Zomato Restaurant Clustering and Sentiment Analysis

Kanike Lakshmi Narayana

Data science Trainee at

AlmaBetter

Abstract

India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. The restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. The growing number of restaurants in every state of India has been a motivation to inspect the data to get some insights, interesting facts and figures about the Indian food industry in each city. So, this project focuses on analysing the Zomato restaurant data for each city in India.

Problem Statement

The Project focuses on Customers and Company, you have to analyse the sentiments of the reviews given by the customer in the data and make some useful conclusions in the form of Visualizations. Also, cluster the zomato restaurants into different segments. The data is visualized as it becomes easy to analyse data in an instant. The Analysis also solves some of the business cases that can directly help the customers find the restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in.

Introduction

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepinder Goyal and Pankaj Chaddah in 2008. Zomato provides information, menus and user reviews of restaurants, and also has food delivery options from partner restaurants in select cities.

India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. The restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. The growing number of restaurants in every state of India has been a motivation to inspect the data to get some insights, interesting facts and figures, about the Indian food industry in each city. So, this project focuses on analysing the Zomato restaurant data for each city in India.

Machine Learning is concerned with computer programs that automatically improve their performance through experience. In machine learning, we have supervised learning, unsupervised learning and reinforcement learning. Again

the supervised learning is further divided into regression and classification. In this project, we are going to look after unsupervised learning.

In unsupervised machine learning, we are not provided with any pre-assigned labels or scores for training the data.

Objective

The main objective is to build a model that can cluster similar restaurants and recommend those to customers.

Dataset Peeping

Owing to the size of the dataset, extensive cleaning of the dataset is not needed. The following steps are performed for the analysis purpose:

- The given dataset has 2 separate data named zomato restaurant names and metadata and zomato restaurant reviews.
- Datetime containing column is converted and extracted some new features from the same.
- The second dataset's columns are having NaN values and which are dropped for analysis.

Data Design

Zomato Restaurant names and Metadata.

- Name: Name of Restaurants
- Links: URL Links of Restaurants
- Cost: Per person estimated Cost of dining
- Collection: Tagging of Restaurants w.r.t. Zomato categories
- Cuisines: Cuisines served by Restaurants
- Timings: Restaurant Timings

Zomato Restaurant reviews.

- Restaurant: Name of the Restaurant
- Reviewer: Name of the Reviewer
- Review: Review Text
- Rating: Rating Provided by Reviewer
- MetaData: Reviewer Metadata - No. of Reviews and followers
- Time: Date and Time of Review
- Pictures: No. of pictures posted with the review

Challenges Faced

The following are the challenges faced in the data analysis:

- Splitting words in features and finding frequency.
- Making insights from the DateTime column.
- Cleaning the NaN values in the dataset.

Approach

As the problem statement says the main objective is to build a model that can show similar restaurants for the customer and the given dataset does not have any labelled data. I have chosen to build a recommendation system from unsupervised machine learning.

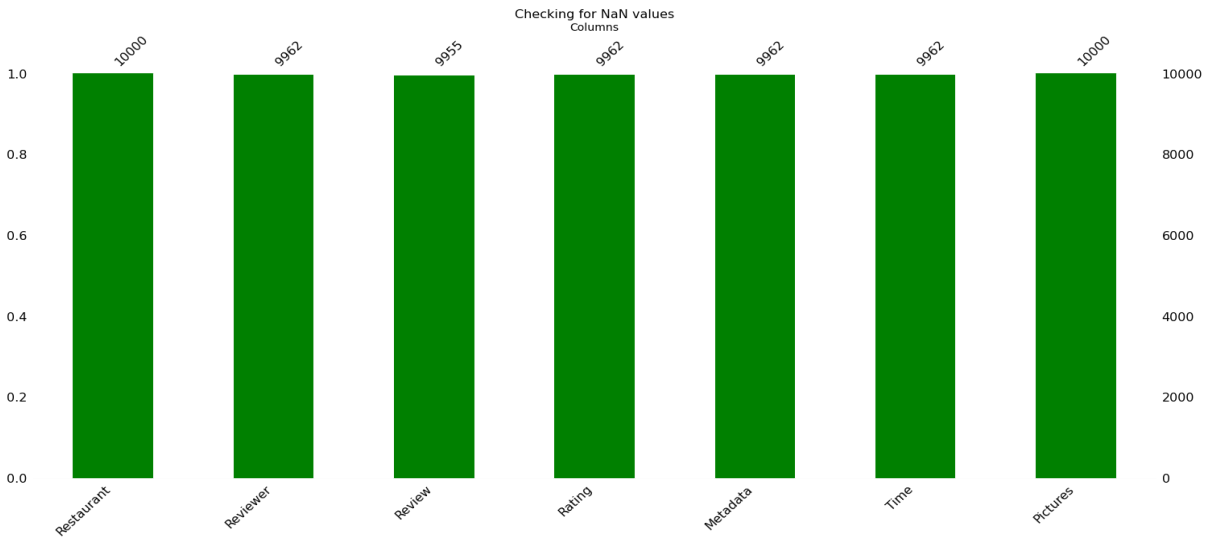
Tools Used

The whole project was done using python, in google collaboratory. Following libraries were used for analysing the data and visualizing:

- Pandas: Extensively used to load and wrangle with the dataset.
- Matplotlib: Used for visualization.
- Seaborn: Used for visualization.
- Warnings: For filtering and ignoring warnings.
- Numpy: For some math operations in predictions.
- Sklearn: For analysis and prediction.

Zomato Restaurant names and Metadata

Visualizing the presence of NaN values



The above bar plot shows that Zomato Restaurant names and Metadata dataset have some null values in the following features: Reviewer, Review, Rating, Metadata and Time.

Pandas DataFrame

Restaurent reviews dataframe:

	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

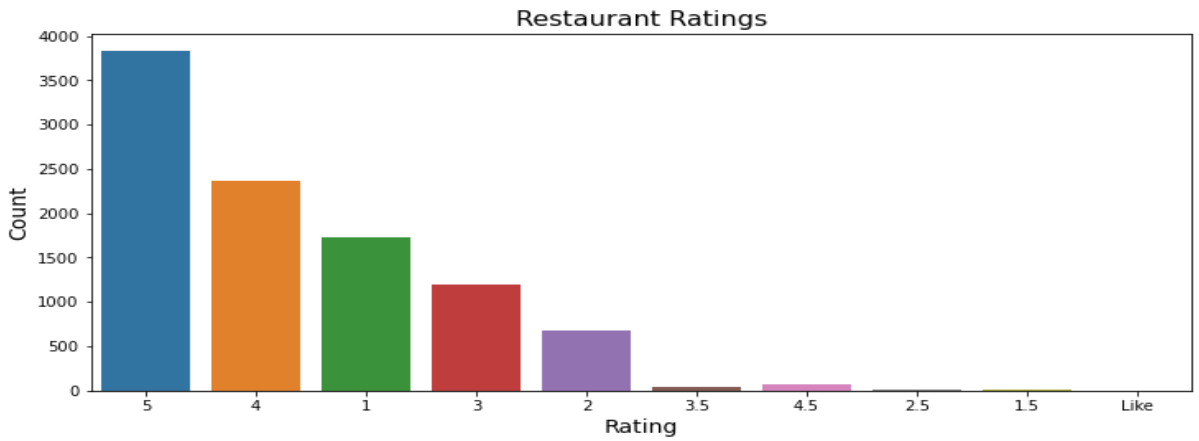
The table shows the Zomato Restaurant names and Metadata dataset in the form of Pandas DataFrame. The dataset has 10000 rows and 7 columns wholly the shape is (10000, 7).

It contains the following columns:

- Restaurant
- Reviewer
- Review
- Rating
- Metadata
- Time
- Pictures

All the null values are dropped and the dataset is cleaned.

Restaurant Rating

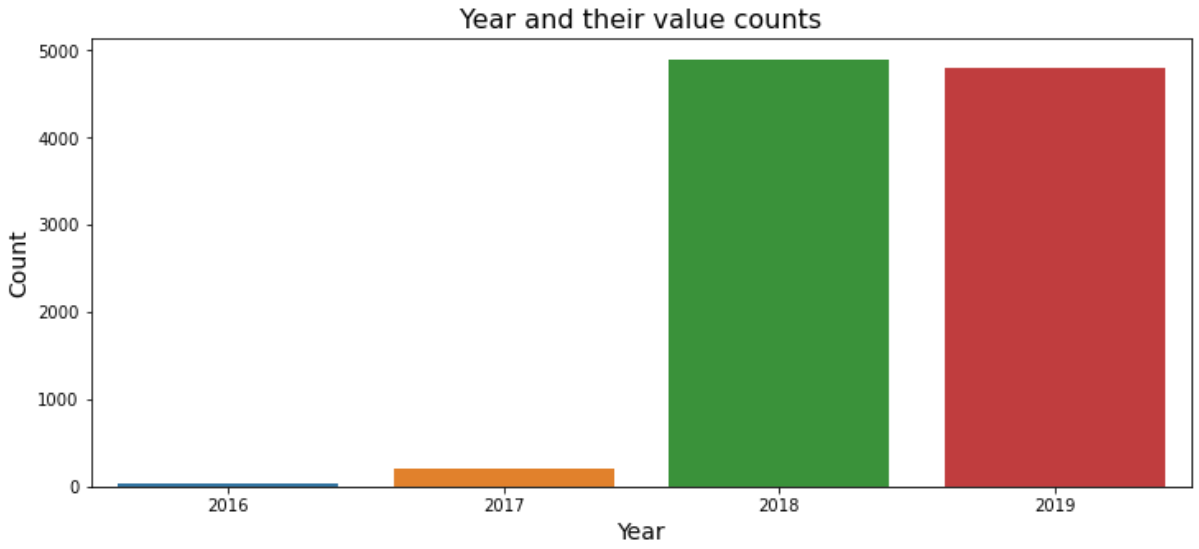


The above figure shows the value counts of records in the rating feature. Like having only one count, it is dropped and converted into a float data type.

Analysis of Metadata

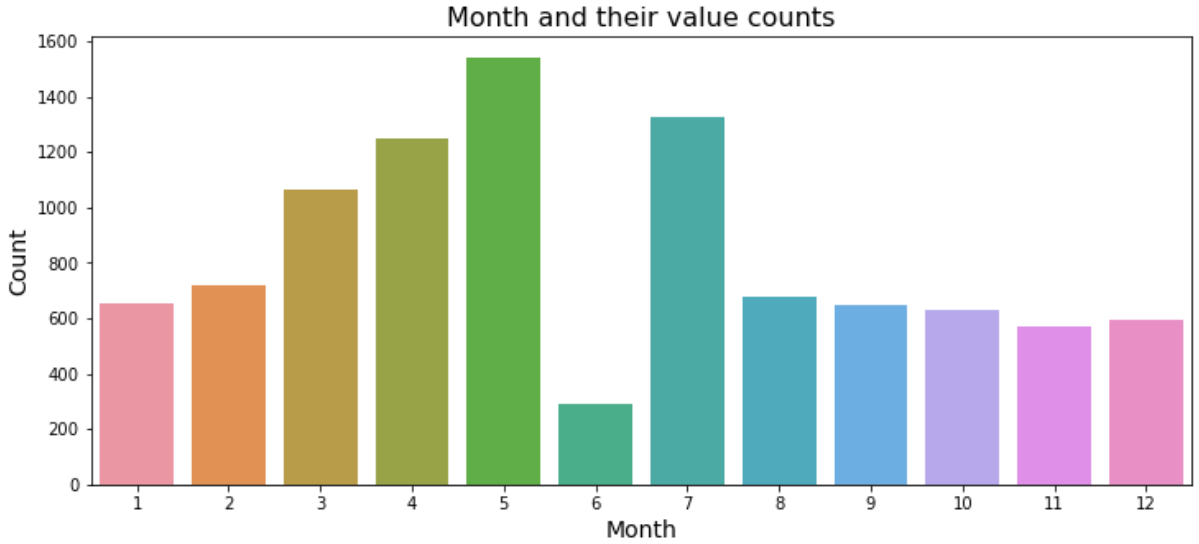
At first, the datatype of the metadata feature was transformed into the string, then it was split with (,) and made into two different columns named Reviewers and Followers. At last, the metadata column was dropped.

Years vs. value counts



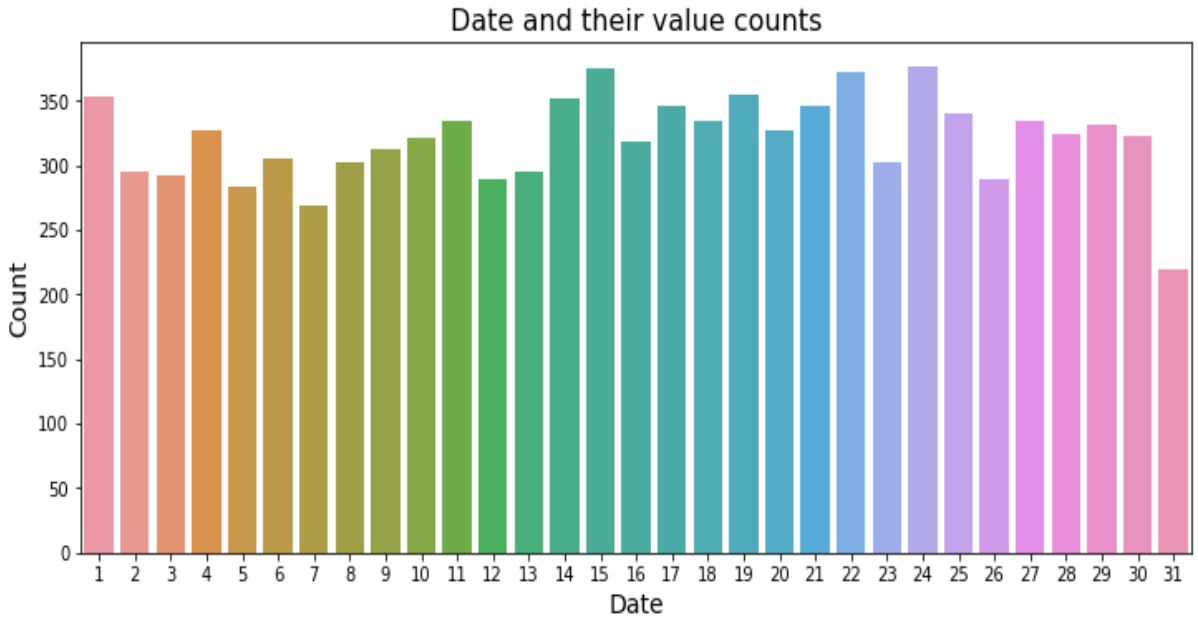
Among all the years we have more orders in the year 2018.

Month vs. value counts

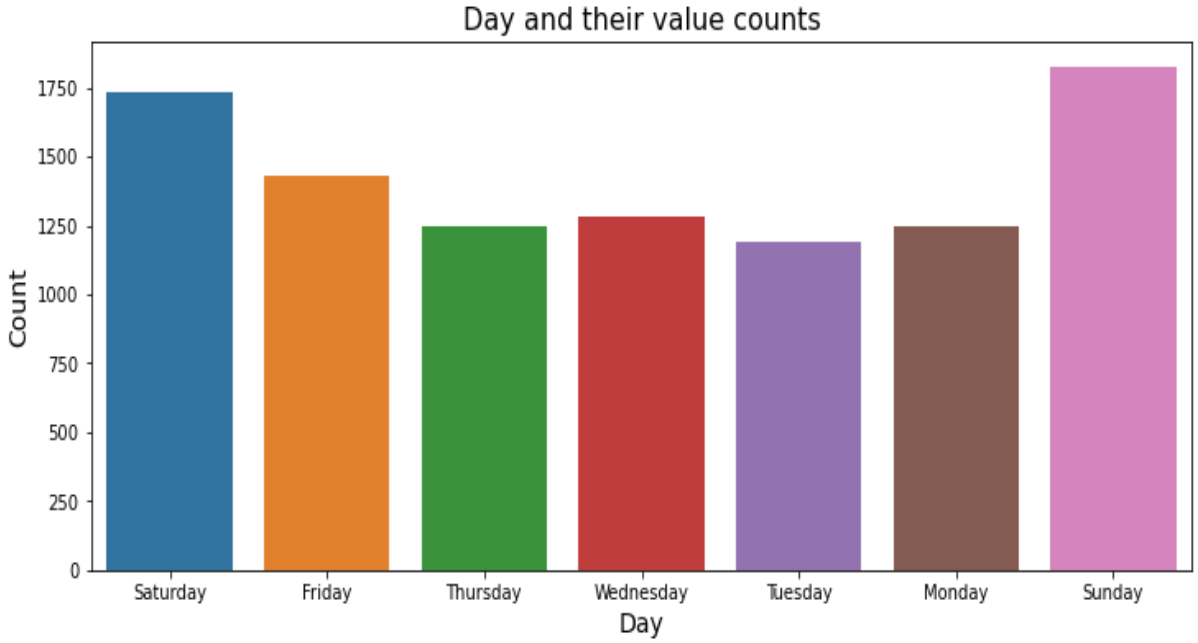


We have more orders in May among all the months and in June is the month having the least orders.

Date vs. value counts

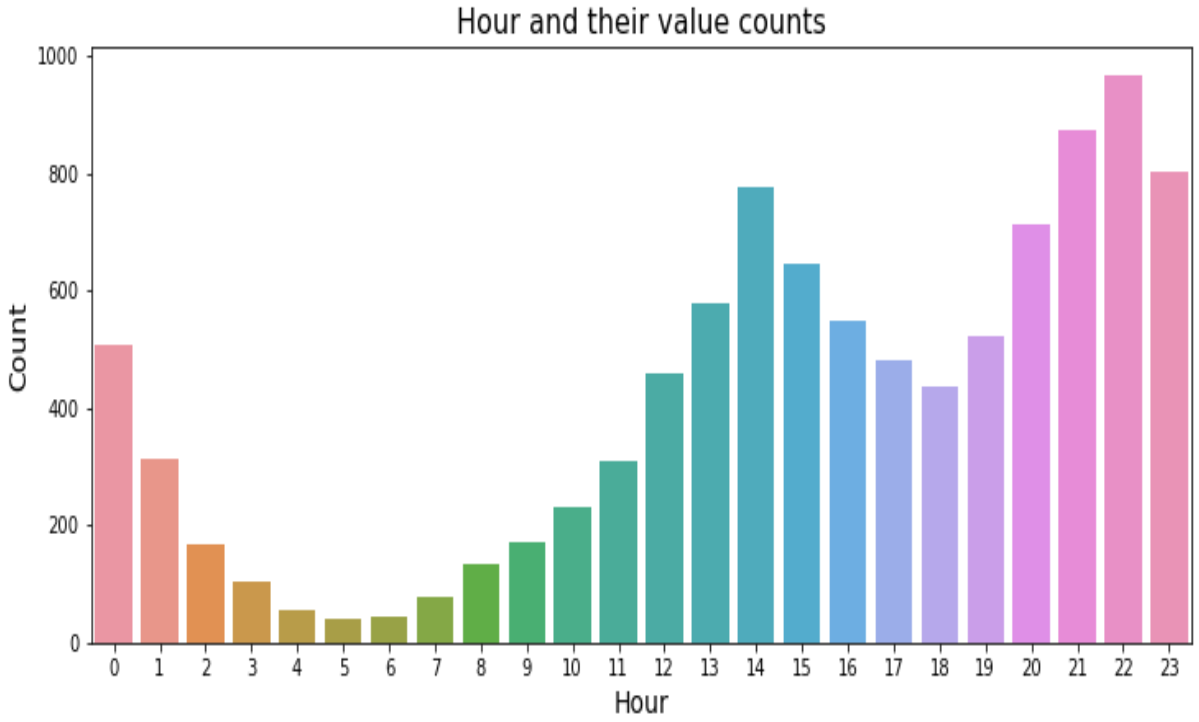


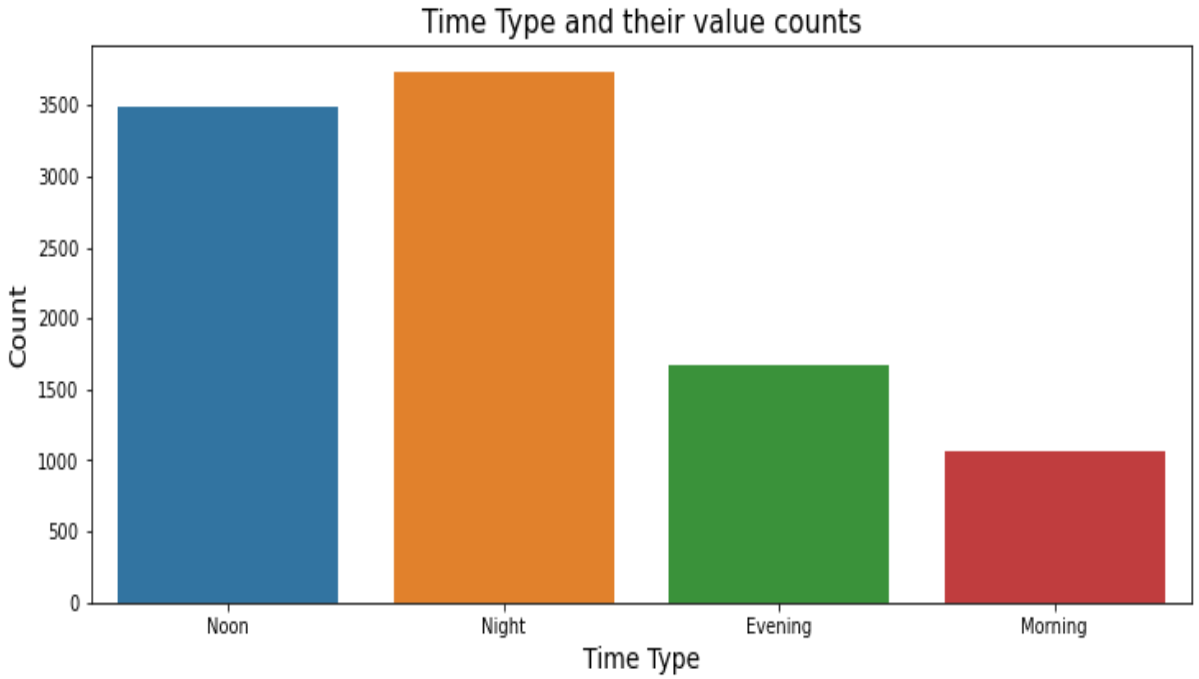
Day vs. value counts



We can see that Saturday and Sunday are having more orders, which means we have more demand during weekends.

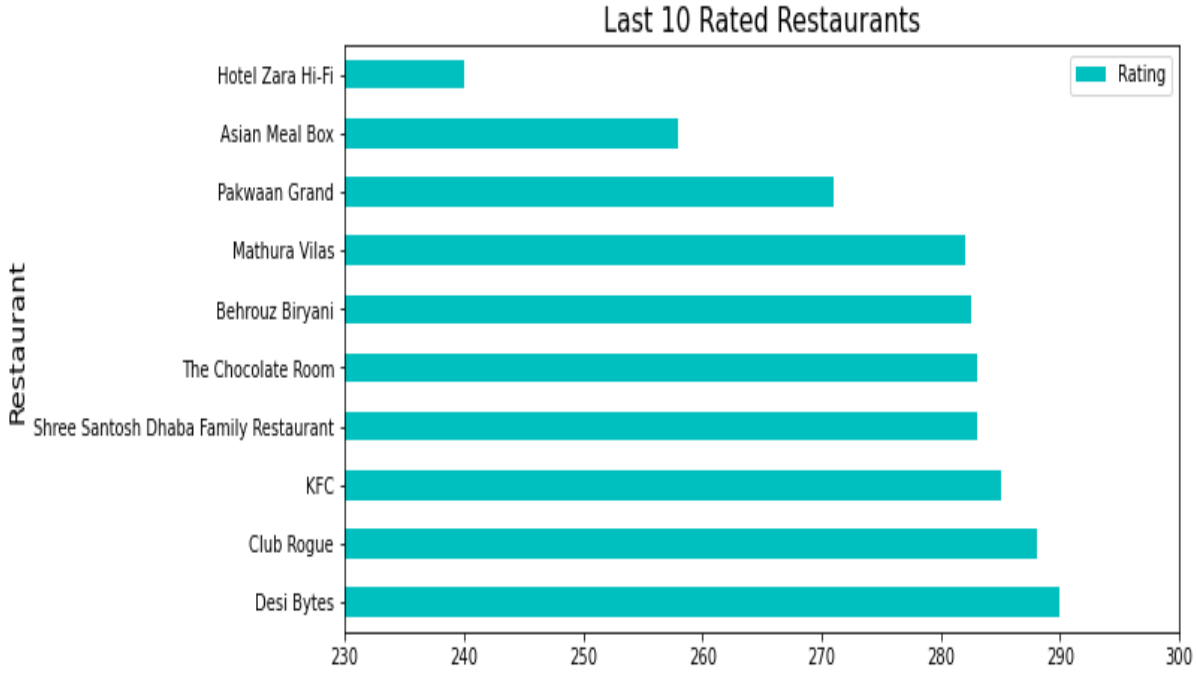
Hour vs. value counts





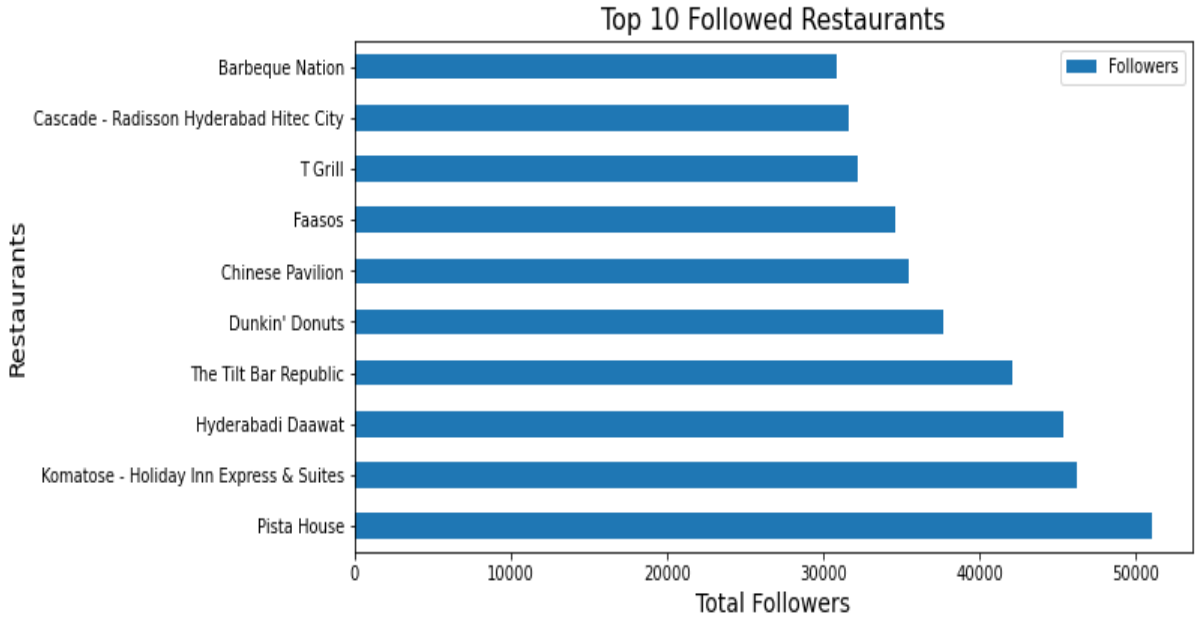
From the above figures, we can say that we have more demand at noon lunch hours and after 8 pm.

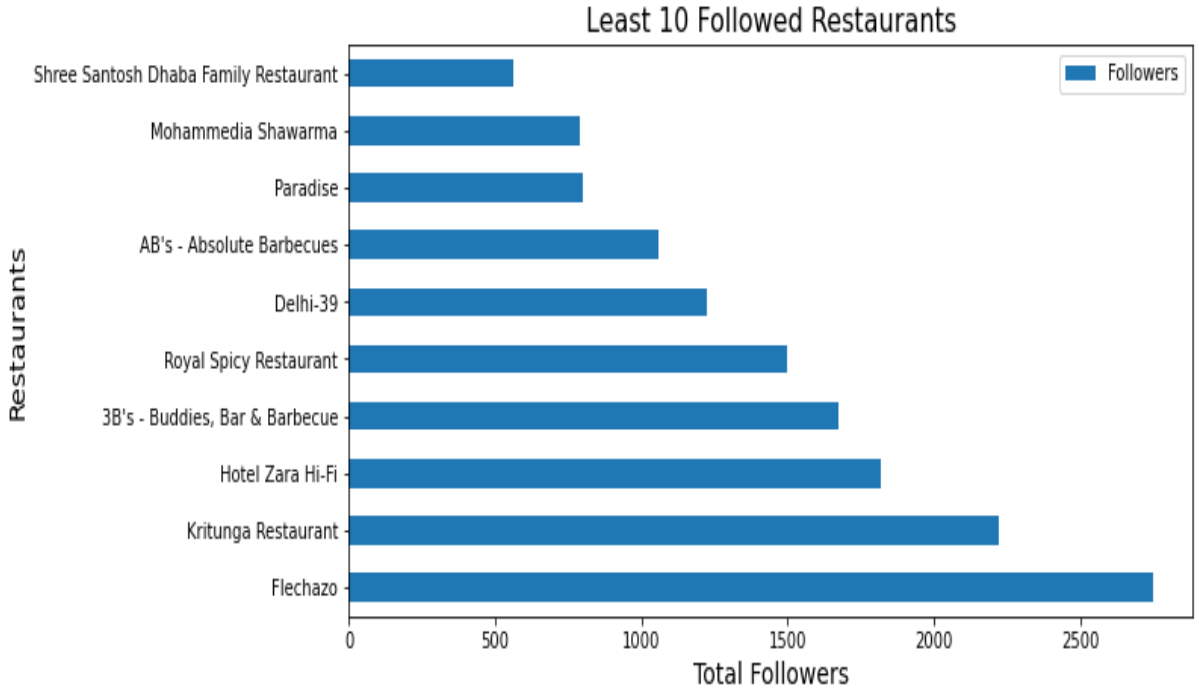
Analysis of Restaurants based on of Ratings



The above figures show the top and least rated restaurants in Hyderabad. In total, we have 105 restaurants in which Absolute Barbecues is ranked first and Hotel Zara Hi-Fi is ranked last, based on ratings given by the users.

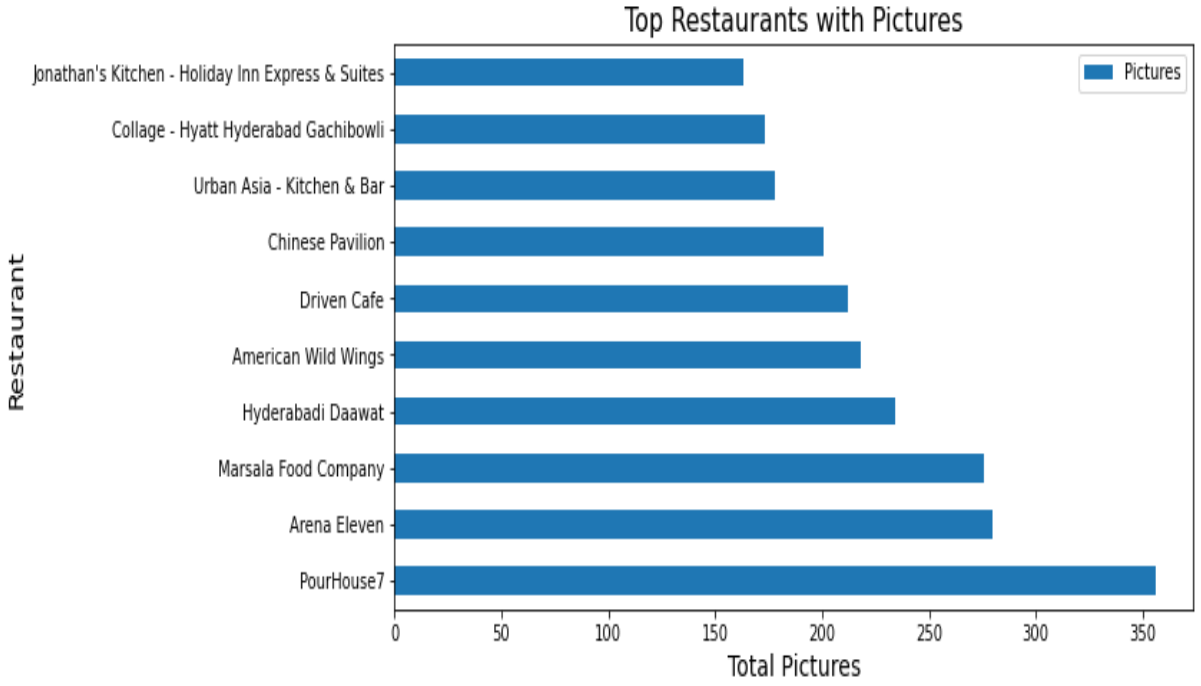
Analysis of Restaurants based on Followers

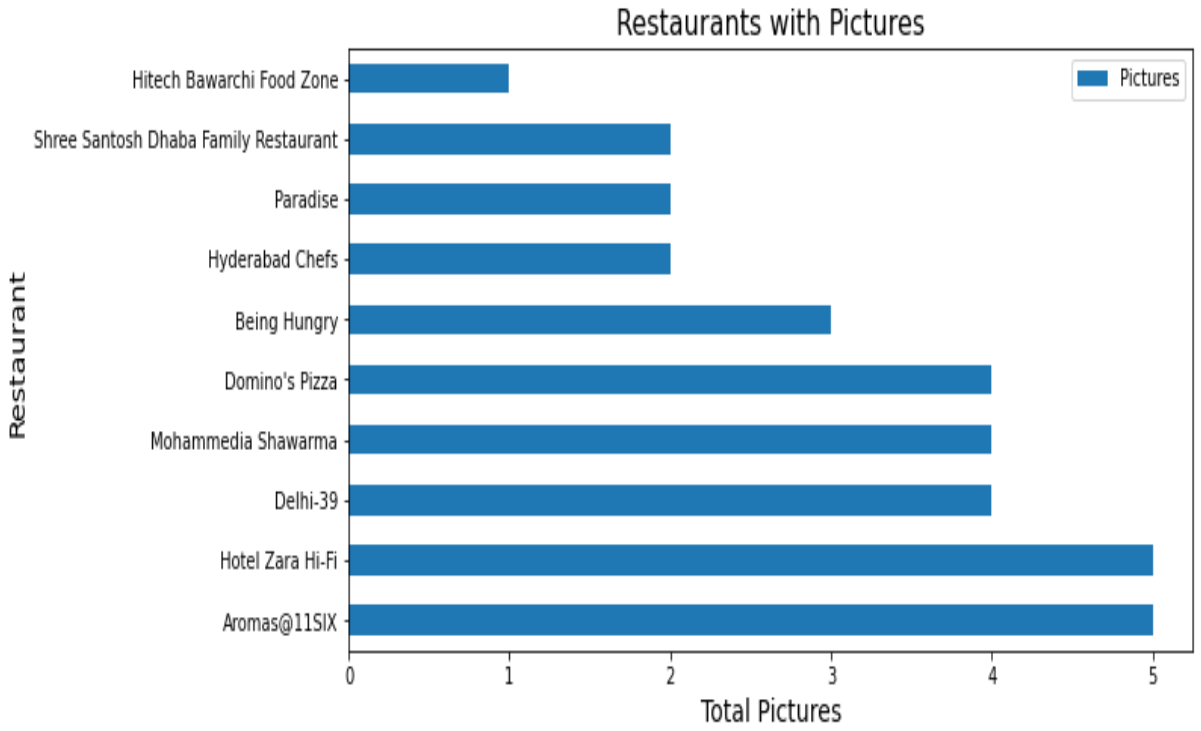




The above figures, show the top and least followed restaurants in Hyderabad. In total we have 105 restaurants among which Pista House is having most of the followers and Shree Santosh Dhaba Family Restaurant is having fewer followers.

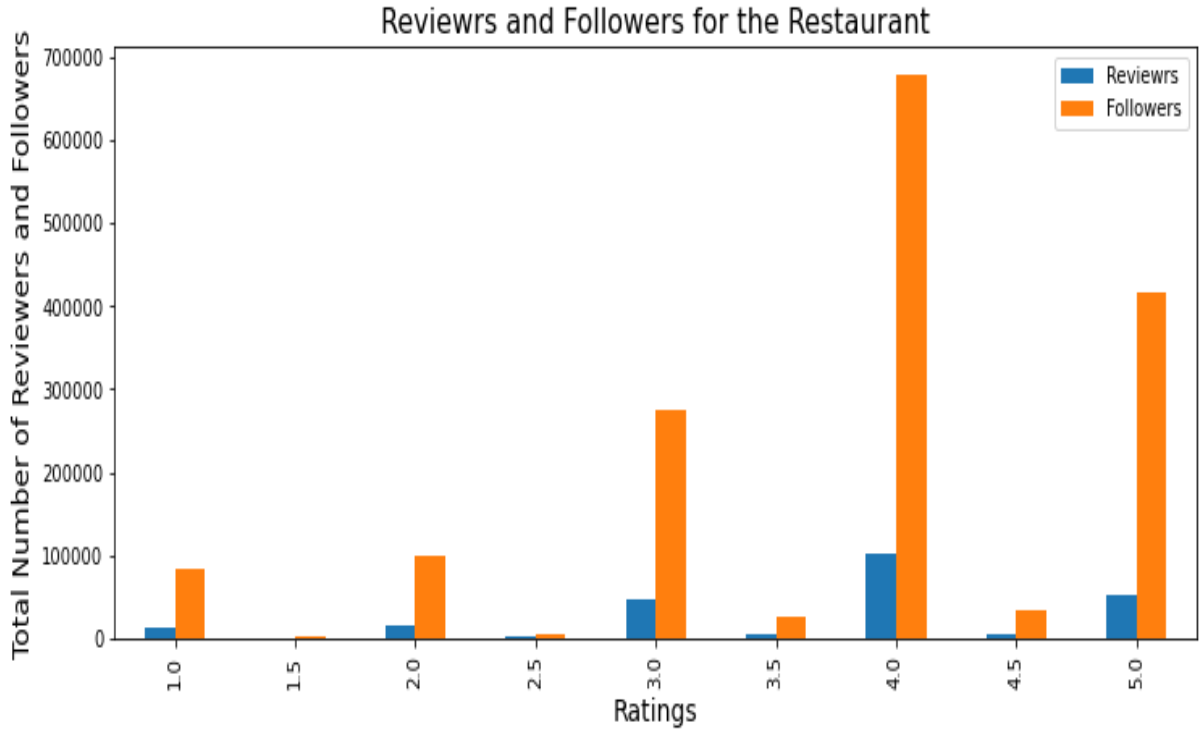
Analysis of Restaurants based on Ratings





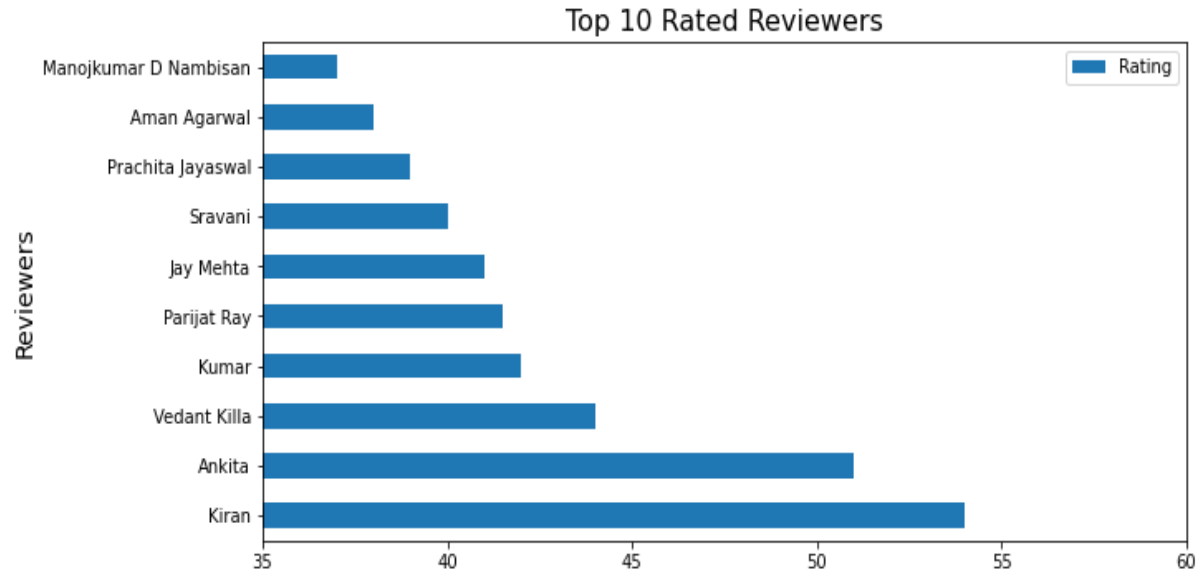
The above figures, show the restaurants having been reviewed by the followers and reviewers with pictures in their reviews.

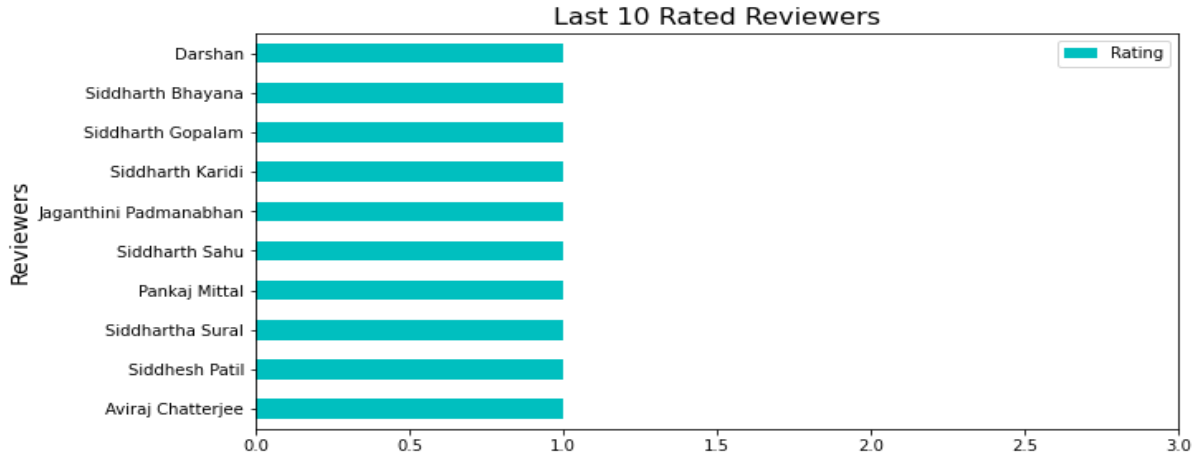
Reviewers and Followers based on Ratings



The above figure shows the number of reviewers and followers having to the restaurant based on ratings to the restaurants. From that, we can say that there are more followers than reviewers. And most of the reviewers and followers are rating 4 stars to restaurants.

Reviewers





The above figure depicts the list of reviewers who write reviews for restaurants.

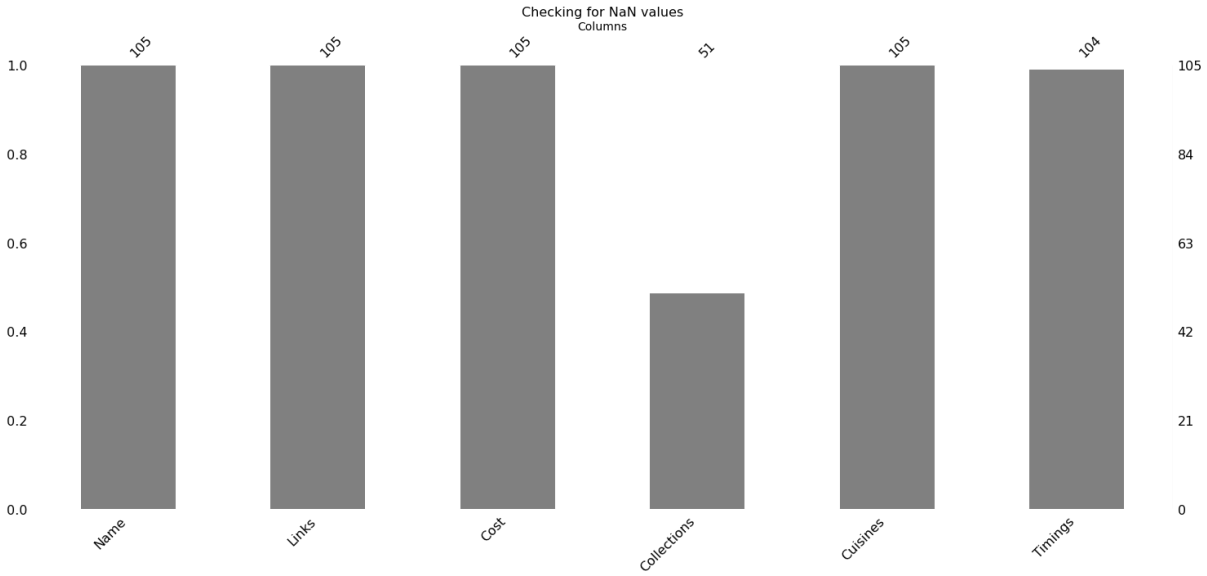
Word Frequency

Review	Count
good	239
Good	50
nice	30
very good	23
excellent	20

After the text in the review was processed we came to know that the above are the most used words in writing the reviews.

Zomato Restaurant reviews

Visualizing the presence of NaN values



The above bar plot shows the null values of the Zomato Restaurant reviews dataset. The dataset collections column is having more null values and it is dropped for analysis purposes as well Timings is having one null value and it is filled with the mode of the feature.

Pandas DataFrame

Restaurent names dataframe:

	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...

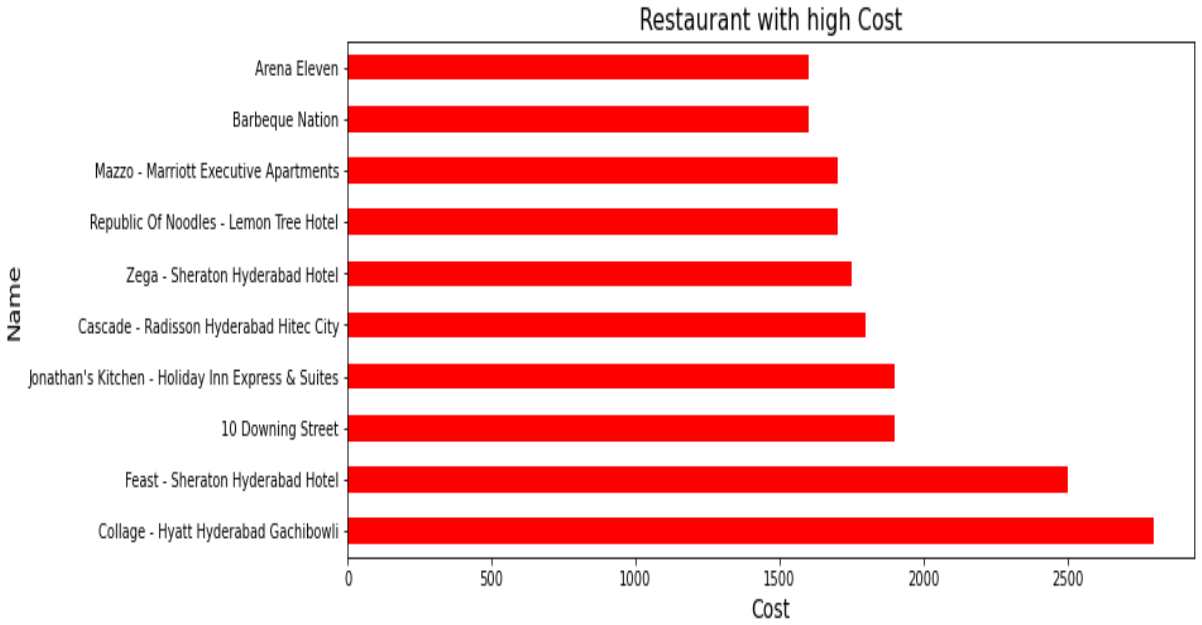
The table shows the Zomato Restaurant names and Metadata dataset in the form of Pandas DataFrame. The dataset has 105 rows and 6 columns wholly the shape is (105, 6).

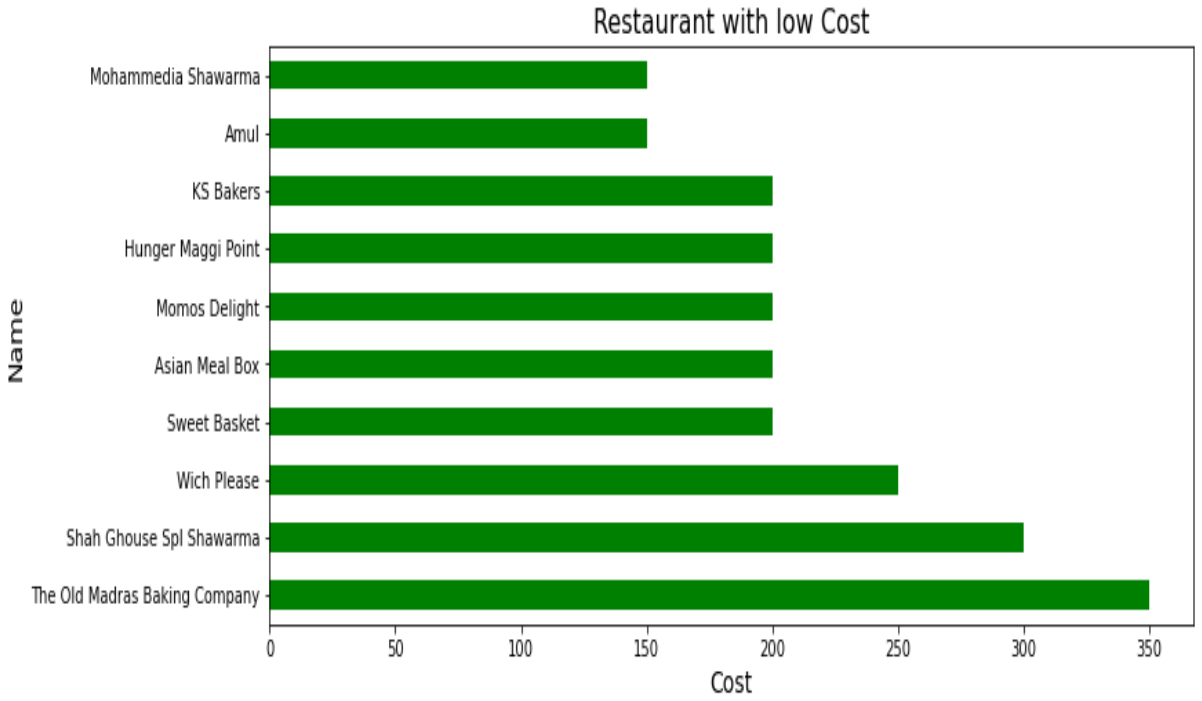
It contains the following columns:

- Restaurant Name
- Link
- Cost
- Collections
- Cuisines
- Timings

All the null values are dropped and the dataset is cleaned.

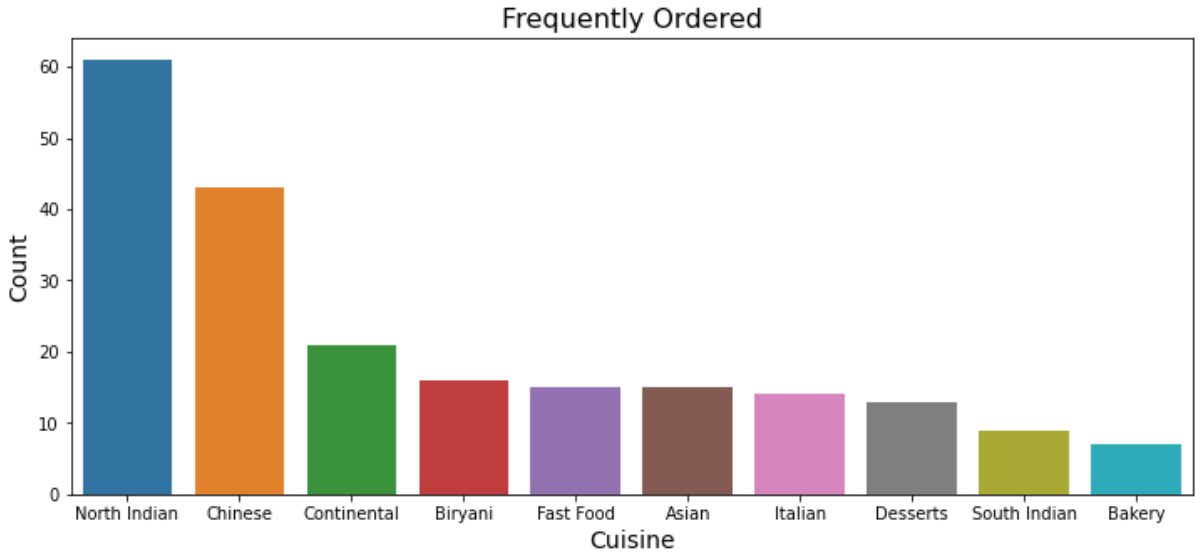
Analysis of Cost

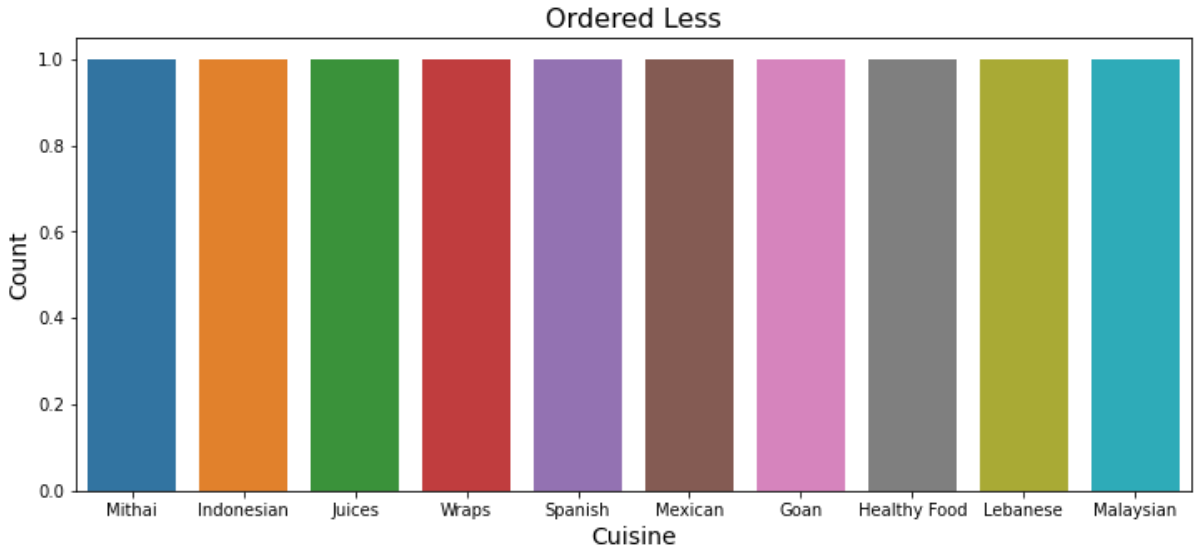




The above figures, show the list of restaurants that are costly and cheap in Hyderabad. Out of 105 restaurants, College Hyatt Hyderabad Gachibowli is the restaurant with high costs and Mohammedia Shawarma is the restaurant with cheap dishes.

Analysis of Cuisines





After the text in the Cuisine was processed we came to know that the above are the top and least varieties of dishes prepared in restaurants.

Merged Dataset

After all the data was processed and exploratory analysis was done the two datasets were merged into one DataFrame and named as df.

TFIDF Vectorizer

The full form of TF-IDF is Term Frequency Inverse Document Frequency. It transforms the text into a meaningful representation of numbers which is used to

fit the machine algorithm for prediction. In simple words, it converts a collection of raw documents to a matrix of TF-IDF features.

Recommendation Engine

A product recommendation engine is essentially a solution that allows organizations to offer their customers relevant product recommendations in real-time. In this recommendation system, we have content-based filtering and collaborative filtering.

Content-based recommendation system

A content-based recommendation system works with data that the user provides, either explicitly (rating) or implicitly (clicking on a link). Based on the data a user profile is generated and which is then used to make recommendations to the user. In this project, we are going to build a model which can recommend restaurants that are similar to the entered restaurant name.

A function is defined to return the list of similar restaurants based on the user preferences. In which a new data frame was built to represent the result. The features it contains are cuisines, rating, cost and timings of the restaurant.

Cosine Similarity

Cosine similarity is a metric used to determine how similar two entities or documents are irrespective of their size. This could be used in building a recommendation system to recommend similar products, movies, shows, books and restaurants. In information retrieval using weighted TF-IDF and cosine similarity is a very common technique to quickly retrieve documents similar to a search query.

Sigmoid Kernel

The sigmoid kernel is the model evaluation metric that ranges its outputs from 0 to 1 in the form of binary form. 0 indicates false and 1 indicates true. The sigmoid kernel is also called an activation function which is commonly used in Artificial Neural Networks (ANN).

Recommendation 1

TOP 10 RESTAURANTS LIKE Pista House WITH SIMILAR REVIEWS:

	Cuisines	Rating	Cost	Timings
Owm Nom Nom	Chinese, Biryani, Andhra, North Indian	5.0	900	12Noon to 11:30PM (Mon-Sun)
The Fisherman's Wharf	Seafood, Goan, North Indian, Continental, Asian	5.0	1500	12Noon to 3:30PM, 4PM to 6:30PM, 7PM to 11:30P...
Hyderabad Chefs	North Indian, Chinese	5.0	600	12 Noon to 10:30 PM
Shah Ghouse Hotel & Restaurant	Biryani, North Indian, Chinese, Seafood, Bever...	5.0	800	12 Noon to 2 AM
Paradise	Biryani, North Indian, Chinese	5.0	800	11 AM to 11 PM
Kritunga Restaurant	Andhra, Biryani, Hyderabadi, North Indian	5.0	500	12 Noon to 4 PM, 7 PM to 11 PM
Hitech Bawarchi Food Zone	Biryani, North Indian, Chinese	5.0	500	12 Noon to 11 PM
Royal Spicy Restaurant	North Indian, South Indian	5.0	700	10:30 AM to 11 PM
Pista House	Bakery, North Indian, Mughlai, Juices, Chinese	4.0	1000	11 AM to 12 Midnight
Hitech Bawarchi Food Zone	Biryani, North Indian, Chinese	4.0	500	12 Noon to 11 PM

The above is the list of restaurants given by our model and the metric used was the linear kernel. The user has given Pista House as his input and the model has shown him the above list of restaurants as similar restaurants to his preference.

Recommendation 2

TOP 10 RESTAURANTS LIKE Paradise WITH SIMILAR REVIEWS:

	Cuisines	Rating	Cost	Timings
Mohammedia Shawarma	Street Food, Arabian	5.0	150	1 PM to 1 AM
Khaan Saab	North Indian, Mughlai	5.0	1100	12 Noon to 3:30 PM, 7 PM to 11:30 PM
Shah Ghouse Hotel & Restaurant	Biryani, North Indian, Chinese, Seafood, Bever...	5.0	800	12 Noon to 2 AM
Shah Ghouse Spl Shawarma	Lebanese	5.0	300	12 Noon to 12 Midnight
Kritunga Restaurant	Andhra, Biryani, Hyderabadi, North Indian	5.0	500	12 Noon to 4 PM, 7 PM to 11 PM
Amul	Ice Cream, Desserts	5.0	150	10 AM to 5 AM
Al Saba Restaurant	North Indian, Chinese, Seafood, Biryani, Hyder...	5.0	750	6 AM to 11:30 PM
Hyderabad Chefs	North Indian, Chinese	5.0	600	12 Noon to 10:30 PM
Paradise	Biryani, North Indian, Chinese	4.0	800	11 AM to 11 PM
Biryanis And More	North Indian, Biryani, Chinese	4.0	500	11 AM to 11 PM

The above is the list of restaurants given by our model and the metric used was the sigmoid kernel. The user has given Paradise as his input and the model has shown him the above list of restaurants as similar restaurants to his preference.

Conclusion

- Among all the years we have more orders in the year 2018.
- We have more orders in May and less in June.
- Restaurants are having more demand during weekends.
- We have more orders from 12 pm - 4 pm and from 8 pm - 11 pm.
- Absolute Barbecues is ranked first and Hotel Zara Hi-Fi is ranked last, based on ratings given by the users.
- Pista House is having most of the followers and Shree Santosh Dhaba Family Restaurant is having fewer followers.
- We have more followers than reviewers.
- Good, good, nice, very good and excellent are the words most used by the reviewers in writing reviews.
- College Hyatt Hyderabad Gachibowli is a restaurant with high costs and Mohammedia Shawarma is a restaurant with cheap dishes.
- North Indian varieties are commonly ordered and Malaysian is the least.
- Linear kernel and sigmoid kernel are used to evaluate the model.