Final Capstone

Sentiment Analysis of New York Restaurants

by

Vera Arrebola Granes

Applied Business Analytics

MDA 720

Long Island University

Prof. Singh

April 30, 2024

Contents

Background

The data for this project was collected from the Yelp website. Yelp is a platform that connects people with great local businesses. It primarily focuses on user-generated reviews of local businesses, such as restaurants, cafes, bars, salons, and more. The platform also provides users with Yelp Fusion which is an Application Programming Interface (API) that allows developers to access Yelp's wealth of local business data and user reviews programmatically. It is very useful since is accessible for anyone that creates an account for free and gives you loads of data to work with. I had to obtain my own API number in order to get the data. The website's link is: https://www.yelp.com/developers.

When you add data from Yelp into python you are able to filter some of the characteristics, so it fits your field of research. In my case, I chose to filter the data to be businesses in the New York City area and the type of business is restaurants. It is important we filter the data, so we only obtain what we need for our research project.

Objective/Goals of the Project

The main goal of this project is to gather information that would be useful for a Consulting firm specializing in the restaurant industry in New York City. This firm would focus on helping entrepreneurs being successful when opening a restaurant thanks to the information obtained and the analysis performed in current restaurants. By leveraging Sentiment Analysis insights, you can identify market opportunities, address customer needs and pain points, and develop innovative solutions that add value to a restaurant.

The analysis performed for this project is only the beginning of a comprehensive approach to understanding the industry and offering consulting services to those restaurants that are in the early stages of their life. A Sentiment Analysis is not enough to make business decisions, that is why further research will be implemented in other areas. However, with the Sentiment Analysis of the customer reviews we can already have insights and hints of where to start.

Data Extraction

As mentioned above, I connected Python to the Yelp API to extract restaurant data, collect relevant information, and store it in a structured format for further analysis. One of the interesting features of the code is that you can easily change the filters and customize them to other types of businesses and location. Moreover, it is important to highlight that when using Yelp API, there is a limit of 50 results that can be extracted. Therefore, I decided to duplicate the code to get 100 results and added an offset of 50 so the instances would not repeat. I believe that 50 points of information was not enough to properly conduct a Sentiment Analysis that could provide useful information.

Once the code is done running a message will appear and inform the user that the data extraction and storage was completed successfully. Finally, the two data sets (each of them with 50 unique results) need to be merged into one before performing any data exploration and analysis. I ended up with a data frame with 100 rows and 6 columns. The following table shows information about each column in the data set:

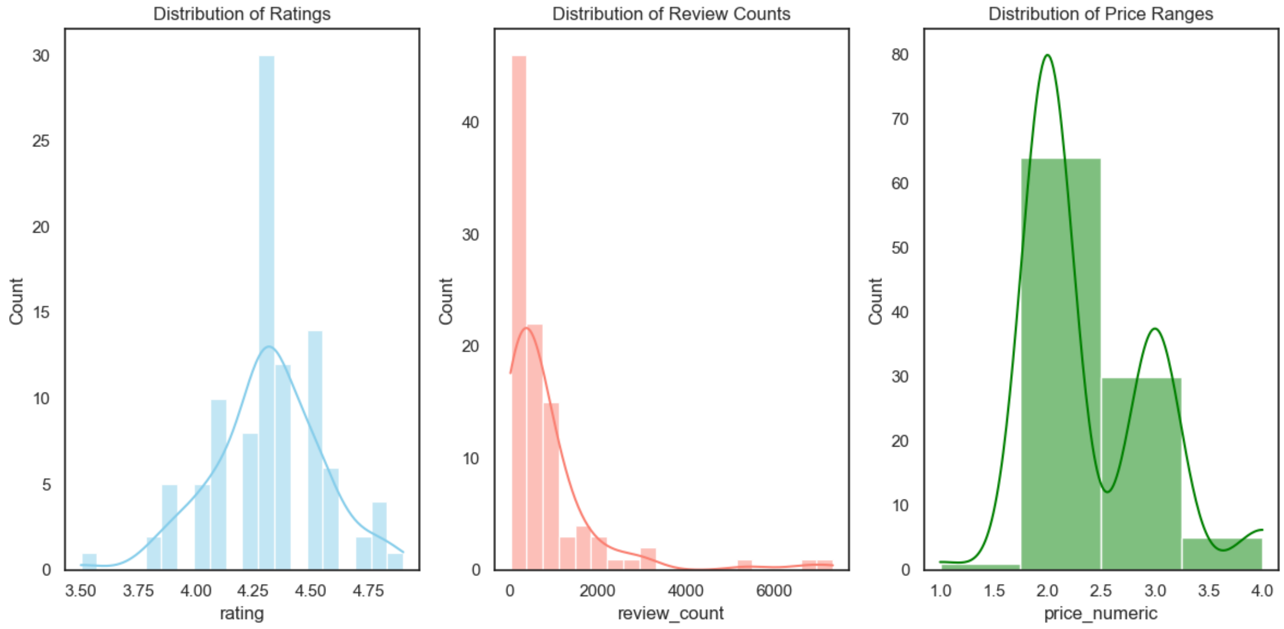
Column Name	***Description***	***Data Type***
Name	Restaurant name	dtype('O')
Price	Price rate using the $ symbol from 1(less expensive) to 4 (more expensive)	dtype('O')
Category	Type of cuisine	dtype('O')
Rating	Customer rating from 0 to 5	dtype('float64')
Review Count	Total number of reviews	dtype('int64')
Reviews	List of three reviews	dtype('O')

Data Exploration/Data Visualization

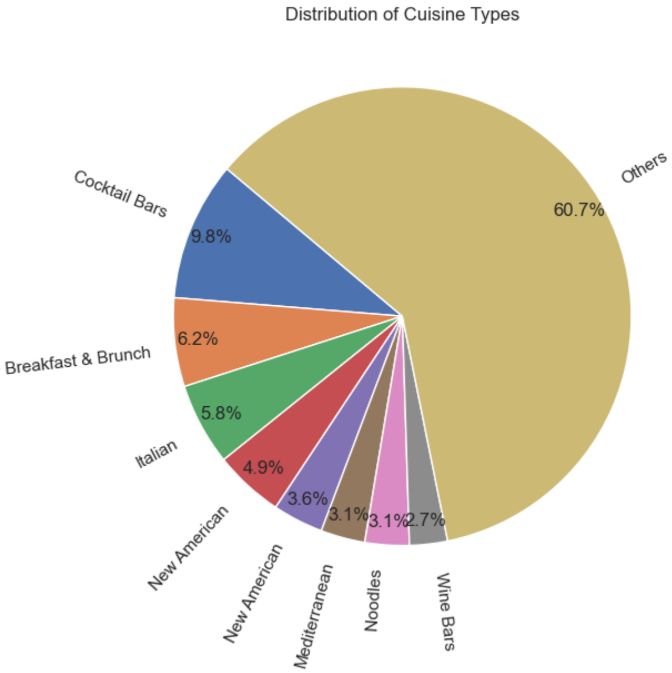
Managing this data set was quite challenging for many reasons and many data cleaning and manipulation techniques had to be implemented. First, I had to create a new column “price_numeric” that served as a dictionary for the “price” column. The problem with “price” was that it used the $ symbol to express how expensive the restaurant is. Therefore, on “price_numeric,” I display the number of $ signs as a way of translating the symbol (1 being cheaper and 4 being more expensive). This process was very important so we can use the information for visualization of the data.

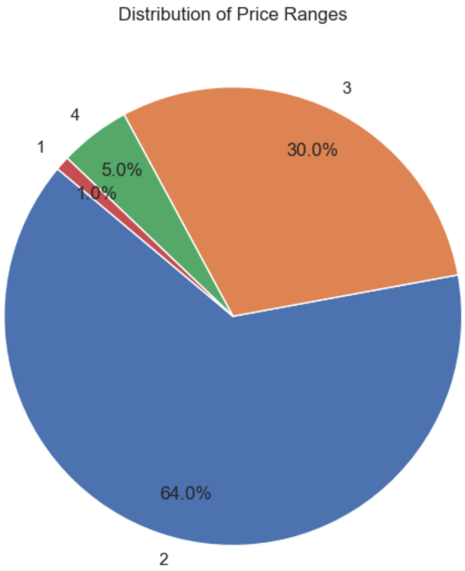
As usual, I checked for missing values and duplicates as these are two techniques that must be done to make sure our data set is as good as possible. There were not any missing values, yet there were some duplicates. However, when I displayed the rows that were supposed to be duplicated, they were not, and I decided to just leave the data frame as it was. Something that is also worth pointing out is that part of the struggle with the data set is the fact that the “reviews” column has lists in each of the rows. This is because from each restaurant I wanted to gather three different reviews. Later, I will explain how it fixed that before conducting the Sentiment Analysis.

Once the data was ready, I created different plots that would give me a deeper understanding of the nature of my data set. Histograms provide a visual representation of the distribution of a dataset. The first one starting from the left shows the frequency of different rating values, where all the ratings are in between 3.5 and 5, having the most frequent around 4.25. The second plot displays the distribution of the number of reviews each restaurant has in total, in which most of them have less than 2000 customer reviews. Finally, the furthest right illustrates the frequency of restaurants categorized by their price levels, the majority of them rated with two- or three-dollar signs.



On to the next set of graphs, I wanted to show the distribution of the different types of restaurants with percentages. The category “Others” is just a combination of all the types of restaurants except the eight most popular. On the right side, there is another pie chart for the distribution of price ranges, and as I discussed earlier, the two and three rates are the most popular.

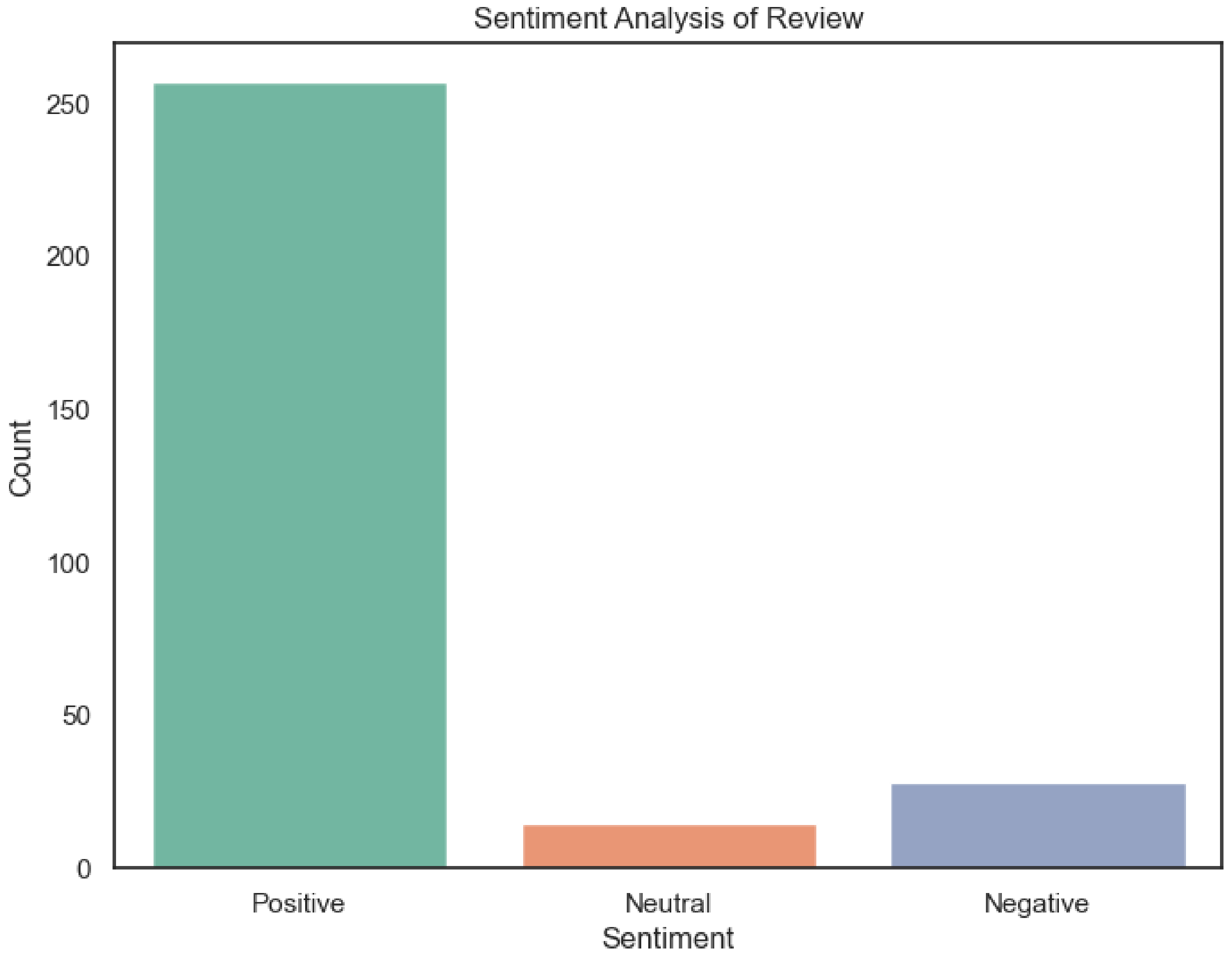




Sentiment Analysis

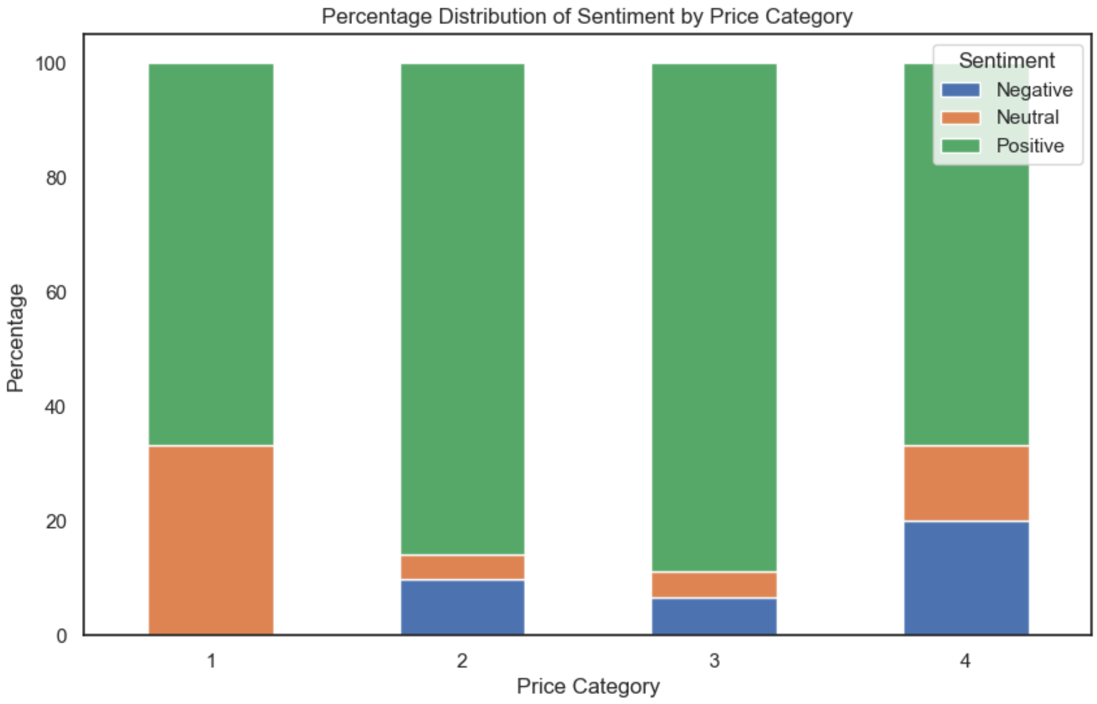
Performing sentiment analysis on text data, such as reviews from customers, involves analyzing the sentiment expressed in each piece of text to determine whether it's positive, negative, or neutral. There are many existing libraries in python that already have Sentiment

Analysis models to be used, in my case I used NLTK and TextBlob. It is crucial to define thresholds or rules to determine the sentiment polarity of each review based on the scores or labels provided by the sentiment analysis tool. Sentiment analysis can reveal the distribution of sentiment across different categories, topics, or segments within a dataset. For example, it can show the proportion of positive, negative, and neutral sentiments for a product, service, or brand.



As you can see in the figure above, most reviews had a positive sentiment. When the sentiment is positive, it means that the overall tone or attitude expressed in the text is favorable, optimistic, or approving. In sentiment analysis, positive sentiment usually indicates satisfaction, enjoyment, approval, or happiness towards a particular subject, product, service, or experience.

Furthermore, I wanted to check how the sentiments were distributed across each price range (one through four) and that is why I made a stacked bar plot. Analyzing sentiment by price category allows us to understand how consumers perceive the value proposition of restaurants at different price points.



It is very interesting to see that specially categories two and three (which are the medium prices) have the highest percentage of positive reviews at almost 90%. Positive sentiment in specific price categories may indicate a competitive advantage or differentiation strategy based on factors such as cuisine, ambiance, service quality, or unique offerings. I also want to highlight the fact that category one has 0% of negative reviews. From the research I have done this can make sense because usually people do not expect as much from those restaurants where they do not spend that much money.

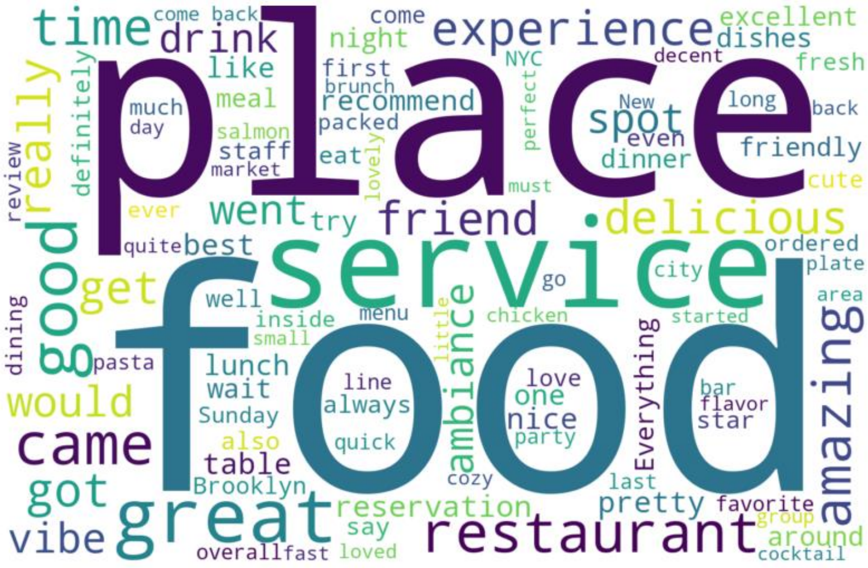
Finally, looking at category four, it is the one with the highest negative sentiment. Customers tend to have higher expectations when dining at more expensive restaurants due to the premium price they pay. This can lead to heightened sensitivity towards any flaws or discrepancies in the dining experience, resulting in a higher likelihood of negative feedback. Moreover, customers dining at more expensive restaurants may be more inclined to provide feedback, both positive and negative, due to the significant investment they make in the dining experience.

Conclusions

Through sentiment analysis of restaurant reviews, we gained valuable insights into customer opinions, preferences, and experiences. By analyzing the sentiment distribution, we can identify areas of strength and areas for improvement in restaurant offerings, service quality, and overall customer satisfaction. As the beginning of my consulting firm journey, Sentiment Analysis is a good start point. The consulting firm can help new or current restaurant owners with both analyzing their own customers' reviews and their competitors' reviews as well. This information can guide decision-making for entrepreneurs looking to enter the restaurant industry in New York City.

With more information such as location, number of employees, hours of operation, total years' operating or average number of customers per day, we can perform a more insightful analysis that provides us with more decision-making solutions for our clients. In addition, with more data we could also create predictive models that we can use with the restaurants and help them create a competitive advantage and understand the restaurant industry better. By leveraging data analytics and industry expertise, the consulting firm can help restaurant owners navigate challenges, capitalize on opportunities, and achieve long-term success. Being able to continuously monitor market conditions, customer feedback, and industry trends, will also help restaurant owners' adapting their strategies and offerings accordingly.

To finalize the project, I want to show a word clow plot with the most repeated words across all the reviews. We can also get interesting information out of this graph. As I said in the previous section, most reviews were positive and that is probably why there are many encouraging words such as "excellent," "great," "amazing," or "good." However, and most importantly, we should pay more attention to words like "place" which means that many reviews (positive, negative, or neutral) mention the location. Or "service," meaning that many customers had something to say about the quality of the service. Obviously, the word "food" being the biggest is not a surprise since the main goal of anyone at a restaurant is to consume good food. Knowing which words or topics are mentioned more frequently in customer reviews is essential for decision making because it shows that clients pay attention to those.


time
come back
drink
come
experience
excellent
dishes
night
like
NYC
decent
fresh
really
definitely
place
first
brunch
perfect
recommend
much
day
meal
long
New
back
packed
spot
salmon
even
review
staff
friendly
eat
lovely
market
dinner
must
cute
good
ever
went
try
friend
delicious
quite
best
ordered
service
get
go
city
plate
well
dining
area
inside
menu
chicken
started
small
little
ambiance
Everything
amazing
pasta
lunch
line
love
bar
flavor
wait
one
food
would
always
nice
star
Sunday
also
quick
party
came
table
got
Brooklyn
cozy
last
reservation
pretty
favorite
group
say
vibe
great
restaurant
around
overall
fast
loved
cocktail