Yelp Review Analysis

BIG DATA (CS GY 6513)

Srividhya Ravichandran (sr5962)
Viswanath Nagarajan (vn2065)
Priyanka Shelar (ps4497)
Anand Pitale (avp5522)

Problem Statement and Motivation:

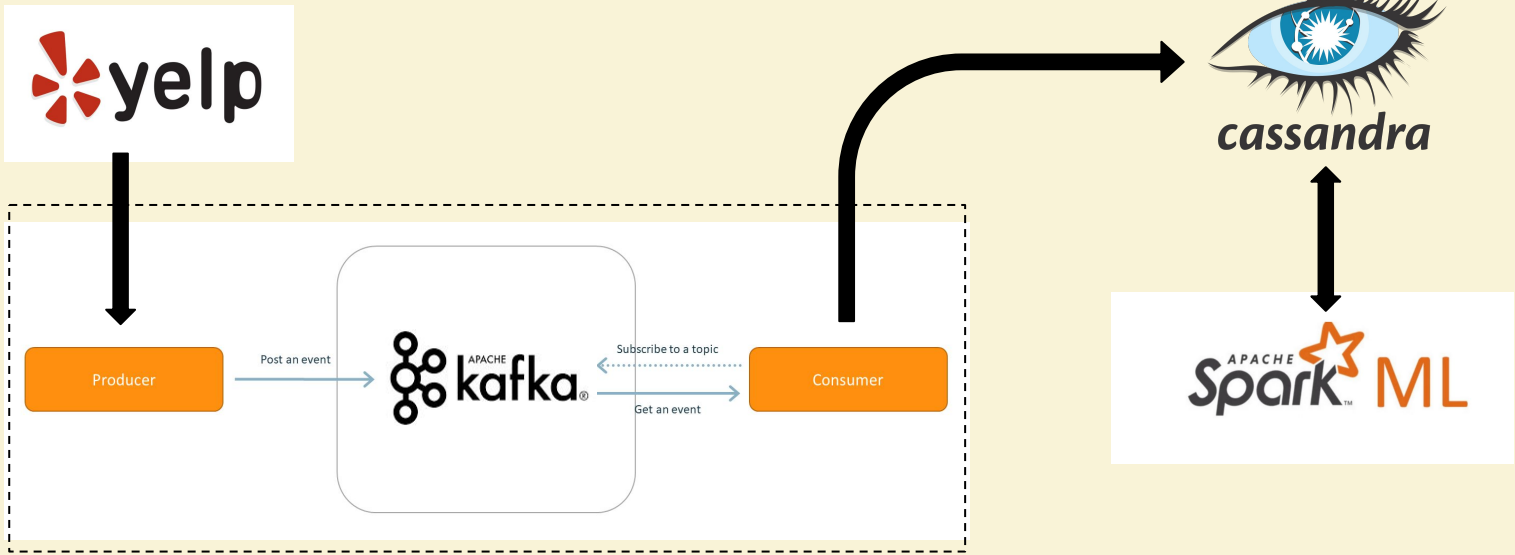
Motivation :

A research indicates that a one-star increase led to 59% increase in revenue of independent restaurants. Therefore, we see great potential of Yelp dataset as a valuable insights repository.

Problem Statement:

- Determine if customers like the food by performing an analysis on different cuisines of restaurants and the reviews they have received on Yelp.
- Recommend restaurants to customers based on their liking.

Architecture / System Design



Why Kafka :

Fast **writes**:

While Kafka persists all data to disk, essentially all writes go to the **page cache** of OS, i.e. RAM.

Fast **reads**:

Very efficient to transfer data from page cache to a network **socket**

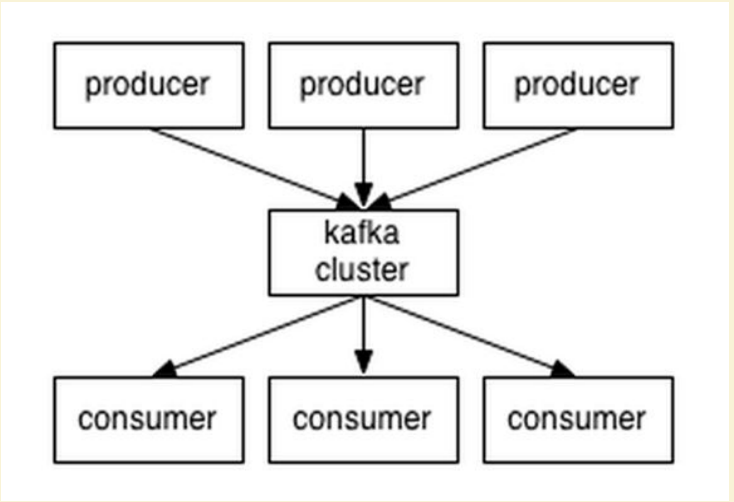
How kafka helps

The who is who

–**Producers** write data to **brokers**.

–**Consumers** read data from **brokers**.

–All this is distributed.



The data

–Data is stored in **topics**.

–**Topics** are split into **partitions**, which are **replicated**.

ELT Process

Data Processing Steps

- **Collection -** Data is fetched using Yelp Fusion API.

  The Yelp dataset is a subset of businesses, reviews, and user data for use in personal, educational, and academic purposes available as JSON data.

address	attributes	business_id	categories	city	hours	is_open	latitude	longitude	name	post
935 Race St	{null, null, u'no...	MTSW4McQd7CbVtyjq...	Restaurants, Food...	Philadelphia	{7:0-21:0, 7:0-20...	1	39.9555052	-75.1555641	St Honore Pastries	
8025 Mackenzie Rd	{null, null, u'fu...	k0hlBqXX-Bt0vf1op...	Pubs, Restaurants...	Affton	null	0	38.5651648	-90.3210868	Tsevi's Pub And G...	
2312 Dickerson Pike	{null, null, u'no...	bBDDEgkFA1Otx9Lfe...	Ice Cream & Froze...	Nashville	{6:0-16:0, 0:0-0:...	1	36.2081024	-86.7681696	Sonic Drive-In	
	{null, null, 'non...	eEOYSgkmpB90uNA7l...	Vietnamese, Food,...	Tampa Bay	{11:0-14:0, 11:0-...	1	27.9552692	-82.4563199	Vietnamese Food T...	
8901 US 31 S	{null, null, 'non...	il_Ro8jwPlHresjw9...	American (Traditi...	Indianapolis	{6:0-22:0, 6:0-22...	1	39.6371332838	-86.127217412	Denny's	
2575 E Bay Dr	{null, null, u'no...	0bPLkL0QhhPO5kt1_...	Food, Delis, Ital...	Largo	{10:0-20:0, 10:0-...	0	27.9161159	-82.7604608	Zio's Italian Market	
205 Race St	{null, null, 'ful...	MUTTqe8uqyMdBl186...	Sushi Bars, Resta...	Philadelphia	{13:30-23:0, null...	1	39.953949	-75.1432262	Tuna Bar	
1224 South St	{null, null, u'no...	ROeacJQwBeh05Rqg7...	Korean, Restaurants	Philadelphia	{11:30-20:30, 11:...	1	39.943223	-75.162568	BAP	
6625 E 82nd St	{null, null, null...	kfNv-JZpuN6TVNSO6...	Steakhouses, Asia...	Indianapolis	{11:0-21:0, 11:0-...	1	39.9043203184	-86.0530799	Hibachi Express	
5505 S Virginia St	{null, null, 'ful...	9OG5YkX1g2GReZM0A...	Restaurants, Italian	Reno	{11:0-21:0, 11:0-...	1	39.4761165	-119.7893392	Romano's Macaroni...	
215 1st Ave S	{null, null, u'fu...	tMkwHmWFUEXrC9Zdu...	Restaurants, Japa...	Nashville	{16:0-23:0, null,...	0	36.1598858	-86.7731974	The Green Pheasant	
767 S 9th St	{null, null, u'fu...	QdN72BWoyFypdGJhh...	Cocktail Bars, Ba...	Philadelphia	{12:0-2:0, 16:0-0...	0	39.9398245705	-75.1574465632	Bar One	
4105 Main St	{null, null, u'no...	Mjboz24M9NlBeiOJK...	Pizza, Restaurant...	Philadelphia	{17:0-0:30, null,...	0	40.0224662	-75.218314	DeSandro on Main	
10 Rittenhouse Pl	{null, null, u'no...	kV_Q1oqis8Qli8dUo...	Pizza, Restaurants	Ardmore	{11:0-1:0, 11:0-0...	1	40.0067071	-75.289671	Ardmore Pizza	
901 N Delaware Ave	{null, null, null...	aPNXGTDkf-4bjhyMB...	Eatertainment, Ar...	Philadelphia	{16:0-19:0, 0:0-0...	1	39.9625821	-75.1356571	Craft Hall	
116 N Pottstown Pike	{null, null, u'fu...	2xVsWBNFwZOxIOdd9...	Restaurants, Burgers	Exton	null	0	40.029962	-75.630607	Cheeseburger In P...	
312 Piasa St	{null, null, u'fu...	ljxNT9p0y7YMPx0fc...	Restaurants, Spec...	Alton	{16:0-22:0, 0:0-0...	1	38.896563	-90.1862032987	Tony's Restaurant...	
1625 W Valencia R...	{null, null, 'bee...	wghnIlMb_i5U46HMB...	Restaurants, Chinese	Tucson	{11:0-21:0, 11:0-...	0	32.1323047	-110.9999851	China Dragon Rest...	
2031 Broadway	{null, null, u'be...	lk9IwjZXqUMqqOhM7...	Coffee & Tea, Res...	Nashville	{7:0-17:0, 7:0-17...	0	36.1483712	-86.7988947	Caviar & Bananas	
10440 N Dale Mabr...	{null, null, u'fu...	uI9XODGY_2_ieTE6x...	Restaurants, Amer...	Tampa	{11:30-22:0, 11:3...	0	28.0462028173	-82.5050526736	Roman Forum	

only showing top 20 rows

Data Processing Steps

- **Cleansing**

  Raw data is checked for any errors. The purpose of this step is to eliminate bad data(redundant,incomplete or incorrect data)

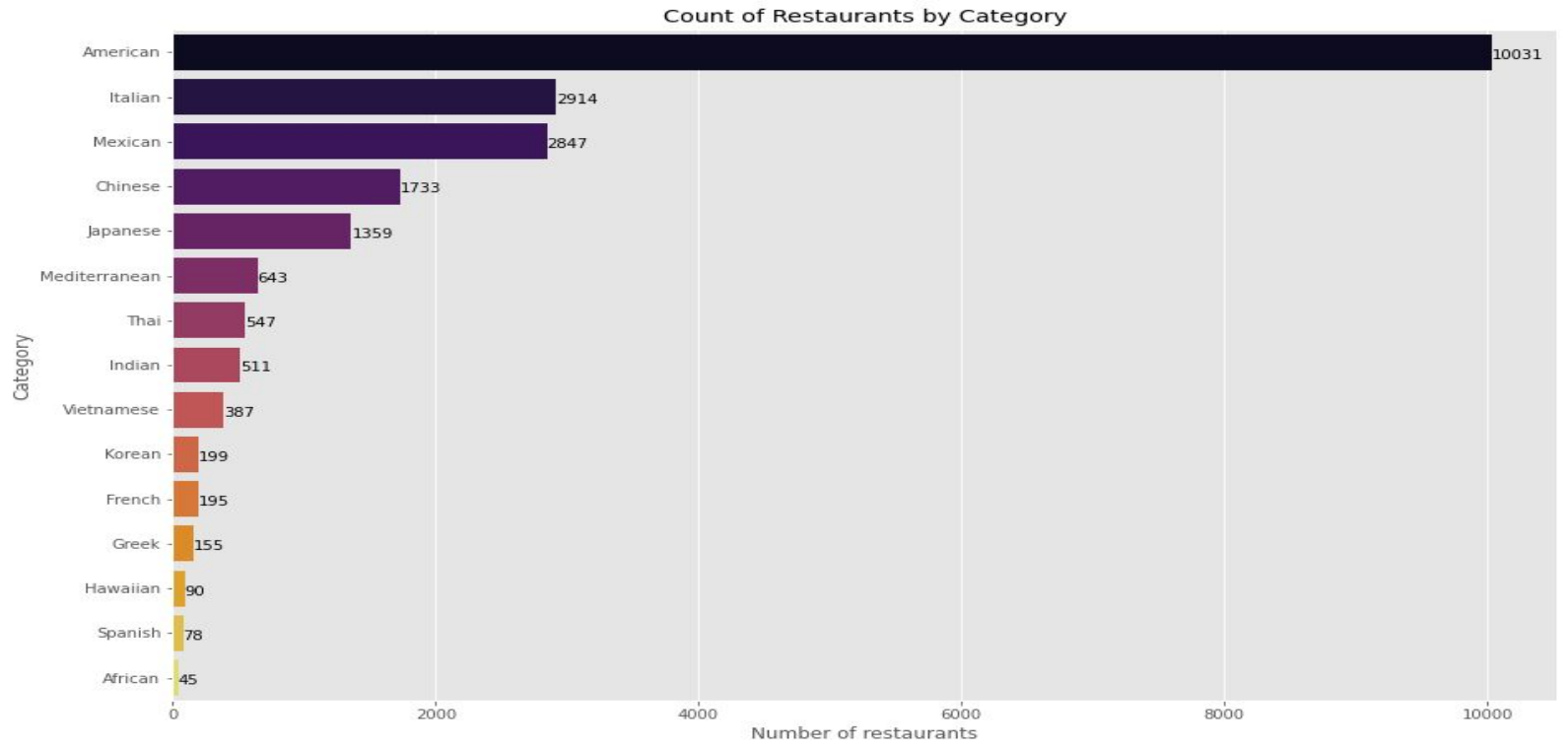
- **Organization & Processing**

  Data require indexing , sorting and then processing.Handling variable categorical / numerical and correlated features.

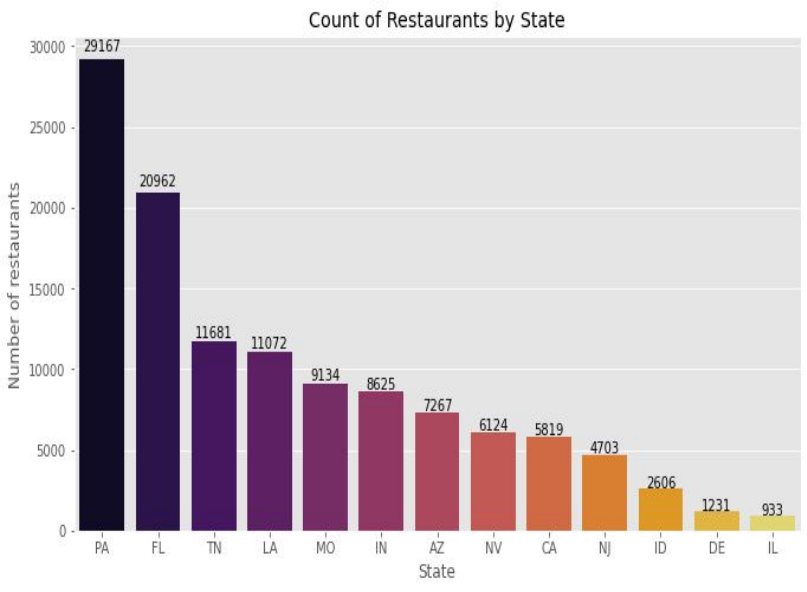
- **Visualization**

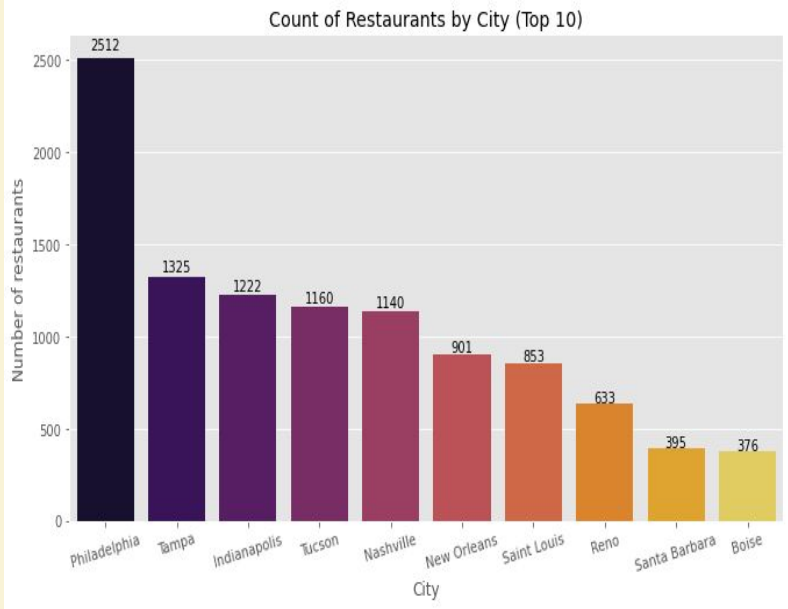
  Gives us a clear idea of what the information means by giving it visual context through graphs.Easier to identify trends, patterns, and outliers within large data sets

Cuisine Categories

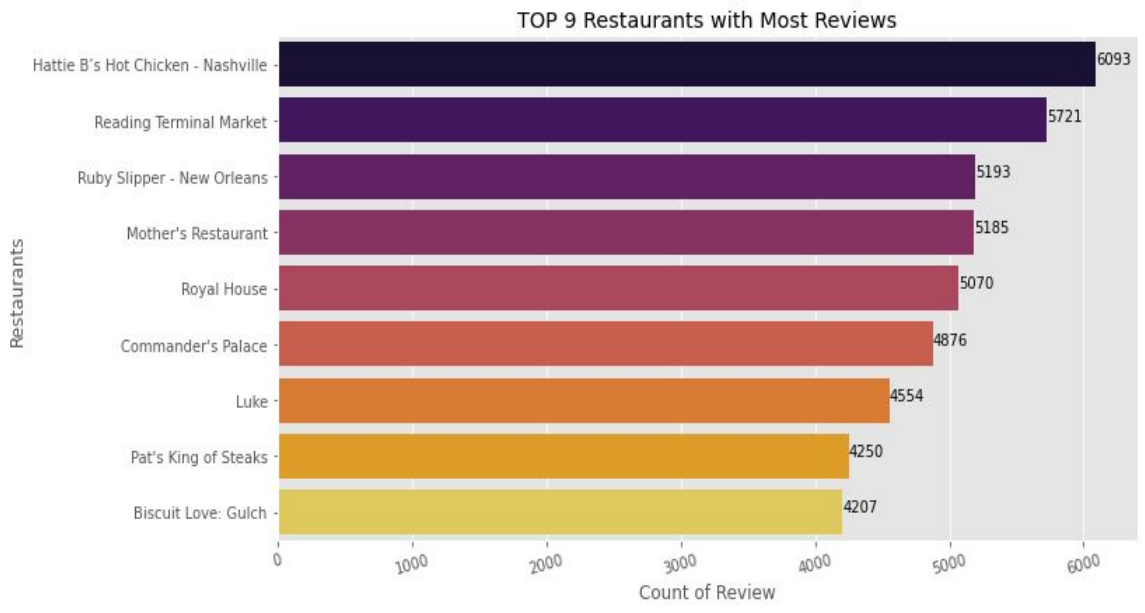


Count of Restaurants by State & City



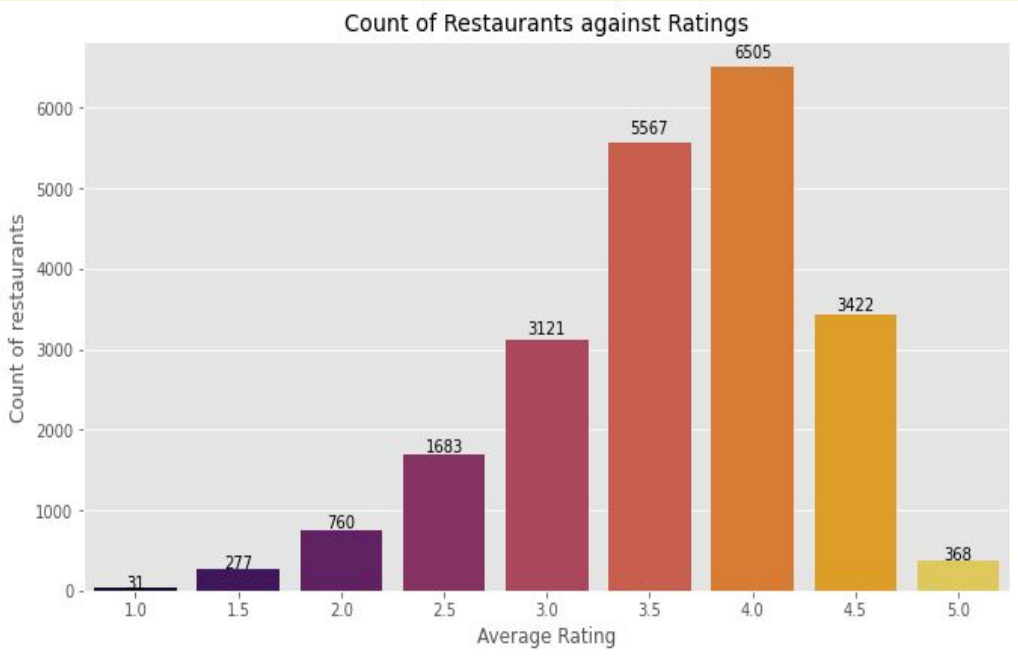


Popular Restaurants



- **Shows Popularity of restaurants**
- **More reviews indicates popularity**

Distribution of Ratings



Distribution of total number of restaurants based on ratings by users

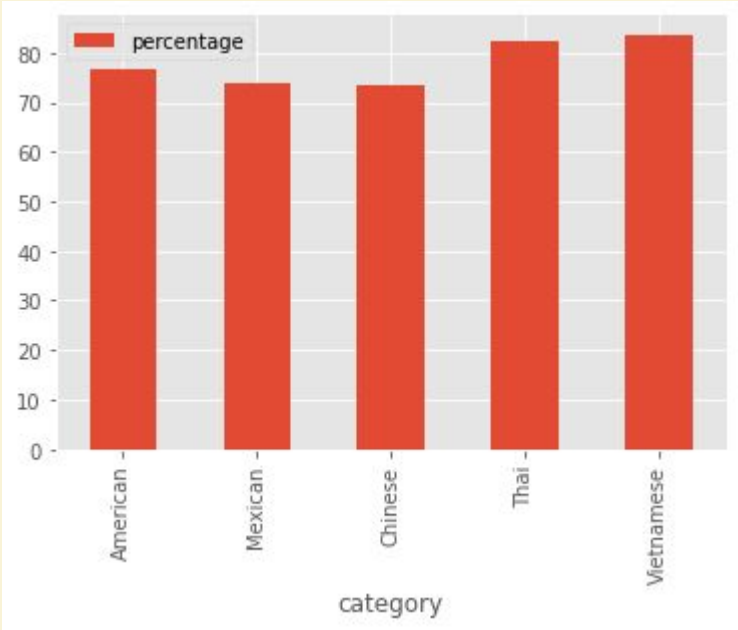
Class distribution

Labels	Count
Positive	91680
Negative	27644

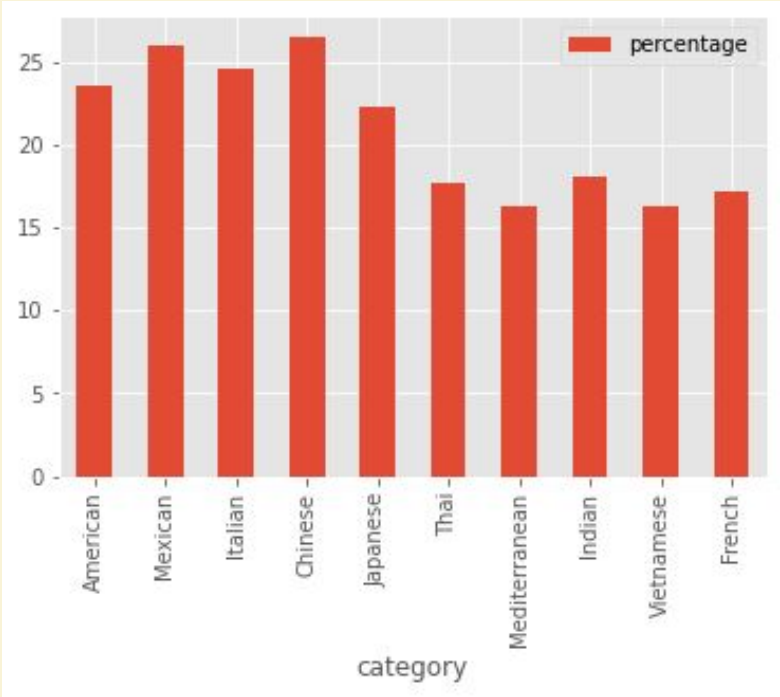
Model specifications

- Ratings > = 4 (Positive)
- Ratings < 4 (Negative)

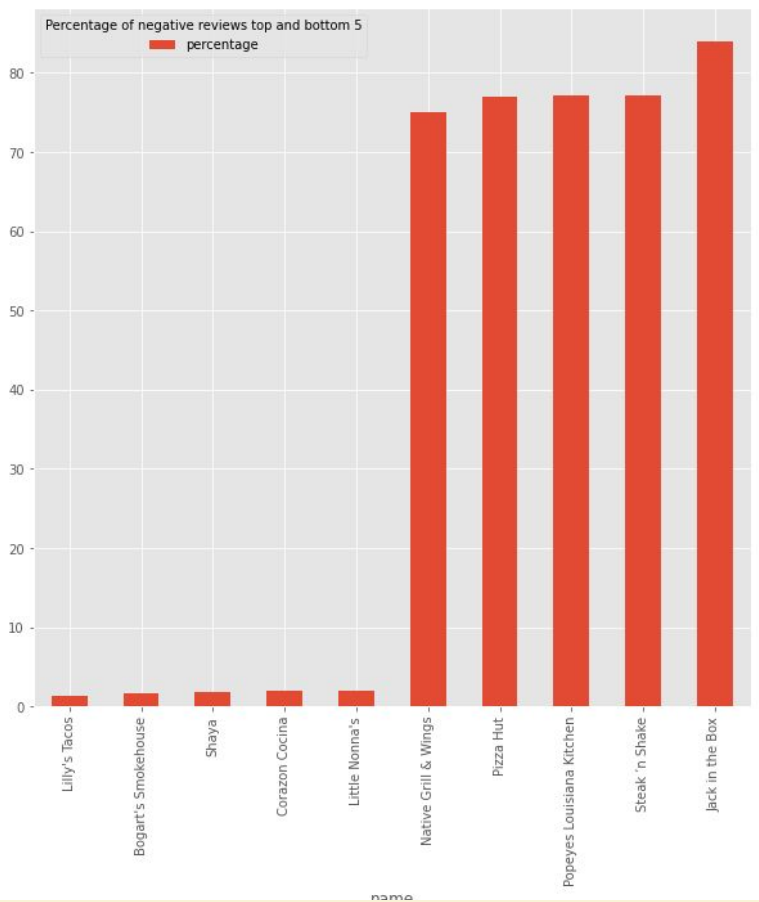
Percent of +ve Reviews/Category



Percentage of -ve Reviews/Category



Higher percentage of negative reviews for cuisines like Chinese and mexican compared to Vietnamese, French

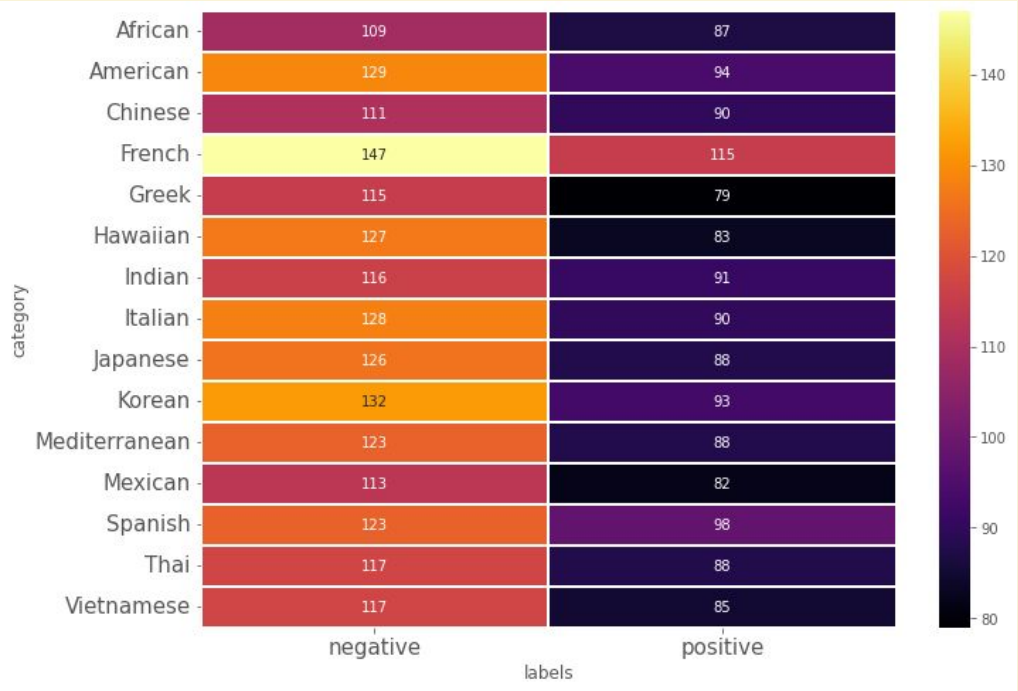


Percentage of negative reviews per restaurant

- Restaurant with Best and Worst respective review %.

- Minimum review count of 100 or more for restaurants.

Analyzing +ive and -ive reviews



Average length of reviews per class.

Positive reviews have a lower word count compared to negative reviews.

Natural Language Processing

Approach :

- Converting text to lowercase
- Removing non Ascii characters, punctuations, stopwords
- Fixing abbreviations

text	labels	Target	lower_text	text_non_asci	fixed_abbrev	removed_features
as a local new or...	positive	1	as a local new or...	as a local new or...	as a local new or...	as a local new or...
this has been my ...	positive	1	this has been my ...	this has been my ...	this has been my ...	this has been my ...
the atmosphere is...	negative	0	the atmosphere is...	the atmosphere is...	the atmosphere is...	the atmosphere is...
cant wait to get ...	positive	1	cant wait to get ...	can not wait to g...	can not wait to g...	can not wait to g...
i am a vegetarian...	negative	0	i am a vegetarian...	i am a vegetarian...	i am a vegetarian...	i am a vegetarian...

only showing top 5 rows

- Stemming
  - finding the root of words
- Lemmatization
  - finding the form of the related word in the dictionary
- Vectorizer (Count and TF-id)
  - Assigning values to words as per count (CV) and their importance (TF-id)

text	labels	words	tf	features	label
zorbas is the bes...	positive	[zorbas, is, the,...	(65536,[338,1578,...	(65536,[338,1578,...	0.0
zona zona zona ha...	positive	[zona, zona, zona...	(65536,[239,513,1...	(65536,[239,513,1...	0.0
zona 78 is my fav...	positive	[zona, 78, is, my...	(65536,[1689,2692...	(65536,[1689,2692...	0.0
zoes is keeping i...	positive	[zoes, is, keepin...	(65536,[1587,1981...	(65536,[1587,1981...	0.0
zinc is the best ...	positive	[zinc, is, the, b...	(65536,[1608,1743...	(65536,[1608,1743...	0.0

only showing top 5 rows

NLP Modeling - Logistic Regression

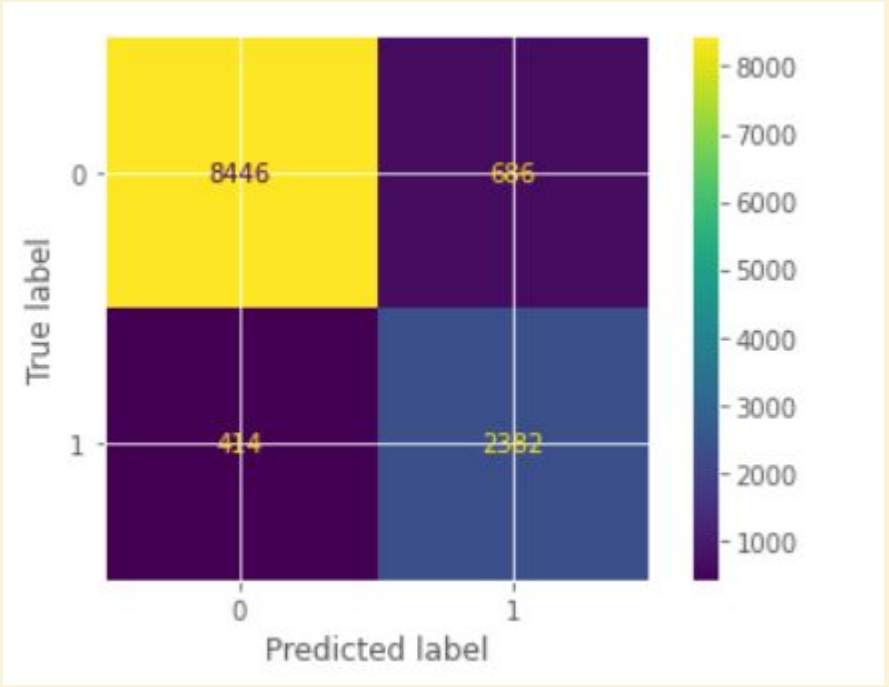
- Logistic Regression is used to classify elements of a set into two groups (binary classification) by calculating the probability of each element of the set.
- It uses the sigmoid function to calculate the probabilities of each class between 0 and 1. If the probability of a class is greater than .5, it will be assigned class 1 (positive) else 0 (negative).

Logistic Regression Results

```
[[8446  686]
 [ 414 2382]]
0.9077800134138162
              precision    recall  f1-score   support

         0.0       0.95      0.92      0.94      9132
         1.0       0.78      0.85      0.81      2796

    accuracy                           0.91     11928
   macro avg       0.86      0.89      0.88     11928
weighted avg       0.91      0.91      0.91     11928
```



Restaurant Recommendations

- Content Based Filtering using K-Nearest Neighbours
- Collaborative Filtering using SVD

Content Based Filtering

KNN:

- It takes similarities between two restaurant based on their features into consideration for recommendation.
- Euclidean dist was taken as selection criteria.

Preprocessing:

- Following features have been used:
  ['index', 'business_id', 'name', 'address', 'categories', 'attributes', 'stars', 'BusinessParking', 'Ambience', 'GoodForMeal', 'Dietary', 'Music']
- For categorical data such as 'GoodForMeal', 'attributes' etc. we created one hot encoding.

Results for Content Based Filtering using KNN

Restaurant indices for restaurants that are similar to 'Adelita Taqueria & Restaurant'

	distance	index	name	stars
0	4.000000	2329	Los Taquitos de Puebla	4
1	4.123106	2312	Yummy Sushi	4
2	4.242641	888	The Flavor Spot	4
3	4.358899	2573	Maker artisan pizza	4
4	4.358899	1488	Mood Indian Restaurant	4

Collaborative Filtering

Singular Value Decomposition:

- We used SVD to generate recommendations based on user's taste and likings.
- Pearson correlation coefficient was used as the selection criteria.

Preprocessing:

- We created a user rating matrices where rows are user_ids and columns are the ratings given to a particular restaurants. We apply SVD to this matrix due to its sparsity.
- We created the correlation matrix from the above matrix. For any restaurant, we create a list of restaurants from the correlation matrix which have high correlation value with the given restaurant.

Results for Collaborative Filtering

Restaurants similar to Reading Terminal Market are:

	corr_val	restaurant_name
0	0.999662	3J's Food Market
1	0.999722	@Ramen
2	0.921199	AmeriThai
3	0.999876	Bistro La Baia
4	0.963554	Café Soho

Future Scope

- Integrate our model with google maps to get location and recommend restaurants taking distance from current location into consideration.
- Fake reviews identification.
- Using Deep Learning techniques to enhance our models (LSTM for Review Analysis).