A Win-Win

for

Customers & Restaurateurs

Final Project: Submitted

by

Rizwan Ahasan Pathan

USC ID: 6909 753 128

Under the Supervision

of

Ulf Hermjakob

Department of Computer Science

Viterbi School of Engineering

University of Southern California (USC)

May, 2025

A Win-Win for Customers & Restaurateurs

The Data-Driven Recipe for Restaurant Success in LA

1. Abstract

This project aims to explore and analyze the various factors that impact restaurant performance and customer satisfaction in the Los Angeles City area. Specifically, it focuses on how health inspection results, Yelp business metadata, and neighborhood demographics influence a restaurant's reputation, rating, and potential for long-term success. Restaurants are often evaluated based on subjective online ratings, but these may not always reflect true service quality or hygiene standards. By integrating objective health data with real restaurant feature and neighborhood profiles, this project will provide a more holistic, data-driven evaluation of restaurants.

2. Motivation

Restaurant choices are often driven by ratings and reputation—but these can overlook critical factors like hygiene and neighborhood dynamics. This project was motivated by the need to uncover the hidden dimensions of restaurant success by integrating regulatory inspection data with business and demographic indicators. It will serve both consumer and investor interests. Consumers will gain clearer insights into which restaurants provide a clean, safe, and valuable dining experience—beyond branding or advertising. Restaurateurs and potential investors, on the other hand, can use the data to identify ideal locations for opening new venues or improving operations in existing ones. Additionally, the project investigates whether economic and demographic disparities across ZIP codes influence restaurant standards and customer experience. Ultimately, this will result in a real-world analytical tool and a compelling, showcasing skills in scraping, API use, data modeling, and business intelligence.

3. Data Sources

This project draws on three primary datasets to assess restaurant performance across Los Angeles, supported by a fourth reference table for ZIP code mapping:

3.1 Health Inspection Records (Scraped)

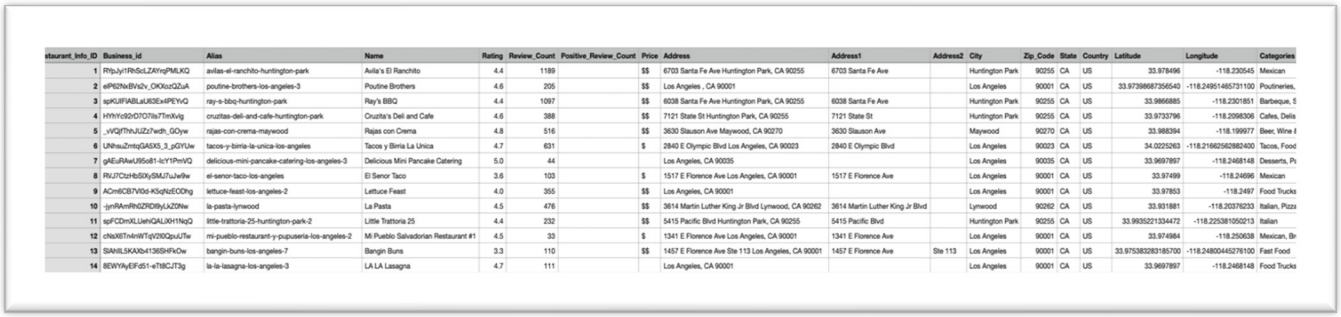
This dataset was scraped from the Los Angeles County Department of Public Health using Selenium and BeautifulSoup. It contains 8,441 entries covering restaurant hygiene metrics including restaurant name, address, ZIP code, inspection score, grade (A/B/C), inspection date. This data provides objective, regulatory insight into restaurant cleanliness and food safety across Los Angeles City.

Restaurant_Info_ID	Inspection_Info_ID	Business_Name	Inspection_date	Score	Grade	Address1	City	Name_Similarity_Percentage	IFrame_Number	Created_At	Is_Processed	Is_Val
8	1	EL SENOR TACO	2025-02-12	97.0	A	1517 E FLORENCE AVE	LOS ANGELES	100.0	904556	2025-04-14 00:50:07	0	
12	2	MI PUEBLO RESTAURANT	2024-10-18	92.0	A	1341 E FLORENCE AVE	LOS ANGELES	72.72	10137390	2025-04-14 00:50:28	0	
13	3	BANGIN BUNS	2025-03-06	93.0	A	1457 E FLORENCE AVE # 113	LOS ANGELES	100.0	17555671	2025-04-14 00:50:50	0	
19	4	BIRRIERIA TLAQUEPAQUE #2	2023-01-10	92.0	A	1734 E FLORENCE AVE	LOS ANGELES	93.33	911598	2025-04-14 00:51:33	0	
20	5	TEQUERIA TIJUANA NUMERO 1 INC	2024-06-01	96.0	A	241 W FLORENCE AVE	LOS ANGELES	66.66	14524932	2025-04-14 00:51:53	0	
21	6	LOBSTA TRUCK	2025-03-26	92.0	A	1542 FISHBURN AVE	LOS ANGELES	74.07	1868843	2025-04-14 00:52:12	0	
27	7	BOYS BURGERS NO1	2025-03-24	95.0	A	1774 E FLORENCE AVE	LOS ANGELES	85.71	21900380	2025-04-14 00:52:33	0	
28	8	BIRRIERIA JALISCO	2023-05-12	95.0	A	7714 S COMPTON AVE	LOS ANGELES	100.0	925295	2025-04-14 00:52:52	0	
36	9	LA ZONA RESTAURANT	2024-11-01	94.0	A	1913 E GAGE AVE	LOS ANGELES	64.51	18311168	2025-04-14 00:54:19	0	
38	10	LA CRAZY CRAB	2024-11-20	96.0	A	1415 E GAGE AVE	LOS ANGELES	92.85	20391215	2025-04-14 00:54:38	0	

Figure 3.1: Sample Records from the LA County Health Inspection Dataset

3.2 Yelp Business Metadata (API)

Collected via the Yelp Fusion API, this dataset includes 20,917 restaurant listings across Los Angeles City area. Each entry contains detailed metadata such as business ID, name, average rating, review count, price level, categories, location coordinates, and status flags like Is_Closed. It offers valuable insights into how restaurants are perceived by the public and indexed by a leading review platform.

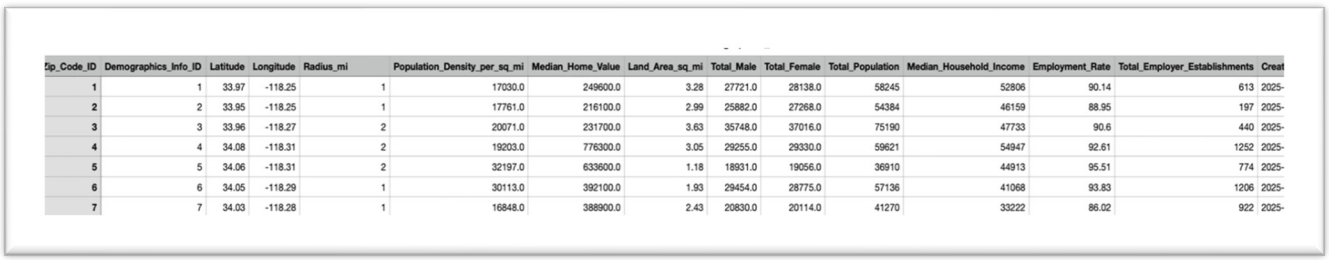


staurant_Info_ID	Business_id	Alias	Name	Rating	Review_Count	Positive_Review_Count	Price	Address	Address1	Address2	City	Zip_Code	State	Country	Latitude	Longitude	Categories
1	RfpJyI1RhScLZAYrqPMLKQ	avilas-el-ranchito-huntington-park	Avila's El Ranchito	4.4	1189		$$	6703 Santa Fe Ave Huntington Park, CA 90255	6703 Santa Fe Ave		Huntington Park	90255	CA	US	33.978496	-118.230545	Mexican
2	ePE7NxBVs2v_OKlozQZuA	poutine-brothers-los-angeles-3	Poutine Brothers	4.6	205		$$	Los Angeles , CA 90001			Los Angeles	90001	CA	US	33.97398687356540	-118.24951465731100	Poutineries,
3	spKUIFIABLaU83Ex4PEYvQ	ray-s-bbq-huntington-park	Ray's BBQ	4.4	1097		$$	6038 Santa Fe Ave Huntington Park, CA 90255	6038 Santa Fe Ave		Huntington Park	90255	CA	US	33.9866885	-118.2301851	Barbeque, S
4	HYhYc92rO7O7Ils7TmXvIg	cruzitas-deli-and-cafe-huntington-park	Cruzita's Deli and Cafe	4.6	388		$$	7121 State St Huntington Park, CA 90255	7121 State St		Huntington Park	90255	CA	US	33.9733796	-118.2098306	Cafes, Delis
5	_vVQiTlYhJUZz7wdh_GQyw	rajas-con-crema-maywood	Rajas con Crema	4.8	516		$$	3630 Slauson Ave Maywood, CA 90270	3630 Slauson Ave		Maywood	90270	CA	US	33.988394	-118.199977	Beer, Wine &
6	UNhsuZmtqGASX5_3_pGYUw	tacos-y-birria-la-unica-los-angeles	Tacos y Birria La Unica	4.7	631		$	2840 E Olympic Blvd Los Angeles, CA 90023	2840 E Olympic Blvd		Los Angeles	90023	CA	US	34.0225263	-118.21662562882400	Tacos, Food
7	gAEuRAwU95o81-IcY1PmVQ	delicious-mini-pancake-catering-los-angeles-3	Delicious Mini Pancake Catering	5.0	44			Los Angeles, CA 90035			Los Angeles	90035	CA	US	33.9697897	-118.2468148	Desserts, Pa
8	RvJTCtzHbSlXySMJ7uJw9w	el-senor-taco-los-angeles	El Senor Taco	3.6	103		$	1517 E Florence Ave Los Angeles, CA 90001	1517 E Florence Ave		Los Angeles	90001	CA	US	33.97499	-118.24696	Mexican
9	ACm6CB7V0d-KSqNzEODhg	lettuce-feast-los-angeles-2	Lettuce Feast	4.0	355		$$	Los Angeles, CA 90001			Los Angeles	90001	CA	US	33.97853	-118.2497	Food Trucks
10	-jynRAmRhSZRDI9yLkZDNw	la-pasta-lynwood	La Pasta	4.5	476		$$	3614 Martin Luther King Jr Blvd Lynwood, CA 90262	3614 Martin Luther King Jr Blvd		Lynwood	90262	CA	US	33.931881	-118.20376233	Italian, Pizza
11	spFCDmXLUehiQALi0H1NqQ	little-trattoria-25-huntington-park-2	Little Trattoria 25	4.4	232		$$	5415 Pacific Blvd Huntington Park, CA 90255	5415 Pacific Blvd		Huntington Park	90255	CA	US	33.9935221334472	-118.22538105021 3	Italian
12	cNsX6Tn4nWTqV2t0QpuUTw	mi-pueblo-restaurant-y-pupuseria-los-angeles-2	Mi Pueblo Salvadorian Restaurant #1	4.5	33		$	1341 E Florence Ave Los Angeles, CA 90001	1341 E Florence Ave		Los Angeles	90001	CA	US	33.974984	-118.250638	Mexican, Br
13	SiAhlL5KAXb413ISHFkOw	bangin-buns-los-angeles-7	Bangin Buns	3.3	110		$$	1457 E Florence Ave Ste 113 Los Angeles, CA 90001	1457 E Florence Ave	Ste 113	Los Angeles	90001	CA	US	33.97538283185700	-118.24800445276100	Fast Food
14	8EWYAyEiFd51-eTt8CJT3g	la-la-lasagna-los-angeles-3	LA LA Lasagna	4.7	111			Los Angeles, CA 90001			Los Angeles	90001	CA	US	33.9697897	-118.2468148	Food Trucks

Figure 3.2: Sample Records from the Yelp Business Metadata Dataset

3.3 Demographics by ZIP Code (Scraped + API)

This dataset combines scraped data from UnitedStatesZipCodes.org and public data retrieved via the U.S. Census Bureau API. It includes demographic and geographic statistics for 182 ZIP codes, with 119 located within the City of Los Angeles. Key fields include median income, population, population density, employment rate, median home value, and gender distribution. These indicators provide essential context for understanding economic and social influences on restaurant success.



Zip_Code_ID	Demographics_Info_ID	Latitude	Longitude	Radius_mi	Population_Density_per_sq_mi	Median_Home_Value	Land_Area_sq_mi	Total_Male	Total_Female	Total_Population	Median_Household_Income	Employment_Rate	Total_Employer_Establishments	Creat
1	1	33.97	-118.25	1	17030.0	249600.0	3.28	27721.0	28138.0	58245	52806	90.14	613	2025-
2	2	33.95	-118.25	1	17761.0	216100.0	2.99	25882.0	27268.0	54384	46159	88.95	197	2025-
3	3	33.96	-118.27	2	20071.0	231700.0	3.63	35748.0	37016.0	75190	47733	90.6	440	2025-
4	4	34.08	-118.31	2	19203.0	776300.0	3.05	29255.0	29330.0	59621	54947	92.61	1252	2025-
5	5	34.06	-118.31	2	32197.0	633600.0	1.18	18931.0	19056.0	36910	44913	95.51	774	2025-
6	6	34.05	-118.29	1	30113.0	392100.0	1.93	29454.0	28775.0	57136	41068	93.83	1206	2025-
7	7	34.03	-118.28	1	16848.0	388900.0	2.43	20830.0	20114.0	41270	33222	86.02	922	2025-

Figure 3.3: Sample Records form ZIP Code Demographics Dataset

3.4 ZIP Code Reference Table (Supporting Dataset)

This reference dataset was compiled from UnitedStatesZipCodes.org with some manual validation. It consists of 528 ZIP codes in total, with 119 specifically flagged as belonging to LA City. Key fields include ZIP Code ID, ZIP Code, a boolean flag for city status (`Is_City_Zip`), and creation timestamp. This table supports consistent geographic joining between datasets and enables focused analysis within the Los Angeles boundary.

la_zip_code

Zip_Code_ID	Zip_Code	Is_City_Zip	Created_At	Is_Processed
1	90001	1	2025-04-13 14:10:08	1
2	90002	1	2025-04-13 14:10:08	1
3	90003	1	2025-04-13 14:10:08	1
4	90004	1	2025-04-13 14:10:08	1
5	90005	1	2025-04-13 14:10:08	1

Figure 3.4: Sample Records form LA ZIP Code Reference Dataset

4. Technical Architecture

This project uses **Python 3.9+** for scraping, API access, data cleaning, and analysis. **Selenium** and **BeautifulSoup** handle dynamic scraping, while the **Yelp Fusion API** and **Census API** supply structured data. **SQLite3** serves as the relational database, with **Pandas** and **Jupyter Notebooks** for data processing and visualization.

4.1 Database Schema

The SQLite database includes four main tables: `restaurant_info`, `inspection_info`, `demographics_info`, and `la_zip_code`. An ERD illustrates table relationships and ensures normalized structure.

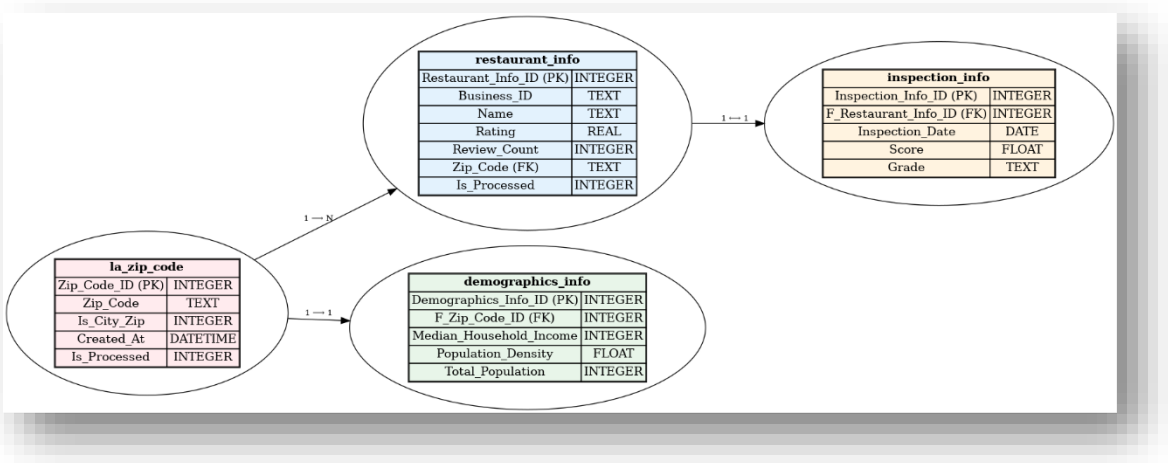


Figure 4.1: Entity Relationship Diagram (ERD)

4.2 Data Integration and Processing

The datasets were integrated using **ZIP codes** and **Restaurant IDs**. Yelp business entries were linked to health inspection records through foreign key of restaurant id, while demographic data was merged via shared ZIP codes to provide neighborhood context.

Data cleaning involved:

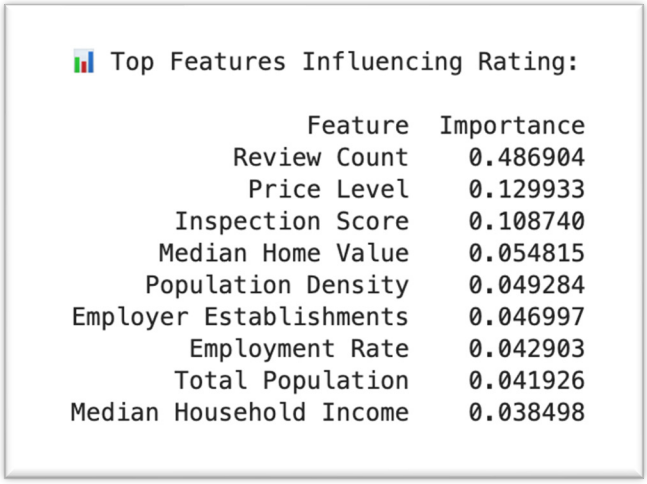
Data cleaning involved several structured steps to ensure accuracy and consistency across sources. First, health inspection records were filtered to include only those from the past 12 months, maintaining relevance to current restaurant operations. All datasets were restricted to valid entries using internal flags, and ZIP codes were normalized to consistent formats for reliable merging.

Deduplication was performed to eliminate redundant rows, and missing values were handled during the view (`Combined_Restaurant_View`) creation stage. Specifically, missing numeric fields were imputed using ZIP-level means and global fallback, while categorical fields like `Price` and `Grade` were handled using mode imputation and group-wise mode imputation.

All data was stored in a normalized schema within `final_project.db`, with four core tables linked via foreign keys and ZIP codes. A consolidated view named `Combined_Restaurant_View` was created to join, clean, and present the data for analysis. This ensured all downstream analysis worked from a single, reliable source. Cleaned outputs were also exported as CSV files for external use, located in the `/csv` folder.

4.3 Feature Importance Analysis

Determine the most influential features in the unified dataset that predict Ratings, serving as a foundation for downstream modeling and strategic interpretation. To achieve this, a **Random Forest Regressor** was trained on selected variables from the integrated dataset. The resulting feature importance scores offer quantitative insight into the primary drivers of rating variability.



📊 Top Features Influencing Rating:

Feature	Importance
Review Count	0.486904
Price Level	0.129933
Inspection Score	0.108740
Median Home Value	0.054815
Population Density	0.049284
Employer Establishments	0.046997
Employment Rate	0.042903
Total Population	0.041926
Median Household Income	0.038498

Figure 4.2: Top Features Influencing Rating

The model reveals that **Review Count**, **Price Level**, and **Inspection Score** are the most influential predictors of a restaurant's rating. These features reflect **customer engagement**, **value perception**, and **cleanliness**, respectively—factors highly visible to consumers and integral to satisfaction. In contrast, **demographic indicators** (e.g., income, population) show limited predictive power, reinforcing the notion that consumer experiences weigh more heavily than location context in shaping ratings.

This following ranked feature importance table ranks the most predictive variables based on their Random Forest importance scores. Review Count alone accounts for nearly 49% of model influence, far outweighing economic or geographic context variables.

📊 Feature Usage Table (For sections 5.1 to 5.4):

	Feature	Importance Score	5.1 Opportunity Score	5.2 Clustering/Correlation	5.3 Composite Success Index	5.4 Success Prediction Model
1	Review Count	0.4869	✅ Must include	✅	✅ Must include	✅
2	Price	0.1299	❌	❌	✅	✅
3	Inspection Score	0.1087	✅	✅	✅	✅
4	Median Home Value	0.0548	Optional	✅	Optional	✅
5	Population Density	0.0493	Optional	✅	Optional	✅
6	Employer Establishments	0.0470	Optional	✅	Optional	✅
7	Employment Rate	0.0429	Optional	✅	Optional	✅
8	Total Population	0.0419	❌ Skip	✅	❌	✅
9	Median Household Income	0.0385	❌ Skip	✅	❌	✅

Figure 4.3: Feature Usage Table Across Sections

This table maps each feature's predictive power to its usage across different analytical sections. Review Count, the most influential feature (48.7%), is included in all sections, while demographic metrics like Median Household Income are omitted due to minimal predictive value. The structured inclusion helps maintain analytical consistency, model clarity, and avoids overfitting across Opportunity Score, Clustering, Success Index, and Regression.

5. Exploratory Analysis & Visualizations

5.1 ZIP-Level Opportunity Score (Multi-Criteria Weighted Ranking)

To uncover the most promising ZIP codes in Los Angeles City for restaurant growth, a composite **Opportunity Score** was developed using a **multi-criteria decision analysis (MCDA)** framework. This approach leverages a weighted scoring model where **five key predictors**—average rating, inspection score, review volume, total employer establishments, and median home value—were selected based on their high feature importance from prior regression analysis. **Price Level** was excluded from the Opportunity Score to avoid biasing the score toward upscale ZIP codes and because it reflects individual restaurant strategy rather than neighborhood potential.

The process began with aggregating ZIP-level data from the unified dataset, grouped by ZIP code and filtered to include only those with at least five restaurant records. Next, each selected feature was **normalized using Min-Max scaling**, ensuring comparability across diverse value ranges. These normalized values were then combined using domain-informed weights:

- ⭐ 30% Rating
- 🧼 25% Inspection score
- 🗣️ 20% Review count
- 🏢 15% Employer establishments
- 🏠 10% Median home value

This formula produced a single **Opportunity Score** (ranging from 0 to 1) for each ZIP code, **economic potential**. The top 10 ZIP codes were ranked by converting opportunity score in percentage accordingly in the bar chart annotated with ranks and score percentages.

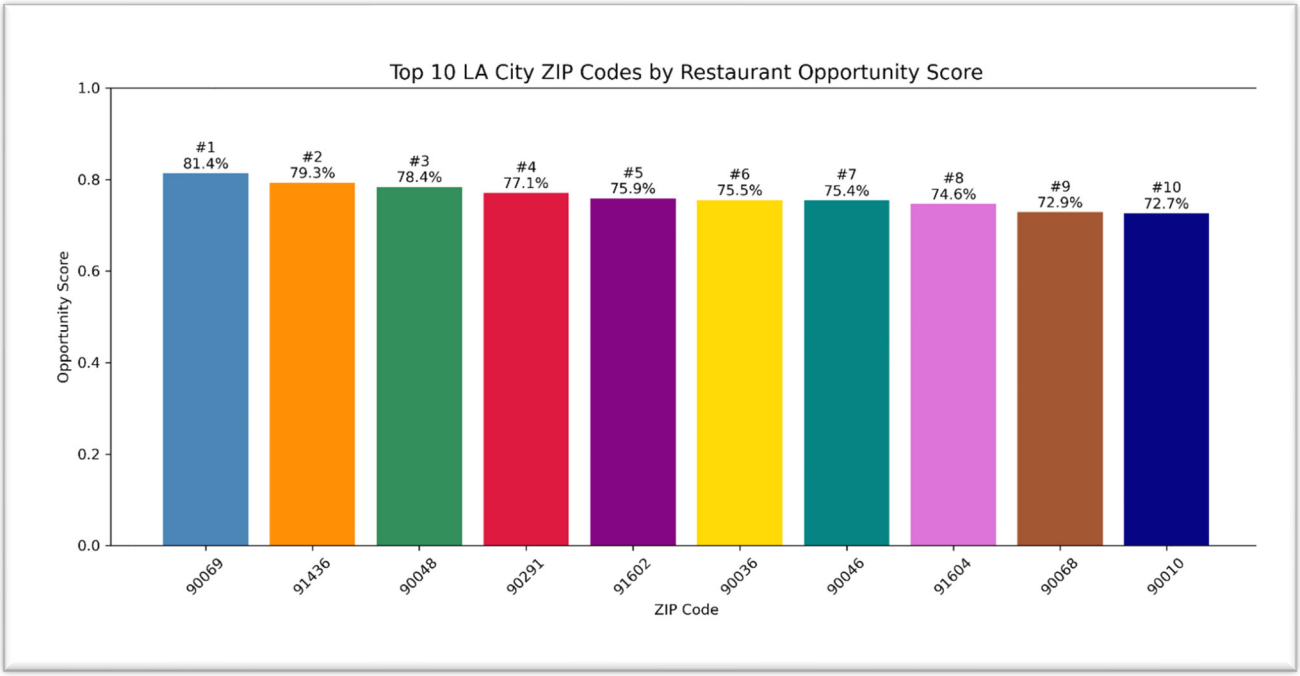


Figure 5.1: Top 10 LA ZIP Codes by Restaurant Opportunity Score

ZIP codes like **90069, 91436, and 90048** stand out as strategic hotspots for restaurant investment—offering the best combination of high ratings, cleanliness, consumer engagement, and local business density. This analysis enables data-driven site selection and supports targeted business expansion within LA City limits.

5.2 Correlation & Clustering Analyses

To better understand restaurant success drivers across Los Angeles City ZIP codes, this section combines correlation analysis and K-Means clustering to reveal relationships between features and group similar neighborhoods.

Correlation Analysis: A Pearson correlation matrix was computed across eight key variables including rating, inspection score, review count, and economic indicators like household income and population density. The dataset was filtered for completeness, and column names were renamed for clarity. The result was visualized using a Seaborn heatmap.

This analysis uncovers weak correlations between perceived quality (e.g., Rating) and official hygiene measures (e.g., Inspection Score at -0.05), while highlighting stronger economic ties like Home Value and Employer Establishments (0.62), emphasizing the role of affluence in restaurant viability.

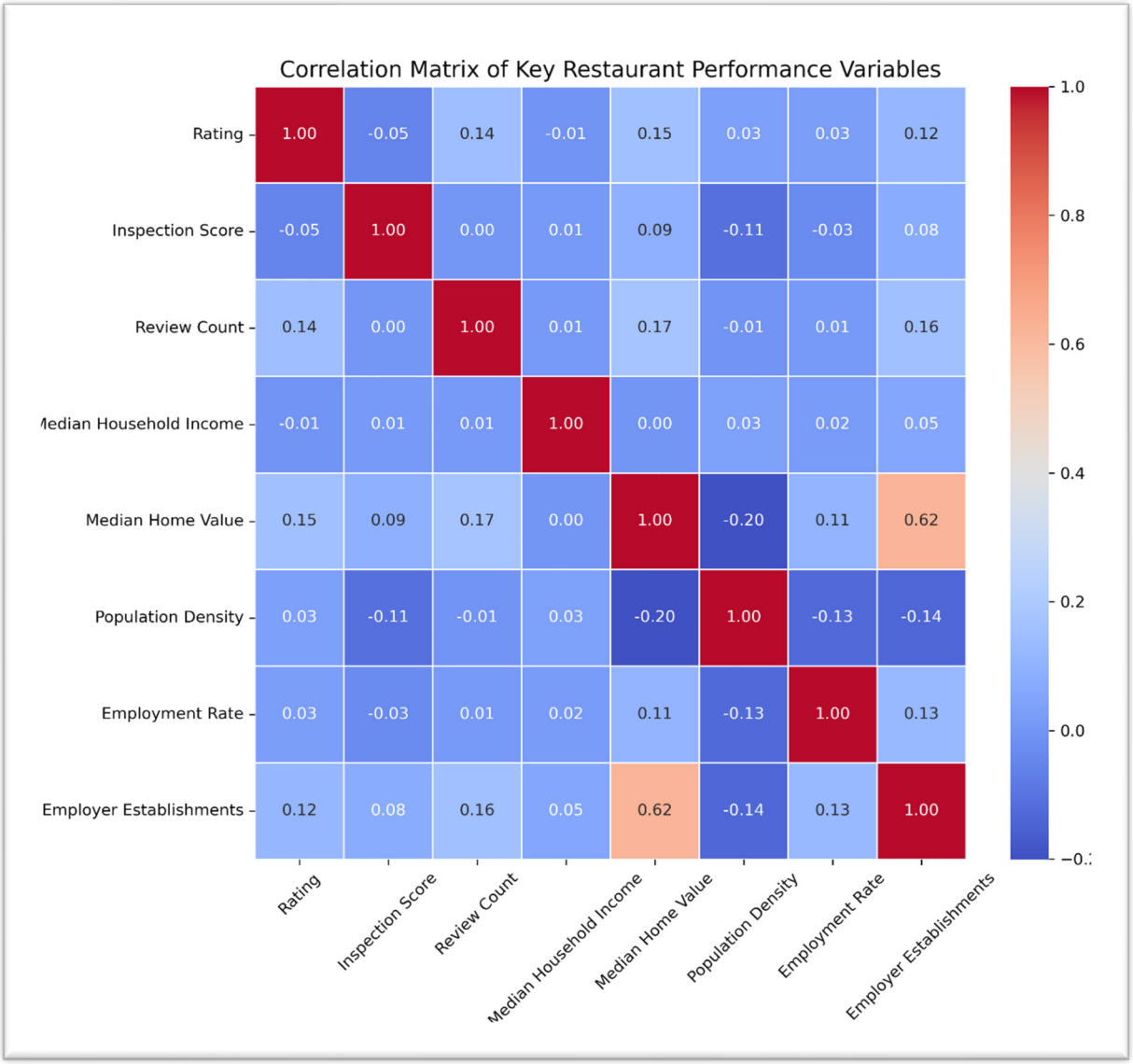


Figure 5.2: Correlation Matrix of Key Restaurant Performance Variables

This matrix helped identify interdependencies between metrics, informing more cohesive clustering logic in the next step.

Clustering Analysis: To segment ZIP codes by restaurant potential, **K-Means clustering (k=3)** was applied to ZIP-level averages of eight normalized features. Prior to clustering, values were scaled using StandardScaler to ensure equal feature influence. The algorithm grouped ZIP codes into three interpretable categories:

- **High Potential**: Affluent ZIPs with strong ratings and dense business presence
- **Emerging**: Moderately performing areas with upward potential
- **Underserved**: Lower-rated ZIPs with limited economic and customer indicators

The distribution chart shows the Emerging category is most common, while High Potential ZIPs like 90036 and 90027 stand out as premium restaurant zones.

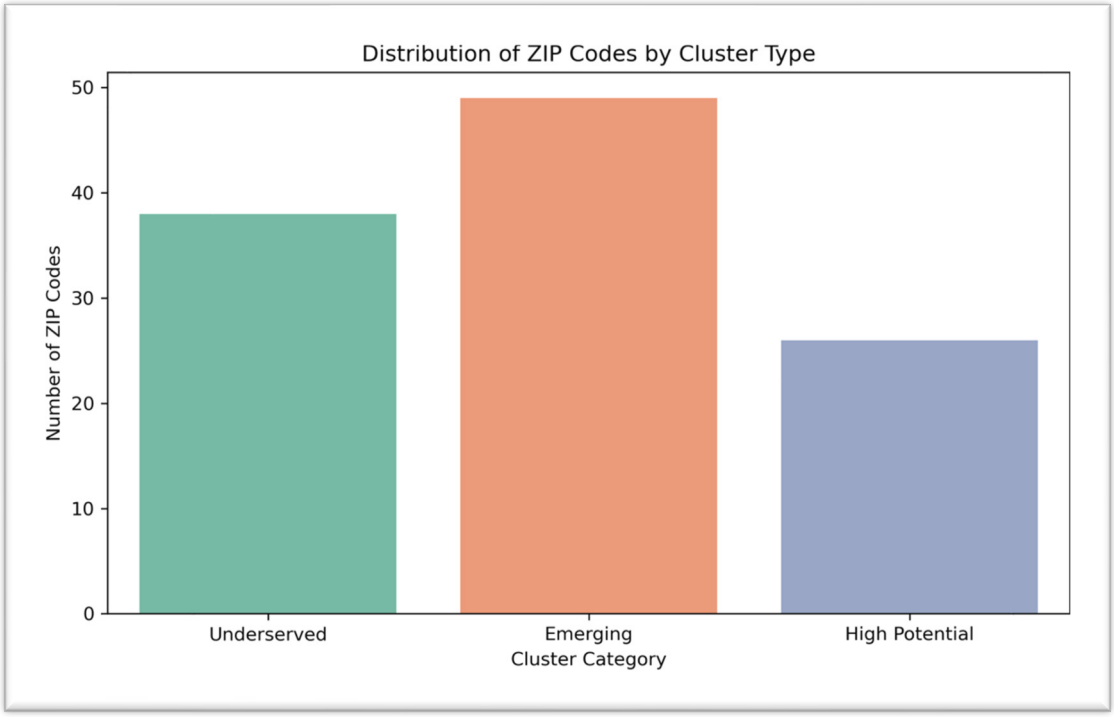


Figure 5.3: ZIP Code Cluster Labels by Opportunity Type

Cluster Category

```
Cluster_Label
Emerging          [90004, 90005, 90006, 90012, 90013, 90014, 900...
High Potential    [90010, 90024, 90025, 90027, 90034, 90036, 900...
Underserved       [90001, 90002, 90003, 90007, 90008, 90011, 900...
Name: Zip_Code, dtype: object
```

Figure 5.4: Distribution of ZIP Codes by Cluster Category

Together, these analyses offer both a micro- and macro-level understanding of how ZIP codes differ in restaurant dynamics—providing critical input for targeting investments and optimizing service strategies.

5.3 Composite Success Index & Restaurant Performance Clustering

To evaluate restaurant performance across Los Angeles City, I developed a **Success Index** using a weighted composite scoring method. Five operational features—**rating, review count, inspection score, price level, and employer density**—were normalized using MinMaxScaler to ensure comparability. These normalized values were then combined via a weighted linear formula (weights: 30%, 25%, 20%, 15%, 10%, respectively) to reflect their influence on perceived restaurant quality and business viability. This resulted in a composite score scaled from 0 to 100, capturing local economic context. The **top 10 restaurants**, based on this Success Index, were visualized in a bar chart with detailed labels including name, rating, price tier, and address.

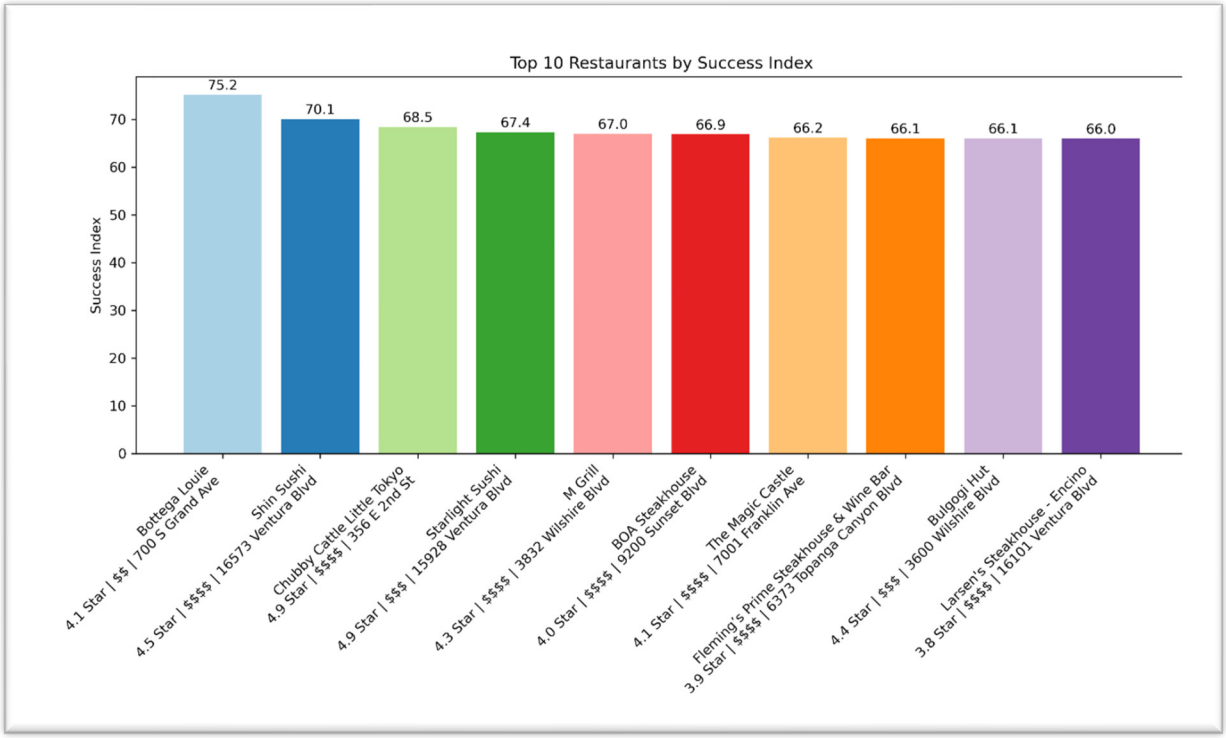


Figure 5.1: Top 10 Restaurants by Success Index

This bar chart reveals the best-performing restaurants, with **Bottega Louie** leading at a score of **75.2**. The visualization synthesizes diverse success indicators—cleanliness, affordability, popularity, and business density—into a clear benchmark for operational excellence.

To complement this individual-level analysis, I applied **K-Means clustering** (with **k=3**) on the same normalized feature set (excluding the manually weighted Success Index) to discover natural groupings among restaurants. This **unsupervised learning** technique grouped restaurants into three clusters: **High**, **Mid**, and **Low** performance tiers, based solely on latent patterns in the data. Final cluster labels were assigned based on average ratings per cluster to maximize interpretability.

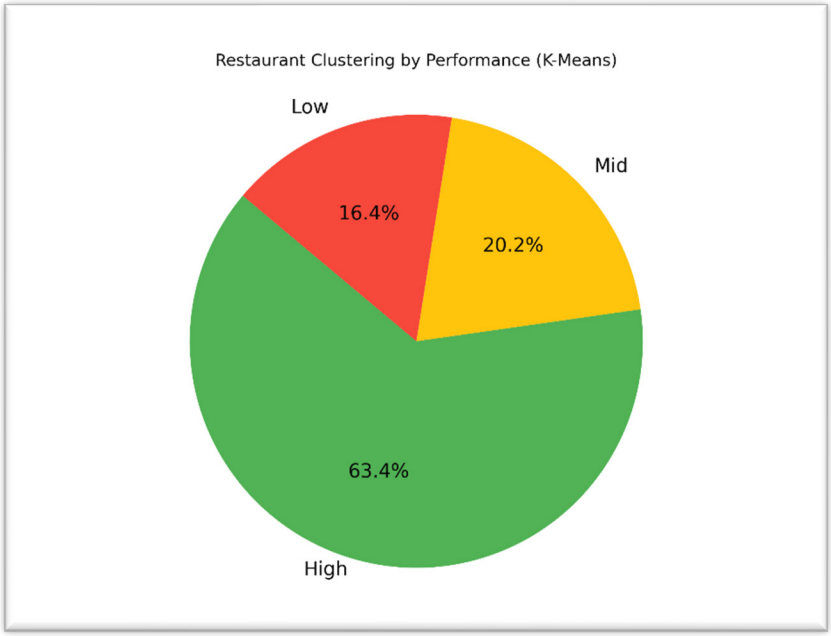


Figure 5.2: Restaurant Clustering by Performance (K-Means)

The pie chart shows that **63.4% of restaurants fall into the High tier**, indicating strong competition and consumer satisfaction citywide. This clustering provides a macro-level segmentation of the LA dining scene and can guide investment, policy, or benchmarking decisions.

5.4 Restaurant Success Prediction Model

To forecast which restaurants are likely to succeed, a supervised machine learning model was developed using a **Random Forest Classifier**. The success label was defined by computing a composite **Success Index**, which aggregated five normalized performance metrics—Rating, Review Count, Inspection Score, Price Level, and Employer Establishments—weighted respectively (30%, 25%, 20%, 15%, 10%). These features were normalized using **MinMaxScaler**, and restaurants ranking in the **top 30% of Success Index scores** were labeled as "successful."

The model was trained on nine normalized predictors: the five core success metrics and four ZIP-level socioeconomic indicators (Median Income, Median Home Value, Population Density, Employment Rate). After a **train-test split**, the Random Forest model was fitted to the training data and evaluated on the test set, achieving an impressive **accuracy of 95.9%** and an **AUC score of 0.99**, as visualized in the ROC curve.

```
Random Forest Accuracy: 0.9590027700831025
Random Forest Confusion Matrix:
 [[1206   34]
 [  40  525]]
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97      1240
           1       0.94      0.93      0.93       565

    accuracy                           0.96      1805
   macro avg       0.95      0.95      0.95      1805
weighted avg       0.96      0.96      0.96      1805
```

Figure 5.1: Random Forest Classification Report

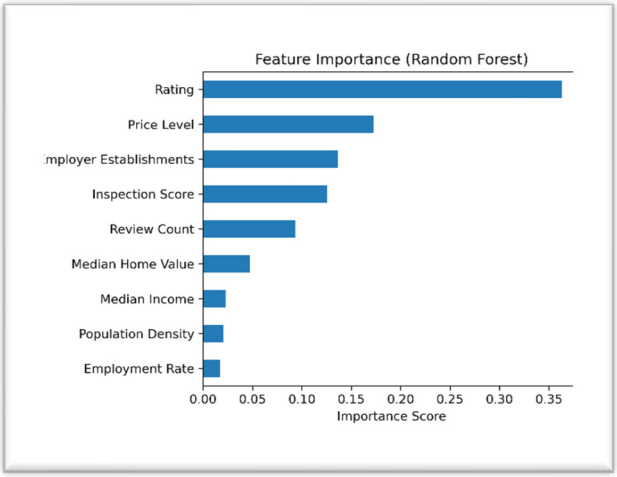


Figure 5.2: Feature Importance for Restaurant Success Classification

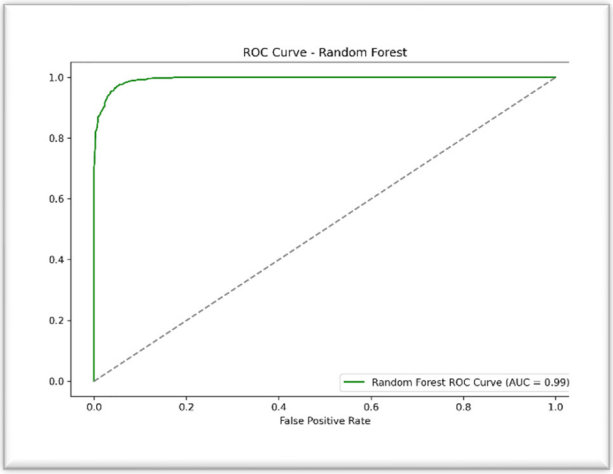


Figure 5.3: ROC Curve for Random Forest Classifier

Key outputs included:

- **Classification Report & Confusion Matrix**: High precision (0.94–0.97) and recall validate the model's reliability.
- **Feature Importance Plot**: Rating, Price Level, and Employer Establishments emerged as top predictors.
- **ROC Curve**: A near-perfect AUC reflects excellent discriminatory power.
- **Boxplots**: A 3x3 grid shows the distribution of key features between high- and low-success restaurants, enhancing model interpretability and supporting validation of feature influence.

This predictive framework enables stakeholders to identify high-potential restaurants and understand key drivers of success within the competitive Los Angeles market.

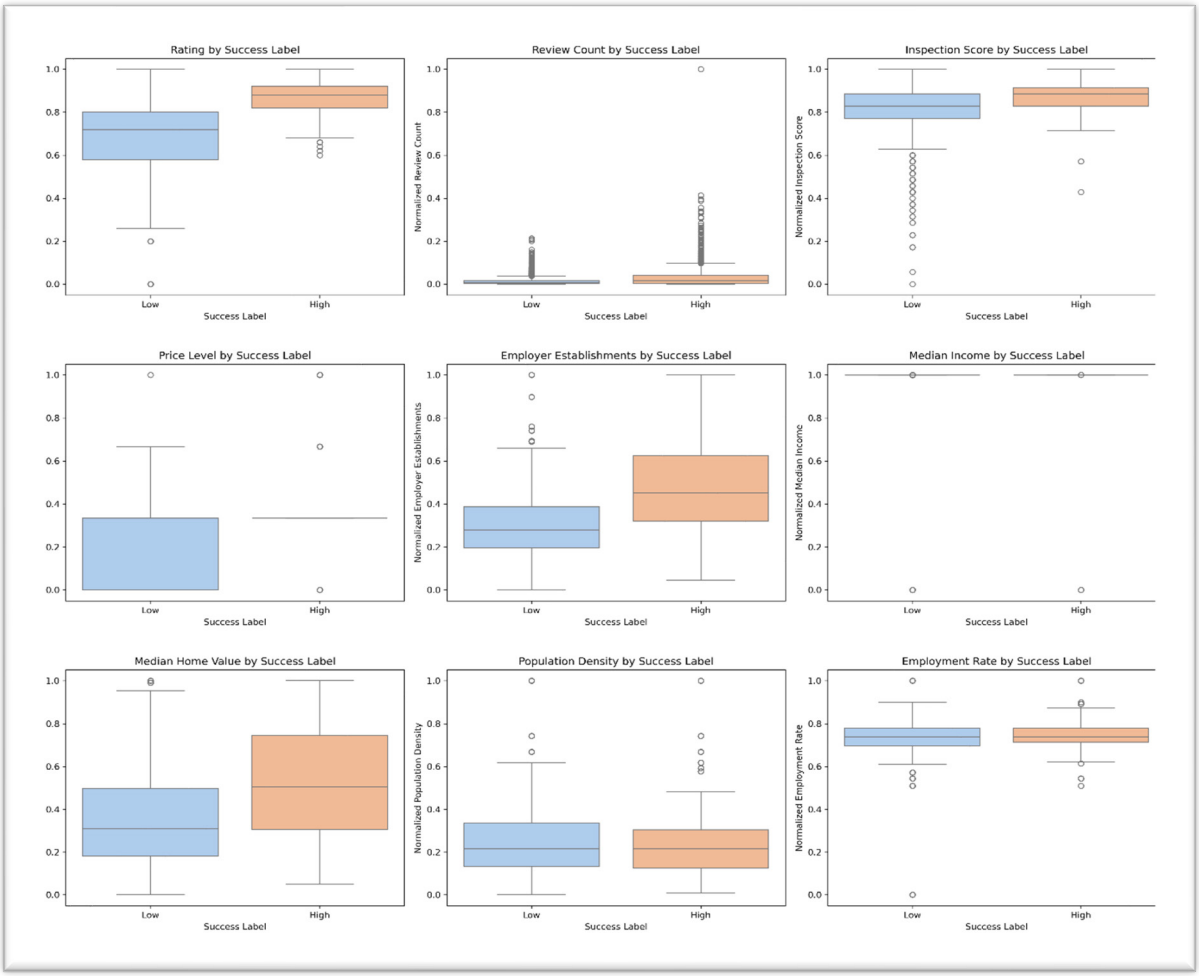


Figure 5.4: Boxplots of Key Features by Restaurant Success Label

6. Key Questions & Data Insights

Q1. Do higher health inspection scores correlate with higher ratings?

Analysis: A Pearson correlation coefficient was computed between inspection scores and star ratings using the combined dataset of matched restaurants. Results were visualized using a correlation heatmap.

Findings: The correlation between Rating and Inspection Score is weakly negative (-0.05), suggesting that public perception does not closely track hygiene outcomes. This reveals a potential disconnect between customer reviews and regulatory cleanliness—highlighting the importance of integrating both views for a full picture of restaurant quality.

Q2: Are lower-income neighborhoods more likely to have poorly rated or less clean restaurants?

Analysis: Restaurants were grouped by ZIP codes and stratified by Median Household Income brackets. The average rating and average inspection score were calculated and compared across these groups using boxplots.

Findings: Lower-income ZIP codes tend to show slightly lower average ratings and inspection scores, although variation exists. This indicates that economic disparity may influence access to high-performing restaurants, underscoring potential equity gaps in LA City's food landscape.

Q3: What features most influence restaurant ratings?

Analysis: A Random Forest Regressor was trained on unified dataset features to predict ratings. Feature importances were extracted and ranked.

Findings: Review Count, Price Level, and Inspection Score emerged as the top predictors of ratings, confirming that consumer engagement, affordability, and hygiene are critical rating drivers—more so than demographics like income or population density.

Q4: Which ZIP codes are most favorable for launching new food businesses?

Analysis: A composite Opportunity Score was developed using a weighted model of 5 ZIP-level metrics: Yelp rating, inspection score, review count, employer establishments, and median home value. Scores were normalized, weighted, and ranked.

Findings: ZIP codes such as 90069, 91436, and 90048 ranked highest in opportunity, offering ideal conditions for new restaurant ventures. These areas balance consumer satisfaction, strong hygiene records, and economic vitality, making them prime targets for investment.

Q5: Which ZIP codes offer the best mix of high customer ratings, cleanliness, and economic opportunity?

Analysis: I filtered the unified dataset to calculate a composite score per ZIP code based on three equally weighted metrics:

- Average Rating
- Average Inspection Score (cleanliness & regulatory compliance)
- Median Household Income (economic opportunity)

Each metric was normalized using Min-Max scaling, and the average of these normalized values was computed to form a balanced ZIP Performance Score. ZIP codes were then ranked and the top performers were identified.

Findings: ZIP codes such as 90024, 90069, and 90036 consistently rank at the top, reflecting strong customer satisfaction, high sanitation standards, and affluent neighborhoods. These ZIPs represent ideal zones for launching or investing in restaurants, offering both demand and quality-of-life infrastructure.

Q6: What demographic factors (i.e.: income, density, housing value) most influence restaurant performance?

Analysis: I explored the influence of demographic variables on restaurant performance using a combination of correlation analysis and K-Means clustering. A Pearson correlation matrix revealed how Yelp Rating relates to ZIP-level factors such as Median Household Income, Median Home Value, Population Density, and Employment Rate. These features were also used in K-Means clustering (after normalization) to group ZIP codes into High Potential, Emerging, and Underserved categories based on overall performance and economic context.

Findings: Among demographic factors, **Median Home Value and Employment Rate** showed stronger associations with higher-rated, higher-performing ZIP codes. In contrast, **Household Income and Population Density** showed weaker correlations with ratings, indicating that economic *stability* and business *vibrancy*—rather than wealth or density alone—are more indicative of restaurant success. These patterns were consistently reflected in both the correlation heatmap and clustering results.

7. Maintainability & Extensibility

This project was architected for scalability and future expansion, with the codebase structured into modular Python classes such as `ZipCodeManager, YelpDataManager, and DemographicsManager` - enabling clean separation of concerns, easier debugging, and targeted updates. In addition, the use of a normalized **SQLite relational database schema**, combined with modular and well-documented Python scripts, ensures that data can be easily updated, extended, or reused. Key maintainability features include the creation of SQL **Views** like `Combined_Restaurant_View`, which encapsulate complex joins and filters. This provides a single source of truth for analysis, reducing code repetition and enabling consistent downstream analytics without altering raw tables.

Python scripts are organized into clear steps: data ingestion, cleaning, merging, modeling, and visualization—each segment is separated by flags and modular functions for reusability. If new restaurant or inspection data is added, the process can be rerun with minimal adjustments.

Extensibility was a core consideration. The same architecture can be scaled to:

- Include **other cities** like San Francisco or New York by swapping ZIP data
- Expand the model with **new APIs** like Google Places, DoorDash, or UberEats
- Add **sentiment analysis** from reviews using NLP
- Integrate **real-time dashboards** using tools like Streamlit or Tableau
- Enrich features with **traffic, zoning, or walkability data**
- Evolve the Success Index into a **Composite Quality Index** by adding cuisine type, service hours.
- Build interactive apps or web based GUIs that take consumer or investor inputs and recommend top restaurants or ideal business locations using the integrated dataset and predictive models.

Moreover, the **opportunity scoring, clustering, and predictive modeling pipelines** are independent and can be fine-tuned or replaced without breaking the rest of the system. Clear documentation and use of open-source libraries (like scikit-learn, matplotlib, seaborn) support long-term code sustainability and team collaboration.

8. Conclusion

This project provides a data-driven perspective on the key factors shaping restaurant success across Los Angeles City. The analysis uncovered that customer engagement (as seen in review count), perceived value (via price level), and hygiene (inspection scores) are the strongest indicators of high ratings—more so than neighborhood wealth or density. Notably, ZIP codes with balanced economic infrastructure and business density, rather than just high income, proved more fertile for restaurant growth.

Clustering models segmented neighborhoods into distinct opportunity tiers, revealing underserved areas in need of quality dining options and highlighting high-potential ZIPs ideal for investment. The Success Index further pinpointed standout restaurants that excel operationally across rating, cleanliness, affordability, and surrounding business conditions. Additionally, predictive modeling achieved over 95% accuracy, empowering stakeholders to anticipate restaurant success based on core performance metrics.

Together, these insights go beyond intuition or online buzz, equipping decision-makers with actionable intelligence to navigate LA City's complex food landscape. By aligning public health, economic context, and user perception, the project delivers a comprehensive framework that can guide smarter choices—from consumers to entrepreneurs and policymakers alike, **making it a win-win for both businesses and consumers.**

9. References

- Yelp Fusion API: https://docs.developer.yelp.com/docs/fusion-intro
- LA County Health Inspection Portal: https://ehservices.publichealth.lacounty.gov
- UnitedStatesZipCodes.org: https://www.unitedstateszipcodes.org
- U.S. Census Bureau API: https://www.census.gov/data/developers/data-sets.html
- Scikit-learn (sklearn): https://scikit-learn.org/stable
- Matplotlib: https://matplotlib.org
- Seaborn: https://seaborn.pydata.org
- Selenium: https://www.selenium.dev
- BeautifulSoup: https://www.crummy.com/software/BeautifulSoup
- SQLite: https://www.sqlite.org
- Python 3.9+: https://www.python.org/downloads/release/python-390
- Jupyter Notebook: https://jupyter.org
- Requests: https://pypi.org/project/requests
- Selenium (PyPI): https://pypi.org/project/selenium
- Pandas: https://pandas.pydata.org
- Webdriver-manager: https://pypi.org/project/webdriver-manager