

Uncovering Insights from the Yelp Dataset: An Analysis of Restaurant Businesses

Project Definition:

This project aims to use the JSON-formatted Yelp Dataset to undertake a thorough analysis using Python Programming Language. It has a massive amount of relevant data that can be pulled from it and analyzed to learn more about businesses and consumer collective feedbacks. The goal of this project is to draw valuable conclusions from a large body of data. These insights can be used for several purposes, such as discovering popular businesses, improving marketing techniques based on user experiences.

YELP DATASET

The Yelp dataset is a collection of data from the Yelp platform, a widely-used online service where individuals can provide reviews and ratings for a wide range of businesses, including restaurants, hotels, and retail stores and others.

The dataset is released as including businesses, reviews, user, tips and check-in parts for use in personal, educational, and academic purposes.

We will use businesses and review datasets for our project.



150,346 businesses

Businesses data provides details about individual businesses including their names, addresses, categories (e.g., restaurants, hotels, bars), geographical coordinates, opening hours, and various other attributes that describe the business.



6,990,280 reviews

Reviews data consists of user reviews for various businesses. It includes the text form of the review, the user's rating (usually on a rating of 1 to 5 stars), the posting date of the review, and additional information related to the review.

The dataset includes reviews collected over a period of almost 17 years: **from February 16, 2005 , to January 19, 2022**. This long time period indicates that people have been using the Yelp platform regularly for many years to share their opinions and experiences about various businesses.

Businesses and Reviews

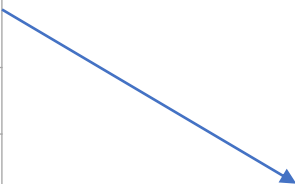
While Business_id is the primary key for business table, it is foreign key for reviews table. There is one-to-many relationship from business to reviews.

Businesses Table

| | Column | Type | Definition |
|----|--------------|---------|-----------------------------------------|
| 1 | business_id | string | business id (22 character unique) |
| 2 | name | string | the business's name |
| 3 | address | string | the full address of the business |
| 4 | city | string | City name |
| 5 | state | string | 2 character state code |
| 6 | postal code | string | the postal code |
| 7 | latitude | float | latitude |
| 8 | longitude | float | longitude |
| 9 | stars | float | star rating, rounded to half-stars |
| 10 | review_count | integer | number of reviews |
| 11 | is_open | integer | 0 or 1 for closed or open, respectively |
| 12 | categories | string | business categories |

Reviews Table

| | Column | Type | Definition |
|---|-------------|---------|---------------------------------|
| 1 | review_id | string | 22 character unique review id |
| 2 | user_id | string | 22 character unique user id |
| 3 | business_id | string | 22 character business id |
| 4 | stars | integer | star rating |
| 5 | date | string | date formatted YYYY-MM-DD |
| 6 | text | string | the review itself |
| 7 | useful | integer | number of useful votes received |
| 8 | funny | integer | number of funny votes received |
| 9 | cool | integer | number of cool votes received |



Businesses.json and Reviews.json

Business data sample:

```
{
  "business_id": "tnhfDv5Il8EaGSXZGiuQGg",
  "name": "Garaje",
  "address": "475 3rd St",
  "city": "San Francisco",
  "state": "CA",
  "postal code": "94107",
  "latitude": 37.7817529521,
  "longitude": -122.39612197,
  "stars": 4.5,
  "review_count": 1198,
  "is_open": 1,
  "categories": [
    "Mexican",
    "Burgers",
    "Gastropubs"
  ]
}
```

Reviews data sample:

```
{
  "review_id": "zdSx_SD6obEhz9VrW9uAWA",
  "user_id": "Ha3iJu77CxlrFm-vQRs_8g",
  "business_id": "tnhfDv5Il8EaGSXZGiuQGg",
  "stars": 4,
  "date": "2016-03-09",
  "text": "Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks.",
  "useful": 0,
  "funny": 0,
  "cool": 0
}
```

TOOLS

We performed analysis on the Yelp dataset using **Python** and utilized the following libraries:



Used to handle JSON files.



A popular library for data analysis and manipulation.



A powerful library for scientific computing.



Used for data visualization and creating graphs.



A visualization library based on matplotlib, known for creating more aesthetically pleasing and informative visualizations.



A language detection library used to identify the language of texts.



A library used for natural language processing (NLP) research and applications.

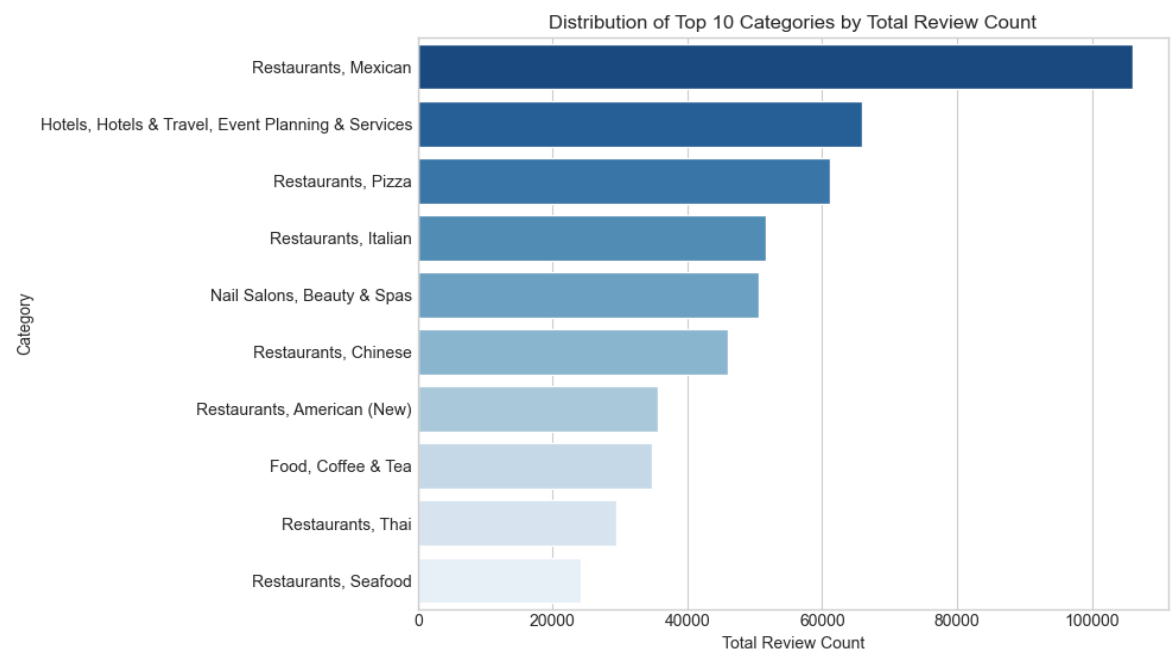


A library used to create a visual representation based on the frequencies of words in text data.

Highly Reviewed Categories Of Business Data

These data demonstrate the popularity of business categories based on customer reviews. Among the top 10 categories, 7 of them represent restaurants from different cuisines.

As we focus on customer reviews, we selected categories that include restaurants in our analysis.



In the business data, there are over 80k different type of businesses.

Restaurants Businesses

Since the Categories column have multiple attributes, we seperated only the businesses which includes 'restaurant' for further analysis. We define a binary column to refer if a business has restaurant attribute or not. After that, we selected if only is_restaurant column has TRUE in the data.

The sample filtered data is shown below;

| business_id | name | city | stars | review_count | categories | is_restaurant |
|------------------------|-------------------------------|---------------|-------|--------------|------------------------------------------------------------------------------------|---------------|
| j9Kaj_6tSeXmVMYCgZithg | Noble Crust | Wesley Chapel | 4.5 | 496 | Restaurants , Italian, Nightlife, Southern, Bars, Breakfast & Brunch, Pizza | TRUE |
| EIfs8kybcG-l60GJjNUIA | Domino's Pizza | Oldsmar | 2.5 | 20 | Pizza, Chicken Wings, Restaurants , Sandwiches | TRUE |
| z9wCTHYI2VZy9YIblwSsgg | The Silo Eatery Coffee Bakery | Tampa | 4 | 5 | Coffee & Tea, Restaurants , Food, Cafes, Bakeries | TRUE |
| 4xhGQGdGqU60BlznBjqnuA | California Tacos and Taproom | Isla Vista | 4 | 49 | Mexican, Beer Bar, Bars, Sports Bars, Nightlife, Restaurants , Tacos | TRUE |

Review Data

For our analysis, we kept specific columns such as star rating, review text, and business ID.

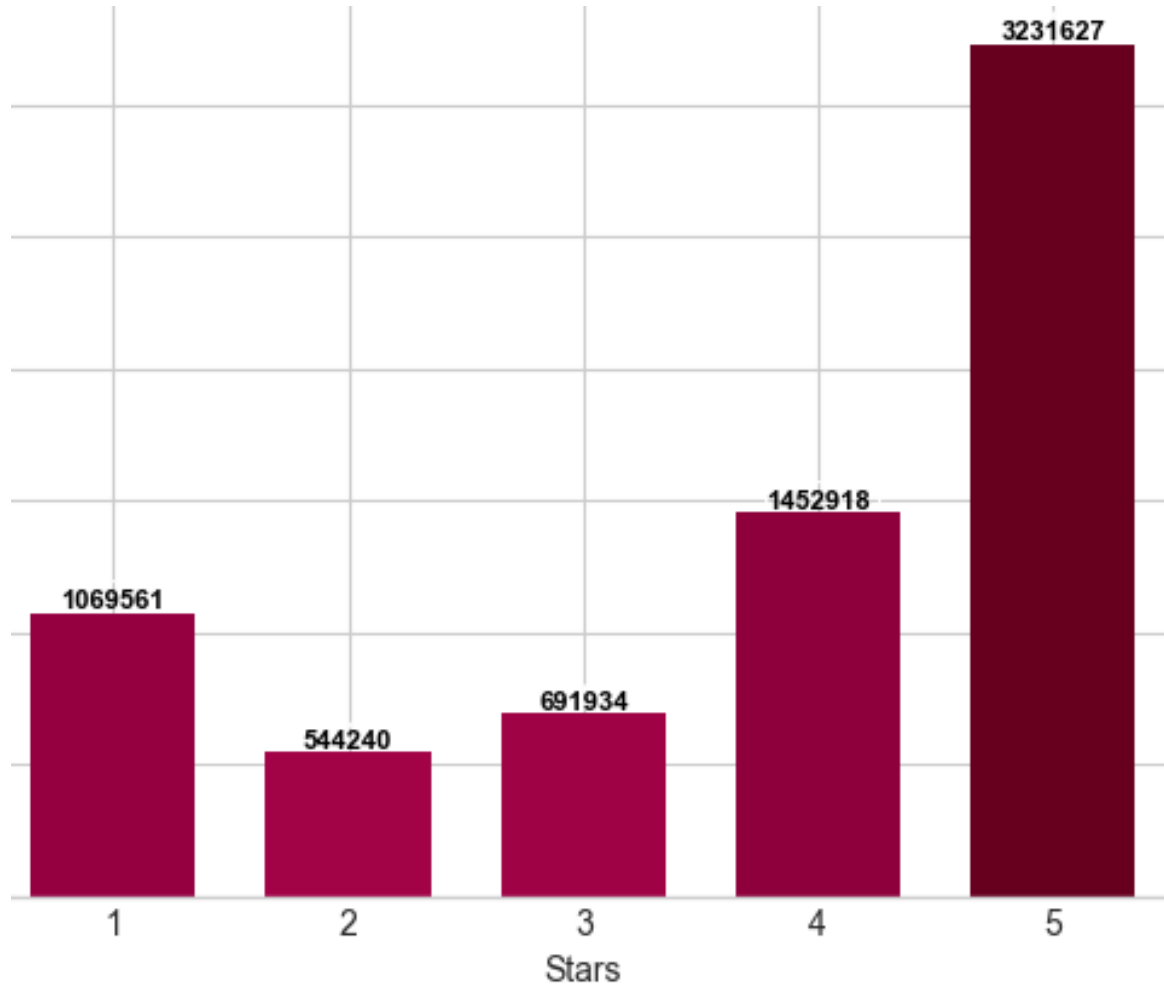
In the data each review is ranging from 1 to 5 stars. Now, let's take a visual look at some of this data:

| stars | text | business_id |
|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| 5 | Great place for breakfast! I had the waffle, which was fluffy and perfect, and home fries which were nice and smashed and crunchy. Friendly waitstaff. Will definitely be back! | BVndHaLihEYbr76Z0CMEGw |
| 5 | Tremendous service (Big shout out to Douglas) that complemented the delicious food. Pretty expensive establishment (40-50\$ avg for your main course), but its definitely backs that up with an atmosphere that's comparable with any of the top tier restaurants across the country. | YtSqYv1Q_pOItsVPSx54SA |

When we filtered restaurant businesses in review data, our data declined to 4.724.471 records.

We can say that the data is mostly composed of restaurant businesses.

The Distribution Of Stars For Restaurant Businesses



We observed that most of the reviews received a **5-star rating**, indicating a high level of satisfaction.

However, it is interesting to see that there are more **1-star** reviews compared to 2- or 3-star reviews.

This suggests that users are more motivated to write a review when they have either had an exceptionally positive experience or a very negative one.

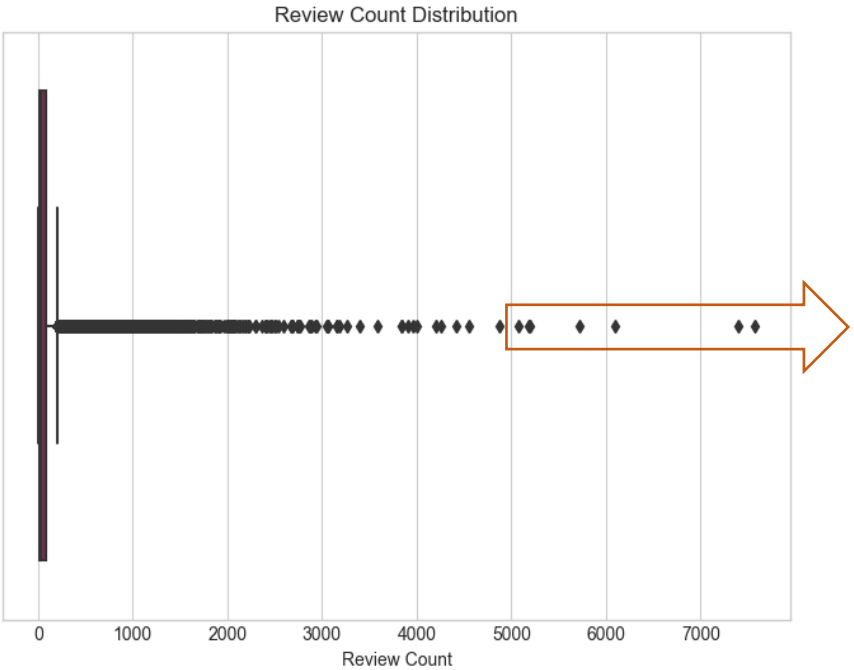
Most Popular Restaurants

Descriptive statistics of the review_count column for Restaurant businesses:

| | | |
|-------|-----|------|
| mean | min | max |
| 87.27 | 5 | 7568 |

Most of the restaurants have fewer than 500 reviews, while a few businesses have more than 1,000 reviews.

| review_count_bins | Business Count | Business Count % |
|-------------------|----------------|------------------|
| (0, 10] | 9784 | 18.7% |
| (10, 20] | 9267 | 17.7% |
| (20, 30] | 5798 | 11.1% |
| (30, 40] | 4077 | 7.8% |
| (40, 50] | 3048 | 5.8% |
| (50, 100] | 8603 | 16.5% |
| (100, 200] | 6243 | 11.9% |
| (200, 500] | 4191 | 8.0% |
| (500, 1000] | 961 | 1.8% |
| (1000, 10000] | 296 | 0.6% |



Since we are interested in most popular restaurants, we filtered the data with a more specific threshold of over 5k reviews.

Most Popular Restaurants-2

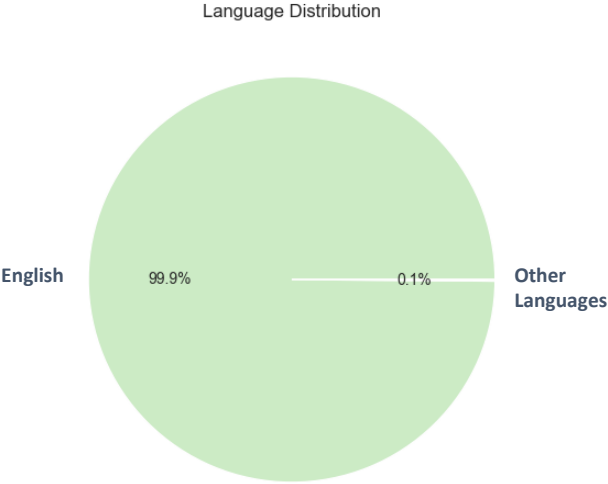
There are only 7 restaurants have at least 5k reviews in the data.

| name | city | stars | review_count | categories |
|-------------------------|--------------|-------|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Acme Oyster House | New Orleans | 4 | 7568 | Live/Raw Food, Restaurants, Seafood, Cajun/Creole |
| Oceana Grill | New Orleans | 4 | 7400 | Restaurants, Seafood, Cajun/Creole, Breakfast & Brunch |
| Hattie B’s Hot Chicken | Nashville | 4.5 | 6093 | American (Traditional), Chicken Shop, Southern, Restaurants, Chicken Wings, American (New), Soul Food |
| Reading Terminal Market | Philadelphia | 4.5 | 5721 | ...Public Markets, Food Court, Wineries, Local Flavor, Ethnic Food, Restaurants... |
| Ruby Slipper | New Orleans | 4.5 | 5193 | Restaurants, American (Traditional), American (New), Cafes, Breakfast & Brunch |
| Mother's Restaurant | New Orleans | 3.5 | 5185 | Cajun/Creole, Restaurants, Event Planning & Services, Southern, Specialty Food, Soul Food, Food, Ethnic Food, American (New), Caterers, Breakfast & Brunch, Sandwiches |
| Royal House | New Orleans | 4 | 5070 | Restaurants, American (New), Sandwiches, Seafood, Cajun/Creole |

As it can be seen from the table, among the restaurants considered as outliers, Mother's Restaurant stands out with an overall rating of 3.5 stars. With 5185 reviews, it is evident that it is a popular restaurant.

We chose Mother’s Restaurant for our analysis to understand why customers hold significantly different reviews about the restaurant.

The Distribution of Languages



We used a language identifier module named ‘Langid’ to detect the language of the reviews. As expected, most of the reviews are in English, and we filtered out reviews in other languages.

Let’s look at the filtered data below;

| text | language |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| Op suggestie van yelp- hier geweest op onze eerste avond in NOLA tijdens onze roadtrip (dag3). It's the best place. Wachttijden kunnen oplopen, maar waar Nederland vaak faliekant in faalt, gaat hier geweldig: er wordt door t personeel vlijmscherp in de gaten gehouden wie er aan de beurt is, zelfs voor de krukken aan de bar. | nl |
| Excelente comida mediterránea! El gyro es buenísimo y la ensalada griega de las mejores q he comido. Las porciones son bastantes generosas. Sitio recomendado al 100% vendré cada vez que pueda | es |
| 点了四个菜 没一个好吃的 不适合中国人吃 口水鸡鸡肉不新鲜 毛血旺的大肠好油 酸豆角炒肉沫做的超级咸 松鼠鱼中规中矩，但卖27块钱也不便宜 总的来说就是又油又咸，吃完了以后要不停喝水。 | zh |
| Wer auf frisch geräucherten Fisch steht, der noch warm aus dem Rauch kommt, der is(s)t hier genau richtig. Dazu gibt es richtig guten deutschen Kartoffelsalat. Die Portionen war sehr groß und der Preis mit 14-18 \$ auch in Ordnung. Die Bedienung war sehr freundlich. | de |
| <u>Wem das Anglerglück hold war, der kann sich seinen Fisch auch direkt vor Ort räuchern lassen.</u> | |

Mother's Restaurant Review Analysis



The first step is to use the NLTK library to identify "stop words" and remove them from the reviews. Stop words are commonly used words in the language (e.g., "the", "and", "in", etc.) that are not relevant for analysis. By identifying and removing these words, we cleaned the text data.

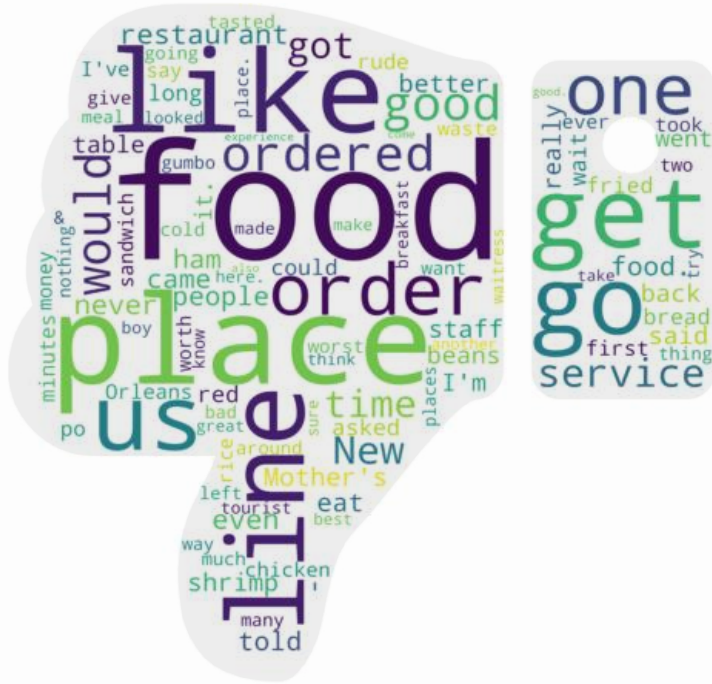


In the next step, we tokenized all the reviews. Tokenization is the process of breaking down the text into individual words or tokens. After tokenization, we counted how many times each word appears across all the reviews.



Finally, we counted each word separately for 5-star and 1-star reviews. This allowed us to understand which words are more frequent in most positive reviews and most negative reviews.

Mother's Restaurant 1-Star Reviews



Customers with negative reviews note that sometimes the **food is served cold** and does not meet their taste expectations. Some of them express that the restaurant **doesn't clean** and encounter **issues with customer service**, including experiencing **rude behavior**. The **taste or appearance of coffee and shrimp**, also fail to satisfy the customers. Additionally, **long waiting lines** at the restaurant contribute to customers' dissatisfaction. These negative reviews indicate that there is a need for improvements in areas like **service quality** and **food quality** to enhance customer satisfaction.

Most Common Word Combinations

| Food | |
|---------------|---------------|
| Food cold | Food service |
| Food bland | Food mediocre |
| Food terrible | |

| Place | Service |
|-------------|------------------|
| Place dirty | Customer service |
| Many places | Service terrible |
| Hype place | Rude service |

| Like | |
|------------------------|--------------------|
| Don't like/Didn't like | Coffee tasted like |
| Food looked like | Shrimp tasted like |
| Like cafeteria food | |

| One | Order | Line |
|---------------|--------------|--------------------|
| One star | Order shrimp | Long line |
| Took one bite | Cold order | Wait line |
| | | Line 20/30 minutes |

Mother's Restaurant 5-Star Reviews



Customers with a positive view of Mother's Restaurant use expressions like **"go back"** and **"come back,"** indicating that the restaurant is a favorite among those who visit New Orleans.

Some customers even mention that it's worth waiting in line for. Most of the high ratings for the restaurant stem from the quality and deliciousness of specific dishes like **Seafood Gumbo** and **Cajun Food**, which are highly appreciated by customers. The phrase "**highly recommend place**" also showcases customers' complete satisfaction with the restaurant.

Most Common Word Combinations

| Food | |
|-----------------------|----------------------|
| Great Food, Good Food | Food Amazing |
| SeaFood Gumbo | Authentic Cajun Food |
| Soul Food | Food Worth Wait |

| Place | |
|-----------------|------------------------|
| Recommend Place | Highly Recommend Place |
| Love Place | Place New Orleans |
| Great Place | Line Place Order |

| New | |
|-------------------|---------------------|
| New Orleans | Back New Orleans |
| Visit New Orleans | New Orleans Mothers |
| Trip New Orleans | New Orleans Food |

| Get | Best | Go & Back |
|----------------|--------------------|-----------|
| Get Line | Best Food | Go Back |
| Get Line Order | Best Fried Chicken | Must Go |
| | Best Bread Pudding | Come Back |