A Data-Driven Analysis of Restaurant Reviews: Clustering and Sentiment Insights

Pasindu Kurukulasooriya (S/19/421)
Harshamala Bandara (S/19/323)
Thineskumar Srirasa (S/19/582)

Supervisor: Dr. Mahasen Dehideniya

DSC 3263 : Independent Study in Data Science
Department of Statistics and Computer Science
Faculty of Science
University of Peradeniya
2024

Declaration

We hereby declare that this project report titled **"A Data-Driven Analysis of Restaurant Reviews: Clustering and Sentiment Insights"** is the result of our independent work carried out under the guidance of Dr. Mahasen Dehideniya, Prof. Y.P.R.D. Yapa and Dr. M.S. Atapattu. The content presented in this report is original and has not been submitted, either in part or in full, for any other academic purpose or examination. We confirm that all sources of information, data, figures, and concepts from external sources have been properly acknowledged and referenced following academic standards. This project was conducted as a part of the requirements of our academic program, and we take full responsibility for the authenticity of the work submitted.

.....	.....	.....
Pasindu Kurukulasooriya	Harshamala Bandara	Thineskumar Srirasa
(S/19/421)	(S/19/323)	(S/19/582)

Certified by:

.....	.....
Dr. M.S. Atapattu	**Prof. Y.P.R.D. Yapa**
Dr. M.S. Atapattu	Professor in Statistics and Computer Science

.....
Dr. Mahasen Dehideniya
Supervisor

Date: _____

Acknowledgments

We would like to extend our sincere gratitude to our supervisor, Dr. Mahasen Dehideniya, who helped us and provided clarifications throughout the project. We are also grateful to Prof. Y.P.R.D. Yapa and Dr. M.S. Atapattu for their guidance and support. In addition, we also extend our gratitude to the Department of Statistics and Computer Science for facilitating and conducting the course module.

Finally, we would like to express our gratitude to one another. This project was a collaborative effort that demanded dedication and teamwork from each group member to achieve its successful completion.

Contents

1 Introduction

The reputation of a restaurant and the decisions made by patrons are greatly influenced by online reviews. Large volumes of insightful feedback are produced daily as a result of patrons sharing their eating experiences online more and more. However, because it is dispersed across numerous platforms and is mainly unstructured, this data frequently goes unutilized. When examined methodically, certain restaurant characteristics like location, food varieties, and ratings can also provide insightful information in addition to reviews. Restaurant owners, managers, and decision-makers can better comprehend customer satisfaction, operational strengths and weaknesses, and new market trends by fusing these structured information with unstructured review content.

This project focuses on restaurants in Kandy, Sri Lanka. By combining clustering and sentiment analysis, this study aims to explore how customer reviews and location-based patterns can be leveraged to support data-driven strategies in the restaurant industry.

2 Problem Statement

Despite the abundance of online reviews, restaurants often lack the tools or expertise to systematically analyze them for meaningful patterns. Without structured analysis, it becomes difficult to understand overall customer sentiment, identify well-performing restaurants, or detect competitive clusters within a city. This project addresses the need for a robust approach to process, analyze, and visualize restaurant review data in a way that highlights trends, uncovers hidden patterns, and supports better decision-making.

3 Objectives

The main objectives of this project are to:

- Analyse restaurant data and customer reviews for restaurants located in Kandy.
- Conduct sentiment analysis to uncover how customers feel about their dining experiences.
- Identify top-performing restaurants and highlight popular cuisines.
- Use clustering techniques to detect natural groupings of restaurants based on location and ratings.

- Develop an interactive dashboard, a simple web application and map visualisation that allows stakeholders to explore these insights intuitively.

4 Methodology

A structured methodology was adopted to achieve the project's objectives. This included comprehensive steps for data collection, cleaning, analysis, modelling, and visualisation.

4.1 Data Collection

Data was gathered from "Tripadvisor" online platform using "APIFY" which provides a full-stack cloud platform for web scraping, browser automation, and data extraction. The data included key attributes such as each restaurant's location ID, name, cuisine types, average ratings, and customer reviews. Reviews were collected with details like review text and review language to support multilingual analysis.

4.2 Data Preprocessing

To ensure the data was ready for analysis, several preprocessing steps were carried out.

- Removed irrelevant columns.
- Renamed columns meaningfully.
- Removed duplicates and checked for null values.
- Validated data types.
- Created new columns for combined analysis

Irrelevant columns were removed to reduce noise in the dataset. Columns were renamed with meaningful, descriptive names for clarity. Duplicate entries were identified and removed to avoid skewing the analysis. Checks for null values were performed, and missing values were handled appropriately. Data types were validated to ensure consistency for numerical and categorical fields.

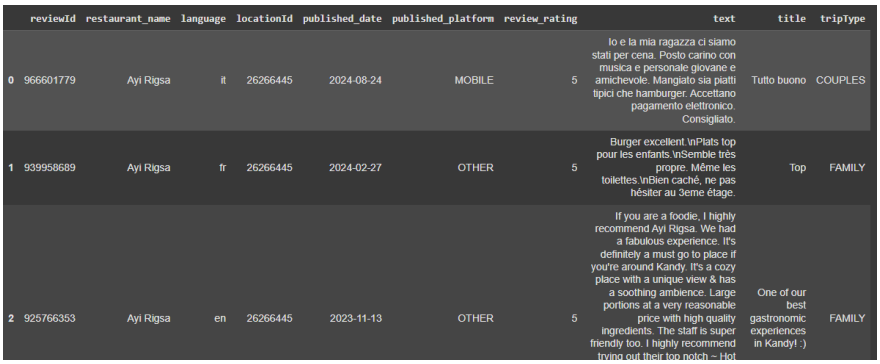
Two main datasets were prepared:

- **restaurant_df**: 239 unique restaurants with 10 variables, including details such as restaurant_name, location coordinates, address, review count and ratings.

- **reviews_df**: 23,371 individual reviews linked to the 239 restaurants, with 9 variables such as review text, review rating, published date and language.

	locationId	restaurant_name	latitude	longitude	address	isLocalChef	isPremium	review_count	rating	cuisines_combined
0	27740176	Grand Sky Lounge	7.285383	80.64447	12 Mahamaya Mawatha, Kandy 20000 Sri Lanka	0	1	157	4.8	Bar, International, Dining bars
1	13551312	Vito Wood Fired Pizza	7.287843	80.64234	56, Saranankara Road, Kandy 20000 Sri Lanka	0	0	1000	4.8	Italian, Pizza, Beer restaurants
2	13427841	Hideout Lounge	7.287413	80.64656	52 Sangaraja Mawatha, Kandy 20000 Sri Lanka	0	0	793	4.6	International, Dining bars
3	10458165	Sulochana's Kitchen	7.290715	80.60727	162/1 Panasara Mawatha,Halloluwa. Bus Junction Halloluwa,kandy, Kandy 20023 Sri Lanka	0	0	89	5.0	Sri Lankan
4	1134977	Sharon Inn	7.287215	80.64206	59 Saranankara Rd Kandy Lake,, Kandy 20000 Sri Lanka	0	0	512	4.4	Asian, Sri Lankan

Figure 1: Restaurant details dataset



	reviewId	restaurant_name	language	locationId	published_date	published_platform	review_rating	text	title	tripType
0	966601779	Ayi Rigsa	it	26266445	2024-08-24	MOBILE	5	Io e la mia ragazza ci siamo stati per cena. Posto carino con musica e personale giovane e amichevole. Mangiato sia piatti tipici che hamburger. Accettano pagamento elettronico. Consigliato.	Tutto buono	COUPLES
1	939958689	Ayi Rigsa	fr	26266445	2024-02-27	OTHER	5	Burger excellent \nPlats top pour les enfants.\nSemble très propre. Même les toilettes.\nBien caché, ne pas hésiter au 3eme étage.	Top	FAMILY
2	925766353	Ayi Rigsa	en	26266445	2023-11-13	OTHER	5	If you are a foodie, I highly recommend Ayi Rigsa. We had a fabulous experience. It's definitely a must go to place if you're around Kandy. It's a cozy place with a unique view & has a soothing ambience. Large portions at a very reasonable price with high quality ingredients. The staff is super friendly too. I highly recommend trying out their top notch ~ Hot	One of our best gastronomic experiences in Kandy! :)	FAMILY

Figure 2: Restaurant reviews dataset

4.3 Exploratory Data Analysis (EDA)

A detailed exploratory data analysis was conducted to understand the structure and key trends in the dataset. Visualisations were created to show the distribution of restaurant ratings, revealing that the majority of restaurants in Kandy are well-rated, with very few having ratings below three. The analysis of cuisine types showed that most restaurants offer multiple cuisines, with three cuisines per restaurant being the most common. Asian and Sri Lankan cuisines emerged as the dominant types, while seafood, fast food, and European cuisines were less common.

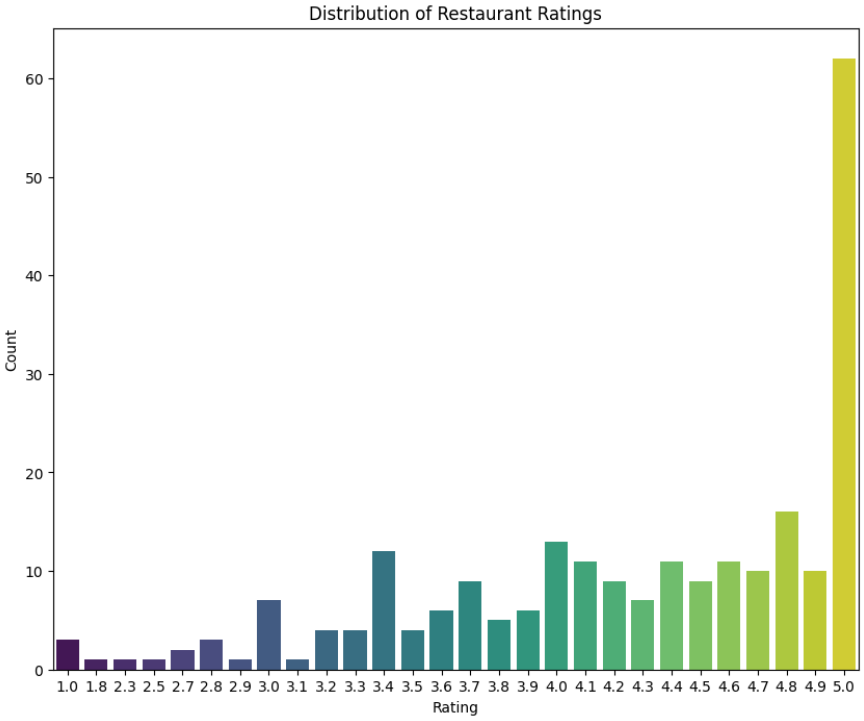


Figure 3: Distribution of restaurant rating

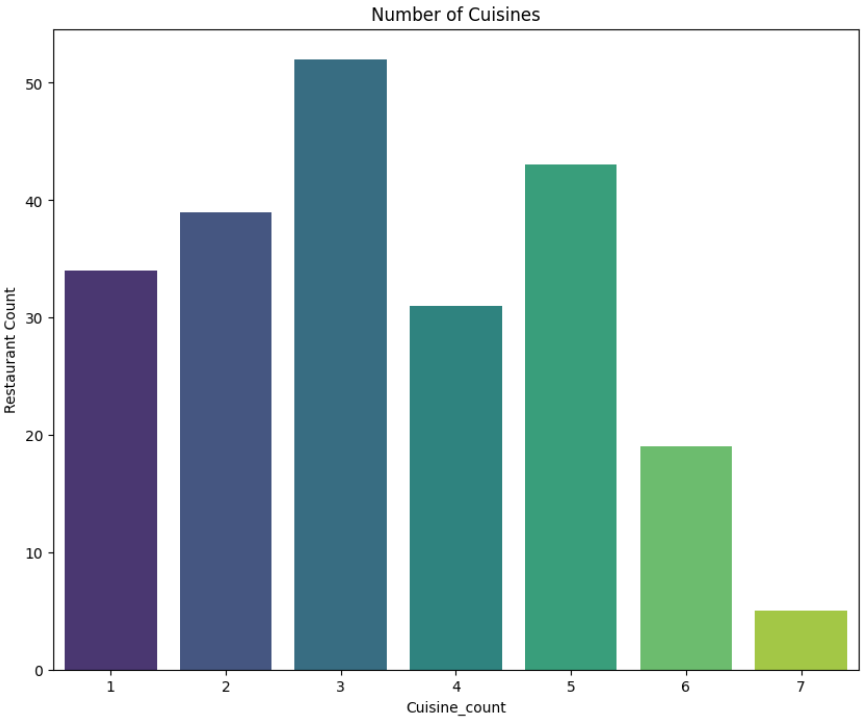


Figure 4: Distribution of cuisine counts offered

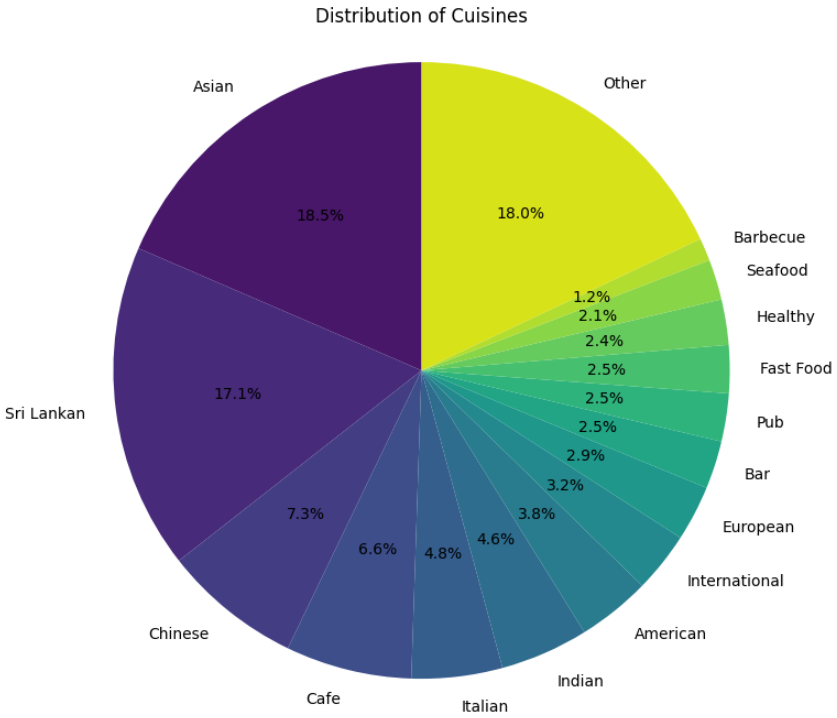


Figure 5: Distribution of cuisine types

Review patterns were also examined, highlighting that customer reviews tend to be more positive than negative.

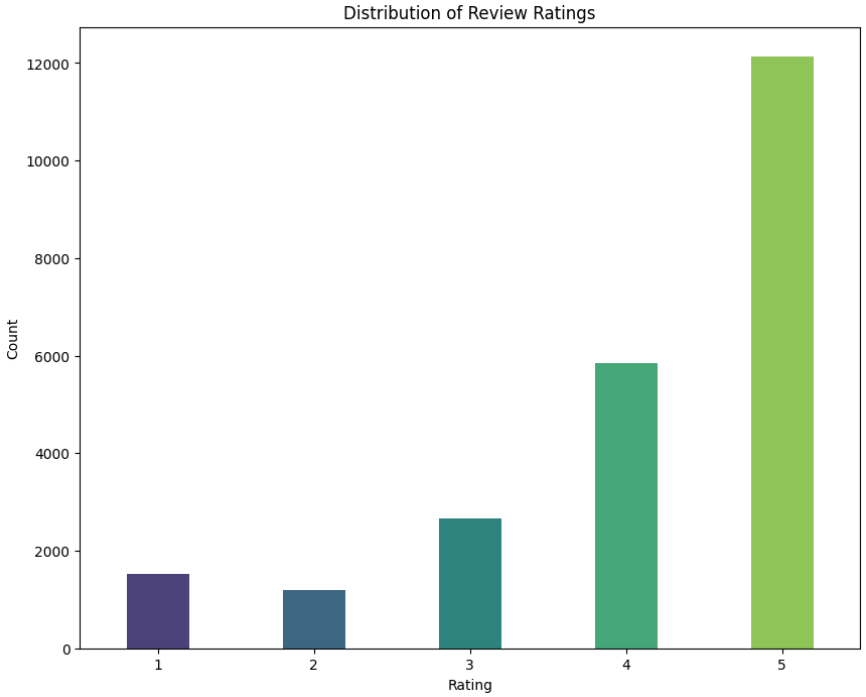


Figure 6: Distribution of restaurant review ratings

The number of reviews submitted each year showed steady growth until 2019, with a noticeable dip in 2020 likely due to COVID-19 restrictions, followed by a recovery in 2021 and beyond.

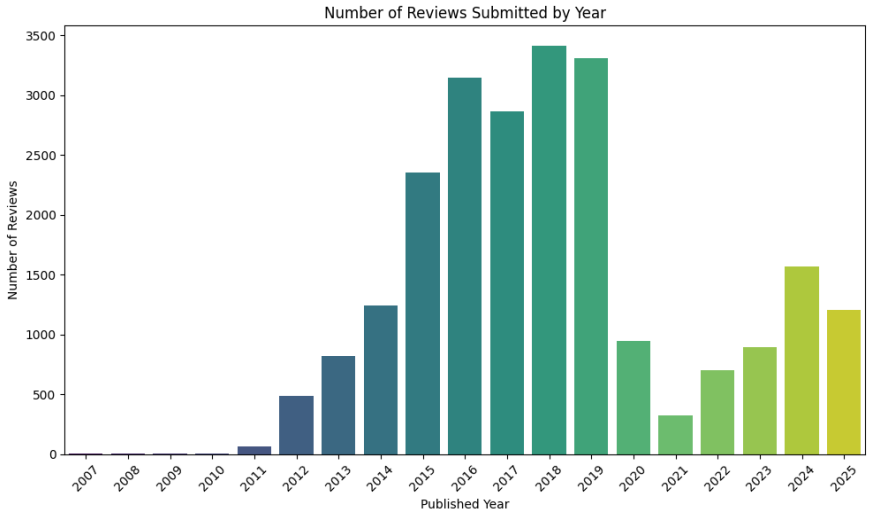


Figure 7: Distribution of review count by year

Top restaurants were identified based on average ratings and total number of reviews, with Vitto Wood Fired Pizza, Sulochana's Kitchen, and Kandyan Rice and Curry standing out

among the top performers. The distribution of reviews by language and a map visualisation of restaurant locations in Kandy provided further insights into the dataset.

4.4 Clustering

To detect patterns in restaurant locations and performance, a clustering was implemented using the K-Means algorithm. The clustering focused on geographic coordinates and ratings to identify groups of similar restaurants. The optimal number of clusters (K) was determined through the Silhouette Score, which measures how well-separated and cohesive the clusters are. For K=5, the Silhouette Score was 0.9078, indicating that the clusters are distinct and meaningful for further analysis.

4.5 Sentiment Analysis

Sentiment analysis was applied to the textual review data to understand how customers feel about the restaurants. The preprocessing steps included removing noise (punctuation, HTML tags, special characters), normalizing the text by converting it to lowercase, tokenizing sentences into words, removing stop words, handling emojis and slang, and applying stemming and lemmatisation to reduce words to their root forms.

The cleaned text data was then converted into numerical representations suitable for machine learning models. Multiple models were tested, including Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machines (SVM). The dataset was split into training and test sets, and the models were evaluated using accuracy, precision, recall, and F1-score. Among the models tested, Naive Bayes achieved the highest accuracy of 76%, demonstrating its effectiveness for classifying the sentiment of short customer reviews.

4.6 Dashboard and Website

To make the results accessible and interactive, a dashboard and a simple web app were developed. The dashboard enables users to filter restaurants based on clusters, sentiment, or other attributes, while the map provides a visual representation of restaurant locations, clusters, and sentiment trends.

5 Results and Discussion

The EDA provided key insights into the restaurant landscape in Kandy. The majority of restaurants maintain good customer ratings, and multi-cuisine menus are common. Asian

and Sri Lankan cuisines dominate the market, while others have smaller representation. Review analysis confirmed that customers are generally positive, with fewer negative experiences being shared. Annual review trends showed growth in customer engagement, with an expected dip during the pandemic period. These insights highlight customer preferences, potential competitive areas, and trends in dining behavior.

The K-Means clustering revealed that restaurants naturally form five distinct clusters based on their geographic proximity and rating similarities. The high Silhouette Score of 0.9078 indicates that the clustering is robust, with restaurants in each cluster sharing more similarities with each other than with restaurants in other clusters. This information can be useful for strategic decision-making, such as identifying competitive zones and planning for expansion or new outlets in underserved areas.

The sentiment analysis showed that the Naive Bayes model was the most effective for classifying customer reviews, achieving an accuracy of 76%. This suggests that simple machine learning models can provide meaningful sentiment insights for short review texts. The majority of reviews were found to be positive, reflecting overall customer satisfaction with restaurants in Kandy. This information can help restaurant owners understand their strengths and areas for improvement.

The interactive dashboard and map serve as a practical tool for stakeholders to explore the project's findings. Users can view restaurant clusters, examine sentiment trends, and drill down into individual restaurant details. By integrating location data and customer feedback into a single interface, the dashboard transforms complex analysis into actionable insights that can be used for marketing strategies, service improvements, and competitive benchmarking.

Dashboard:

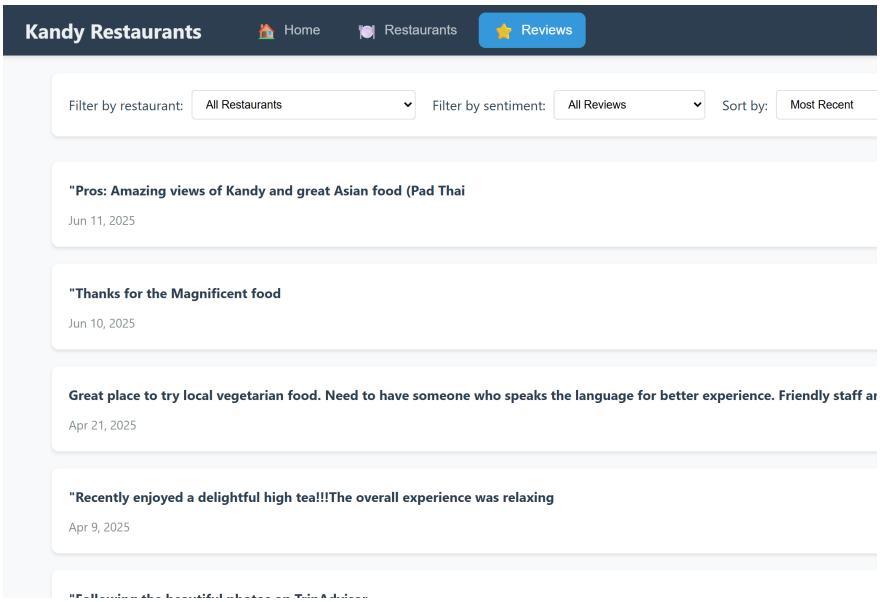


Figure 8: Dashboard

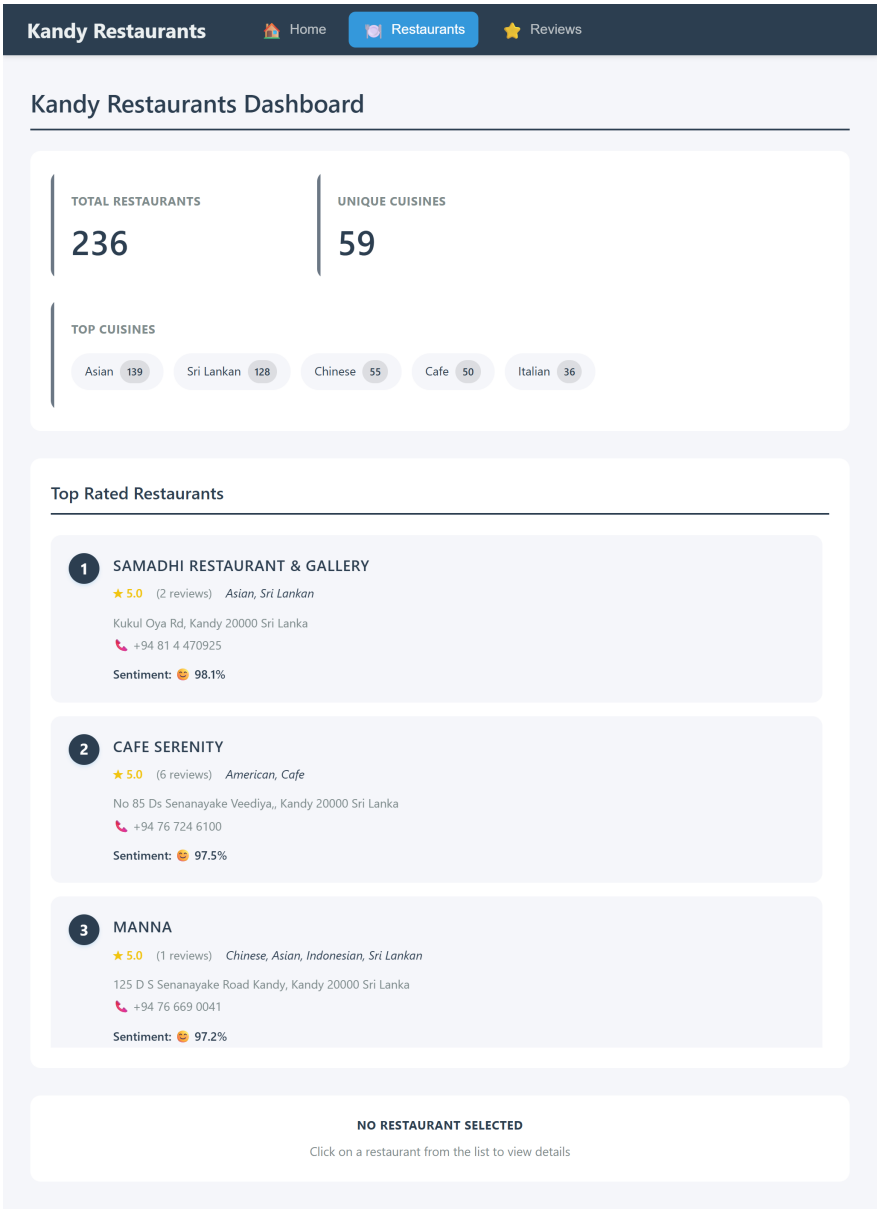


Figure 9: Dashboard

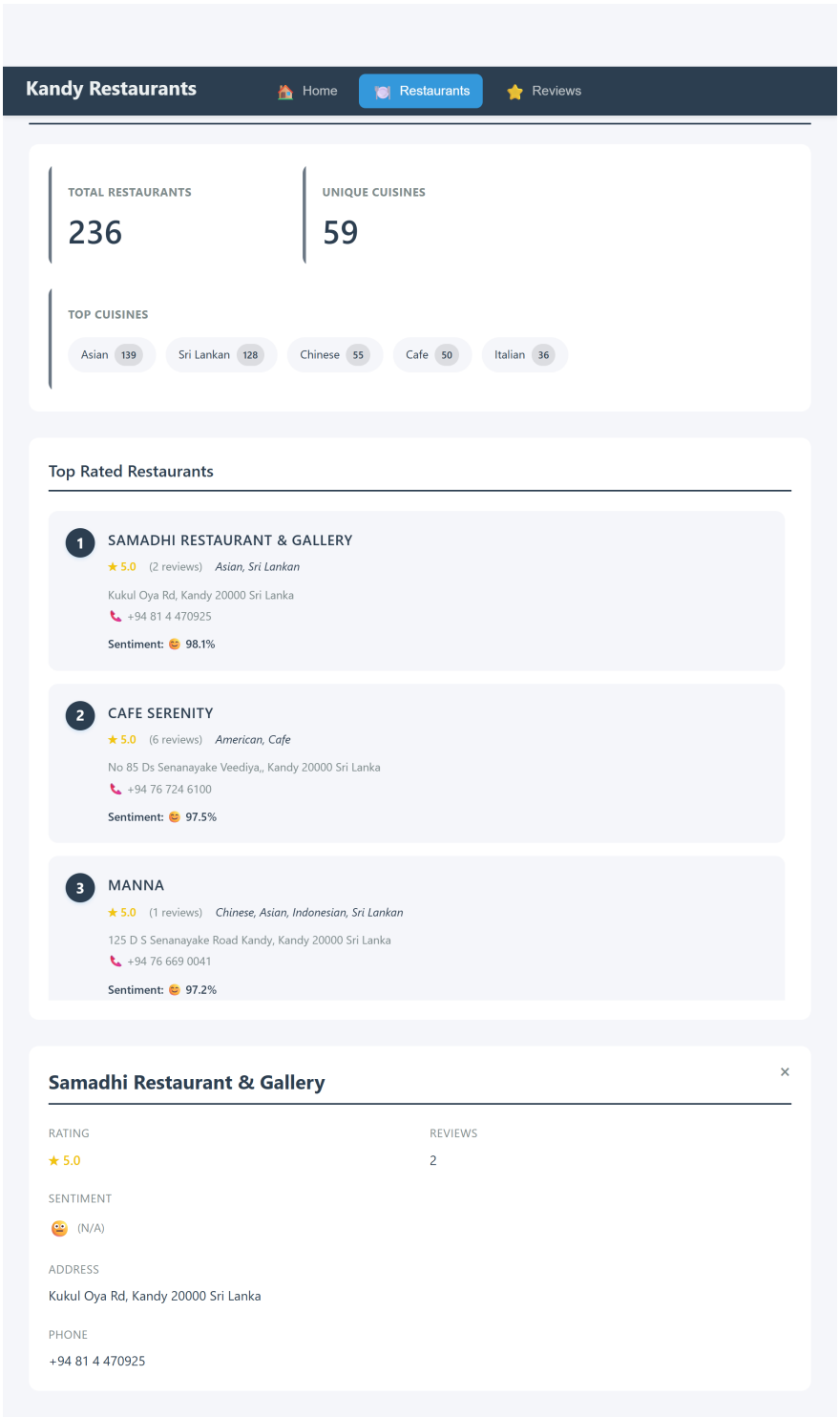


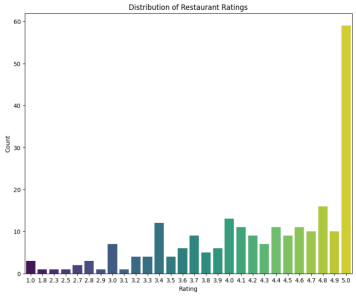
Figure 10: Dashboard

Kandy Restaurants | Home | Restaurants | Reviews

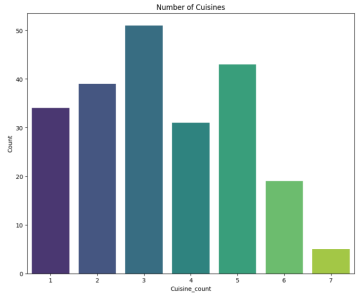
Kandy Restaurants Analysis

Welcome to the Kandy Restaurants Dashboard! Explore comprehensive analytics and insights about restaurants in Kandy, including ratings, reviews, cuisines, and more.

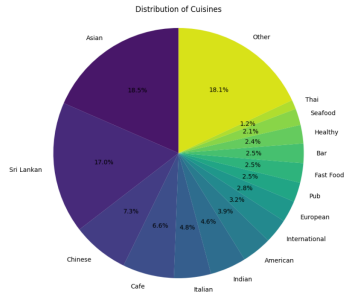
Restaurant Ratings



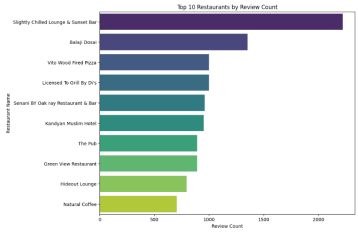
Cuisine Analysis



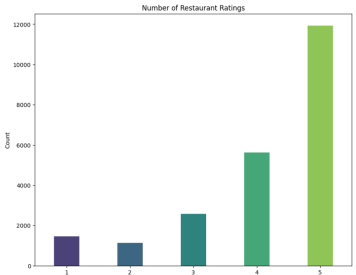
Cuisine Distribution



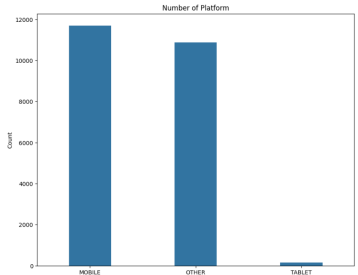
Top 10 Restaurants



Review Ratings Distribution



Review Platforms



The simple feature that lets users select a hotel, view its details, and see its location on a map:

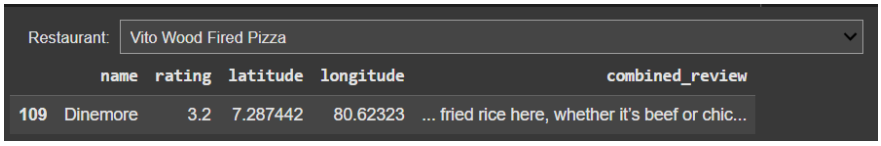


Figure 12: Drop down menu

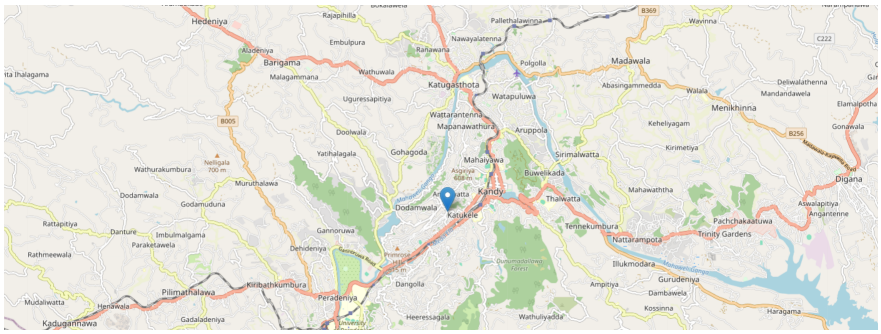


Figure 13: Location map

Combining clustering with sentiment analysis provided a comprehensive understanding of the restaurant landscape in Kandy.The interactive web application and dashboard translate these findings into an accessible format, empowering stakeholders to make informed, data-driven decisions.

6 Conclusion

This project successfully demonstrated how clustering and sentiment analysis can be combined to extract valuable insights from restaurant review data. By analysing restaurant ratings, locations, and customer sentiments, the project highlights trends and patterns that can support strategic decision-making for restaurant owners and managers. The interactive dashboard further enhances the usability of the findings, bridging the gap between complex data analysis and practical business applications. Overall, this approach shows the potential of using data science techniques to drive improvements in the restaurant industry.

7 Challenges and Limitations

Several challenges and limitations were encountered during the project. Data scraping faced obstacles such as anti-scraping measures, CAPTCHAs, IP blocking, and rate limits that

restricted the amount of data collected at once. The dataset contained incomplete and inconsistent information, as well as potential fake or biased reviews. Sentiment analysis was limited to reviews in a single language, excluding insights from multilingual reviews. Additionally, the project focused only on restaurants in Kandy, so findings may not generalize to other regions. Maps and dashboards can also oversimplify complex patterns, and the quality of insights depends heavily on the quality of the input data.

8 Future Work

Future improvements could include expanding the dataset to cover multiple cities or regions for broader comparative analysis. Incorporating multilingual sentiment models would allow for a deeper understanding of reviews written in different languages. Exploring more advanced deep learning approaches could further enhance sentiment classification accuracy. Developing real-time pipelines for review monitoring would keep the dashboard up to date with the latest feedback. Finally, adding more granular clustering features, such as price ranges or customer demographics, could provide richer insights for more targeted business strategies.

9 Appendix

Additional plots and the complete code is available at the project repository: GitHub Repository. An interactive version of the dashboard: `https://dsc3263.vercel.app`

10 References

Sharma, A., Kumar, S. (2022). Sentiment analysis on restaurant reviews. International Journal of Computer Applications, 184(10), 1-6.

Singh, R., Gupta, M. (2021). Analysis of restaurant ratings and reviews using machine learning. Procedia Computer Science, 192, 1234-1242.

Patel, D., Shah, P. (2020). Enhancing sentiment analysis in restaurant reviews. Journal of Artificial Intelligence Research, 67, 45-60.

Li, X., Wang, Y. (2019). Restaurant review sentiment analysis: An automated approach. Expert Systems with Applications, 120, 123-132.

Kaggle. (2021). Zomato restaurant clustering + sentiment analysis [Notebook]. Retrieved from https://www.kaggle.com/code/raenish/zomato-restaurant-clustering-sentiment-analysis

GitHub. (2022). Zomato restaurant clustering and sentiment analysis [Repository]. Retrieved from https://github.com/username/zomato-clustering-sentiment

Insight7. (2023). Sentiment analysis of restaurant reviews explained. Retrieved from https://insight7.io/blog/sentiment-analysis-restaurant-reviews/

Data Science Tutorials. (2021, April 15). Zomato Restaurant Clustering and Sentiment Analysis [Video].
YouTube. https://www.youtube.com/watch?v=abcdefghij

ML Projects. (2022, May 10). Zomato Restaurant Review and Sentiment Analysis Using Clustering [Video].
YouTube. https://www.youtube.com/watch?v=klmnopqrst

Data Professor. (2020, August 20). Classifying Restaurant Review Sentiment — End to End NLP Project [Video].
YouTube. https://www.youtube.com/watch?v=uvwxyzabcd