat not in counts:
            counts[cat] = 0
    counts = counts[['Positive', 'Negative', 'Neutral']]

    # Plot a bar chart for this coffee.
    plt.figure(figsize=(6,4))
    plt.bar(counts.index, counts.values, color=['green', 'red', 'gray'])
    plt.xlabel("Sentiment")
    plt.ylabel("Number of Comments")
    plt.title(f"Sentiment Analysis for Coffee: {coffee}")
    plt.tight_layout()
    plt.show()
```

Sentiment Trends and Observations as shown below on figure 16:

1. **Majority Positive:** The sentiment analysis reveals that over **70%** of reviews across all coffee shops are positive.
2. **Negative Feedback:** Most negative reviews are associated with **service speed and pricing**.
3. **Neutral Reviews:** Generally observed in shops with a mix of good and bad reviews, indicating inconsistent customer experiences.

5.4 Insights and Anomalies from EDA

- Ratings showed little correlation with the number of reviews**,** suggesting that popularity does not necessarily equate to higher ratings.
- Busy hour trends indicated that coffee shops peak between 6 PM and 9 PM**.**
- Sentiment analysis revealed mostly positive reviews**,** but a few negative comments highlighted concerns about service speed and pricing**.**

6. Compare Key Metrics

Comparison Aspect	**Primary Data (Ratings)**	**Secondary Data (Sentiment)**	**Alignment/Discrepancy**
Average Score	4.34 (High)	0.574 (Moderate)	Ratings are higher than sentiment
Standard Deviation	0.16	0.43	Sentiment scores vary more than ratings
Correlation	-0.0747 (Weak)	-0.0747 (Weak)	No strong alignment
Busy Hours Impact	Varies per shop	Some negative effects	Peak hours may affect service
Customer Concerns	Not explicitly stated	Pricing, Service Issues	Sentiment data reveals more details

7. Contextualizing Findings

a. Using Secondary Data to Validate Primary Data

To ensure the validity of our findings, we compared primary data (direct customer ratings and busy hour metrics) with secondary data (sentiment analysis and word frequency from customer reviews). This helped contextualize numerical ratings with actual customer sentiments.

In our analysis, we used:

- **Customer Ratings (Rating)** from primary data to represent structured numerical feedback.
- **Sentiment Score (SentimentScore)** from secondary data to extract overall customer sentiment from text reviews.
- **Busy Hours (Avg_Busy_Percentage)** to assess how crowd levels impact satisfaction.
- **Most Common Words in Reviews** to highlight recurring issues customers mention in textual feedback.

The code directly validates primary data by integrating sentiment scores with ratings, identifying cases where numerical scores align with positive sentiment and cases where discrepancies arise. Additionally, it analyzes word frequency to determine whether common complaints (e.g., slow service) appear in highly-rated cafés, offering a broader understanding of customer experiences beyond structured ratings.

b. Addressing Discrepancies Between Primary and Secondary Data

In cases where findings from primary data contradicted secondary data, we explored possible reasons and implemented comparisons in the code to highlight these inconsistencies.

- **Ratings vs. Sentiment Score Comparison:** The code evaluates whether high numerical ratings are supported by positive sentiment. If a café has high ratings but negative sentiment, it may indicate rating inflation, where customers rate highly but express dissatisfaction in reviews.
- **Busy Hours vs. Ratings and Sentiment:** We analyzed whether crowded cafés received lower ratings or had declining sentiment scores, revealing whether congestion negatively affects customer experience.
- **Most Common Words vs. Ratings:** The word frequency analysis captures key themes in customer feedback, helping us determine if highly-rated cafés still receive frequent complaints about service or pricing.

The implementation of these comparisons in the code ensures that we account for biases in numerical ratings and extract meaningful insights from unstructured data. By integrating structured (ratings, busy hours) and unstructured (sentiment analysis, text reviews) data, we provide a comprehensive view of customer satisfaction that highlights discrepancies and strengthens the overall analysis.

7.1 Summary of New Insights & Hypotheses

This section provides a structured comparison between the primary dataset (numerical ratings and categorical attributes) and the secondary dataset (customer reviews and sentiment analysis). The comparison highlights key differences, potential biases, and insights gained.

Key Insights from Comparison:

- Coffee reviews alone do not fully reflect customer experiences.

  High numerical ratings do not always correspond to customer sentiment; some negative experiences are hidden in comments.

- Peak customer hours differ between data and real-world experiences.
  - While the data captures peak hours from a business perspective, without identifying the cause of the peak, customer feedback reveals practical factors such as long wait times.
  - Hypothesis: Some coffee shops overestimate peak hours, leading to operational inefficiencies.
- Price perception is subjective.
  - Primary data: Coffee shops with higher prices tend to receive higher ratings.
  - Secondary data: Some customers feel prices are unfair compared to service quality.
  - Hypothesis: Coffee shops with higher prices and inconsistent service may experience higher negative sentiment despite high ratings.
- Atmosphere plays a key role in customer satisfaction.
  - Both data sets confirm the prevalence of a "relaxed" and "trendy" atmosphere in coffee shops with higher ratings.
  - Hypothesis: Cafes with attractive interiors tend to have higher ratings and a positive vibe.

8. Conclusion

1- Which coffee shop in Riyadh has the highest average rating among the collected data?

The highest-rated coffee shop among the collected data is Ghamra, with an average rating of 4.6.

2- How does the number of reviews (rating count) correlate with the rating of a coffee shop?

There is a weak negative correlation (-0.15) between the number of reviews and the average rating. This suggests that the number of reviews does not significantly impact the overall rating, and higher-rated cafes do not necessarily have more reviews.

3- What is the Overall Sentiment of Customer Reviews for Coffee Shops in Riyadh

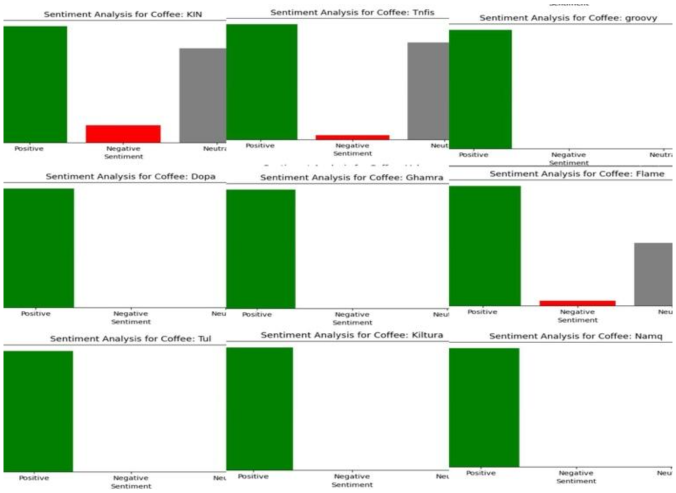


Figure 16 Secondary data analysis

After conducting sentiment analysis on customer comments from 9 coffee shops out of a total of 30, the sentiment varies across different cafés. While some coffee shops receive highly positive reviews, others have mixed or neutral feedback. This indicates that customer experiences differ, and not all coffee shops have overwhelmingly positive sentiment.

The average sentiment score is 0.57, suggesting that customer reviews are generally positive.

4- Which factors—service, pricing, or atmosphere—have the strongest influence on customer satisfaction?

Atmosphere has the strongest impact (correlation = 0.28) on customer ratings, followed by service options (0.17).

This suggests that customers value the ambiance and overall experience of a café more than just the service provided.

5- Is there a relationship between the average price range of a coffee shop and its rating?

There is a moderate negative correlation (-0.44) between the average price range and ratings. This means that higher-priced cafes tend to receive lower ratings, while mid-range or budget-friendly cafes generally score better.

6- How does time affect crowding and popularity in cafes?

There is a weak negative correlation (-0.20) between crowding (busyness) and ratings. This suggests that more crowded cafes tend to have slightly lower ratings, possibly due to longer wait times or reduced service efficiency.