Modeling Phase 3 – LOGBOOK

King Saud University

College of Computer and Information Sciences

Department of Information Technology

LOGBOOK OF COFFEE SHOP DATA ANALYSIS

IT 326 Course Project

Semester-2, 1446H

Prepared by:

<Lamya Alnahdi, 443202647>

<Norah Alfaheed, 444200779>

<Leen Alqahtani, 443200591>

<Hissah Alotaibi, 444200349>

Date	Task	Tools/Libraries Used	Actions/Analysis	Findings/Insights
3-03-2025	Hypothesis 1 – Sentiment vs Ratings	Pandas, Scipy	Compared average sentiment between highly-rated (≥ 4.5) and low-rated (≤ 4.0) shops.	No statistically significant difference was found.
3-03-2025	Hypothesis 2 – Busy Hours Impact	Pandas, Seaborn	Investigated whether sentiment scores vary during peak hours.	Slight variation observed, not statistically decisive.
3-03-2025	Hypothesis 3 – Service Options vs Ratings	Pandas, Statsmodels	Tested correlation between number of service options and shop ratings.	Weak and non-significant relationship.
3-03-2025	Hypothesis 4 – Atmosphere + Services	Pandas, Statsmodels	Used regression model to analyze combined effect of atmosphere	R^2 was weak (0.14); no statistically significant

			and service.	coefficients.
3-03-2025	Sentiment Analysis	TextBlob, Pandas	Calculated sentiment polarity scores and statistics.	Average = 0.574, Std Dev = 0.435
3-03-2025	Comment Length Analysis	Pandas, Seaborn	Analyzed the average number of characters per comment.	Most comments range between 66–132 characters.
3-03-2025	Word Frequency	Regex, Pandas	Identified most frequent words in positive and negative reviews.	Frequent words: service, ambiance, price
3-03-2025	Rating Distribution	Pandas	Computed mean and standard deviation of ratings across shops.	Average = 4.34, Std Dev = 0.161
3-03-2025	Price Distribution	Seaborn, Matplotlib	Used boxplot to visualize price variation across shops.	Average price: 25.3 SAR, with upscale café outliers.
3-03-2025	Correlation – Sentiment & Rating	Pandas, Seaborn	Explored scatter plot between sentiment and rating scores.	Weak correlation: r = -0.0747
3-03-2025	Busy Hours vs Sentiment	Pandas, Seaborn	Compared review sentiment during peak vs. non-peak hours.	Slightly lower sentiment during peak hours.
3-03-2025	Baseline Model	Scikit-learn	Used mean rating as a baseline predictive model.	RMSE = 0.165
3-03-2025	Linear Regression Model	Scikit-learn	Built a linear regression model for predicting ratings.	Moderate performance; better than baseline.
3-03-2025	Random Forest Model	Scikit-learn	Trained Random Forest Regressor to predict ratings.	Best performance overall, though still limited.
3-03-2025	OLS Regression Summary	Statsmodels	Applied OLS model: Rating ~ Sentiment +	R^2 = 0.14; all predictors were non-significant (p

			CommentLength + ServiceOptions + Atmosphere.	> 0.05).

Key Findings

• Most coffee shops in Riyadh have a high average rating (~4.34), indicating overall customer satisfaction.

• Sentiment scores are generally positive (avg. 0.574), but do not strongly correlate with rating values.

• Longer and more detailed comments tend to reflect extreme experiences (very positive or negative).

• Word frequency shows that customers value service, ambiance, and price the most.

• Price level does not significantly affect ratings; consistent service is more valued.

• Busy hours slightly reduce review sentiment, suggesting pressure on service quality.

• Random Forest outperformed other models in rating prediction but lacked strong accuracy.

• OLS Regression showed low R^2 (0.14) and no statistically significant predictors.

Challenges and Solutions

Problems Encountered	Solutions Applied
Some comments were in Arabic, requiring translation.	Used googletrans to translate Arabic to English.
Sentiment scores were inconsistent due to limitations of TextBlob.	Applied normalization and weighted averaging.
Outliers in prices and ratings affected visualizations.	Handled outliers using boxplot-based filtering.
Some reviews were extremely short or vague.	Filtered short reviews to maintain insight quality.
Imbalanced data with far more high-rated shops.	Used balanced sampling during hypothesis testing.
Busy hours were unclear due to limited time data.	Estimated patterns using available review timestamps.