Capstone Project 4

Zomato Restaurant Clustering and Sentiment Analysis Unsupervised Machine Learning

Submitted by:
Name: Rahul Gupta
Email: rahulgupta41298@gmail.com
Education: PG NIT TRICHY(IM)

Content

- Problem Statement
- Business Problem Analysis
- Data Summary
- Methodology
- Exploratory Data Analysis
  - Best Restaurants in the City
  - The Most Popular Cuisines in Hyderabad
  - Restaurants and their Costs
  - Cost-Benefit Analysis
  - Critics in the Industry
  - Positive and Negative Word Clouds
- Restaurant Clustering
- Sentiment Analysis
- Conclusion and Recommendations
- References

Problem Statement

The Project focuses on analyzing the Zomato restaurant data. You have to analyze the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of Visualizations. Also, cluster the zomato restaurants into different segments. The Analysis also solves some of the business cases that can directly help the customers finding the Best restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in.

This could help in clustering the restaurants into segments. Also the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis

Data could be used for sentiment analysis. Also the metadata of reviewers can be used for identifying the critics in the industry.

Business Problem Analysis

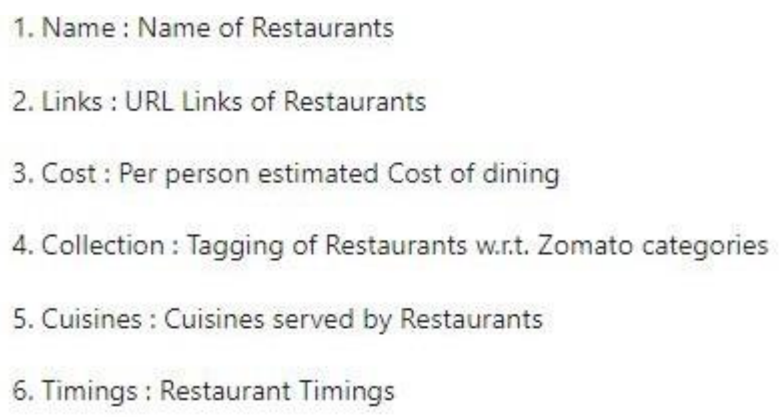
- To assure Zomato's success it is important for the company to analyze its datasets and make appropriate strategic decisions.

- The problem statement here asks us to cluster the restaurants to help customers find the best restaurants in their city and according to their taste and requirement. This will help Zomato in building a good recommendation system for their customers. Do a cost-benefit analysis using the cuisines and costs of the restaurants.

- It is important to do sentiment analysis to get an idea about how people really feel about a particular restaurant and understand the fields they are lagging in. To identify the industry critics and especially work on their reviews to build a reputation worth praising.

Data Summary

Restaurant Names and Metadata



1. Name : Name of Restaurants

2. Links : URL Links of Restaurants

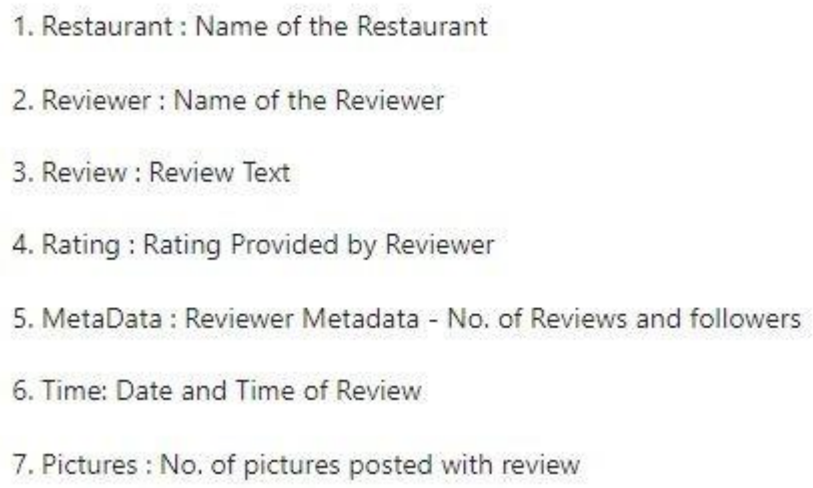
3. Cost : Per person estimated Cost of dining

4. Collection : Tagging of Restaurants w.r.t. Zomato categories

5. Cuisines : Cuisines served by Restaurants

6. Timings : Restaurant Timings

Restaurant Reviews



1. Restaurant : Name of the Restaurant

2. Reviewer : Name of the Reviewer

3. Review : Review Text

4. Rating : Rating Provided by Reviewer

5. MetaData : Reviewer Metadata - No. of Reviews and followers

6. Time: Date and Time of Review

7. Pictures : No. of pictures posted with review

Methodology

- Business Problem Analysis
- Data Collection
- Data Cleaning and Preprocessing
- Feature Engineering
- Exploratory Data Analysis
    - Best Restaurants in the City
    - The Most Popular Cuisines in Hyderabad
    - Restaurants and their Costs
    - Cost-Benefit Analysis
    - Hypotheses Generation on visualized data for Clustering
- Restaurant Clustering
    - K means Clustering on Cost and Ratings
    - Multi-Dimensional K means Restaurant Clustering
        - Principal Component Analysis
        - Silhouette Score
        - K means Clustering
        - Cluster Exploration
- Sentiment Analysis
    - Exploratory Data Analysis
        - Critics in the Industry
    - Text Pre-Processing and Text Visualization
    - Modeling
- Conclusion

Data Cleaning and Processing

Date set 1 which contains restaurant name and meta data having a total 6 features and 105 rows

Data set- feature name collection is having around 50% null data that’s why we are going to drop that column

Second data set having a 10000 row and 7 columns with no null value.

The "Review" column has text that needs to be analyzed to understand the sentiments and without it, the analysis cannot be done. It can also be seen that most of the null values in the review column also have nulls in other corresponding columns such as Reviewer, Rating, Metadata, and Time. These instances should be dropped.

Feature Engineering

The restaurants dataset has columns such as Links, Cuisine, and Timings which aren't directly interpretable. The location of the restaurant can be extracted by the Links column. Cuisines can be clubbed and categorized into a few categories and a total number of cuisines served by a particular restaurant. Timings can be categorized into three categories to make analysis a little simpler.

cuisines

it is observed that various cuisines served by every restaurant are in the form of strings and it's important to categorize and create dummy variables for all the cuisines served. The procedure followed in doing this is as follows:
First, strings are split to get the cuisines in the list datatype.
A frequency dictionary is created to understand the unique cuisines and the frequency in which the cuisine occurs.
An attempt is made to the club and categoriz various misspelled cuisines and get a minimized number of unique cuisines.
Next, we need these cuisines in the one-hot encoded form. To get these a data frame is created with the unique cuisines as columns and if a particular restaurant has this cuisine available we get a positive.

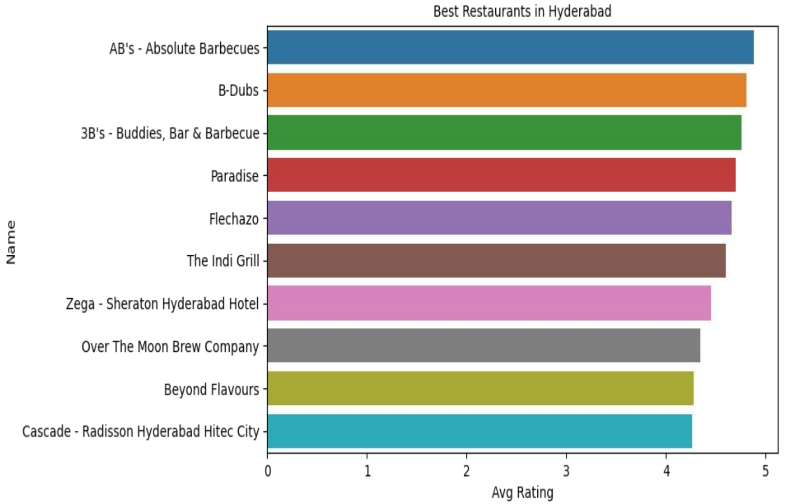
It is observable that many of the cuisines are misspelled in terms of an extra space added at the beginning of the string.
For example, there are two categories for North Indian food - 'North Indian' and ' North Indian'.
Another point to note is there are various unnecessary categories made. For example, there are 'Chinese' and ' Momos' both in the dataset as different cuisines. Let's try to club and correct them.

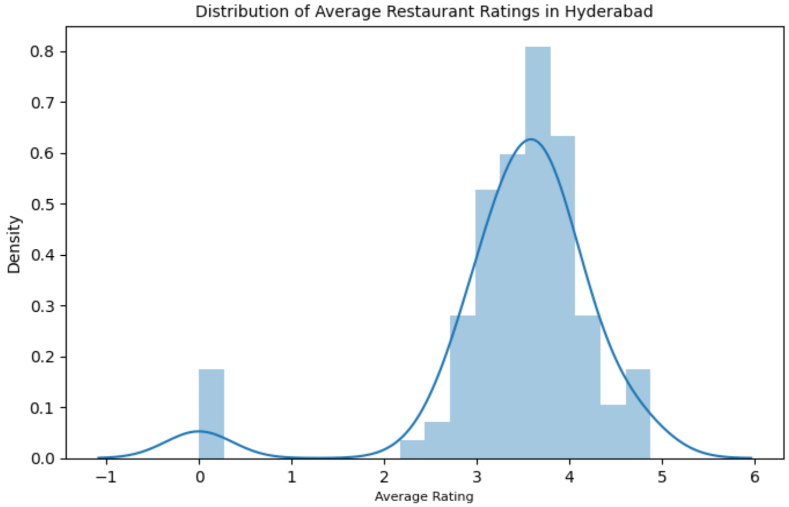
The next step is to create column features for the unique cuisines and assign values according to the row values available.

Upon analyzing the unique values in the timings columns, it can be concluded that the restaurants are more or less open at the same timings and don't really provide a considerable variation in order to cluster the restaurants.

Exploratory Data Analysis

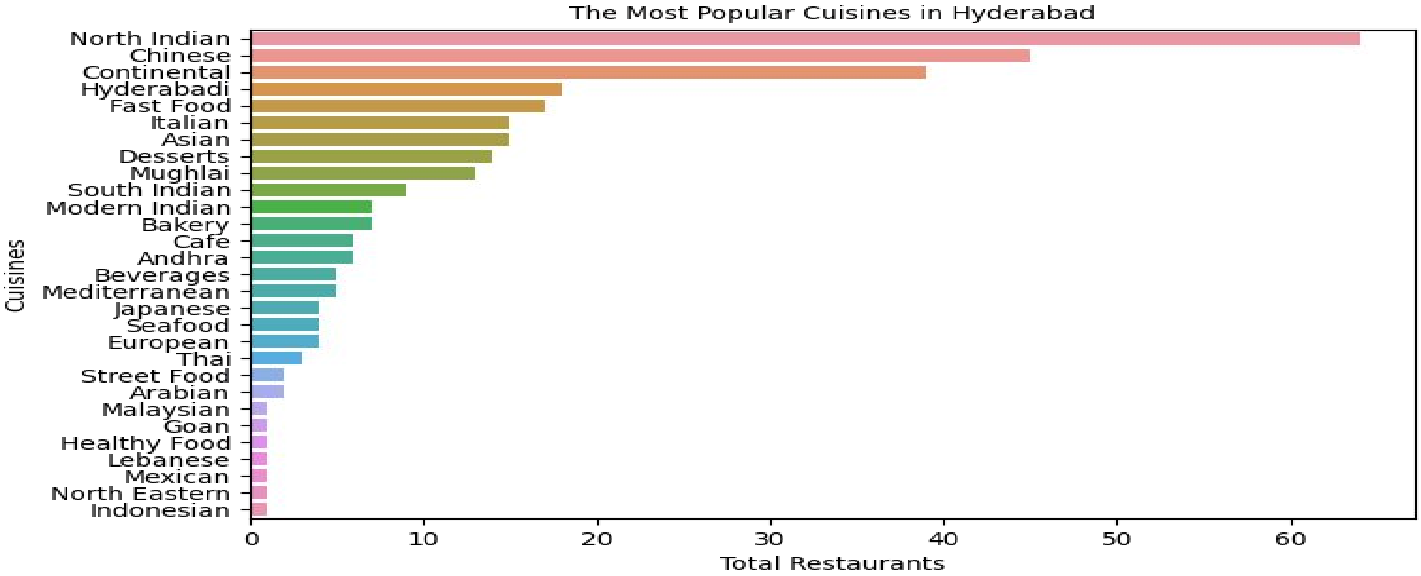
Best Restaurants in the City





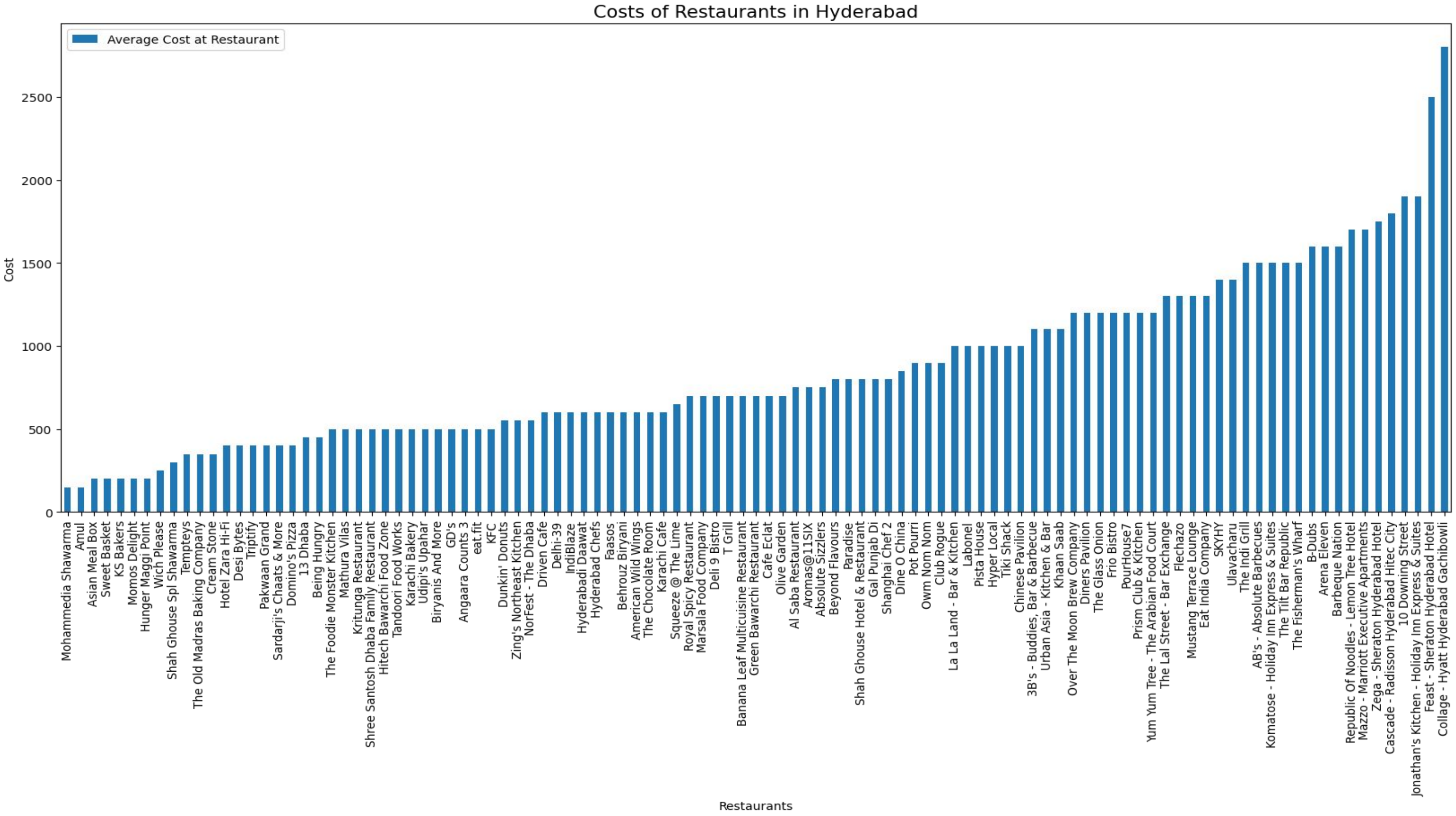
Few restaurants in the original restaurant dataset have not been rated by the people yet, most restaurants have ratings between 3.5 and 4. Efforts should be made by the company to improve the existing restaurants by pushing them to act on the reviews and to include restaurants with better services in the future to improve overall rating distribution.

The Most Popular Cuisines in Hyderabad



Although located in South India, North Indian food is dominating in the restaurants followed by Chinese, and Continental. The number of cuisines shows the diverse food options available in Hyderabad.

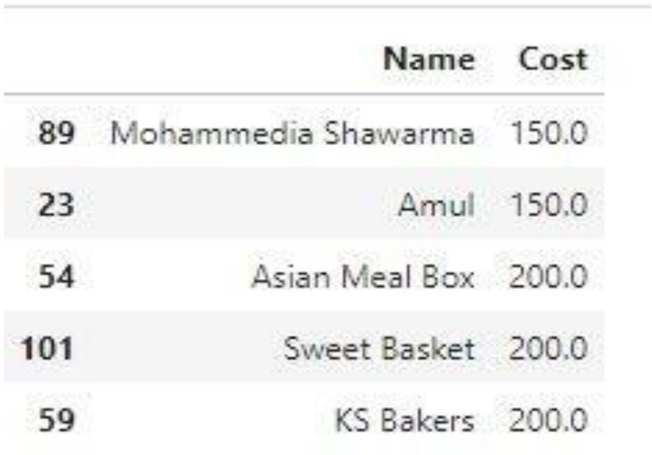
Restaurants and their Costs



Costs of Restaurants in Hyderabad

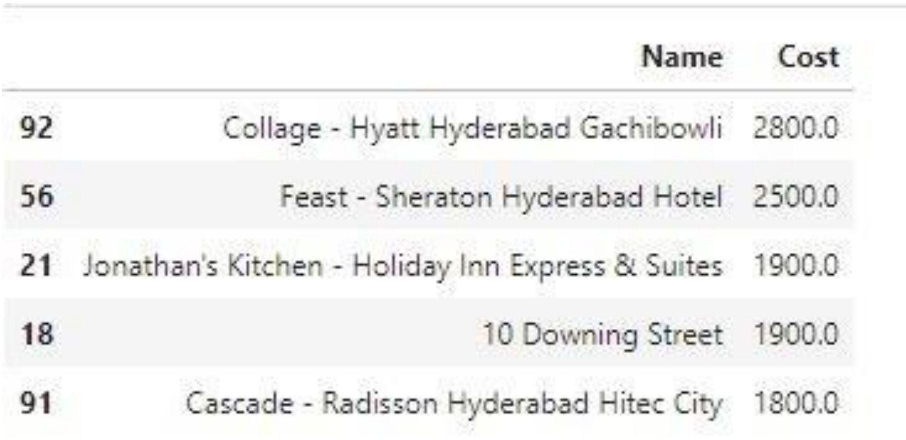
Restaurants and their Costs

Top 5 Cheapest Restaurants



	Name	Cost
89	Mohammedia Shawarma	150.0
23	Amul	150.0
54	Asian Meal Box	200.0
101	Sweet Basket	200.0
59	KS Bakers	200.0

Top 5 Costliest Restaurants



	Name	Cost
92	Collage - Hyatt Hyderabad Gachibowli	2800.0
56	Feast - Sheraton Hyderabad Hotel	2500.0
21	Jonathan's Kitchen - Holiday Inn Express & Suites	1900.0
18	10 Downing Street	1900.0
91	Cascade - Radisson Hyderabad Hitec City	1800.0

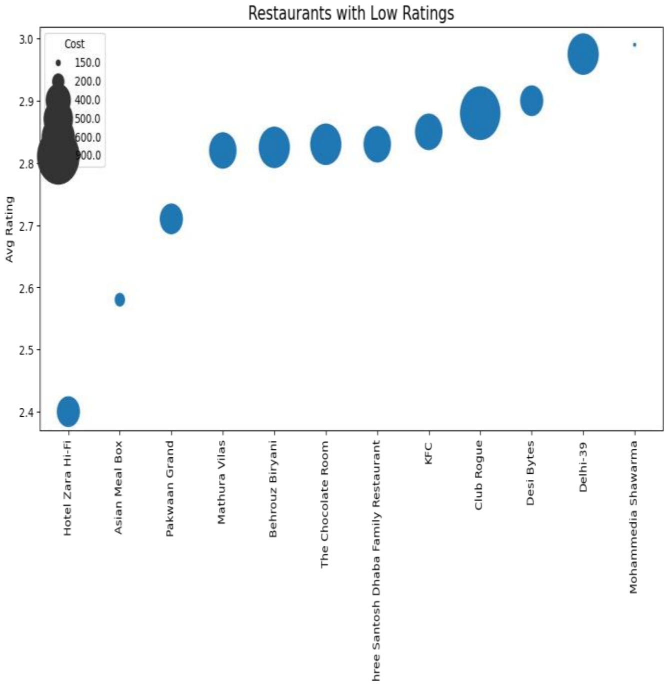
The cheapest restaurants in the dataset are basically small food joints and bakeries.
The most expensive restaurants in the dataset are restaurants by 4 star above hotels

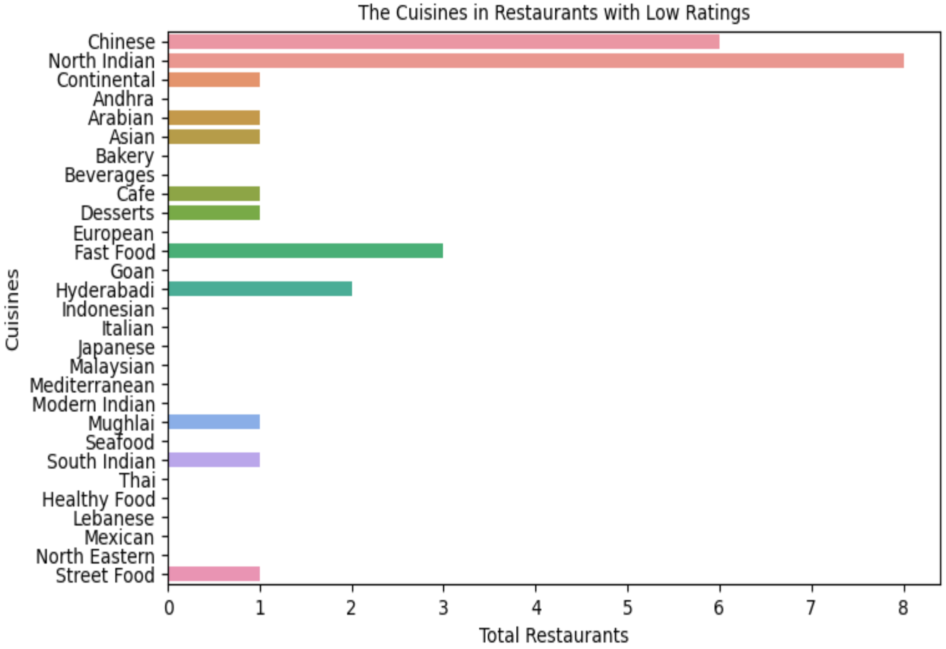
Cost-Benefit Analysis

- A Cost-Benefit Analysis is a process of analyzing the worth of a decision by estimating the costs incurred in implementing that decision and comparing them with the benefits of that decision. If the projected benefits outweigh the costs, you'll be making money out of that decision and if not, it's important to strategize a better plan.

- The data that we have consists of per-person cost, cuisines available at the restaurant, and an average rating of the restaurant. If a restaurant isn't performing well in terms of rating and has a high per-person cost and a low number of popular cuisines, this is going to be a problem for Zomato. Since negative reviews would be an intangible cost to the company and with that the company will start to lose daily application users. The application users are an asset to the company, Zomato gets advertising by different restaurants because of the large audience they have.

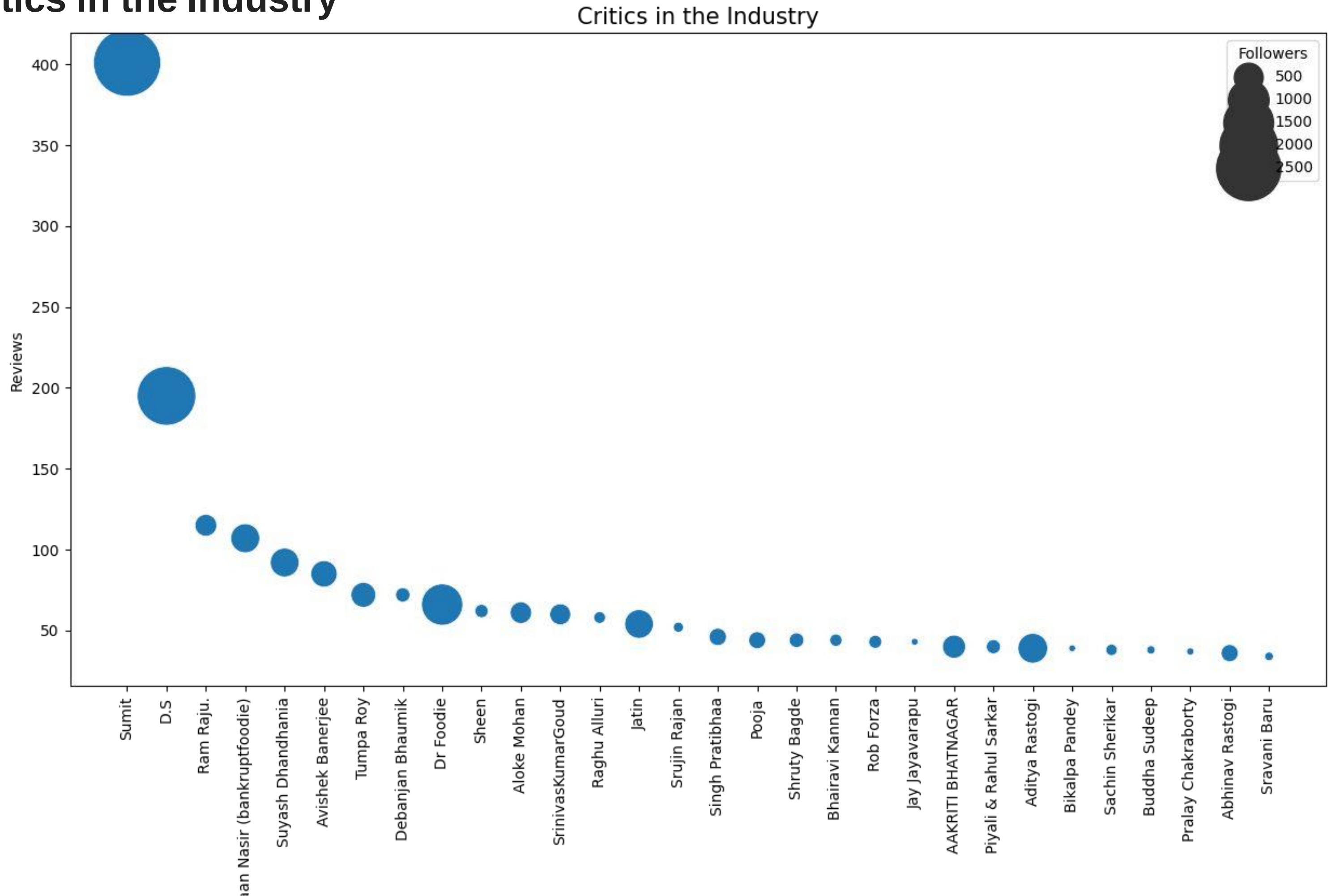
- All in all, it is important to separate out the restaurants that Zomato needs to work on in order to improve its overall customer experience and if improvement strategies don't work out, they need to delist those restaurants themselves.





These restaurants are basically small food joints or restaurants with high prices according to the food they are serving. Efforts should be made to advertise more and analyze the reviews, especially for these restaurants, and work on them. Mohammedia Shawarma has the highest rating with the lowest cost. It seems it is doing well in its capacity.

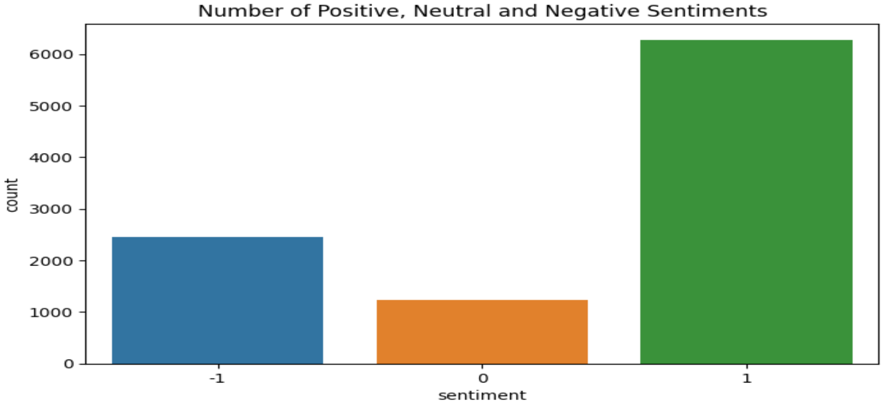
Critics in the Industry



Positive and Negative Word Clouds

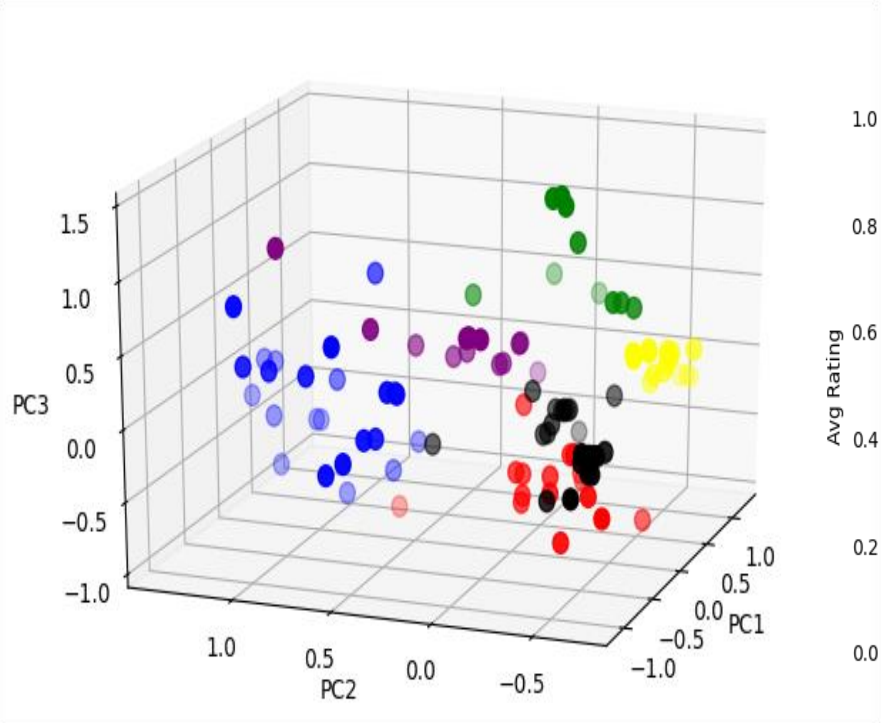
service
great
ambience
taste
best
time
nice
one
also
like
awesome
well
visit
must
amazing
delicious
hyderabad
would
restaurant
loved
price
pizza
cake
dishe
starter
main
menu
much
friendly
biryani
little
perfect
quite
really
try
tried
staff
friend
order
served
tasty
people
excellent
made
buffet
indian
love
item
definitely
overall
experience
quantity
recommend
quality
lunch
thank
spicy
dessert
veg
bit
music
option
paneer
make
fried
fish
course
better
worth
drink
go
rice
went
every
us
could
tasted
dinner
get
visited
decent
special
gachibowli
chocolate
burger
kebab
delivery
thing
lot
day
ambiance

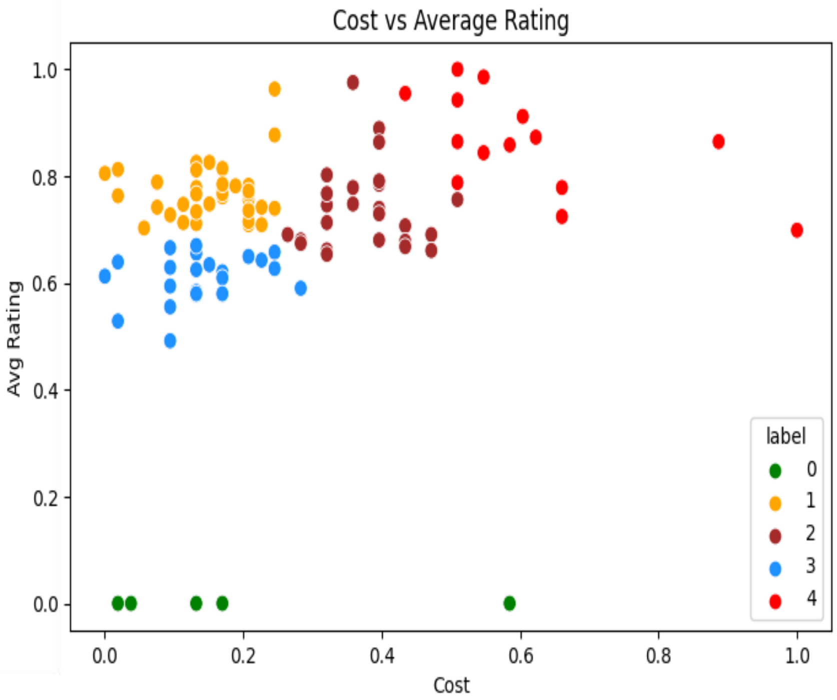
service
order
bad
taste
worst
experience
ambience
served
restaurant
us
pathetic
rating
disappointed
horrible
eat
buffet
waiter
tasted
star
veg
pizza
well
call
paneer
menu
went
piece
back
asked
one
got
get
need
review
took
time
staff
quality
first
main
table
never
zomato
quantity
much
made
visit
poor
great
better
come
friend
even
also
like
money
customer
know
drink
starter
didnt
would
item
manager
thing
cooked
worth
give
biryani
dont
ever
please
say
guy
curry
lunch
delivery
really
spicy
min
people
could
lot
dish
price
recommend
day
going
make
noodle
time
said
waste
average
rice
bill
cold
take
ok
soup
course
fried
less
shawarma
ice
late
delivered
half
received
served
high
given



Looking at the words involved in the reviews of the restaurants with low ratings it can be understood that the customers didn't like the taste, good worst experience, or maybe the order packaging wasn't good. As earlier stated some of these restaurants had medium-ranged prices.

Restaurant Clustering

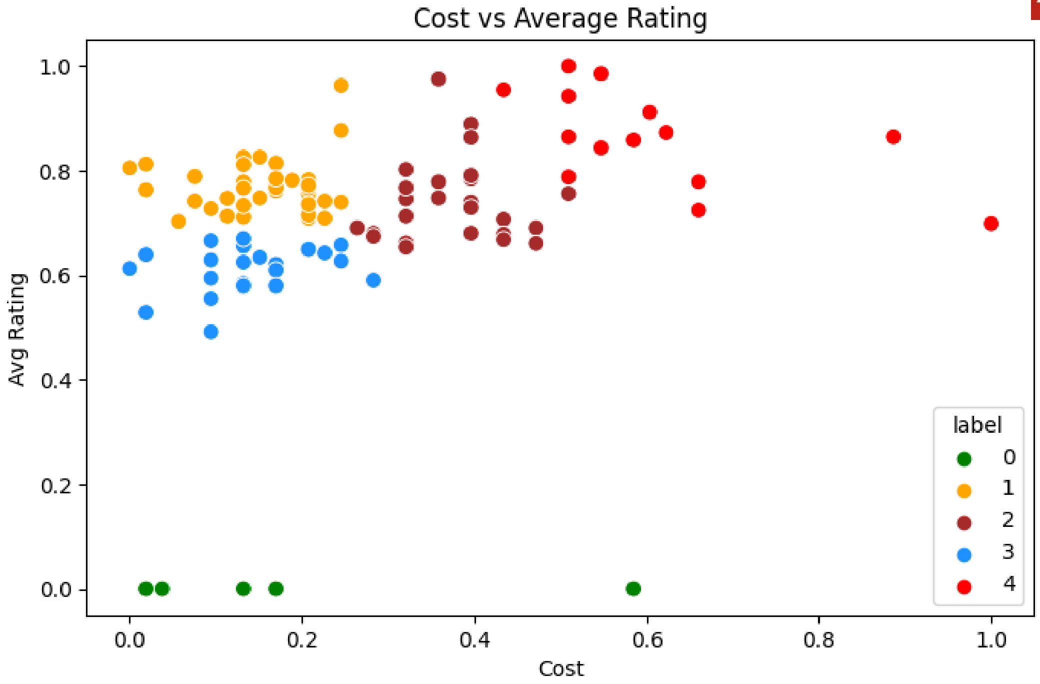




Cluster 0:
Color: Purple
Cuisines: Fast food and Continental
Average Rating: 3.42
Average Cost: 942 INR
Median Cost: 600 INR

Cluster 1:
Color: Red
Cuisines: North Indian and Complimentary
Average Rating: 3.63
Average Cost: 823 INR

Cluster 2:
Color: Blue
Cuisines: North Indian, Chinese and Continental
Average Rating: 3.77
Average Cost: 1331 INR

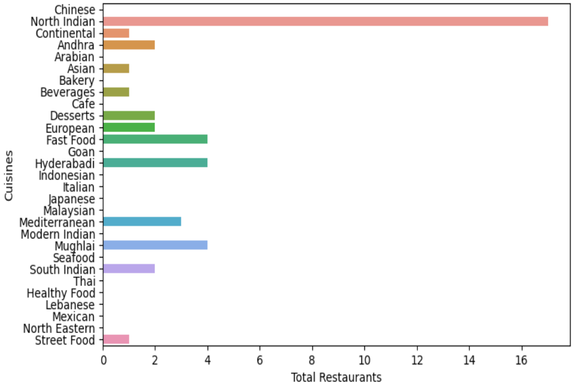


Cluster 3:
Color: Green
Cuisines: Chinese, Thai, Asian, Malaysian etc
Average Rating: 3.18
Average Cost: 890 INR

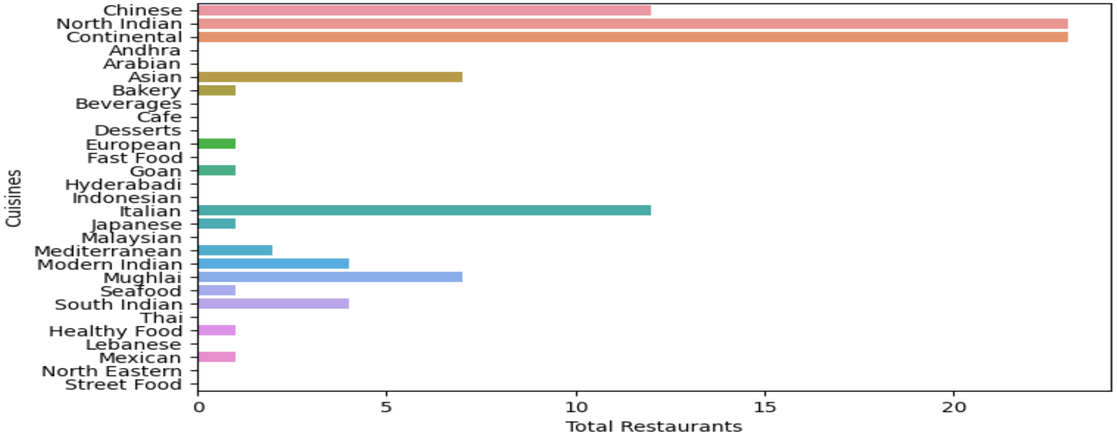
Custer 4:
Color: Yellow
Cuisines: Cafe, Bakeries, Desserts, etc
Average Rating: 3.14
Average Cost: 406 INR

Custer 5:
Color: Black
Cuisines: North Indian, Chine
Hyderabadi
Average Rating: 3.24
Average Cost: 674 INR

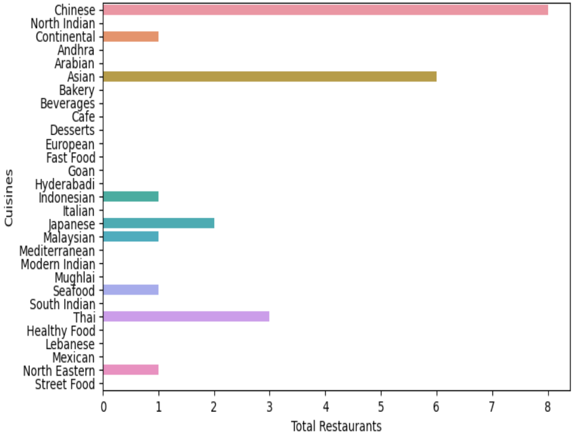
The Most Popular Cuisines in Cluster 1



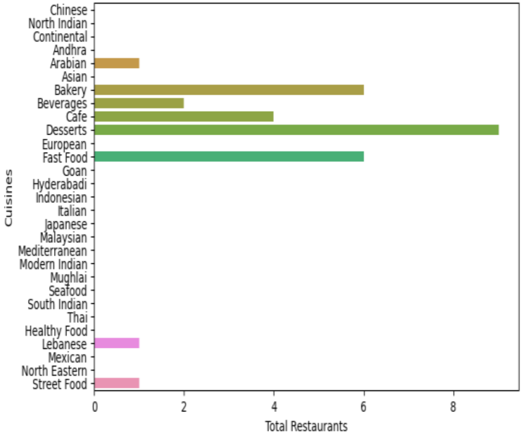
The Most Popular Cuisines in Cluster 2



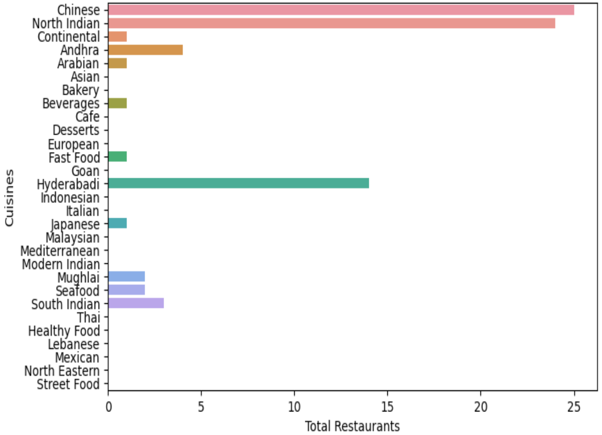
The Most Popular Cuisines in Cluster 3



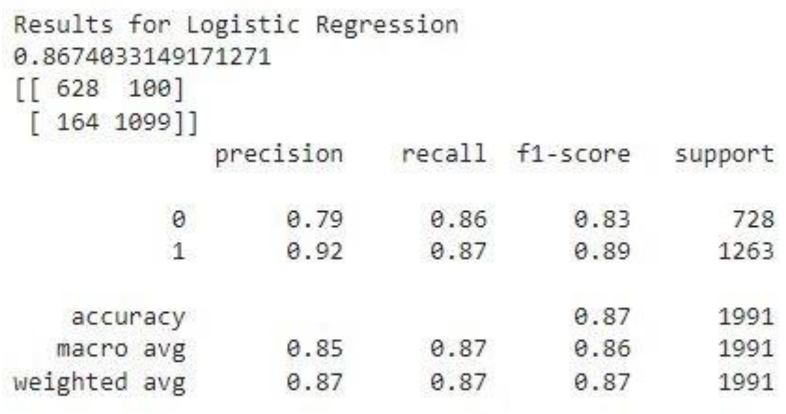
The Most Popular Cuisines in Cluster 4



The Most Popular Cuisines in Cluster 5



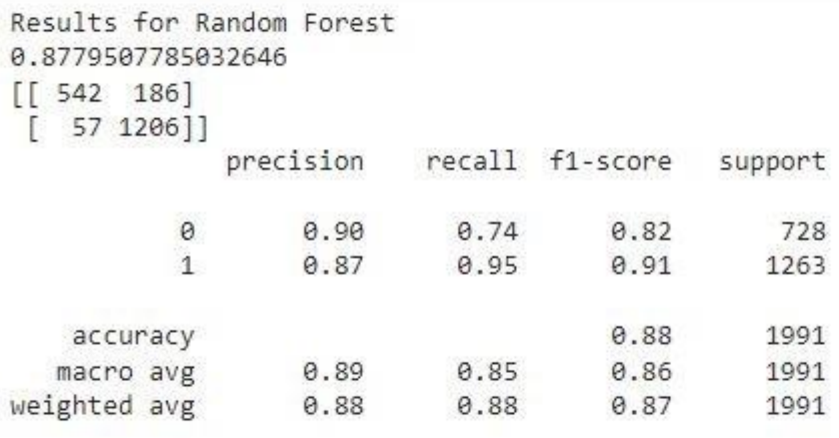
Sentiment Analysis:



```
Results for Logistic Regression
0.8674033149171271
[[ 628  100]
 [ 164 1099]]
              precision    recall  f1-score   support

           0       0.79      0.86      0.83       728
           1       0.92      0.87      0.89      1263

    accuracy                           0.87      1991
   macro avg       0.85      0.87      0.86      1991
weighted avg       0.87      0.87      0.87      1991
```



```
Results for Random Forest
0.8779507785032646
[[ 542  186]
 [  57 1206]]
              precision    recall  f1-score   support

           0       0.90      0.74      0.82       728
           1       0.87      0.95      0.91      1263

    accuracy                           0.88      1991
   macro avg       0.89      0.85      0.86      1991
weighted avg       0.88      0.88      0.87      1991
```

Evaluation:

- In the business problem, predicting the negative sentiments correctly is really important but is more important for the models to reduce the number of false positives.
- False positives indicate that the reviews were actually negative but they were categorized as positive and this will lead to missing a complaint to work on.
- Even though the number of false negatives is higher in the case of Logistic Regression than Random Forest, it is performing better in terms of reducing False positives. This indicates that Logistic Regression is penalizing False positives more just as we want.

Conclusion and Recommendations:

Some important conclusions drawn from the analysis are as follows:

- The best restaurants in Hyderabad are AB's - Absolute Barbecues, B-Dubs, and 3B's - Buddies, Bar & Barbecue..
- The most popular cuisines are the cuisines which most of the restaurants are willing to provide. The most popular cuisines in Hyderabad are North Indian, Chinese, Continental, and Hyderabadi.
- The restaurants in Hyderabadi have a flexible per person cost of 150 INR to 2800 INR. The cheapest is the food joint called Mohammedia Shawarma and the costliest restaurant is Collage - Hyatt Hyderabad Gachibowli.
- Restaurant Clustering was done in two approaches. First with just two features and then with all of them. K means Clustering worked well in the first approach but as we increase the dimensions, it isn't able to distinguish the clusters hence principal component analysis was done and then clustered into 6 clusters. The similarities in the data points within the clusters were pretty great.
- Even though the number of false negatives is higher in the case of Logistic Regression than Random Forest, it is performing better in terms of reducing False positives. This indicates that Logistic Regression is penalizing False positives more just as we want.

Recommendations:

- Restaurants with negative reviews should be worked with in order to arrive at a win-win situation.
- Ratings should be collected on a category basis such as rating for packaging, delivery, taste, quality, quantity, service, etc. This would help in targeting specific fields that are lagging.

References:

- Machine Learning Mastery
- GeeksforGeeks
- Analytics Vidhya Blogs
- Towards Data Science Blogs
- Built in Data Science Blogs
- Scikit-Learn Org

THANK YOU