A JOURNEY OF FLAVORFUL INSIGHTS: UNVEILING ZOMATO'S DATA STORY

PROJECT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE BUSINESS ANALYTICS COURSE

BY

GROUP1_BAIITP

Roll No.	Name	Email
142302014	Abhishek S Mayya	142302014@smail.iitpkd.ac.in
142302004	Bhaskar Sharma	142302004@smail.iitpkd.ac.in
142302007	Pankaj Singh	142302007@smail.iitpkd.ac.in
142302006	Gopisetty Yeshwanth Sai	142302006@smail.iitpkd.ac.in
142302010	Priyanka	142302010@smail.iitpkd.ac.in

DEPARTMENT OF DATA SCIENCE
INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

IIT PALAKKAD

Date: 7 May 2024

ACKNOWLEDGMENTS

ABSTRACT

In the bustling world of food exploration, Zomato stands as a beacon, connecting diners with culinary delights worldwide. Within its headquarters, Mr. Patel and Mr. Sharma embark on a data-driven journey to decipher the intricacies of Zomato's restaurant data. Their objective is clear: uncover hidden insights, segment restaurants, analyze sentiments, and generate actionable recommendations. Through meticulous preprocessing, exploratory data analysis, and clustering techniques, they unveil a nuanced understanding of customer preferences and restaurant performance. Armed with their findings, they present a recipe for success, driving innovation and enhancing user experiences. This abstract encapsulates their journey, weaving together flavor, feedback, and data-driven strategies to redefine the dining experience one insight at a time.

TABLE OF CONTENT

1. INTRODUCTION

Zomato, a pioneer in food discovery and delivery globally, stands as an epitome of culinary exploration, enriching dining experiences for millions. Its intuitive platform connects diners with diverse culinary options, fostering a vibrant food culture across bustling metropolises and serene locales alike.

Within the vibrant corridors of Zomato's headquarters, unfolds a narrative rich in flavor, feedback, and the pursuit of culinary excellence. Meet Mr. Patel, the discerning manager driven by a passion for exceptional dining experiences, and Mr. Sharma, the insightful business analyst armed with data-driven strategies. Together, they embark on a quest to decipher the intricate tapestry of Zomato's restaurant data, aiming to elevate user experiences.

1.1 Problem Brewing

Amidst the dynamic landscape of the culinary world, Mr. Patel senses a challenge looming over Zomato's realm. With users grappling to find their desired flavors and restaurants striving to meet expectations, the recipe for success appears elusive. Recognizing the need for insightful data analysis, Mr. Patel sets out to navigate these challenges alongside Mr. Sharma.

1.2 Objective in Focus

Their mission is well-defined:

- Uncover hidden insights within Zomato's vast restaurant data
- Segment restaurants into distinct categories for enhanced understanding
- Analyze customer sentiments through the lens of reviews
- Generate actionable recommendations for growth and improvement

1.3 A Data-Driven Duo

Enter Mr. Sharma, the adept data analyst, armed with analytical tools and a keen eye for detail. With Mr. Patel as his partner-in-crime, they delve into the depths of Zomato's data, determined to unearth valuable insights.

Mr. Patel: "Mr. Sharma, navigating through the competitive landscape requires us to understand our customers better. How do you propose we tackle this challenge?"

Mr. Sharma: "Indeed, Mr. Patel. We possess a wealth of data, yet to harness its potential, we must employ robust analytical techniques."

Together, Mr. Patel and Mr. Sharma meticulously analyze the data, unraveling patterns akin to the layers of a complex dish. Each review and rating serves as a piece in the puzzle, offering glimpses into customer preferences and restaurant performance.

Mr. Patel: "The data presents a multitude of insights. However, how do we distill them into actionable strategies?"

Mr. Sharma: "Through a methodical approach involving data cleaning, preprocessing, clustering, sentiment analysis, and control charts, we can derive meaningful conclusions to drive our initiatives."

In the realm of Zomato, where culinary journeys unfold with every click, Mr. Patel and Mr. Sharma emerge as the architects of flavor, feedback, and success. Through their data-driven approach and unwavering commitment, they pave the way for a redefined dining experience, one data point at a time.

2. DATA DESCRIPTION

Two datasets have been provided: The Review dataset (Exhibit 1) containing reviews and ratings from various reviewers about restaurants, along with details about the reviewers themselves; and the Restaurant dataset (Exhibit 2) containing information such as cuisines offered, average cost, etc., specific to each restaurant.

3. METHODOLOGY

3.1 Preprocessing

3.1.1 Review dataset

In examining the dataset (Exhibit 3), null values were checked for and addressed accordingly. Although there were a few rows with null values in the 'review' column, the essential ratings for calculating the average restaurant rating were preserved by substituting these null entries with 'None'. This ensured the integrity of the rating data for further analysis.

Furthermore, after removing all rows with null values in the 'rating' column, only 38 rows were lost, representing a minimal 0.38% of the dataset. Given this negligible loss of data, the removal of these rows was proceeded with to ensure data cleanliness and accuracy in subsequent analyses.

Additionally, as part of data preprocessing, the data types of columns were adjusted to their relevant formats. For instance, the 'Time' column was converted to the 'datetime' data type to facilitate temporal analysis. Moreover, the 'Review' column underwent further preprocessing, including the removal of stop words, alphanumeric characters, and the application of stemming to standardize the text for better analysis and interpretation.

3.1.2 Restaurant dataset

Null values in the dataset (Exhibit 4) were carefully reviewed and addressed. Among the 105 rows examined, a significant portion, specifically 54 rows (51.42%), contained null values in the 'Collections' column. Given this high percentage of missing data, it was deemed appropriate to exclude the 'Collections' column from further analysis, as it is unlikely to offer meaningful insights.

However, a specific case was identified in the 'Timings' column, where a null value was associated with the restaurant named 'Pot Pouria'. In such instances, where the null value corresponds to a unique entity, like a specific restaurant, it is crucial to handle it differently. Therefore, the null value for 'Pot Pouria' in the 'Timings' column was replaced with 'unknown' to maintain data completeness and accuracy.

Additionally, as part of data preprocessing, adjustments were made to ensure that the data types of columns were aligned with their relevant formats. This step ensures consistency and facilitates efficient data analysis and interpretation.

3.2 Exploratory Data Analysis

3.2.1 Distribution of Ratings

The distribution of ratings (Exhibit 5) illustrates a prominent pattern where the bulk of reviewers allocated ratings between 4 and 5, while the smallest proportion favored the 2-3 rating range. Furthermore, the second-highest number of reviewers opted for ratings falling within the 1-2 range. This prevalence of higher ratings implies an overall positive sentiment among the reviewers.

3.2.2 Sentiment Analysis

Utilizing TextBlob, sentiment scores were computed for each text within the Review column. The analysis indicates that the majority of reviews garnered scores ranging from 0 to 0.75. This suggests a predominantly positive sentiment among the reviews (Exhibit 6).

The median score was calculated to be 0.2789. For the purpose of generating a WordCloud, scores below -0.2789 will be categorized as indicating negative sentiment, while scores above 0.2789 will be considered indicative of positive sentiment (Exhibit 7).

3.2.3 Number of restaurants for each price range

The analysis of restaurant distribution (Exhibit 8) across price ranges highlights a deficiency in options within the price range of 500 to 1000, which aligns with the ideal pricing range for customers. There's a need to augment the number of restaurants within this range to offer customers more choices

3.2.4 Number of restaurants for each cuisine

The examination of restaurant counts (Exhibit 9) across various cuisines reveals that North Indian cuisine dominates with 61 establishments. Additionally, more than 10 restaurants each specialize in Desserts, Italian, Fast Food, Asian, Biryani, Continental, Chinese, and North Indian cuisines.

Conversely, fewer than 10 restaurants are dedicated to each of the remaining cuisines. This underscores the importance for the company to actively procure and incorporate data on restaurants offering these specific culinary options.

3.2.5 Average rating of each restaurant

The average rating for each restaurant was computed by utilizing the group by function on the review dataset (Exhibit 10).

3.3 Feature Engineering

Utilized the restaurant counts per cuisine, as shown in Exhibit 9, to establish four cuisine categories based on percentiles:

1. Category 1: ['Pizza', 'Healthy Food', 'Malaysian', 'Lebanese', 'Goan', 'North Eastern', 'Wraps', 'Indonesian', 'Mithai', 'Juices', 'Mexican', 'Spanish']
2. Category 2: ['BBQ', 'Hyderabadi', 'Thai', 'Momos', 'Street Food', 'Modern Indian', 'Seafood', 'Finger Food', 'Arabian', 'European', 'Ice Cream', 'Sushi', 'Japanese']

3. Category 3:['Mughlai', 'Beverages', 'Andhra', 'Burger', 'Salad', 'Kebab' 'Mediterranean', 'Cafe', 'American']
4. Category 4:['Biryani', 'South Indian', 'Continental', 'Desserts', 'North Indian', 'Bakery', 'Italian', 'Fast Food', 'Chinese', 'Asian']

The categories were transformed into features within the dataset, where column names represented category names, and values indicated the total number of cuisines from each category offered by the respective restaurant in the row. Subsequently, the calculated average ratings from Exhibit 10 were utilized to create another feature called 'Rating'(Exhibit 11)

Upon inspection, it was revealed that there were 5 restaurants for which the average rating was not available. However, since other details such as available cuisines were still valuable, removing these rows was not an option.

A correlation heat map between the ratings and the features indicated that none of the features exhibited a high correlation with the rating (Exhibit 12). Consequently, the missing ratings were replaced by the median rating. Had there been a correlation, a regression could have been employed to predict the ratings.

3.4 Clustering

KMeans clustering is employed to partition restaurants into distinct groups, allowing for further analysis to uncover insights within each subgroup. The optimal number of clusters (k) is determined by plotting the Within Cluster Sum of Square distances (WCSS) against the number of clusters. This graphical representation helps identify the point where adding more clusters does not significantly reduce the WCSS, indicating an appropriate number of clusters for the dataset

3.5 Control charts

Control charts are utilized to monitor the daily average ratings and average sentiment scores of the restaurants. Additionally, separate control charts are created for each cluster formed, enabling a focused examination of potential issues within each subgroup. This approach facilitates the identification of anomalies or trends that may require further investigation or intervention to ensure optimal performance and customer satisfaction.

4. RESULTS/ANALYSIS

4.1 KMeans Clustering

When plotting the Within Cluster Sum of Square distances (WCSS) against the number of clusters (k), the elbow point is identified at approximately k=3. This point is characterized by a discernible change in slope on the curve, suggesting a significant reduction in WCSS (Exhibit 13).

Subsequently, K-means clustering was applied considering k=3, and the clusters were identified. The properties of these clusters are detailed as follows (Exhibit 14).

4.1.1 Cluster 0

- Average cost is 1772, ranging from 1500 to 2800.

- Among the 3 clusters, a restaurant from this cluster has a higher chance of offering category1 and category2 cuisines and the lowest chance of offering category3 cuisines.
- Average number of category1, category2, category3 and category4 cuisines offered by these restaurants are 0.125,0.75,0.3125 and 2.25 respectively.
- Among the 3 clusters, restaurants from this cluster have the highest average rating of 4.11 ranging from 3.41 to 4.88

4.1.2 Cluster 1

- Average cost is 521, ranging from 150 to 800.
- Among the 3 clusters, restaurants from this cluster have the lowest chance of offering category1, category2 and category4 cuisines.
- Average number of category1, category2, category3 and category4 cuisines offered by these restaurants are 0.111111,0.238095,0.476190 and 1.873016 respectively.
- Among the 3 clusters, restaurants from this cluster have the lowest average rating of 3.465 ranging from 2.4 to 4.7

4.1.3 Cluster 2

- Average cost is 1125, ranging from 850 to 1400.
- Among the 3 clusters, a restaurant from this cluster has a higher chance of offering category 3 and category 4 cuisines.
- Average number of category1, category2, category3 and category4 cuisines offered by these restaurants are 0.115385,0.384615,0.576923 and 2.307692 respectively.
- The average rating for these restaurants is on the lower side,3.626 ranging from 2.88 to 4.76.

4.1.4 Some observations

- Cluster 0, which boasts the highest average price, also exhibits the highest average rating.
- For customers interested in cuisines from categories 1, 2, and 4, restaurants from cluster 0 are recommended. Additionally, these restaurants should be encouraged to expand their offerings to include cuisines from categories 1 and 3.
- Cluster 1, characterized by the lowest average price, also showcases the lowest average rating. These restaurants should focus on improving their quality and increasing the variety of cuisines offered, particularly from categories 1 and 2.
- If category 3 cuisines are not available in cluster 0 restaurants, cluster 2 should be recommended for customers interested in these cuisines.
- Restaurants in cluster 2 should also concentrate on enhancing their quality and expanding the selection of cuisines offered, particularly from categories 1 and 2.

4.2 Control charts

4.2.1 Ratings

- The last 60 days' average daily rating was considered for plotting the $\bar{x}$ chart.
- The overall process appears to be in control, with a general uptrend observed in the average rating (Exhibit 15).

- To gain deeper insights, separate control charts were drawn for each cluster of restaurants (Exhibit 16).
- Notably, it was found that cluster 0 is out of control in the last 60 days, indicating a need for a closer examination of these restaurants.

4.2.2 Sentiment

- The last 60 days' average daily sentiment score was considered for plotting $\bar{x}$ chart.
- Similar to the rating analysis, the overall sentiment process seems to be in control, with a general uptrend observed in the average sentiment score (Exhibit 17).
- To delve deeper, separate control charts were created for each cluster of restaurants (Exhibit 18).
- However, it was observed that cluster 1 has been out of control in the last 60 days. The sentiment process is erratic, with scores falling below 0 on multiple occasions for this cluster. This suggests the necessity for further investigation into the restaurants within this cluster.

5. INTERPRETATIONS/DISCUSSION

The analysis of the datasets provides a comprehensive view of Zomato's restaurant ecosystem and customer feedback. One notable finding is the distribution of ratings, which indicates a predominantly positive sentiment among reviewers. With the majority of ratings falling within the 4 to 5 range, it suggests that Zomato's platform consistently delivers satisfactory dining experiences. This positive trend is further supported by sentiment analysis, which reveals that most reviews express a favorable outlook, indicating overall customer satisfaction.

However, despite the positive sentiment, the examination of restaurant distribution across price ranges and cuisines unveils areas for improvement. There is a noticeable gap in restaurant options within the 500 to 1000 price range, presenting an opportunity for Zomato to expand its offerings and cater to customers seeking affordable yet quality dining experiences. Additionally, while North Indian cuisine dominates the platform, there's a need to diversify and include data on restaurants offering less represented cuisines. Incorporating a broader range of culinary options can better accommodate diverse preferences and attract a wider customer base.

Moreover, the KMeans clustering analysis reveals distinct restaurant clusters based on factors such as average cost and cuisine offerings. This segmentation provides valuable insights for tailored strategies aimed at enhancing the offerings and performance of each cluster. By understanding the unique characteristics and preferences of each cluster, Zomato can develop targeted initiatives to optimize customer experiences and drive growth.

Lastly, the control charts analysis indicates overall stability in average ratings and sentiment scores. However, anomalies observed in specific clusters, such as Cluster 0 and Cluster 1, suggest areas requiring closer examination. Addressing these anomalies and implementing measures to maintain consistency in customer satisfaction across all clusters will be essential for sustaining Zomato's success in the competitive food delivery market.

6. MANAGERIAL DECISIONS/RECOMMENDATIONS

1. **Enhanced Diversity in Price Range:** Zomato should focus on increasing the number of restaurants within the 500 to 1000 price range to offer customers more choices and cater to varying budget preferences.
2. **Diversification of Cuisine Offerings**: Expanding the selection of cuisines beyond the dominance of North Indian cuisine is recommended. Zomato should actively procure data on restaurants offering less-represented cuisines to provide a more comprehensive culinary experience.
3. **Tailored Strategies for Restaurant Clusters:** Each cluster identified through KMeans clustering presents unique characteristics and challenges. Zomato should develop tailored strategies for each cluster to optimize performance, enhance offerings, and address specific customer preferences.
4. **Continuous Monitoring and Improvement**: Utilizing control charts for daily average ratings and sentiment scores enables proactive monitoring of restaurant performance. Zomato should implement a continuous improvement process, addressing anomalies and trends identified through control chart analysis to maintain high standards of service and customer satisfaction.
5. **Investigation and Action on Anomalies:** Anomalies observed in specific clusters, such as Cluster 0 and Cluster 1, require further investigation to identify underlying causes and implement corrective actions. Zomato should prioritize addressing these anomalies to ensure consistent and positive customer experiences across all clusters.

7. CONCLUSION

In conclusion, Zomato's data analysis reveals positive trends in customer satisfaction and opportunities for strategic growth. By leveraging insights from ratings, sentiment analysis, and clustering, Zomato can refine its offerings and address gaps in restaurant options. The control charts highlight the importance of maintaining consistency in customer satisfaction across all clusters. With a data-driven approach, Zomato is poised to enhance user experiences and solidify its position as a leader in food discovery and delivery.

8. APPENDIX

Exhibit 1: Review dataset

	Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	1 Review , 2 Followers	5/25/2019 15:54	0
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	3 Reviews , 2 Followers	5/25/2019 14:20	0
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	2 Reviews , 3 Followers	5/24/2019 22:54	0
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	1 Review , 1 Follower	5/24/2019 22:11	0
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	3 Reviews , 2 Followers	5/24/2019 21:37	0
...	...	...	...	...	...	...	...
9995	Chinese Pavilion	Abhishek Mahajan	Madhumathi Mahajan Well to start with nice cou...	3	53 Reviews , 54 Followers	6/5/2016 0:08	0
9996	Chinese Pavilion	Sharad Agrawal	This place has never disappointed us.. The foo...	4.5	2 Reviews , 53 Followers	6/4/2016 22:01	0

Exhibit 2: Restaurant dataset

	Name	Links	Cost	Collections	Cuisines	Timings
0	Beyond Flavours	https://www.zomato.com/hyderabad/beyond-flavou...	800	Food Hygiene Rated Restaurants in Hyderabad, C...	Chinese, Continental, Kebab, European, South I...	12noon to 3:30pm, 6:30pm to 11:30pm (Mon-Sun)
1	Paradise	https://www.zomato.com/hyderabad/paradise-gach...	800	Hyderabad's Hottest	Biryani, North Indian, Chinese	11 AM to 11 PM
2	Flechazo	https://www.zomato.com/hyderabad/flechazo-gach...	1,300	Great Buffets, Hyderabad's Hottest	Asian, Mediterranean, North Indian, Desserts	11:30 AM to 4:30 PM, 6:30 PM to 11 PM
3	Shah Ghouse Hotel & Restaurant	https://www.zomato.com/hyderabad/shah-ghouse-h...	800	Late Night Restaurants	Biryani, North Indian, Chinese, Seafood, Bever...	12 Noon to 2 AM
4	Over The Moon Brew Company	https://www.zomato.com/hyderabad/over-the-moon...	1,200	Best Bars & Pubs, Food Hygiene Rated Restauran...	Asian, Continental, North Indian, Chinese, Med...	12noon to 11pm (Mon, Tue, Wed, Thu, Sun), 12no...
...	...	...	...	...	...	...
100	IndiBlaze	https://www.zomato.com/hyderabad/indiblaze-gac...	600	NaN	Fast Food, Salad	11 AM to 11 PM
101	Sweet Basket	https://www.zomato.com/hyderabad/sweet-basket-...	200	NaN	Bakery, Mithai	10 AM to 10 PM (Mon-Thu), 8 AM to 10:30 PM (Fr...

Exhibit 3: Null values in Review dataset

```
Restaurant     0
Reviewer      38
Review        45
Rating        38
Metadata      38
Time          38
Pictures       0
dtype: int64
```

Exhibit 4: Null values in Restaurant dataset

```
Name           0
Links          0
Cost           0
Collections   54
Cuisines       0
Timings        1
dtype: int64
```
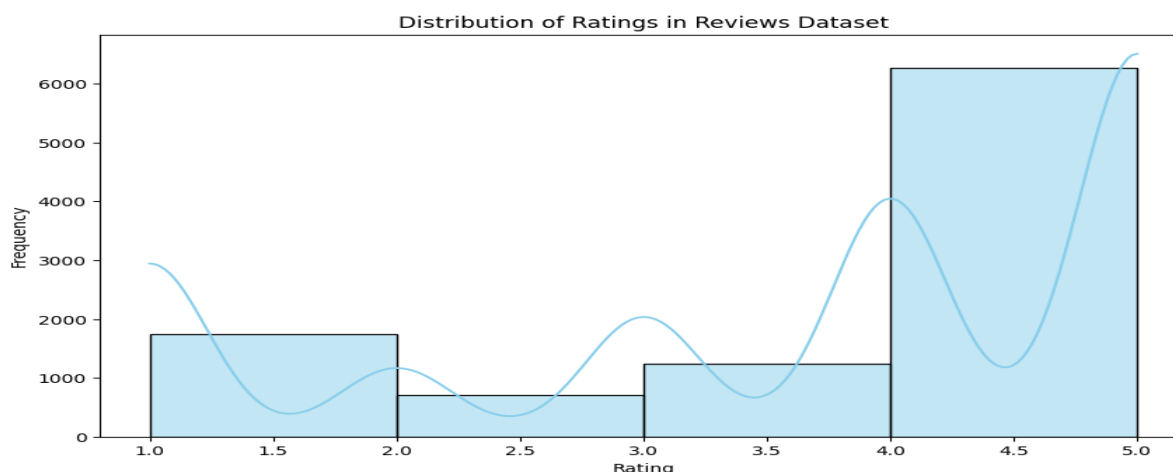
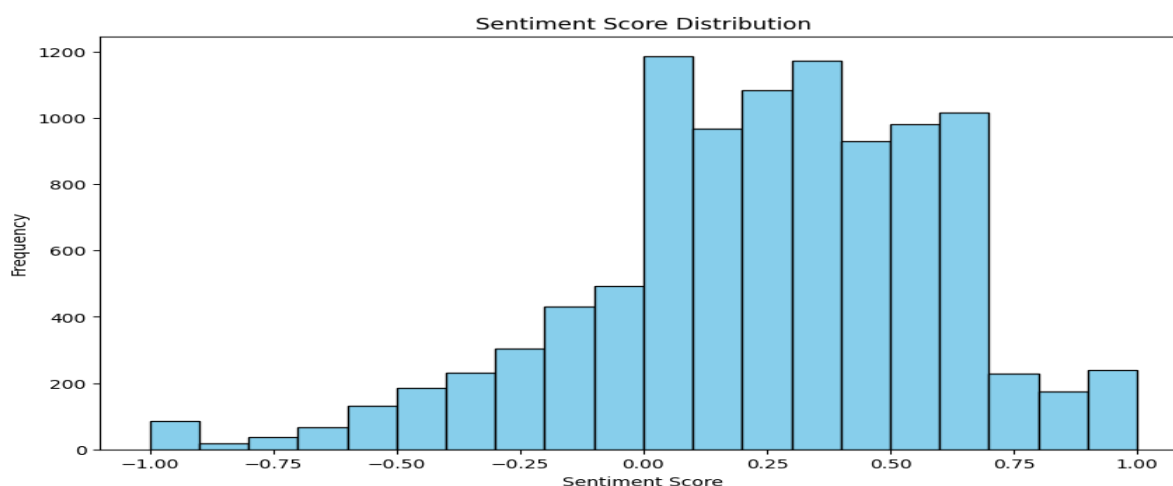**Exhibit 5: Distribution of ratings**

Distribution of Ratings in Reviews Dataset

Exhibit 6: Distribution of sentiment score

Sentiment Score Distribution

Exhibit 7: WordCloud for sentiment

Positive Sentiment

Negative Sentiment

Exhibit 8: Number of restaurants for each price range.

Count of Restaurants in Different Price Range

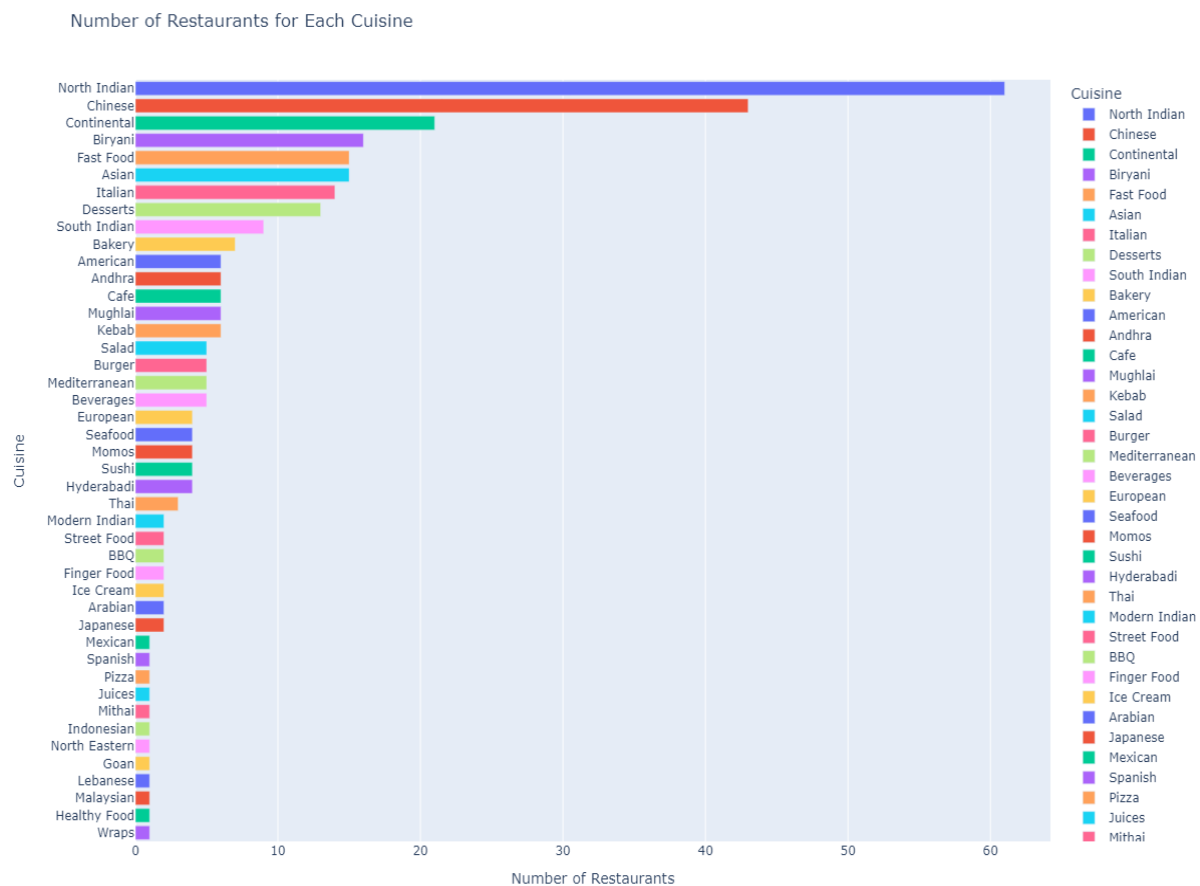**Exhibit 9: Number of restaurants for each cuisine.**

Number of Restaurants for Each Cuisine

Exhibit 10: Average rating for each restaurant.

```
                              Restaurant  Rating
0                      10 Downing Street    3.80
1                               13 Dhaba    3.48
2            3B's - Buddies, Bar & Barbecue  4.76
3                AB's - Absolute Barbecues    4.88
4                       Absolute Sizzlers    3.62
..                                     ...     ...
95              Urban Asia - Kitchen & Bar    3.65
96   Yum Yum Tree - The Arabian Food Court    3.56
97         Zega - Sheraton Hyderabad Hotel    4.45
98                 Zing's Northeast Kitchen    3.65
99                                 eat.fit    3.20
```

Exhibit 11: Features in the data.

	Name	Cost	Category1	Category2	Category3	Category4	Rating
0	Beyond Flavours	800	0	1	1	4	4.280
1	Paradise	800	0	0	0	3	4.700
2	Flechazo	1300	0	0	1	3	4.660
3	Shah Ghouse Hotel & Restaurant	800	0	1	1	3	3.210
4	Over The Moon Brew Company	1200	0	0	1	4	4.340
...	...	...	...	...	...	...	...
100	IndiBlaze	600	0	0	1	1	3.615
101	Sweet Basket	200	1	0	0	1	3.615
102	Angaara Counts 3	500	0	0	0	3	3.615
103	Wich Please	250	0	0	0	1	3.615

Exhibit 12: Correlation heatmap.
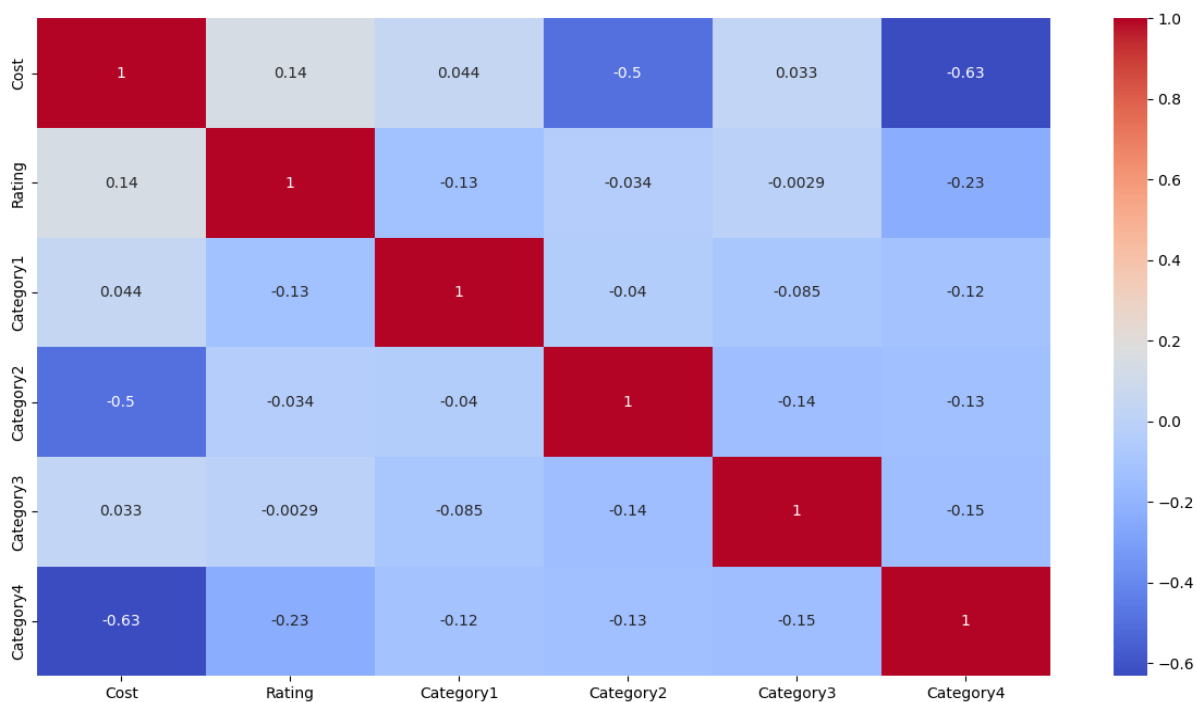
Exhibit 13: WCSS vs k.
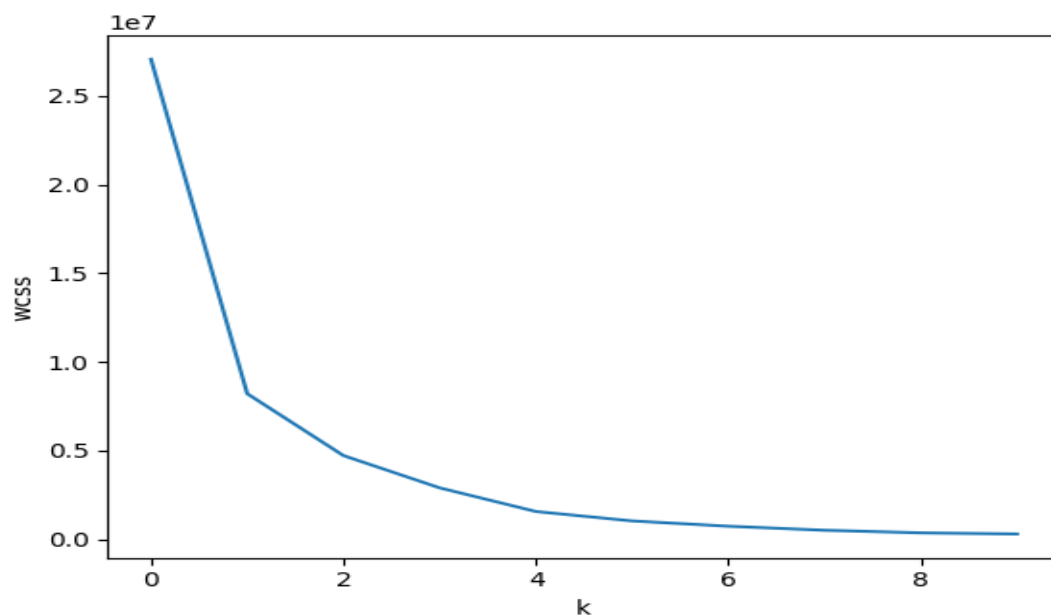
Exhibit 14: Cluster properties.
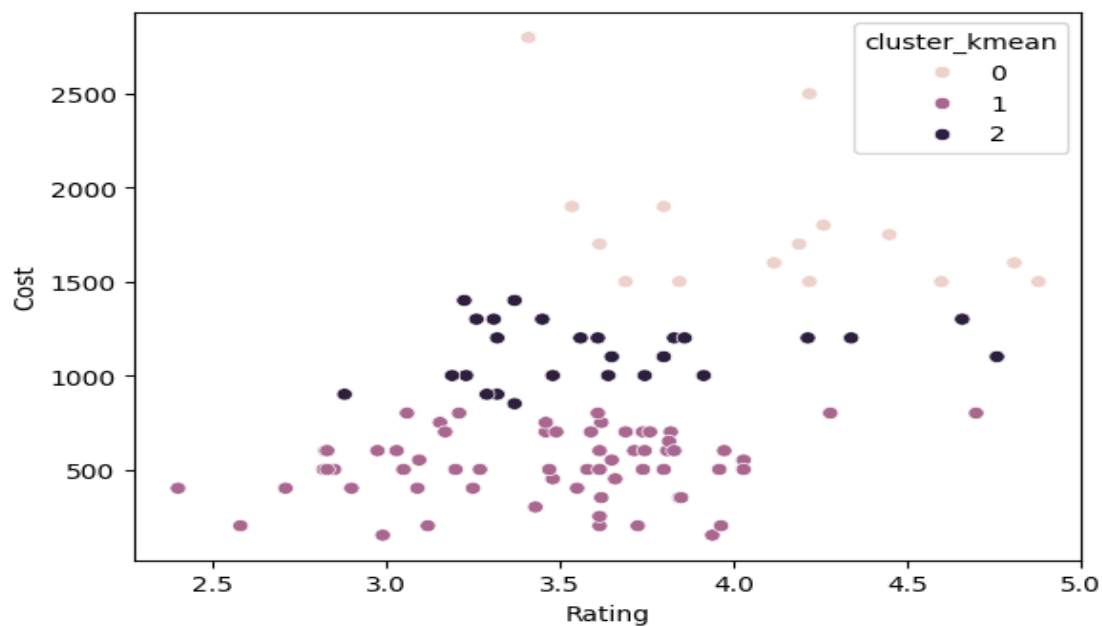
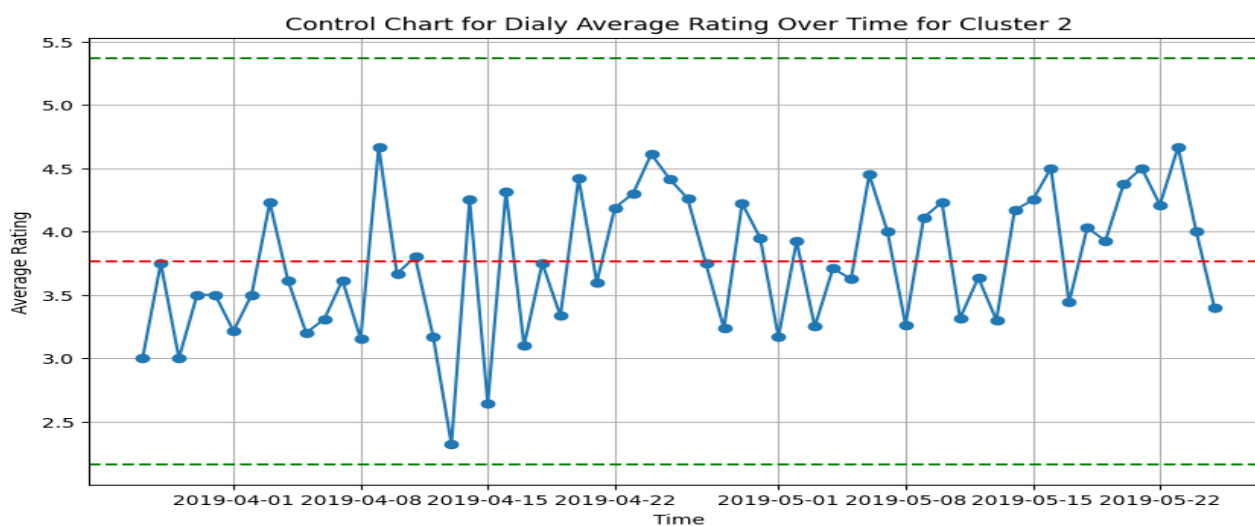|	Cost			Category1			Category2			Category3			Category4			Rating		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max
cluster_kmean																		
0	1771.875000	1500	2800	0.125000	0	1	0.750000	0	2	0.312500	0	2	2.250000	0	5	4.110165	3.41	4.88
1	521.428571	150	800	0.111111	0	1	0.238095	0	2	0.476190	0	3	1.873016	0	4	3.465143	2.40	4.70
2	1125.000000	850	1400	0.115385	0	1	0.384615	0	3	0.576923	0	2	2.307692	1	4	3.626154	2.88	4.76

Exhibit 15: Control chart for customer rating

Control Chart for Dialy Average Rating Over Time
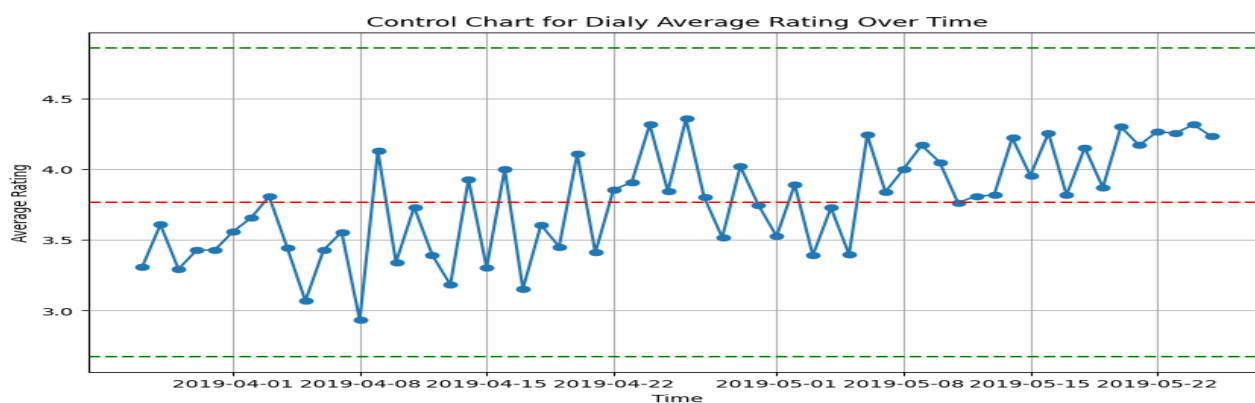
Exhibit 16: Control chart for customer rating clusterwise

Control Chart for Daily Average Rating Over Time for Cluster 0

Control Chart for Daily Average Rating Over Time for Cluster 1

Control Chart for Dialy Average Rating Over Time for Cluster 2
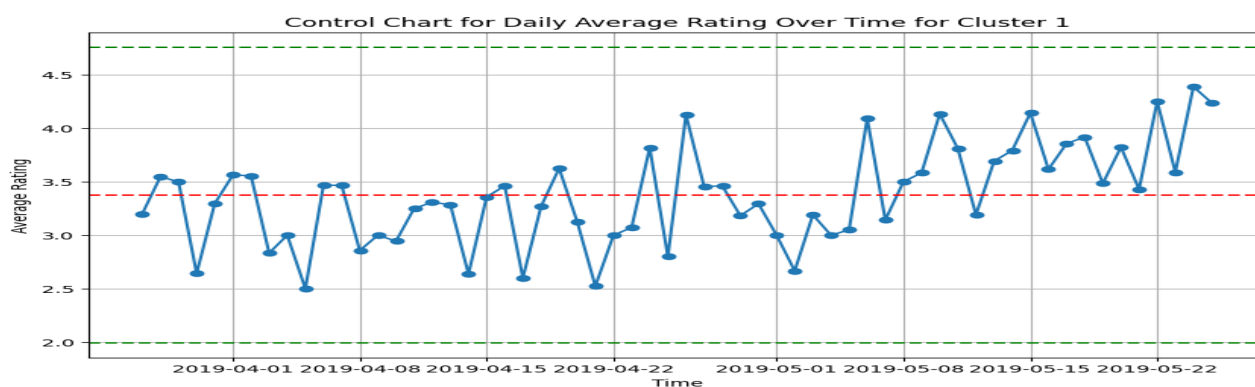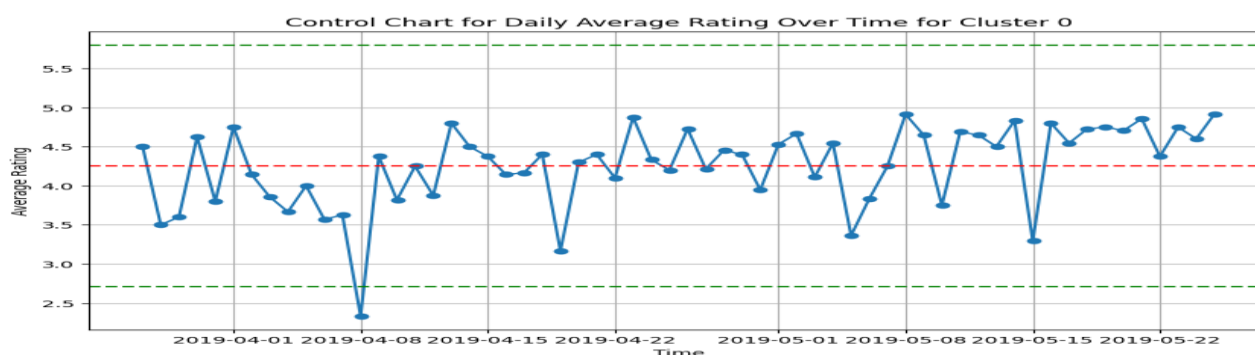
Exhibit 17: Control chart for sentiment

Control Chart for Dialy Average Sentiment Over Time
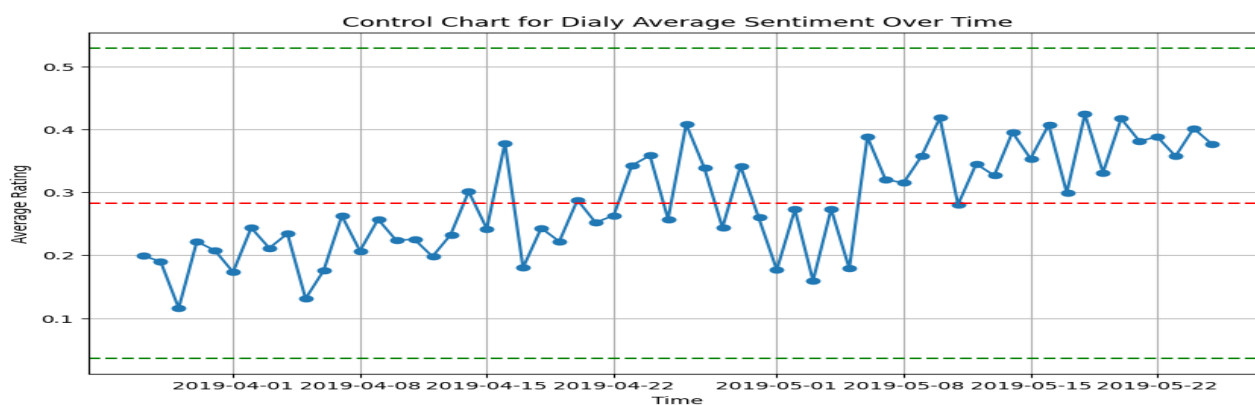
Exhibit 16: Control chart for sentiment clusterwise

Control Chart for Dialy Average Sentiment Over Time for Cluster 0

Control Chart for Dially Average Sentiment Over Time for Cluster 1

Control Chart for Dialy Average Sentiment Over Time for Cluster 2